mathematics

MDPI

Article

A MCDM-Based Framework for Quantifying Reproducibility Readiness in Machine Learning Research

Paulina Leščinskaitė, Remigijus Paulavičius and Ernestas Filatovas *

Institute of Data Science and Digital Technologies, Faculty of Mathematics and Informatics, Vilnius University, Akademijos Str. 4, LT-08412 Vilnius, Lithuania; lesci.paulina@gmail.com (P.L.); remigijus.paulavicius@mif.vu.lt (R.P.)

* Correspondence: ernestas.filatovas@mif.vu.lt

Abstract

Machine Learning (ML) is increasingly used across scientific domains, raising concerns about the reproducibility of published results. Reproducibility is a fundamental principle of scientific research, yet many ML research works remain difficult to reproduce due to missing artifacts and insufficient reporting. This study addresses the lack of practical quantitative methods for assessing reproducibility in ML research by proposing a paper-level evaluation framework based on Multi-Criteria Decision Making (MCDM). Through a synthesis of theoretical and data-driven analyses, we identified seven key reproducibility criteria: data and code availability, the inclusion of a `README` and trained models, hyperparameter and training descriptions, and paper readability. These criteria are then aggregated into a unified quantitative score for 'R1' level reproducibility readiness using the Weighted Sum Model (WSM) and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). In the baseline evaluation, equal criterion weights were employed, while the Analytic Hierarchy Process (AHP) was demonstrated as a practical approach for deriving context-dependent weights tailored to diverse stakeholder priorities. The framework was applied to an annotated set of 139 randomly sampled ML research papers and benchmarked against an existing reproducibility label, achieving an accuracy of 0.64, a precision of 0.40, and a recall of 0.70. These results demonstrate that MCDM provides a feasible, interpretable, and flexible foundation for quantifying reproducibility readiness, allowing the assessment to be adapted to different decision-making contexts through customized criterion weighting.

Keywords: machine learning; computational reproducibility; quantitative assessment; reproducibility readiness; AHP; MCDM; TOPSIS; WSM

MSC: 90B50; 90C29; 68T05; 68T09

Academic Editors: Gabriella Marcarelli, Pietro Amenta and Antonio Lucadamo

Received: 20 March 2026
Revised: 23 April 2026
Accepted: 28 April 2026
Published: 1 May 2026

1. Introduction

Machine learning (ML) has seen significant growth in research over the past decade. Recent advances demonstrate the widespread adoption of ML across diverse fields, including cybersecurity [1], bioinformatics [2], healthcare [3,4], materials science [5], remote sensing [6,7], and quantum computing [8]. Illustrating this trend, the Stanford AI Index reports that the prevalence of ML in research papers has increased fourfold since 2015 [9].

With the growing use of ML in research, concerns about the reproducibility of published results have become increasingly prominent. Gundersen and Kjensmo [10] reported that fewer than one third of artificial intelligence and ML research papers were reproducible,

with commonly cited causes including incomplete code and data sharing [11,12], insufficient reporting of experimental procedures [13], unaccounted sources of randomness [14], and methodological issues such as data leakage [15]. Collectively, these factors contribute to the widely discussed "reproducibility crisis" [15], raising concerns for scientific integrity and the trustworthiness of ML-based evidence [16]. Reproducibility is essential to verify that findings are robust and can reliably support scientific claims; conversely, irreproducible results can misdirect subsequent work and waste research effort as the community builds on prior studies.

Despite numerous initiatives aimed at enhancing transparency and reproducibility in machine learning (ML) research [13,17,18], universally accepted standardized practices have yet to be established. Current efforts primarily rely on reproducibility checklists [19], publication guidelines [20], open data and code policies [21,22], and standardized documentation protocols [23,24].

While these measures have significantly raised awareness and improved reporting standards, they remain predominantly qualitative and often depend on binary or subjective assessments. Consequently, there is a persistent lack of practical, quantitative methodologies for measuring the reproducibility of individual ML research papers.

This study addresses this gap by proposing a quantitative reproducibility assessment framework grounded in Multi-Criteria Decision-Making (MCDM). MCDM provides a rigorous mathematical foundation to aggregate heterogeneous reproducibility criteria into a single, interpretable score. Furthermore, it offers the flexibility to adjust criterion weights to reflect diverse evaluation objectives, such as prioritizing artifact availability over reporting quality.

Importantly, the proposed framework evaluates reproducibility readiness rather than guaranteeing actual reproducibility. Specifically, it assesses whether a research paper provides sufficient artifacts and methodological transparency to enable reproduction without executing the experiments themselves. While a higher score indicates better preparedness and increases the likelihood of successful reproduction, it does not ensure that reproduction will succeed, as this may depend on factors not directly observable from the paper, such as hidden dependencies, environment configurations, or incomplete implementations.

Building on this distinction, the novelty of the proposed framework lies not in introducing new MCDM methods, but in integrating established decision-making techniques into a structured approach for assessing reproducibility readiness. By transforming reproducibility-related criteria into a quantitative and interpretable score, the framework enables systematic comparison and flexible adaptation to different evaluation objectives.

The main contributions of this study are as follows:

- A comprehensive review of reproducibility in ML, including key definitions, primary sources of irreproducibility, and current mitigation strategies.
- The formulation of a set of quantifiable reproducibility criteria for ML research papers, derived from both theoretical considerations and empirical analysis.
- The introduction of a MCDM-based framework for the systematic quantitative assessment of reproducibility in ML research.
- An empirical evaluation of the proposed framework using a manually annotated dataset of 139 ML research papers to demonstrate its applicability and effectiveness.

The remainder of this paper is organized as follows. Section 2 reviews prior work on reproducibility in machine learning, including key definitions, common sources of irreproducibility, and existing interventions. Section 3 presents the proposed framework, including the derivation of quantifiable reproducibility criteria, the ordinal scoring rubric, and the MCDM-based score computation, together with the datasets and the preprocessing, cleaning, sampling, and annotation procedures. Section 4 reports the empirical evaluation

results on an annotated dataset of ML research papers. Section 5 discusses the main findings, implications, limitations, and future research directions. Finally, Section 6 concludes the paper.

2. Background & Related Work

This section defines the reproducibility notion adopted in this work, summarizes major sources of irreproducibility in ML, reviews commonly used interventions, and surveys checklist-based and quantitative assessment approaches.

2.1. Reproducibility and Replicability in Machine Learning

The terms *reproducibility* and *replicability* are used inconsistently across scientific communities, and no universally accepted distinction exists [25]. While some fields treat them as synonyms, others adopt conflicting conventions regarding the reuse of original code and data [25]. Within the ML domain, reproducibility is typically defined by whether published results can be consistently re-obtained under specified experimental conditions [26].

Barba [25] summarizes three frequent conventions:

- The terms are used interchangeably;
- Reproducibility means obtaining the same results using the original code and data, while replicability refers to reaching similar conclusions with different data;
- The inverse convention, where replicability relies on original data and reproducibility allows alternative datasets.

To avoid ambiguity, this study adopts Gundersen's degree-based framework [26]. At level R1, reproducibility refers to obtaining identical outcomes using the same data, code, model, and computational environment. Level R2 permits independent implementations while retaining the original dataset. Level R3 evaluates reproducibility at the level of interpretation, where similar conclusions are reached using different datasets and modeling approaches.

This work focuses on R1 reproducibility. This restriction is intentional, as R1 reproducibility is the most clearly defined and objectively assessable level, relying on the availability of original artifacts. In contrast, higher levels (R2 and R3) require independent implementation and validation under varying conditions, introducing significant methodological variability and typically requiring experimental reproduction. As the proposed framework is designed for paper-level assessment without executing experiments, focusing on R1 reproducibility provides a well-defined and practically applicable scope. Accordingly, an ML study is treated as reproducible if its reported results can be reproduced using the original artifacts (e.g., code, datasets, model specifications, parameters, and computational setup). This operationalization is consistent with the definition adopted by the National Academies of Sciences, Engineering, and Medicine [27]. Importantly, the goal here is to assess *reproducibility readiness* based on what a paper discloses and provides (artifacts and reporting), rather than to reproduce experiments or evaluate generalization.

2.2. Sources of Irreproducibility in Machine Learning

Empirical evidence indicates that a substantial fraction of ML results are difficult to reproduce. For example, Gundersen and Kjensmo [10] report that fewer than one-third of AI/ML research papers were reproducible, motivating analysis of common failure modes.

Code, environment, and model specification. Unshared or partially shared code restricts independent verification [11]. Even when code is shared, missing dependency information (e.g., library versions) can prevent successful execution. Insufficient reporting of model architectures and training procedures is also a frequent barrier [13].

Data accessibility and dataset specification. Private or inaccessible datasets can directly prevent reproduction. Heller et al. [12] found that more than half of the analyzed MICCAI studies relied on private data, limiting verification. In addition, undocumented preprocessing and methodological issues such as data leakage can yield overly optimistic performance estimates and discrepancies between original and reproduced results [15].

Stochasticity and uncontrolled randomness. ML pipelines typically include stochastic components (e.g., initialization, sampling), and repeated runs may yield different outcomes if randomness is not controlled. Bouthillier et al. [14] emphasize that randomness control is necessary for reproducibility, although it does not guarantee robustness across configurations.

Reporting and evaluation practices. Selective reporting of only successful experiments biases the evidence base and obscures experimental context [28]. Inconsistent metrics, incorrect statistical testing, and missing uncertainty reporting further complicate reproduction and comparison [13].

Bridge to criteria formulation. These sources collectively indicate that R1 reproducibility depends on (i) the availability of concrete artifacts required to rerun the work (e.g., data, code, trained models, and usage instructions) and (ii) sufficient methodological transparency (e.g., training setup, hyperparameters, and overall clarity of presentation). This observation directly motivates the artifact- and reporting-centric criteria defined and quantified in this study.

2.3. Existing Approaches to Improving Reproducibility

A wide range of initiatives aim to improve reproducibility and transparency in ML, yet no universal standard has emerged. Reviewing these approaches clarifies why a complementary *quantitative, per-paper* assessment is still needed.

Sharing infrastructure and environment capture. Code hosting platforms and artifact repositories support version control and dissemination, while containerization enables distribution of complete computational environments [21,22]. These tools can reduce environment-related discrepancies, but they do not, by themselves, provide a quantitative assessment of whether a given paper is reproducible.

Integrity and provenance of research artifacts. Beyond hosting and packaging, ensuring that shared artifacts are trustworthy and traceable has motivated approaches based on tamper-evident provenance and auditability. For example, blockchain-based mechanisms have been proposed to strengthen artifact integrity and reproducibility workflows [18], complementing repository- and container-based sharing.

Randomness control and determinism. Seeding and deterministic settings can reduce variance induced by stochastic components; however, reproducibility can still be affected by nondeterministic GPU operations, parallelism, and floating-point differences across hardware/software configurations. As a result, randomness management is helpful but insufficient on its own and must be complemented by artifact availability and clear reporting.

Improving data accessibility. Dataset availability is a key limiting factor for reproducibility and scientific progress [29]. While policies such as mandatory open data statements have been adopted [20], empirical evidence suggests that policy introduction alone may not substantially improve reproducibility outcomes [20].

Publication and documentation standards. Reproducibility checklists and documentation practices have become increasingly common. For example, major venues require checklist-style reporting of datasets, code, and experimental procedures [19]. More broadly, conferences and journals increasingly request reproducibility statements during submis-

sion [23,24]. However, interpretation and enforcement vary across venues, and resulting assessments are often qualitative or binary.

Overall, existing interventions primarily target improving practice (sharing and reporting). They rarely provide an interpretable quantitative score that can be computed for a single paper without rerunning experiments, which motivates the MCDM-based framework proposed in this study.

2.4. Checklist-Based Criteria and Their Limitations

Reproducibility checklists implicitly encode criteria by specifying what information should be reported to enable reproduction. They are therefore a useful starting point for identifying candidate criteria and for designing an appropriate scoring scale.

A widely cited example is the *Machine Learning Reproducibility Checklist* [13], which covers model specification, datasets, code, and experimental results. While comprehensive, such checklists are often lengthy and use binary responses (e.g., "Yes/No"), making it difficult to represent partial fulfillment and potentially over-penalizing papers that provide incomplete yet meaningful information.

The *NLP Reproducibility Checklist* allows authors to justify unmet criteria via short explanations [30], supporting more nuanced reporting and reducing burden compared to longer binary checklists. Nevertheless, textual justification does not directly yield a quantitative measure. For a paper-level score, an *ordinal* scale (from fully satisfied to omitted) is more suitable, because it encodes degrees of compliance and can be mapped to numerical values for aggregation. Binary checklist designs may also distort *perceived reproducibility* by conflating partial versus missing reporting; an ordinal rubric explicitly captures intermediate cases [30].

Magnusson et al. [30] further provide empirical evidence (based on large-scale checklist responses) that checklist compliance correlates with perceived reproducibility and acceptance rates, identifying code availability as particularly influential. These insights support the inclusion of artifact-centric criteria in the proposed framework.

2.5. Quantitative Studies of Reproducibility

Quantitative reproducibility research can be broadly grouped into *prediction-based* and *measurement-based* approaches.

Prediction-based approaches treat reproducibility as a classification task and use textual, bibliographic, or structural features to predict whether a study will replicate. NLP-based studies report that cues related to dataset accessibility and statistical reporting can be predictive, with reported accuracies ranging from approximately 65% to 87% [31–33]. Other models incorporate structural or social features (e.g., citations, author attributes, venues) [34,35]. While informative, these approaches primarily optimize predictive performance and typically do not yield an explicit, criteria-based score that is directly interpretable for a single paper.

Measurement-based approaches are more aligned with the objective of this work when "measurement" is understood as producing a criteria-grounded score. Belz et al. [36] propose Quantified Reproducibility Assessment (QRA), comparing reported and replicated performance across runs; however, it requires rerunning experiments and is task-dependent. Gundersen and Kjensmo [10] operationalize reproducibility by defining variables that should be documented to enable reproduction across reproducibility levels [26], and analyze how often papers satisfy these requirements.

Table 1 summarizes how major approach families differ in output type, interpretability, and practical cost.

Table 1. Comparison of prominent approaches for improving or assessing reproducibility in ML.

Approach Family	Primary Output	Requires Rerun	Quantitative Score	Partial Compliance	Representative Examples
Sharing infrastructure & containers	Artifacts/environment	No	No	N/A	Code hosting, containers [21,22]
Provenance/integrity mechanisms	Traceability/auditability	No	No	N/A	Blockchain-based workflows [18]
Policies & disclosure requirements	Statements/mandates	No	Indirect	Limited	Open-data statements [20]
Checklists & structured reporting	Item responses	No	Often binary	Weak–moderate	ML checklists [13,19]
Prediction models	Probability/label	No	Model-dependent	Indirect	NLP/social predictors [31,35]
Rerun-based metrics	Metric differences	Yes	Yes	N/A	QRA [36]
Criteria audits (descriptive)	Coverage statistics	No	Aggregate rates	Moderate	Variable reporting analysis [10]
This work (criteria + MCDM)	Per-paper score	No	Yes	Yes	Ordinal criteria + aggregation

In summary, prior work motivates an artifact- and reporting-centric view of reproducibility, but existing quantitative approaches either prioritize prediction without producing an explicit, criteria-based score, or require running replications. This gap motivates the MCDM-based framework introduced in this study, which aggregates multiple explicit criteria into a single interpretable score for individual papers.

3. Materials and Methods

This section presents the proposed framework for computing a quantitative reproducibility score for an ML research paper and the methods used to construct, operationalize, and evaluate it. The framework combines a set of quantifiable reproducibility criteria, an ordinal rubric for criterion scoring, normalization, and MCDM-based aggregation into a single reproducibility score.

3.1. Framework Overview

Figure 1 summarizes the proposed paper-level framework for quantifying R1 reproducibility readiness. The framework consists of two phases:

Phase I: Framework construction.

- Criteria derivation: define a set of reproducibility criteria based on theoretical grounding (e.g., checklist analysis) and empirical evidence (e.g., feature screening on annotated datasets);
- Rubric design: specify an ordinal scoring rubric for each criterion (capturing missing vs. partial vs. complete fulfillment) and a normalization rule to map criterion scores to $[0, 1]$.

Phase II: Framework application.

- Criterion scoring: assign ordinal criterion levels for a given research paper according to the rubric (e.g., via manual annotation or automated extraction);
- Normalization: map criterion scores to the $[0, 1]$ range;
- Weighting: specify a non-negative weight vector $\mathbf{w}$ for the criteria, with $\sum_{i=1}^{n} w_i = 1$;
- Aggregation: combine the weighted criterion scores into a single reproducibility score using an appropriate MCDM aggregation method (in this study, WSM and TOPSIS are used as representative methods).

Instantiation in this study. Criterion scores were obtained via dual annotation by the study authors and a large language model (ChatGPT (OpenAI), version GPT-5) and combined by

averaging (Section 3.6). For weighting, equal weights were used in the baseline evaluation and AHP-derived weights were used in the scenario analysis (Section 3.8).

The output is a single numerical score in $[0, 1]$, where higher values indicate stronger reproducibility readiness under the adopted R1-oriented definition, and the score can be computed without rerunning experiments.

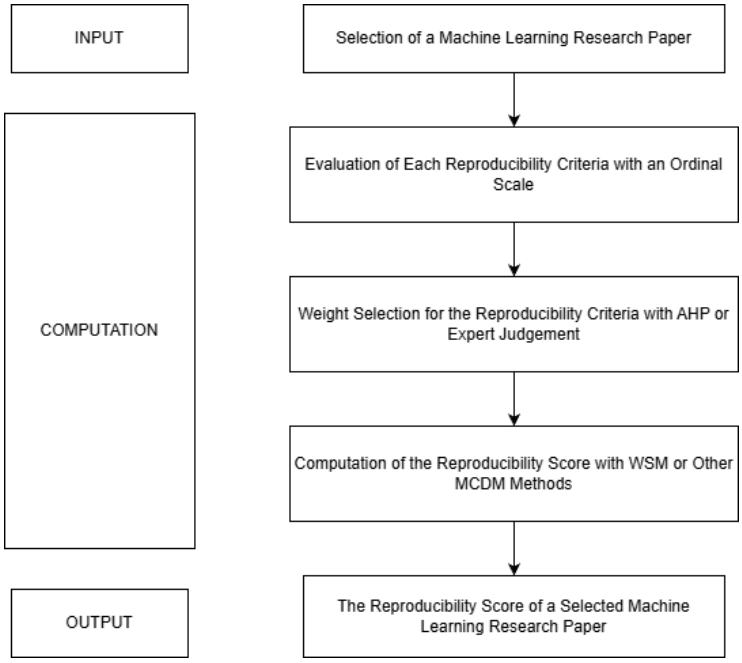

Figure 1. Proposed framework for computing a quantitative reproducibility score for an ML research paper using ordinal criteria and MCDM aggregation. Arrows indicate the sequential flow from paper selection and criterion scoring to weighting, aggregation, and the final reproducibility-readiness score.

3.2. Data Sources

This study uses two publicly available datasets from prior reproducibility studies. These datasets serve two roles within the proposed framework: supporting the empirical screening of candidate criteria and enabling evaluation of the computed reproducibility score against an existing paper-level reproducibility label. Constructing an original dataset would require reproducing a substantial number of papers and aggregating assessments across multiple independent annotators to reduce bias, which is costly in terms of expert time and computational resources; therefore, leveraging existing datasets provides a practical basis for both criteria selection and evaluation.

The first dataset originates from Olszewski et al. [37], who investigated reproducibility of ML papers in top-tier security venues (NDSS, IEEE S&P, CCS, and USENIX) between 2013 and 2022. The dataset contains 745 papers and includes a binary paper-level reproducibility label together with 19 annotated features capturing reproducibility aspects. These features cover, among others, code and experiment availability, data availability, documentation quality, model and hyperparameter specification, and train/test split descriptions. The dataset also enables analysis of reproducibility before and after the introduction of *Artifact Evaluation Committees* (AECs). Due to its open availability and alignment with paper-level assessment, this dataset was adopted as the primary basis for sampling and MCDM evaluation.

The second dataset is from Raff [38], who analyzed factors influencing reproducibility via statistical analysis. This dataset contains 255 papers (primarily NeurIPS and ICML) published between 1984 and 2017 and includes a binary reproducibility label together

with 28 manually annotated features. The features span artifact- and reporting-oriented indicators (e.g., code and data availability, hyperparameter specification, algorithm description quality) and structural/presentation variables (e.g., paper readability, length, and counts of figures or equations). These factors complement the first dataset by explicitly capturing paper structure and clarity, which have been associated with reproducibility prediction [32,34,35]. This dataset was therefore used to enrich the empirical screening stage.

3.3. Data Cleaning and Pre-Processing

Data cleaning and pre-processing were performed separately for each dataset. The datasets were not merged due to non-overlapping feature sets, as merging would introduce substantial missingness and distort feature-based analyses.

First dataset [37]**.** Cleaning focused on removing incomplete and non-informative entries. Since features were categorical or textual, missing values were not imputed to avoid introducing bias; rows with missing values were removed, reducing the dataset from 745 to 292 papers. Several variables that were not universally applicable or analytically meaningful (e.g., paper title, author information, publication venue identifiers, links) were excluded, together with redundant or purely textual fields. After filtering, eight reproducibility-related features remained. During pre-processing, categorical variables were converted to binary indicators to reflect whether a factor was present or absent. Additionally, the reproducibility label was restricted to papers labeled as either fully reproducible or irreproducible, yielding a final cleaned dataset of 215 papers.

Second dataset [38]**.** Only minimal missing-data handling was required: one missing value was corrected and one sparsely populated feature was removed. Venue- and topic-specific variables were excluded to retain only features that can be applied broadly across ML research, resulting in 20 retained features. Categorical variables were transformed to binary indicators. Numerical variables describing paper structure and complexity were retained; for logistic regression analyses, numerical features were scaled to $[0, 1]$ via min–max normalization to facilitate coefficient comparability, while for tree-based models these variables were used without scaling.

3.4. Criteria Derivation

The development of quantifiable reproducibility criteria followed a three-stage procedure combining theoretical grounding, empirical screening, and rubric design.

3.4.1. Stage 1: Theoretical Grounding

Candidate criteria were first derived from the reproducibility background in Section 2, with emphasis on checklist-based reporting practices (Section 2.4) and the dominant sources of irreproducibility summarized in Section 2.2. Checklists specify what information and artifacts authors are expected to provide to enable reproduction, making them a structured basis for paper-level criteria aligned with R1 reproducibility readiness. In this work, the theoretical grounding was used to ensure coverage of both major components of R1: the availability of runnable artifacts (e.g., data, code, trained models, and minimal usage instructions such as a `README`) and the clarity of methodological reporting required to correctly execute the workflow (e.g., training setup and hyperparameters). This stage, therefore, constrains the selection of criteria to factors that can be assessed from the paper and its associated artifacts without rerunning experiments.

3.4.2. Stage 2: Empirical Screening Using Supervised Models

An empirical analysis was performed on both datasets to identify factors that consistently relate to reproducibility labels. Logistic regression was used as an interpretable

baseline (results reported in Appendix A), followed by tree-based classifiers (Random Forest and XGBoost) to capture non-linear relationships and interactions between factors. Models were trained separately for each dataset using an 80/20 train–test split. Class imbalance was handled via balanced class weights for Random Forest and via the `scale_pos_weight` parameter for XGBoost.

Feature importance for XGBoost was computed using gain-based feature importance:

$$\mathrm{Imp}(j) = \sum_{\text{splits using } j} \Delta\mathcal{L}_{\text{split}}, \tag{1}$$

where $\Delta\mathcal{L}_{\text{split}}$ is the reduction in objective loss due to a split using feature j [39]. Importance scores were min–max normalized for comparability. Because the two datasets differ in coverage and feature definitions, feature-importance results were used as supportive empirical evidence rather than as a strict selection rule. The empirical results were interpreted together with the theoretical grounding from Stage 1 to obtain a compact criteria set suitable for paper-level scoring.

To interpret whether a factor contributes positively or negatively to reproducibility, we performed a feature effect analysis by toggling binary features between 0 and 1 while keeping all remaining features fixed for each paper [40]. For feature x_j, we compute the average change in predicted probability as:

$$\Delta_j = \frac{1}{N}\sum_{i=1}^{N}\left[P\left(y=1 \mid x_j^{(i)}=1, \mathbf{x}_{-j}^{(i)}\right) - P\left(y=1 \mid x_j^{(i)}=0, \mathbf{x}_{-j}^{(i)}\right)\right], \tag{2}$$

where N is the number of papers, $y = 1$ denotes the reproducible class, $x_j^{(i)}$ is the value of feature j for paper i, and $\mathbf{x}_{-j}^{(i)}$ denotes all other feature values for paper i.

3.4.3. Stage 3: Final Criteria Set and Descriptions

Stage 3 consolidates the candidate factors from Stage 1 and the influential variables identified in Stage 2 into a compact set of criteria for paper-level scoring. Factors were retained when they were relevant to R1 reproducibility readiness, assessable at the paper and artifact levels without rerunning experiments, and supported by either theoretical grounding or empirical screening. Most criteria are supported by the empirical screening results, while checklist-based evidence is used to ensure coverage of core artifact-sharing requirements such as code availability. The resulting seven criteria are summarized in Table 2, and the overall criteria-derivation workflow is illustrated in Figure 2.

Table 2. Reproducibility criteria used in this study.

Criterion	Description
Data availability	Availability of datasets used in the paper (public, partial, sensitive, or available on request).
Code availability	Availability of code required to run experiments (public, partial/incomplete, or available on request).
Inclusion of a README	Presence of usage documentation describing environment, dependencies, and steps to run experiments (partial vs. complete).
Trained model inclusion	Availability of trained model artifacts/weights enabling reuse without retraining.
Hyperparameter description	Reporting of key hyperparameters used for training (partial vs. complete).
Training description	Reporting of preprocessing, sampling, splits, and training procedure (partial vs. complete).
Paper readability	Clarity and organization of the methodological exposition (structure, consistency, and comprehensibility).

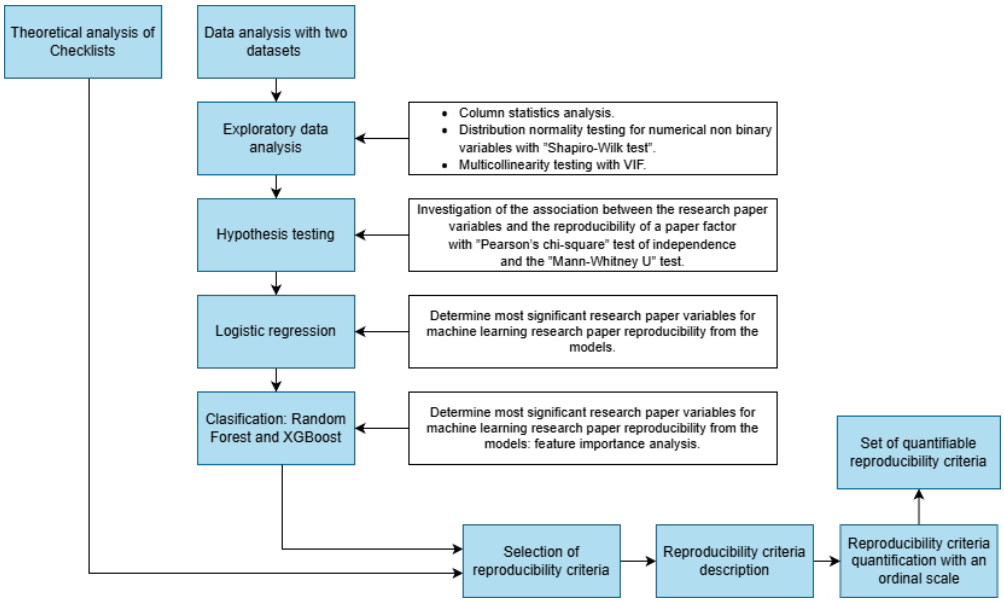

Figure 2. Workflow for creating a set of quantifiable reproducibility criteria for machine learning research papers. Arrows indicate the sequential flow from theoretical grounding and empirical screening to the final criteria set and rubric design.

While the resulting criteria set captures the most influential and consistently observable reproducibility factors, it is not exhaustive. Certain potentially relevant aspects—such as the availability of automated tests, consistency of software licenses, or the presence of structured dataset documentation (e.g., data cards)—were not included, as they were not consistently available across the analyzed datasets or could not be reliably assessed at the paper level without additional verification. Therefore, the framework prioritizes criteria that are both empirically supported and broadly applicable, while acknowledging that future extensions may incorporate additional dimensions of reproducibility.

Compared to traditional checklist-based approaches, the proposed framework enables quantitative aggregation of multiple criteria into a single interpretable score, allowing systematic comparison across studies. In addition, weighting schemes enable adaptation to different priorities and decision-making contexts, which is not supported by standard checklist methods.

3.5. Ordinal Rubric and Normalization

To support quantitative scoring while capturing partial compliance, each criterion was encoded using an ordinal scale rather than a binary indicator. The ordinal levels were defined to reflect practically meaningful distinctions between missing information (lowest level), partial fulfillment, and complete fulfillment (highest level), consistent with the checklist limitations discussed in Section 2.4. This enables the framework to differentiate between papers that provide some reproducibility-enabling information and those that omit it entirely.

Because different criteria have different numbers of ordinal levels (e.g., data availability uses four levels, whereas code availability uses three), criterion scores were normalized to a common $[0, 1]$ range before aggregation. For criterion k with maximum ordinal level L_k, the normalized score is computed as:

$$x_k = \frac{s_k}{L_k}, \qquad x_k \in [0, 1], \tag{3}$$

where $s_k \in [0, L_k]$ is the assigned ordinal level for criterion k after annotation aggregation and L_k is the maximum possible level for that criterion. For example, a paper with $s_k = 2$

for data availability and $L_k = 3$ yields $x_k = 2/3$. This linear mapping assumes monotonic utility and treats adjacent ordinal levels as equal steps, providing a simple, transparent normalization suitable for aggregation; more complex (non-linear) utility mappings can be explored in future work.

Subjective criteria, such as training description and paper readability, are evaluated using a structured rubric with predefined ordinal levels, as shown in Table 3. Although these assessments inherently involve some degree of subjectivity, the use of consistent evaluation guidelines and normalization helps improve comparability across papers.

Table 3 summarizes the rubric used for annotation and subsequent MCDM aggregation.

Table 3. Ordinal rubric for quantifying reproducibility criteria.

Criterion	Level	Description
Data availability	0	Dataset is publicly unavailable.
	1	Dataset is specified as sensitive.
	2	Dataset is partially available or available on request.
	3	Dataset is publicly available.
Code availability	0	Code is not publicly available.
	1	Code is partially available or incomplete (gaps prevent full rerun).
	2	Code is publicly available and sufficiently complete.
Inclusion of a README	0	No README/usage instructions.
	1	README present but incomplete (missing key steps/dependencies).
	2	README present and sufficiently complete to run the workflow.
Trained model inclusion	0	Trained model artifacts are not provided.
	1	Trained model artifacts are provided.
Hyperparameter description	0	Hyperparameters are not described.
	1	Hyperparameters are partially described.
	2	Hyperparameters are fully described.
Training description	0	Training/preprocessing/splits are not described.
	1	Training/preprocessing/splits are partially described.
	2	Training/preprocessing/splits are fully described.
Paper readability	0	Structure is unclear; key methodological steps are difficult to follow.
	1	Overall structure is logical but contains unclear or inconsistent parts.
	2	Well organized, clear structure, and consistent methodological exposition.

3.6. Data Sampling and Annotation

To evaluate the proposed framework, a subset of the cleaned Olszewski et al. dataset (215 papers; Section 3.3) was sampled and annotated using the rubric in Table 3.

A representative sample size was determined using Cochran's formula with finite-population correction [41]. Using a 95% confidence level, a conservative population proportion ($p = 0.5$), and a 5% margin of error, a sample size of 139 papers was obtained and randomly selected.

Each sampled paper was independently scored on all seven criteria by the study authors and by the ChatGPT model [42]. The final criterion score was computed as the mean of the two annotations:

$$S_{\text{final}} = \frac{A_{\text{study}} + A_{\text{ChatGPT}}}{2}, \tag{4}$$

where A_{study} is the author annotation and A_{ChatGPT} is the model-generated annotation.

To assess annotation reliability, inter-annotator agreement between human and ChatGPT-assisted annotations was evaluated using weighted Cohen's kappa [43]. The results indicated very high agreement overall (mean $\kappa = 0.96$), with somewhat lower agreement for more subjective criteria such as paper readability. These results support the internal consistency of the scoring process, although the annotation design remains limited by the use of a small number of annotators.

3.7. MCDM Score Computation

Multi-Criteria Decision Making aggregates multiple criteria into a single interpretable outcome [44,45]. In this work, the objective is a *single-paper* reproducibility score in $[0, 1]$, computed from normalized criterion values without rerunning experiments. Two representative and interpretable aggregation methods were applied: the Weighted Sum Model (WSM) and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS).

WSM. This aggregates criteria via a weighted average:

$$S = \sum_{i=1}^{n} w_i x_i, \qquad \sum_{i=1}^{n} w_i = 1, \tag{5}$$

where $x_i \in [0, 1]$ is the normalized score of criterion i (Equation (3)), $w_i \geq 0$ is the weight of criterion i, n is the number of criteria, and $S \in [0, 1]$ is the resulting reproducibility score.

It should be noted that WSM assumes additive independence among criteria, which may not fully reflect the logical structure of reproducibility. In practice, the absence of critical components, such as code or data, can render reproduction infeasible regardless of performance on other criteria. Therefore, the linear aggregation used in this study should be interpreted as an approximation that provides a transparent and interpretable score, rather than a strict representation of dependency relationships between reproducibility factors.

TOPSIS. This computes the relative closeness of an alternative to an ideal solution. After normalization, the ideal profile corresponds to all criteria at their best values, and the anti-ideal profile corresponds to all criteria at their worst values:

$$v_j^+ = 1, \qquad v_j^- = 0. \tag{6}$$

For a paper i and criterion j, let v_{ij} denote the normalized criterion value (here, $v_{ij} = x_j$ for the paper being scored). The distances to the ideal and anti-ideal are:

$$D_i^+ = \sqrt{\sum_{j=1}^{n} w_j (v_{ij} - v_j^+)^2}, \qquad D_i^- = \sqrt{\sum_{j=1}^{n} w_j (v_{ij} - v_j^-)^2}, \tag{7}$$

and the TOPSIS score (relative closeness to the ideal) is:

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-}, \tag{8}$$

where $C_i \in [0, 1]$ and higher values indicate closer proximity to the ideal profile.

3.8. Evaluation Protocol and Weighting Scenarios

This subsection describes the methodology used to evaluate the proposed quantitative reproducibility measure and to demonstrate its behavior under different criteria weighting strategies. The evaluation consists of two complementary parts: a quantitative comparison between the derived reproducibility score and an existing reproducibility label, and an

illustrative example analysis showing how alternative weighting objectives influence the resulting score. The corresponding workflows are shown in Figures 3 and 4.

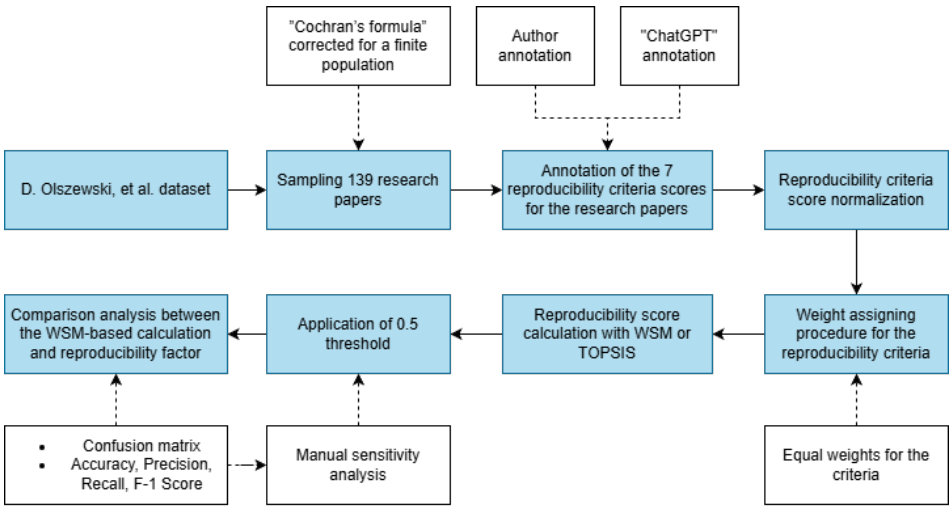

Figure 3. Workflow schema for the evaluation analysis. Arrows indicate the sequential flow from dataset sampling and annotation to score computation, thresholding, and comparison with the reference reproducibility label [37].

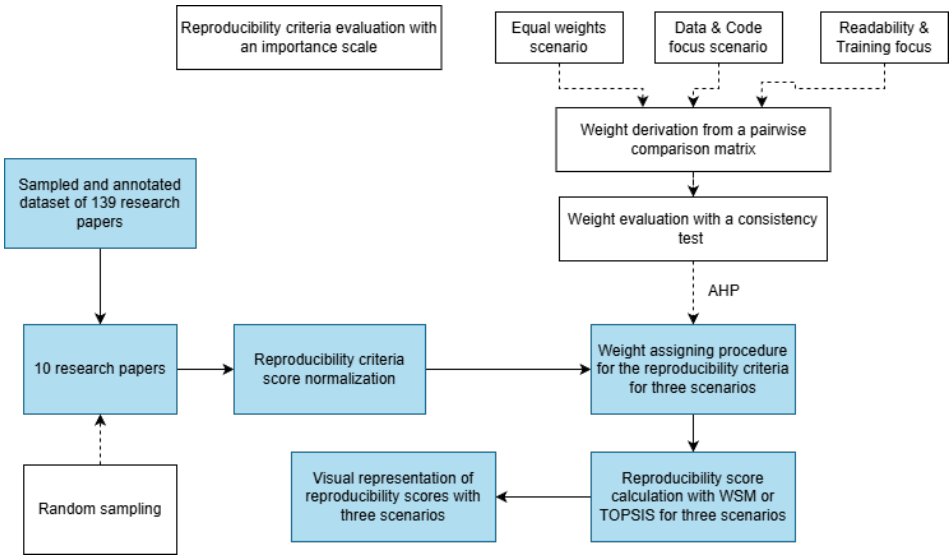

Figure 4. Workflow schema for the example analysis. Arrows indicate the sequential flow from paper selection and criteria normalization to AHP-based weight derivation, scenario-specific score computation, and visual comparison of reproducibility scores.

Baseline score evaluation (equal weights). For the baseline evaluation, the reproducibility scores derived using WSM or TOPSIS were compared with the binary reproducibility label provided in the dataset by Olszewski et al. [37]. The sampled papers were annotated using the seven selected reproducibility criteria and scored with equal weights to isolate the effects of the criteria and the rubric design, independent of subjective prioritization. The resulting reproducibility scores lie in $[0, 1]$ and were converted into binary predictions by thresholding. A manual threshold sensitivity analysis was performed by sweeping thresholds from 0 to 1 in increments of 0.1; a threshold of 0.5 provided the best overall balance of accuracy, precision, and recall in this study. Performance was assessed using a confusion matrix and standard classification metrics (accuracy, precision, recall, and F1 score).

Example analysis (AHP weighting scenarios). The example analysis illustrates how different weighting objectives affect reproducibility scores. From the annotated dataset, ten research papers were selected and evaluated under three weighting scenarios: equal weighting, emphasis on artifact availability (data and code), and emphasis on reporting quality (paper readability and training transparency). Scenario-specific weights were derived using the Analytic Hierarchy Process (AHP). Criteria importance was specified using Saaty's 1–9 scale [46] and pairwise comparison matrices were constructed accordingly. Weights were obtained through normalization of the comparison matrices, and judgment coherence was assessed using the Consistency Ratio (CR) [47]. The resulting weights were then applied within WSM and TOPSIS to compute scenario-specific reproducibility scores. This analysis demonstrates the sensitivity of the proposed measure to varying assumptions about criterion importance and highlights the framework's flexibility.

This procedure enables the estimation of score variability and ranking stability under plausible variations in decision-maker preferences, providing a quantitative assessment of the robustness of the proposed scoring framework.

Sensitivity and robustness analysis. To assess the robustness of the proposed framework with respect to uncertainty in criterion weights, a Monte Carlo-based sensitivity analysis [48] was conducted. For each AHP weighting scenario, the original weight vector was independently perturbed by up to $\pm 20\%$ for each criterion and subsequently renormalized to ensure that the weights sum to one. For each scenario, 500 perturbed weight vectors were generated, and reproducibility scores were recomputed using both WSM and TOPSIS. This procedure was used to evaluate score variability and ranking stability under plausible variations in criterion weights.

Together, the baseline evaluation and the example analysis provide complementary evidence: the former assesses alignment of the proposed score with an established reproducibility label, while the latter illustrates interpretability under alternative weighting objectives.

4. Results

This section reports the empirical results obtained by applying the proposed framework. We first summarize the empirical screening outcomes that support the selected criteria set, then evaluate the baseline equal-weight MCDM scores against the reference reproducibility label, and finally present an example analysis illustrating how alternative weighting objectives affect the resulting scores.

4.1. Empirical Screening Results Supporting Criteria Selection

Supervised models were used to screen reproducibility-related factors in the two datasets described in Section 3.2. Logistic regression served as an interpretable baseline (Tables A1 and A2), while Random Forest and XGBoost were used to capture potential non-linear relationships and interactions (Random Forest: Tables A3 and A4; XGBoost: Tables A5 and A8). Feature importance rankings used for criteria screening are reported for XGBoost in Tables A6 and A9, and directional feature effects are summarized in Tables A7 and A10.

Across both datasets, the most influential factors corresponded primarily to artifact availability and methodological reporting. These empirical results support the final criteria set reported in Table 2. Overall, the screening results align with the R1-oriented view that reproducibility readiness depends on accessible artifacts and sufficient methodological transparency.

4.2. Baseline Evaluation Against the Reference Reproducibility Label

The baseline evaluation compares the framework-derived reproducibility scores against the binary reproducibility label provided by Olszewski et al. [37] (0 = non-reproducible, 1 = reproducible). A randomly sampled subset of 139 papers was annotated using the ordinal rubric (Table 3). Reproducibility scores were computed using WSM and TOPSIS under equal weights and converted into binary predictions using the threshold of 0.5 selected in the threshold sweep described in Section 3.8. In the sampled set, the reference label distribution is imbalanced (102 negatives and 37 positives).

Table 4 reports the confusion matrix, and Table 5 reports the corresponding classification metrics. Under equal weighting, WSM and TOPSIS produced identical classification outcomes for this dataset.

Table 4. Confusion matrix comparing WSM- and TOPSIS-based predictions with the reference reproducibility label from Olszewski et al. [37] (0 = non-reproducible, 1 = reproducible; identical results under equal weighting).

Reference Label	Predicted Label	
	0	**1**
0	63	39
1	11	26

Table 5. Baseline classification metrics for WSM and TOPSIS (identical results under equal weighting).

Metric	Value
Accuracy	0.64
Precision	0.40
Recall	0.70
F1-score	0.51

The baseline results indicate moderate agreement with the reference label (accuracy 0.64). The relatively high recall (0.70) suggests that the framework identifies a substantial portion of papers labeled reproducible, whereas the lower precision (0.40) indicates false positives relative to the reference label. This behavior is consistent with a readiness-oriented assessment: a paper may satisfy several disclosure-based criteria yet still fail reproduction in practice due to unobserved execution barriers. The comparison with simple baseline models further suggests that aggregating multiple criteria provides a more balanced approximation of reproducibility readiness than relying on individual artifact-based heuristics alone.

To contextualize these findings, we introduce a set of simple baseline models based on individual reproducibility criteria. Specifically, four scenarios are considered: (i) code availability only, (ii) data availability only, (iii) code OR data availability, and (iv) code AND data availability. Each scenario is converted into a binary prediction using the same thresholding procedure applied to the framework-based scores. These baselines represent minimal heuristics that rely solely on the presence of key artifacts, without accounting for their quality or completeness.

The comparison with single-criterion baselines in Table 6 highlights the limitations of simple artifact-based heuristics. While the Code OR Data condition achieves the highest recall (0.86), it does so at the expense of precision, indicating a high number of false positives. In contrast, the proposed framework (WSM/TOPSIS) achieves a more balanced performance, with higher precision (0.40) and the best overall F1-score (0.51). This suggests that aggregating multiple criteria provides a more reliable approximation of reproducibility than relying on the presence of individual artifacts alone.

Table 6. Comparison of framework-based scores (WSM/TOPSIS) with single-criterion baseline models.

Model	Precision	Recall	F1-Score
Code only	0.27	0.62	0.38
Data only	0.33	0.73	0.45
Code OR Data	0.30	**0.86**	0.45
Code AND Data	0.30	0.49	0.37
WSM/TOPSIS (equal weights)	**0.40**	0.70	**0.51**

Note: Bold values indicate the best value in each metric column.

It is important to interpret these results in light of the framework's scope: it evaluates the presence and quality of artifacts and reporting rather than the successful execution of experiments. As a result, practical reproduction may still fail due to factors not captured by the criteria, such as incomplete implementations, hidden dependencies, or environment-specific issues. Consequently, the observed discrepancies are consistent with the intended role of the framework as a measure of reproducibility readiness. Since criterion weights can be specified to reflect different evaluation objectives, the following subsections examine both the effect of alternative weighting schemes and the robustness of the resulting scores under moderate weight perturbations.

4.3. Scenario Analysis Under Alternative Weighting Objectives

To illustrate how different weighting objectives affect reproducibility scores, a scenario analysis was conducted on ten papers selected from the annotated subset. Three weighting scenarios were considered: **Scenario 1** (Equal weights), **Scenario 2** (Data & Code focus), emphasizing artifact availability, and **Scenario 3** (Training & Readability focus), emphasizing reporting quality. For illustrative purposes, scenario-specific weights were constructed using AHP.

Table 7 reports the AHP-derived weights using the same canonical criterion order as in Section 3.4.3 and Table 2. Table 8 reports the corresponding consistency checks. Reproducibility scores computed under each scenario using WSM and TOPSIS are shown in Table 9.

Table 7. Scenario-based weights for the reproducibility criteria (AHP-derived), reported in canonical criterion order.

Criterion	Scenario 1 (Equal Weights)	Scenario 2 (Data & Code Focus)	Scenario 3 (Training & Readability Focus)
Data availability	0.143	0.241	0.107
Code availability	0.143	0.241	0.107
Inclusion of a README	0.143	0.103	0.107
Trained model inclusion	0.143	0.103	0.036
Hyperparameter description	0.143	0.103	0.143
Training description	0.143	0.103	0.250
Paper readability	0.143	0.103	0.250
Sum	1.000	1.000	1.000

Table 8. Consistency ratio (CR) results for the weighting scenarios.

Scenario	λ_{max}	CR
Scenario 1 (Equal weights)	7.0	0.0
Scenario 2 (Data & Code focus)	7.0	0.0
Scenario 3 (Training & Readability focus)	7.0	0.0

Table 9. Reproducibility scores for selected papers under three weighting scenarios using WSM and TOPSIS.

Paper No.	Paper Title	WSM			TOPSIS		
		S1	**S2**	**S3**	**S1**	**S2**	**S3**
1	WTAGRAPH: Web Tracking and Advertising Detection using Graph Neural Networks	0.286	0.207	0.446	0.277	0.210	0.490
2	DeepSteal: Advanced Model Extractions Leveraging Efficient Weight Stealing in Memories	0.429	0.448	0.500	0.370	0.442	0.484
3	Model Stealing Attacks Against Inductive Graph Neural Networks	0.452	0.420	0.589	0.375	0.385	0.570
4	Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages	0.643	0.603	0.643	0.649	0.575	0.627
5	Cheetah: Lean and Fast Secure Two-Party Deep Neural Network Inference	0.524	0.609	0.554	0.424	0.555	0.501
6	DEEPCASE: Semi-Supervised Contextual Analysis of Security Events	0.643	0.741	0.714	0.478	0.625	0.658
7	Khaleesi: Breaker of Advertising and Tracking Request Chains	0.595	0.592	**0.750**	0.447	0.498	**0.712**
8	AI/ML for Network Security: The Emperor Has No Clothes	**0.786**	**0.845**	0.679	**0.782**	**0.837**	0.622
9	Improving Password Guessing via Representation Learning	0.524	0.471	0.714	0.402	0.401	0.638
10	Privacy Risks of Securing Machine Learning Models against Adversarial Examples	0.595	0.661	0.696	0.431	0.549	0.611

Note: bold values denote the highest reproducibility score within each scenario.

To aid interpretation, the weighting scenarios can be understood intuitively as reflecting different evaluation priorities. For example, in Scenario 2, higher weights assigned to artifact-related criteria (e.g., data and code availability) result in higher scores for papers with strong artifact disclosure. In contrast, Scenario 3 emphasizes reporting quality (e.g., training description and readability), favoring papers with more detailed and clear methodological descriptions. These differences illustrate how the proposed framework adapts to varying decision-making objectives.

The AHP-derived weights used in this study are illustrative and were constructed to demonstrate how different evaluation priorities influence the resulting scores. They do not represent weights elicited from real stakeholders.

The scenario analysis shows that weight selection directly affects the resulting reproducibility scores and, in several cases, the relative ordering of papers. Under Scenario 2, papers that provide stronger artifact availability tend to receive higher scores. For example, Papers 6, 8, and 10 increase under WSM when moving from Scenario 1 to Scenario 2, indicating that emphasizing data and code availability strengthens their assessment. Conversely, papers that are comparatively weaker in terms of artifact availability do not benefit from this weighting objective (e.g., Paper 1 decreases under WSM from 0.286 to 0.207).

Under Scenario 3, papers with clearer methodological exposition and more detailed training descriptions tend to receive higher scores. This effect is visible for Papers 1, 3, 6, 7, and 9, which increase under WSM relative to Scenario 1, consistent with the increased weights assigned to training description and paper readability in Table 7.

Differences between WSM and TOPSIS are observed across scenarios. Although both methods use the same criterion weights, WSM aggregates criteria using a linear additive model, whereas TOPSIS evaluates a paper based on its relative closeness to an ideal profile. Consequently, TOPSIS may yield different score scaling and, in some cases, different sensitivity to strong or weak performance on individual criteria. Overall, these results confirm that weighting is a critical component of MCDM-based reproducibility assess-

ment. By allowing criterion importance to be adjusted according to different evaluation objectives and the preferences of decision makers (e.g., emphasizing artifact availability versus reporting quality), the proposed framework supports a flexible and context-sensitive assessment of reproducibility readiness. This adaptability increases the practical applicability of the quantitative reproducibility measure across diverse research contexts and assessment settings.

4.4. Robustness Analysis Under Weight Perturbations

Table 10 summarizes the results of the Monte Carlo-based sensitivity analysis described in Section 3.8, in which the AHP-derived weights were moderately perturbed and the resulting WSM and TOPSIS scores were recomputed. Mean reproducibility scores and 5th–95th percentile intervals are reported across all weighting scenarios and aggregation methods.

Table 10. Sensitivity analysis of reproducibility scores across weighting scenarios and aggregation methods. Values are reported as mean scores with 5th–95th percentile intervals obtained from 500 Monte Carlo perturbations of the criterion weights within ±20% of their original values.

Paper	WSM			TOPSIS		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**
1	0.29 [0.26–0.31]	0.21 [0.18–0.23]	0.45 [0.42–0.48]	0.19 [0.16–0.22]	0.14 [0.12–0.16]	0.38 [0.34–0.41]
2	0.43 [0.40–0.46]	0.45 [0.41–0.49]	0.50 [0.48–0.52]	0.30 [0.26–0.34]	0.39 [0.34–0.44]	0.37 [0.34–0.40]
3	0.45 [0.42–0.48]	0.42 [0.39–0.45]	0.59 [0.56–0.62]	0.31 [0.28–0.35]	0.33 [0.29–0.37]	0.49 [0.44–0.53]
4	0.64 [0.62–0.67]	0.60 [0.57–0.64]	0.64 [0.62–0.67]	0.59 [0.55–0.62]	0.51 [0.46–0.56]	0.56 [0.52–0.60]
5	0.53 [0.50–0.55]	0.61 [0.58–0.64]	0.55 [0.52–0.59]	0.43 [0.39–0.47]	0.59 [0.54–0.63]	0.44 [0.40–0.48]
6	0.64 [0.62–0.67]	0.74 [0.72–0.77]	0.71 [0.69–0.73]	0.49 [0.46–0.53]	0.65 [0.60–0.69]	0.62 [0.58–0.66]
7	0.60 [0.57–0.62]	0.59 [0.58–0.61]	**0.75** [0.73–0.77]	0.44 [0.40–0.47]	0.49 [0.47–0.52]	**0.66** [0.63–0.70]
8	**0.79** [0.77–0.80]	**0.84** [0.83–0.86]	0.68 [0.66–0.70]	**0.77** [0.74–0.80]	**0.83** [0.81–0.85]	0.60 [0.56–0.63]
9	0.52 [0.49–0.56]	0.47 [0.44–0.51]	0.71 [0.69–0.74]	0.33 [0.30–0.37]	0.34 [0.30–0.38]	0.55 [0.51–0.60]
10	0.60 [0.57–0.62]	0.66 [0.64–0.69]	0.70 [0.67–0.72]	0.41 [0.37–0.45]	0.56 [0.52–0.61]	0.55 [0.50–0.59]
Avg. variation (Δ)	**0.09**	**0.09**	**0.08**	**0.11**	**0.12**	**0.13**

Note: bold values denote the highest reproducibility score within each scenario.

The results indicate that the framework remains robust under moderate perturbations of the criterion weights. WSM scores exhibit relatively low variability across all scenarios, with average score variations of 0.091, 0.089, and 0.083 for Scenarios 1–3, respectively. TOPSIS scores show slightly higher variability (0.114, 0.116, and 0.127), reflecting the method's greater sensitivity to distance-based normalization, but the observed variation remains limited relative to the $[0, 1]$ score scale.

Importantly, the relative ranking of papers remains largely consistent across perturbations. Papers that achieve the highest scores under the original weighting schemes remain among the top-performing alternatives across most simulations. When ranking changes do occur, they are primarily associated with differences in weighting objectives between scenarios rather than with random perturbations within a given scenario.

These findings suggest that the proposed MCDM-based reproducibility score is not overly sensitive to moderate variations in criterion weights and should not be interpreted as an arbitrarily tunable ranking mechanism. Instead, the results are primarily driven by the underlying criterion evaluations, while the weights provide controlled and interpretable adjustments reflecting different decision-making priorities.

Taken together, the results show that the proposed framework provides a feasible quantitative assessment of reproducibility readiness. The baseline evaluation demonstrates moderate agreement with the reference reproducibility label, the scenario analysis shows that the framework adapts to different weighting objectives, and the robustness analysis indicates that the resulting scores remain stable under moderate perturbations of the criterion weights.

5. Discussion

This section discusses the implications of the proposed MCDM-based framework, interprets the empirical results in light of the adopted R1 reproducibility readiness notion, and outlines the main limitations and directions for future work.

5.1. Interpretation of the Baseline Evaluation

The baseline evaluation (Section 4.2) shows moderate agreement between the framework-derived reproducibility scores and the reference label from Olszewski et al. [37] (accuracy 0.64, precision 0.40, recall 0.70). A key observation is the asymmetry between recall and precision. The relatively high recall indicates that the proposed criteria and rubric capture many papers that are labeled reproducible in the reference dataset. In contrast, the lower precision suggests that a non-trivial fraction of papers that appear "ready" under disclosure-based criteria are labeled non-reproducible in practice.

This behavior is consistent with the scope of this work. The score is designed to quantify *readiness* for R1 reproduction based on artifacts and reporting, without rerunning experiments. Therefore, false positives relative to the reference label may arise when barriers that are not directly observable from the paper and linked artifacts prevent successful reproduction (e.g., hidden environment incompatibilities, missing dependencies, non-deterministic execution, incomplete scripts despite nominal availability, or data access that is claimed but not practically usable). Conversely, false negatives can occur when reproduction is possible even though the paper omits some rubric elements or provides them in non-standard forms.

Overall, the baseline results support the feasibility of a criteria-grounded quantitative score while highlighting that a readiness score and an observed reproduction outcome are related but not equivalent constructs. It is also important to note that the evaluation relies on binary reproducibility labels, which may oversimplify reproducibility outcomes and fail to capture partial or conditional reproducibility. As a result, discrepancies between the proposed score and the reference labels may arise not only from limitations of the proposed approach, but also from the coarse granularity of the ground-truth data. Therefore, the evaluation results should be interpreted in light of these constraints.

5.2. Role of Weighting and Decision-Maker Objectives

The scenario analysis (Section 4.3) demonstrates that criterion weighting is not merely a technical detail but a substantive modeling choice. When decision makers prioritize artifact availability (Scenario 2), papers with stronger data and code disclosure increase in score, whereas papers with weaker artifact disclosure decrease. When decision makers prioritize reporting quality (Scenario 3), papers with clearer training descriptions and overall exposition benefit.

This behavior is a desirable property of the framework: different communities and stakeholders legitimately adopt different evaluation objectives. For example, an artifact evaluation committee may emphasize runnable artifacts, while reviewers in domains with sensitive data may place more emphasis on methodological transparency and clarity. By making these objectives explicit through weights (e.g., equal weighting for a neutral baseline

and AHP-derived weights for scenario-based preferences), the framework provides an interpretable mechanism for adapting the score to context while maintaining a consistent rubric and normalization scheme. Because weights encode decision-maker preferences, scenario-based conclusions are conditional on the adopted weighting scheme.

In practical deployments, simple sensitivity checks (e.g., small perturbations of the AHP pairwise judgments or reporting weight intervals) can help assess how stable scores and rankings are under plausible preference variation. The sensitivity analysis further supports that the proposed framework is not overly sensitive to the choice of weights. Score variability remains limited under moderate perturbations and ranking structures are largely preserved, indicating that the framework does not allow arbitrary manipulation of results through small changes in criterion weights.

Differences between WSM and TOPSIS observed in the scenario analysis further emphasize that the aggregation rule influences score behavior. WSM corresponds to a linear additive utility model, whereas TOPSIS is based on relative closeness to an ideal profile. In settings where decision makers care about balancing weaknesses across criteria versus rewarding strong performance on a subset of criteria, the two aggregation families may behave differently. In practice, this suggests reporting both scores, or selecting one aggregation method depending on the intended decision context.

5.3. Validity of the Criteria Set and Connection to Prior Work

The final criteria set (Table 2) aligns with prior checklist-based recommendations and empirical findings that artifact availability and reporting quality are primary drivers of reproducibility [10,13,30]. The empirical screening results (Section 4.1) provide complementary support: influential variables in the analyzed datasets are largely consistent with artifact-centric factors (data, code, README, trained model) and reporting-centric factors (hyperparameters, training description, readability).

At the same time, the criteria set is intentionally compact to remain practical for research paper-level assessment. This compactness imposes a trade-off: some known sources of irreproducibility (e.g., randomness control, evaluation leakage, hardware/software determinism) are not explicitly represented as standalone criteria. In the present framework, such aspects are partially captured indirectly via training description and documentation quality, but not fully. Extending the criteria set while retaining usability is therefore an important direction for future work.

It is also important to note that reproducibility readiness, as defined in this framework, is not necessarily aligned with the scientific impact or novelty of a study. In some cases, highly influential research may not fully disclose all implementation details due to practical, commercial, or strategic considerations. The proposed framework does not aim to evaluate the scientific merit of such work, but rather to assess the extent to which it supports reproducibility. Therefore, a lower reproducibility readiness score should not be interpreted as an indicator of lower research quality, but rather as a reflection of limited transparency or accessibility of artifacts.

5.4. Limitations

Several limitations should be considered when interpreting the results.

Dataset scope and generalizability. The evaluation relies on two existing datasets with specific coverage (security venues for Olszewski et al. [37] and primarily NeurIPS/ICML for Raff [38]). These corpora may not fully represent other subfields of ML or applied domains with different disclosure norms. In addition, the datasets provide binary labels that collapse multiple degrees of reproduction difficulty into a single outcome.

Label–score construct mismatch. The proposed score measures disclosure-based R1 readiness, whereas the reference label reflects observed reproducibility outcomes. Disagreement is expected when practical reproduction failures arise from factors not visible in the paper/artifacts. This mismatch limits the maximum achievable agreement without expanding the framework to include execution-verified criteria.

Manual and LLM-assisted annotation. Criterion scoring was obtained via dual annotation (authors and ChatGPT) and averaged (Equation (4)). While inter-annotator agreement was assessed and found to be high, the annotation process still relied on a limited number of annotators and did not include a multi-human consensus or adjudication procedure. This may introduce bias, particularly for more subjective criteria such as paper readability or partial completeness of documentation.

Rubric and normalization assumptions. The ordinal rubric and linear normalization (Equation (3)) assume monotonic utility and treat adjacent ordinal levels as equally spaced once mapped to $[0, 1]$. This provides transparency and simplicity, but it may not reflect how practitioners value transitions between levels (e.g., moving from "partial" to "complete" documentation may yield a larger practical benefit than moving from "missing" to "partial" for some criteria). Using a more granular rubric (i.e., more ordinal levels) can partially reduce discretization effects, but it does not resolve the possibility of non-uniform utility gaps between levels. Non-linear utility mappings and sensitivity analyses could better characterize and mitigate this effect.

Scope of the Framework. The proposed framework should be interpreted as a supplementary tool for assessing reproducibility readiness rather than a definitive measure of reproducibility or research quality. While it provides a structured and interpretable approximation based on observable artifacts and reporting practices, it cannot fully capture all factors influencing successful reproduction. Therefore, it is best suited as an initial screening or comparative analysis tool, complementing more rigorous reproduction-based evaluations.

5.5. Future Work

Future research can extend the framework along several directions.

Broader and execution-grounded evaluation. Constructing new datasets that include execution outcomes (e.g., containerized reruns, verified dependency resolution, and artifact completeness checks) would allow direct validation of the readiness score against practical reproduction barriers.

Richer criteria and multi-level reproducibility. The framework can be extended with additional criteria targeting known sources of irreproducibility (e.g., randomness control, environment specification, or evaluation leakage checks), or adapted to use reduced/domain-specific criteria subsets depending on the assessment objective. It can also be generalized to support multi-level reproducibility notions (R2/R3) by revising the criteria set and rubric for those settings. This direction directly addresses the current limitation of focusing on R1 reproducibility and extends the applicability of the framework toward more comprehensive reproducibility assessment.

Improved annotation and automation. Scaling the approach benefits from partially automated extraction of artifact links, documentation completeness, and structured reporting elements. Human-in-the-loop protocols and inter-annotator agreement analysis would improve scoring reliability, especially for subjective criteria.

Aggregation and uncertainty reporting. In this study, WSM and TOPSIS were used as well-known, representative, and interpretable MCDM aggregation methods. Future work could compare their behavior with alternative aggregation operators and incorporate explicit uncertainty quantification (e.g., score intervals induced by annotation variance and weight uncertainty) to strengthen interpretability and robustness for decision making.

Future work may explore non-linear aggregation strategies or constraint-based models (e.g., conjunctive approaches) to better capture dependencies between criteria.

Alternative quantitative modeling approaches. Future research could also investigate alternative approaches for deriving quantitative reproducibility measures, including statistical models and machine learning-based methods, and compare their performance and interpretability with the proposed MCDM-based framework. In addition, alternative MCDM aggregation techniques, such as LINMAP [49], as well as other families of methods including outranking approaches (e.g., PROMETHEE [50]) and compromise-based methods (e.g., VIKOR [51]), could be explored and compared with WSM and TOPSIS to better understand how different aggregation principles influence score behavior, interpretability, and sensitivity to weights. Moreover, future research could explore the integration of the proposed framework with existing reproducibility and research-evaluation methodologies. For example, criteria-based scoring could be combined with non-compensatory screening mechanisms to enforce minimum requirements on key criteria [52]. Reproducibility-readiness scores could also be examined alongside broader article-level indicators of research uptake and attention [53]. Similarly, alignment with established open-science and reproducibility infrastructures, such as the Open Science Framework (OSF) [54] or ACM artifact review and badging standards [55], could further enhance the framework's practical applicability.

In summary, the proposed framework provides a practical and interpretable approach for quantifying disclosure-based R1 reproducibility readiness at the research paper level. The results indicate that the approach is feasible and that weighting meaningfully adapts the score to different evaluation objectives, while the identified limitations motivate further methodological and empirical development.

6. Conclusions

This study addressed the lack of practical quantitative approaches for assessing reproducibility in machine learning research papers. To this end, it proposed an MCDM-based framework for quantifying disclosure-based R1 reproducibility readiness at the paper level. The framework combines theoretically and empirically derived reproducibility criteria with an ordinal rubric, normalization to the $[0, 1]$ range, and MCDM-based aggregation into a single interpretable score that can be computed without rerunning experiments.

A set of seven reproducibility criteria was established by combining theoretical grounding from prior reproducibility literature and checklist-based practices with empirical screening on two existing reproducibility datasets. The resulting criteria capture two central components of R1 reproducibility readiness: availability of runnable artifacts and adequacy of methodological reporting. To operationalize these criteria, an ordinal rubric was introduced in order to capture partial fulfillment more effectively than binary checklist-style assessment, and the normalized criterion values were aggregated using the WSM and the TOPSIS as representative and interpretable MCDM methods.

The empirical evaluation showed that the proposed framework is feasible in practice. In the baseline equal-weight setting, the derived reproducibility scores achieved moderate agreement with the reference reproducibility label, with an accuracy of 0.64, a precision of 0.40, and a recall of 0.70. These results indicate that the framework captures a meaningful part of paper-level reproducibility readiness, while also reflecting the expected difference between disclosure-based readiness and observed reproduction outcomes. The example analysis further showed that weighting has a direct effect on resulting scores and can adapt the assessment to different evaluation objectives, such as emphasizing artifact availability or reporting quality. A sensitivity analysis further indicates that the proposed scoring framework is robust to moderate variations in criterion weights, with stable score distributions and largely consistent ranking structures across scenarios.

Overall, the study demonstrates that MCDM provides a suitable foundation for quantitative reproducibility assessment in machine learning research papers. The proposed framework offers a structured, transparent, and adaptable way to convert multiple reproducibility-related criteria into a single score while preserving interpretability for different decision contexts. At the same time, the results highlight that reproducibility readiness and actual reproducibility are related but distinct concepts. Although further refinement and broader validation are needed, the results show that this direction is promising for supporting more systematic and context-aware assessment of reproducibility in machine learning research.

Author Contributions: Conceptualization, P.L., R.P. and E.F.; Methodology, P.L. and E.F.; Software, P.L.; Validation, P.L., R.P. and E.F.; Formal analysis, P.L., R.P. and E.F.; Investigation, P.L., R.P. and E.F.; Data curation, P.L.; Writing—original draft, P.L. and E.F.; Writing—review & editing, R.P. and E.F.; Visualization, P.L.; Supervision, E.F.; Project administration, R.P.; Funding acquisition, R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Research Council of Lithuania under the Program "University Excellence Initiatives" of the Ministry of Education, Science and Sports of the Republic of Lithuania (Measure No. 12-001-01-01-01 "Improving the Research and Study Environment"), Project No. S-A-UEI-23-11.

Data Availability Statement: The complete codebase, datasets, and a README file, detailing each data and code file, are available at: https://github.com/l3paulina/Quantifying-reproducibility-of-ML-papers-using-MCDM.git (accessed on 22 April 2026). The methods and metrics used in this study were implemented using `Python 3.13.3`.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ACL	Annual Meeting of the Association for Computational Linguistics
AEC	Artifact Evaluation Committee
AHP	Analytic Hierarchy Process
CCS	ACM Conference on Computer and Communications Security
CR	Consistency Ratio
EMNLP	Empirical Methods in Natural Language Processing
GPU	Graphics Processing Unit
ICML	International Conference on Machine Learning
IEEE S&P	IEEE Symposium on Security and Privacy
LLM	Large Language Model
MCDM	Multi-Criteria Decision Making
MICCAI	International Conference on Medical Image Computing and Computer-Assisted Intervention
ML	Machine Learning
NAACL	North American Chapter of the Association for Computational Linguistics
NDSS	Network and Distributed System Security Symposium
NeurIPS	Conference on Neural Information Processing Systems
NLP	Natural Language Processing
QRA	Quantified Reproducibility Assessment

R1	Reproducibility Level 1
R2	Reproducibility Level 2
R3	Reproducibility Level 3
README	README file
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
USENIX	USENIX Security Symposium
WSM	Weighted Sum Model

Appendix A. Supplementary Computational Results

This appendix reports the supplementary computational results referenced in Section 4.1. These results support the empirical screening stage used to identify candidate reproducibility criteria. Logistic regression was used as an interpretable baseline, while Random Forest and XGBoost were used to capture non-linear relationships and interactions.

Appendix A.1. Logistic Regression Results

Tables A1 and A2 report the logistic regression results for the two datasets described in Section 3.2. In both cases, the models converged successfully and are included primarily for interpretability rather than for predictive performance.

For the dataset of Olszewski et al. [37], the pseudo R^2 value is low (0.054), indicating limited explanatory power. Among the retained variables, trained model inclusion and data availability show positive and statistically significant associations with the reproducible class. For the Raff dataset, the model fit is stronger (pseudo $R^2 = 0.33$), although still limited for classification purposes. In this dataset, paper readability and hyperparameter specification are positively associated with reproducibility, whereas the difficulty of the used algorithm is negatively associated with the reproducibility class.

Table A1. Logistic regression results for the dataset of Olszewski et al. [37].

Variable	Coef.	Std. Error	z	P > |z|	95% Low	95% High
Constant	−2.196	0.492	−4.458	0.000	−3.161	−1.230
Trained model inclusion	0.882	0.330	2.670	0.008	0.235	1.529
Data availability	1.106	0.510	2.167	0.030	0.106	2.105
Pseudo R-squared	0.054					
Converged	True					

Table A2. Logistic regression results for the dataset of Raff [38].

Variable	Coef.	Std. Error	z	P > |z|	95% Low	95% High
Constant	−1.611	0.489	−3.294	0.001	−2.570	−0.653
Paper readability	2.764	0.367	7.531	0.000	2.045	3.484
Used algorithm diff.	−0.963	0.360	−2.676	0.007	−1.669	−0.258
Hyperparameter spec.	1.183	0.402	2.942	0.003	0.395	1.972
Pseudo R-squared	0.33					
Converged	True					

Appendix A.2. Classification Results

Tables A3–A8 summarize the classification performance of the Random Forest and XGBoost models used during empirical screening. These models were applied separately to the two datasets in order to identify variables that are consistently associated with reproducibility labels.

The Random Forest model achieved moderate performance on the Olszewski et al. dataset (Table A3) and stronger performance on the Raff dataset (Table A4). The XGBoost model slightly improved performance on the Olszewski et al. dataset (Table A5) and also performed strongly on the Raff dataset (Table A8). Because XGBoost provided the most informative and stable feature-importance outputs for the purposes of criteria screening, its importance and feature-effect results are reported in Tables A6, A7, A9 and A10.

For the dataset of Olszewski et al. [37], the most influential variables according to normalized XGBoost importance were inclusion of a `README`, data availability, trained model inclusion, training description, and hyperparameter specification. For the Raff dataset, paper readability was the most influential variable by a substantial margin. In the empirical screening stage described in Section 3.4, normalized importance values were used as supportive empirical evidence for criteria selection rather than as a strict cutoff-based rule.

Table A3. Random Forest results for the dataset of Olszewski et al. [37].

Class	Precision	Recall	F1-Score	Support
0	0.79	0.71	0.75	31
1	0.40	0.50	0.44	12
Accuracy		0.65		43
Macro Avg	0.59	0.60	0.60	43
Weighted Avg	0.68	0.65	0.66	43

Table A4. Random Forest results for the dataset of Raff [38].

Class	Precision	Recall	F1-Score	Support
0	0.83	0.79	0.81	19
1	0.88	0.91	0.89	32
Accuracy		0.86		51
Macro Avg	0.86	0.85	0.85	51
Weighted Avg	0.86	0.86	0.86	51

Table A5. XGBoost results for the dataset of Olszewski et al. [37].

Class	Precision	Recall	F1-Score	Support
0	0.80	0.77	0.79	31
1	0.46	0.50	0.48	12
Accuracy		0.70		43
Macro Avg	0.63	0.64	0.63	43
Weighted Avg	0.71	0.70	0.70	43

Table A6. Feature importance results for the dataset of Olszewski et al. [37].

Feature	Importance	Normalized Importance
Inclusion of a README	0.242	0.279
Data availability	0.233	0.269
Trained model inclusion	0.121	0.140
Training description	0.100	0.116
Hyperparameter specification	0.090	0.103
Train/test split specification	0.078	0.090
Experiment availability	0.000	0.000

Table A7. Feature-effect analysis for the dataset of Olszewski et al. [37].

Feature	Class 0	Class 1
Inclusion of a README	−0.300	0.300
Trained model inclusion	−0.190	0.190
Data availability	−0.108	0.108
Training description	0.004	−0.004
Hyperparameter specification	0.050	−0.050

Table A8. XGBoost results for the dataset of Raff [38].

Class	Precision	Recall	F1-Score	Support
0	0.88	0.79	0.83	19
1	0.88	0.94	0.91	32
Accuracy		0.88		51
Macro Avg	0.88	0.86	0.87	51
Weighted Avg	0.88	0.88	0.88	51

Table A9. Feature importance results for the dataset of Raff [38].

Feature	Importance	Normalized Importance
Paper readability	5.185	0.288
Data availability	1.128	0.062
Number of conceptualization figures	1.018	0.056
Hyperparameter specification	0.917	0.051
Number of tables and figures	0.866	0.048
Example of the problem given	0.838	0.046
Code availability	0.788	0.043
Number of authors	0.788	0.043
Number of other figures	0.776	0.043
Number of proofs	0.762	0.042
Inclusion of appendix	0.697	0.038
Used algorithm difficulty	0.681	0.037
Number of references	0.659	0.036
Number of equations	0.553	0.030
Paper intimidation outlook	0.514	0.028
Algorithm written in pseudo code	0.501	0.027
Number of pages	0.464	0.025
Number of tables	0.454	0.025
Number of graphs/plots	0.348	0.019

Table A10. Feature-effect analysis for the dataset of Raff [38].

Feature	Class 0	Class 1
Paper readability	−0.6198	0.6198

References

1. Casey, B.; Santos, J.C.S.; Perry, G. A Survey of Source Code Representations for Machine Learning-Based Cybersecurity Tasks. *ACM Comput. Surv.* **2025**, *57*, 1–41. [CrossRef]
2. Jiang, J.; Li, Y.; Cao, S.; Shan, Y.; Liu, Y.; Fei, T.; Yu, Y.; Feng, Y.; Li, Y.; Li, Y.; et al. Artificial Intelligence in Bioinformatics: A Survey. *Brief. Bioinform.* **2025**, *26*, bbaf576. [CrossRef]
3. Mizna, S.; Arora, S.; Saluja, P.; Das, G.; Alanesi, W.A. An Analytic Research and Review of the Literature on Practice of Artificial Intelligence in Healthcare. *Eur. J. Med. Res.* **2025**, *30*, 382. [CrossRef]
4. Donaire, L.M.; Ortega, G.; Orts, F.; Martín Garzón, E.; Filatovas, E. A hybrid quantum-classical approach for liver disease detection using quantum machine learning. *Eng. Appl. Artif. Intell.* **2026**, *164*, 113240. [CrossRef]

5. Peivaste, I.; Belouettar, S.; Mercuri, F.; Fantuzzi, N.; Dehghani, H.; Izadi, R.; Ibrahim, H.; Lengiewicz, J.; Belouettar-Mathis, M.; Bendine, K.; et al. Artificial Intelligence in Materials Science and Engineering: Current Landscape, Key Challenges, and Future Trajectories. *Compos. Struct.* **2025**, *372*, 119419. [CrossRef]
6. Gudžius, P.; Kurasova, O.; Darulis, V.; Filatovas, E. Deep learning-based object recognition in multispectral satellite imagery for real-time applications. *Mach. Vis. Appl.* **2021**, *32*, 98. [CrossRef] [PubMed]
7. Gudzius, P.; Kurasova, O.; Darulis, V.; Filatovas, E. AutoML-Based Neural Architecture Search for Object Recognition in Satellite Imagery. *Remote Sens.* **2023**, *15*, 91. [CrossRef]
8. Marcozzi, M.; Filatovas, E.; Paulavičius, R. A Review of Quantum Machine Learning Applications. In *Data Science in Applications: Towards AI-Driven Approaches*; Studies in Computational Intelligence; Dzemyda, G., Bernatavičienė, J., Kacprzyk, J., Eds.; Springer: Cham, Switzerland, 2025; Volume 1206, pp. 143–163. [CrossRef]
9. Maslej, N.; Fattorini, L.; Brynjolfsson, E.; Etchemendy, J.; Ligett, K.; Lyons, T.; Manyika, J.; Ngo, H.; Niebles, J.C.; Parli, V.; et al. Artificial Intelligence Index Report 2023. 2023. Available online: https://hai.stanford.edu/ai-index/2023-ai-index-report (accessed on 22 April 2026).
10. Gundersen, O.E.; Kjensmo, S. State of the Art: Reproducibility in Artificial Intelligence. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32, pp. 1644–1651. [CrossRef]
11. Collberg, C.; Proebsting, T.A. Repeatability in Computer Systems Research. *Commun. ACM* **2016**, *59*, 62–69. [CrossRef]
12. Heller, N.; Rickman, J.; Weight, C.; Papanikolopoulos, N. The Role of Publicly Available Data in MICCAI Papers from 2014 to 2018. *arXiv* **2019**, arXiv:1908.06830. [CrossRef]
13. Pineau, J.; Vincent-Lamarre, P.; Sinha, K.; Larivière, V.; Beygelzimer, A.; d'Alché Buc, F.; Fox, E.; Larochelle, H. Improving Reproducibility in Machine Learning Research: A Report from the NeurIPS 2019 Reproducibility Program. *J. Mach. Learn. Res.* **2021**, *22*, 1–20.
14. Bouthillier, X.; Laurent, C.; Vincent, P. Unreproducible Research Is Reproducible. In *Proceedings of the 36th International Conference on Machine Learning*; Proceedings of Machine Learning Research; PMLR: New York, NY, USA, 2019; Volume 97, pp. 725–734.
15. Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* **2023**, *4*, 100804. [CrossRef]
16. Heil, B.J.; Hoffman, M.M.; Markowetz, F.; Lee, S.I.; Greene, C.S.; Hicks, S.C. Reproducibility Standards for Machine Learning in the Life Sciences. *Nat. Methods* **2021**, *18*, 1132–1135. [CrossRef]
17. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Daumé, H., III; Crawford, K. Datasheets for Datasets. *Commun. ACM* **2021**, *64*, 86–92. [CrossRef]
18. Filatovas, E.; Stripinis, L.; Orts Gomez, F.J.; Paulavičius, R. Advancing Research Reproducibility in Machine Learning through Blockchain Technology. *Informatica* **2024**, *35*, 227–253. [CrossRef]
19. Pham, H.V.; Qian, S.; Wang, J.; Lutellier, T.; Rosenthal, J.; Tan, L.; Yu, Y.; Nagappan, N. Problems and Opportunities in Training Deep Learning Software Systems: An Analysis of Variance. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 771–783. [CrossRef]
20. Hardwicke, T.E.; Mathur, M.B.; MacDonald, K.; Nilsonne, G.; Banks, G.C.; Kidwell, M.C.; Hofelich Mohr, A.; Clayton, E.; Yoon, E.J.; Tessler, M.H.; et al. Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal Cognition. *R. Soc. Open Sci.* **2018**, *5*, 180448. [CrossRef]
21. Chacon, S.; Straub, B. Basic Branching and Merging. 2014. Available online: https://git-scm.com/book/en/v2/Git-Branching-Basic-Branching-and-Merging (accessed on 22 April 2026).
22. Docker, Inc. *What Is Docker?* Docker, Inc.: Palo Alto, CA, USA, 2026. Available online: https://docs.docker.com/get-started/docker-overview/ (accessed on 22 April 2026).
23. Neural Information Processing Systems Foundation. *NeurIPS 2025 Call for Papers*; Neural Information Processing Systems Foundation: San Diego, CA, USA, 2025. Available online: https://neurips.cc/Conferences/2025/CallForPapers (accessed on 22 April 2026).
24. International Conference on Learning Representations. ICLR 2024 Author Guide. 2024. Available online: https://iclr.cc/Conferences/2024/AuthorGuide (accessed on 22 April 2026).
25. Barba, L.A. Terminologies for Reproducible Research. *arXiv* **2018**, arXiv:1802.03311. [CrossRef]
26. Gundersen, O.E. The Fundamental Principles of Reproducibility. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **2021**, *379*, 20200210. [CrossRef]
27. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*; National Academies Press: Washington, DC, USA, 2019. Available online: https://nap.nationalacademies.org/read/25303/chapter/3 (accessed on 22 April 2026).
28. Neves, K.; Amaral, O.B. Addressing Selective Reporting of Experiments through Predefined Exclusion Criteria. *eLife* **2020**, *9*, e56626. [CrossRef]

29. Paullada, A.; Raji, I.D.; Bender, E.M.; Denton, E.; Hanna, A. Data and Its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns* **2021**, *2*, 100336. [CrossRef] [PubMed]
30. Magnusson, I.; Smith, N.A.; Dodge, J. Reproducibility in NLP: What Have We Learned from the Checklist? In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 12710–12731. [CrossRef]
31. Luo, T.; Meng, R.; Wang, X.E.; Liu, Y. Interpretable Research Replication Prediction via Variational Contextual Consistency Sentence Masking. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 3959–3974. [CrossRef]
32. Yang, Y.; Wu, Y.; Uzzi, B. Estimating the Deep Replicability of Scientific Findings Using Human and Artificial Intelligence. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 10762–10768. [CrossRef] [PubMed]
33. Luo, T.; Li, X.; Wang, H.; Liu, Y. Research Replication Prediction Using Weakly Supervised Learning. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1464–1474. [CrossRef]
34. Altmejd, A.; Dreber, A.; Forsell, E.; Huber, J.; Imai, T.; Johannesson, M.; Kirchler, M.; Nave, G.; Camerer, C. Predicting the Replicability of Social Science Lab Experiments. *PLoS ONE* **2019**, *14*, e0225826. [CrossRef]
35. Wu, J.; Nivargi, R.; Lanka, S.S.T.; Menon, A.M.; Modukuri, S.A.; Nakshatri, N.; Wei, X.; Wang, Z.; Caverlee, J.; Rajtmajer, S.M.; et al. Predicting the Reproducibility of Social and Behavioral Science Papers Using Supervised Learning Models. *arXiv* **2021**, arXiv:2104.04580. [CrossRef]
36. Belz, A.; Popović, M.; Mille, S. Quantified Reproducibility Assessment of NLP Results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 16–28. [CrossRef]
37. Olszewski, D.; Lu, A.; Stillman, C.; Warren, K.; Kitroser, C.; Pascual, A.; Ukirde, D.; Butler, K.R.B.; Traynor, P. "Get in Researchers; We're Measuring Reproducibility": A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, Copenhagen, Denmark, 26–30 November 2023; pp. 3433–3459. [CrossRef]
38. Raff, E. A Step Toward Quantifying Independently Reproducible Machine Learning Research. In *Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32, pp. 5486–5496. Available online: https://papers.nips.cc/paper_files/paper/2019/hash/c429429bf1f2af051f2021dc92a8ebea-Abstract.html (accessed on 22 April 2026).
39. Kampakis, S. XGBoost: A Mathematical Explanation. 2025. Available online: https://www.mathstoml.com/xgboost/ (accessed on 22 April 2026).
40. Molnar, C. *Interpretable Machine Learning*, 2nd ed.; Lulu Press, Inc.: Durham, NC, USA, 2022. Available online: https://christophm.github.io/interpretable-ml-book/ (accessed on 22 April 2026).
41. Chan, J. STAT3014 Tutorial Solutions. 2015. Available online: https://www.maths.usyd.edu.au/u/jchan/STAT3014/sur15_1_sol.pdf (accessed on 22 April 2026).
42. OpenAI. Shared ChatGPT Conversation Used for Annotation. Annotation of the Seven Quantitative Reproducibility Criteria Scores for a Research Paper. 2025. Available online: https://chatgpt.com/share/6954075d-6134-8005-ab71-6a1e51bad9fa (accessed on 22 April 2026).
43. Vieira, S.M.; Kaymak, U.; Sousa, J.M. Cohen's kappa coefficient as a performance measure for feature selection. In Proceedings of the International Conference on Fuzzy Systems, Barcelona, Spain, 18–23 July 2010; pp. 1–8. [CrossRef]
44. Sotoudeh-Anvari, A. The Applications of MCDM Methods in COVID-19 Pandemic: A State-of-the-Art Review. *Appl. Soft Comput.* **2022**, *126*, 109238. [CrossRef]
45. Filatovas, E.; Marcozzi, M.; Mostarda, L.; Paulavičius, R. A MCDM-based framework for blockchain consensus protocol selection. *Expert Syst. Appl.* **2022**, *204*, 117609. [CrossRef]
46. Saaty, T.L. Decision Making with the Analytic Hierarchy Process. *Int. J. Serv. Sci.* **2008**, *1*, 83–98. [CrossRef]
47. Mitra, S.; Goswami, S.S. Application of Simple Average Weighting Optimization Method in the Selection of Best Desktop Computer Model. *Adv. J. Grad. Res.* **2019**, *6*, 60–68. [CrossRef]
48. de Lataillade, A.; Blanco, S.; Clergent, Y.; Dufresne, J.L.; El Hafi, M.; Fournier, R. Monte Carlo Method and Sensitivity Estimations. *J. Quant. Spectrosc. Radiat. Transf.* **2002**, *75*, 529–538. [CrossRef]
49. Srinivasan, V.; Shocker, A.D. Linear Programming Techniques for Multidimensional Analysis of Preferences. *Psychometrika* **1973**, *38*, 337–369. [CrossRef]
50. Brans, J.P.; Vincke, P. A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria Decision-Making). *Manag. Sci.* **1985**, *31*, 647–656. [CrossRef]
51. Opricovic, S.; Tzeng, G.H. Compromise Solution by MCDM Methods: A Comparative Analysis of VIKOR and TOPSIS. *Eur. J. Oper. Res.* **2004**, *156*, 445–455. [CrossRef]

52. Greco, S.; Ishizaka, A.; Tasiou, M.; Torrisi, G. The ordinal input for cardinal output approach of non-compensatory composite indicators: The PROMETHEE scoring method. *Eur. J. Oper. Res.* **2021**, *288*, 225–246. [CrossRef]
53. Fenner, M. What Can Article-Level Metrics Do for You? *PLoS Biol.* **2013**, *11*, e1001687. [CrossRef] [PubMed]
54. Foster, E.D.; Deardorff, A. Open Science Framework (OSF). *J. Med. Libr. Assoc.* **2017**, *105*, 203–206. [CrossRef]
55. Association for Computing Machinery. Artifact Review and Badging—Current. 2020. Available online: https://www.acm.org/publications/policies/artifact-review-and-badging-current (accessed on 16 April 2026).