

An Automated and Explainable Framework for Verifying Reproducibility in Machine Learning Research

V Pavan Krishna

Dept. of Computer Science
Engineering

Velagapudi Ramakrishna Siddhartha
Engineering College
Vijayawada, India

Jhansi Thota

Dept. of Artificial Intelligence and
Data Science

Velagapudi Ramakrishna Siddhartha
Engineering College Vijayawada,
India

Nidha Ahmed Md

Dept. of Artificial Intelligence and
Data Science

Velagapudi Ramakrishna Siddhartha
Engineering College
Vijayawada, India

Abstract - Reproducibility is a crucial aspect when it comes to validating scientific studies. However, most machine learning experiments are poorly documented, have missing information, have inconsistent methodologies, and lack proper documentation. Thus, manually validating these studies is often lengthy, prone to errors, and cannot be scaled. This paper seeks to introduce an automated and explainable approach to validate the reproducibility of machine learning experiments. This solution works by analyzing papers, code repositories, and datasets that provide essential experimental artifacts. These include the machine learning model used, the hyperparameters of the model, the training and test datasets, and the evaluation metrics used. Furthermore, the proposed framework validates the experiment using a rule-based approach and determines any problems, including missing information, inconsistency, and potential data leakage problems. The proposed framework assigns a reproducibility score and presents explainable feedback to help researchers improve their experiments.

Keywords - Reproducibility, Machine Learning Research, Natural Language Processing, Automated Verification, Explainable Artificial Intelligence, Data Leakage Detection.

I. INTRODUCTION

Machine learning research has experienced rapid growth in recent years, driving innovations across domains such as healthcare, finance, natural language processing, and computer vision. However, a critical challenge that has emerged alongside this progress is the lack of reproducibility in many published machine learning studies. Reproducibility issues arise due to incomplete reporting of experimental setups, missing details about datasets and preprocessing steps, unavailability of source code, inconsistent evaluation protocols, and lack of clarity regarding hyperparameters, random seeds, and model configurations. These limitations make it difficult for researchers and practitioners to validate results, compare methods fairly, or build upon prior work effectively. Additionally, existing efforts to address reproducibility often rely on manual inspection, which is time-consuming, subjective, and not scalable. Current tools also struggle with automatically extracting structured information from research papers, identifying potential risks such as data leakage, and providing clear explanations for reproducibility failures.

To address these challenges, our major project proposes an automated and explainable framework for verifying reproducibility in machine learning research. The system leverages natural language processing and rule-based analysis to extract critical experimental details from research papers, including models, datasets, hyperparameters, and evaluation metrics. It further integrates repository analysis to examine code quality, dependency specifications, and configuration completeness, along with dataset verification to assess availability, splits, and potential leakage risks. Based on these factors, the framework computes a reproducibility confidence score and provides interpretable feedback highlighting missing or inconsistent information, enabling researchers to improve transparency and reliability in their work.

a) Contributions

The major contributions of this paper are listed here:

- Developed an automated framework to extract critical experimental details (models, datasets, hyperparameters, metrics) from machine learning research papers using NLP and rule-based techniques.
- Implemented a reproducibility verification engine that analyzes research papers, code repositories, and datasets to detect missing information, inconsistencies, and potential risks such as data leakage.
- Designed an explainable scoring system with a user-friendly interface that provides a reproducibility confidence score along with actionable feedback to improve transparency and reliability of research work.

II. LITERATURE SURVEY

Building on the growing concern of reproducibility in machine learning research, T. Ross-Hellauer et al. [1] provided a comprehensive conceptual analysis highlighting key barriers, drivers, and classifications of reproducibility issues. Their work establishes a strong theoretical foundation but lacks automated solutions for practical implementation. In a domain-specific

approach, H. S. Irwansyah et al. [2] applied structured methodologies such as CRISP-DM and reproducibility checklists for EEG-based mental workload classification, demonstrating improved experimental transparency, though their approach remains manual and limited to a specific domain. Similarly, C. Merkle et al. [3] conducted a large-scale empirical evaluation of 100 biodiversity-related deep learning papers, identifying reproducibility gaps through manual auditing, which highlights scalability limitations. Efforts toward standardization were further explored by H. Wallach et al. [4], who introduced checklist-based policies in the NeurIPS community to enforce better reporting practices, significantly improving transparency but still relying on author compliance rather than automation. In applied industrial settings, S. K. Sharma et al. [5] examined reproducibility in HVAC fault diagnosis systems through artifact availability audits, demonstrating practical relevance but limited generalizability. Addressing one of the most critical issues, J. Kapoor et al. [6] analyzed 294 papers across multiple domains and identified data leakage as a major contributor to irreproducibility, providing a detailed taxonomy but focusing primarily on leakage alone. In the life sciences domain, M. Wilkinson et al. [7] proposed tier-based reproducibility standards, offering clear guidelines for experimental reporting, though their approach requires significant manual effort for compliance. Lastly, E. Topol et al. [8] discussed the challenges of reproducibility in healthcare machine learning, emphasizing real-world implications and risks, but without providing a concrete automated framework. Collectively, these studies highlight the importance of reproducibility while revealing a clear gap in scalable, automated, and explainable solutions—motivating the need for the proposed system.

III. PROPOSED METHODOLOGY

This section provides an overview of the methodology employed in this research work.

a) Process Flow Diagram

Figure 1 presents the workflow of the proposed automated reproducibility verification framework for machine learning research. The process begins with the input of a research paper in PDF format, which serves as the primary source of information. In Module 1 (Data Collection and Preparation), the system extracts relevant content from the paper using parsing tools and NLP techniques, identifying key experimental components such as models, datasets, hyperparameters, and evaluation metrics. This structured information is then passed to Module 2 (Reproducibility Score Calculation), where a rule-based engine evaluates the completeness, consistency, and transparency of the extracted details. The system also analyzes associated code repositories and datasets to detect missing configurations, dependency issues, and potential risks such as data leakage, ultimately generating a reproducibility confidence score. In Module 3 (Conversational Explainable AI Interface), the results are presented through an interactive interface that explains the reproducibility score and highlights gaps or inconsistencies in a user-friendly manner. The system provides actionable insights and suggestions to improve the reproducibility of the research work. Based on this feedback,

users can refine their paper or code and re-upload an updated version, enabling an iterative improvement cycle. This approach ensures not only automated evaluation but also interpretability and continuous enhancement of research quality.

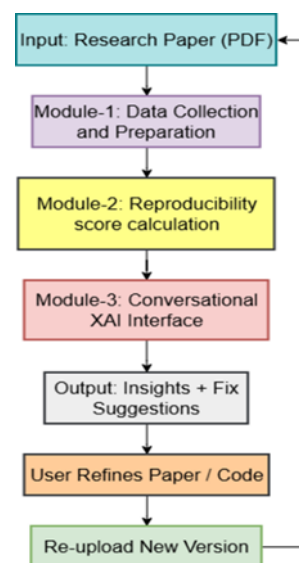


Fig. 1: Process Flow Diagram of Proposed System

b) Methodology

a) *Data Collection and Processing:* In the initial phase of the proposed framework, research papers are collected from sources such as arXiv, Papers with Code, or through manual uploads. These documents are converted into structured TEI XML format using GROBID, enabling systematic extraction of key sections like title, abstract, and methodology. The extracted content is cleaned and tokenized using regular expressions and NLP techniques. Semantic segmentation groups the content based on headings to focus on experimental details, while normalization ensures consistency through unit standardization and parameter structuring. The processed information is then stored in a JSON-based knowledge base for further analysis.

b) *Reproducibility Score Calculation:* In the second phase, the system evaluates reproducibility by computing a structured confidence score using the extracted experimental details. A rule-based engine analyzes four key aspects: documentation, code availability, dataset transparency, and evaluation clarity. It detects elements such as hyperparameters, hardware configurations, and repository links while identifying missing information, inconsistencies, and potential data leakage risks. The validation results are aggregated and processed through a weighted scoring mechanism to generate a final reproducibility score, which is stored as a structured JSON report.

c) *Explainable AI (xAI) Framework:* The final phase focuses on generating interpretable insights from the reproducibility analysis. The JSON report is processed and transformed into user-friendly explanations using an LLM-

based system. It highlights gaps, inconsistencies, and risks while providing actionable suggestions such as adding missing details or improving dataset and code transparency. The results are presented through an interactive dashboard, allowing users to visualize outcomes and refine their work. The system supports an iterative process where updated versions can be re-uploaded, enabling continuous improvement and tracking of reproducibility over time.

IV. RESULTS AND ANALYSIS

This chapter presents the results of the proposed automated and explainable framework for verifying reproducibility in machine learning research. The system evaluates research papers by analyzing experimental details, code repositories, and datasets, and generates a reproducibility confidence score along with explainable insights. The results demonstrate the system’s effectiveness in identifying missing information, inconsistencies, and potential reproducibility issues through score-based analysis and interpretable outputs.

a) Reproducibility Analysis Results

This section presents the results of the proposed automated reproducibility verification framework applied to machine learning research papers. The system evaluates papers by analyzing experimental details, code repositories, and datasets, and generates a reproducibility confidence score along with explainable insights. The GUI of the Auditor Pro system provides an interface for performing and visualizing this analysis as shown in Figure 2.

The reproducibility evaluation is carried out based on four key dimensions:

- Documentation
- Code and Environment
- Dataset
- Evaluation

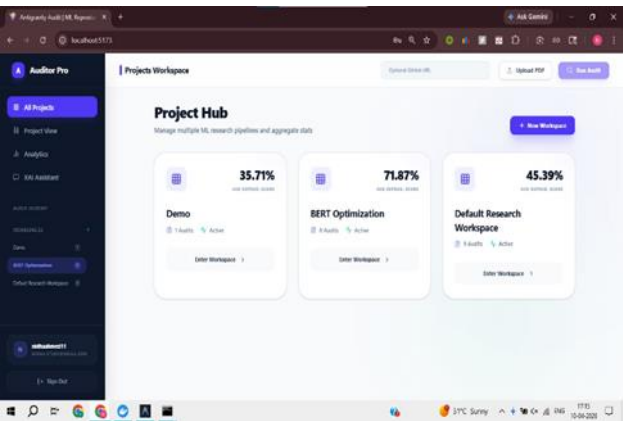


Figure 2: Project Workspace

Figure 3 illustrates the dashboard workspace presenting the analysis of a sample research paper (BERT). The interface displays the reproducibility confidence score along with key validation details such as extracted experimental information and availability of hyperparameters. An XAI summary is also

provided, offering a clear and concise explanation of the audit results.

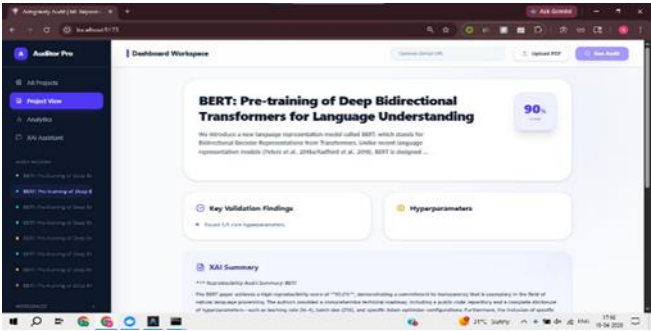


Figure 3: Reproducibility Score

Figure 4 illustrates the analytics workspace, providing a detailed breakdown of reproducibility across different dimensions. A radar chart visualizes performance in documentation, code, dataset, and evaluation, while a line chart shows reproducibility improvement across multiple iterations. This enables easy identification of weaker components affecting overall reproducibility.

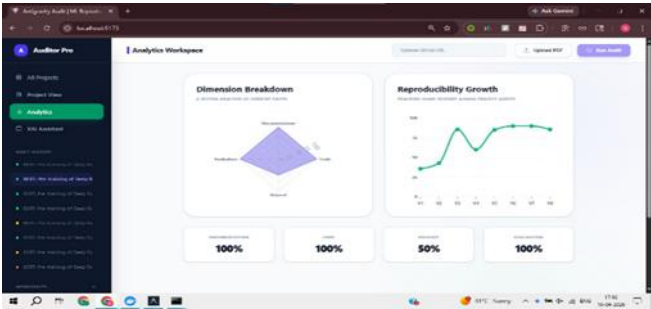


Figure 4: Analysis of Reproducibility

Figure 5 presents the explainable AI assistant interface, where users can interact with the system to understand reproducibility issues. The assistant provides clear explanations and suggests actionable improvements, such as dataset versioning and proper documentation of experimental settings.

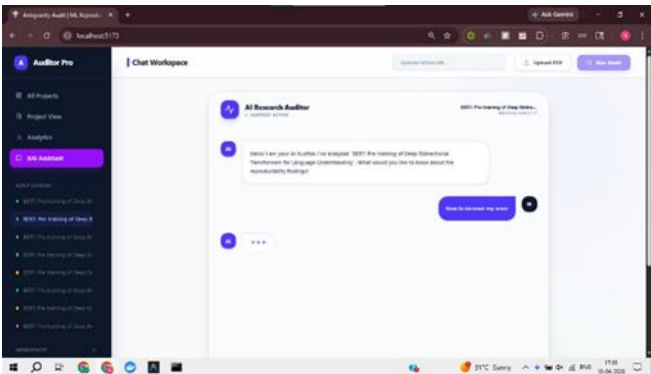


Figure 5: Explainable AI

Figure 6 shows the sign-up page of the system, which includes fields for username, email address, and password. These details are securely stored in a MongoDB database, where both the

username and email must be unique, and the password is stored in a hashed format for security purposes.

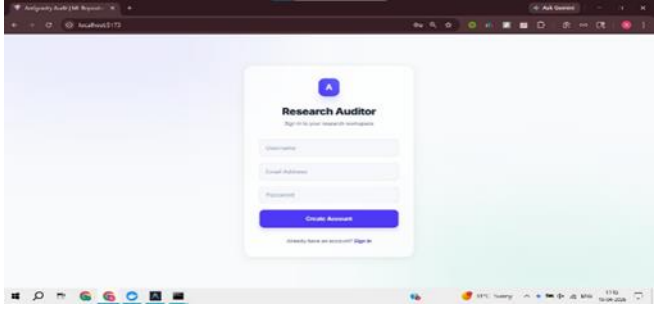


Figure 6: Sign-Up Page

Figure 7 shows the sign-in page of the system, which includes fields for email address and password. These credentials are validated against records stored in the MongoDB database, where the email address is unique and the password is securely stored in hashed form.

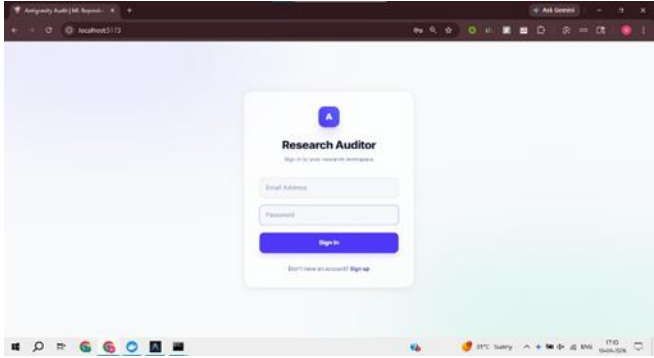


Figure 7: Sign-In Page

b) Performance Metrics

The performance of the proposed reproducibility verification framework is evaluated based on operational efficiency, extraction capability, and consistency in decision-making. The key metrics are described below:

a) Audit Latency This metric measures the computational efficiency of the complete audit pipeline using precise time tracking.

- **Total Audit Time:** Time from PDF upload to final score generation.
- **Parser Latency:** Time taken by the GROBID microservice to convert the document into structured format.
- **Target Performance:** The system aims to complete processing within 30 seconds for a responsive user experience.

b) Entity Identification Count (EIC) EIC evaluates the system's ability to identify essential experimental parameters required for reproducibility.

- **Hyperparameter Count:** Detection of key parameters such as learning rate, batch size, optimizer, epochs, and loss function.
- **Audit Logic:** Scoring is based on whether these parameters are identified as verified or missing (e.g., 3/5 parameters found).

c) Section Awareness Coverage

This metric assesses how effectively the system segments a research paper into meaningful sections.

- **Segment Detection:** Identification of important sections such as Abstract, Methodology/Architecture, and Experimental Results.
- **Fallback Triggering:** Indicates whether structured parsing was successful or fallback methods were required.

d) Ingestion Success Rate (ISR)

ISR measures the robustness and reliability of the document parsing process.

- **GROBID Success Rate:** Percentage of documents successfully converted into structured TEI XML format.
- **AI Fallback Dependency:** Percentage of cases where fallback methods (e.g., alternative PDF parsing) were required due to parsing failures.

e) Multi-Dimensional Reproducibility Score

This is the final quantitative metric representing overall reproducibility quality. It is computed across four dimensions:

- Documentation
- Code Availability
- Dataset Transparency
- Evaluation Clarity

The reproducibility score is calculated using a weighted formula:

$$Score = D_1 \times 0.40 + D_2 \times 0.30 + D_3 \times 0.20 + D_4 \times 0.10$$

Where D_1, D_2, D_3, D_4 represent the respective evaluation dimensions.

Table 3 shows that the Comparison of proposed Reproducibility system with other- works

TABLE III
COMPARATIVE ANALYSIS OF DIFFERENT STUDIES

STUDY	METHODOLOGY	AUTOMATION	SCORING	LIMITATIONS
SEMMEL ROCK ET AL. [1]	CONCEPTUAL SURVEY + TAXONOMY	No	No	NO IMPLEMENTATION; THEORETICAL ANALYSIS ONLY
DEMIREZEN ET AL. [2]	CRISP-DM + CHECKLIST	PARTIAL	No	DOMAIN-SPECIFIC; MANUAL EVALUATION
AHMED ET AL. [3]	VARIABLE EXTRACTION + MANUAL AUDIT	No	No	MANUAL PROCESS; LIMITED SCALABILITY
KAPOOR ET AL. [4]	LEAKAGE ANALYSIS + CASE STUDY	No	No	FOCUSED ONLY ON DATA LEAKAGE
PROPOSED WORK	NLP + RULE-BASED + XAI FRAMEWORK	YES	YES (0–100 SCORE)	DEPENDS ON EXTRACTION ACCURACY

V. CONCLUSION AND FUTURE WORK

The proposed framework is designed to evaluate the reproducibility of machine learning research papers in an automated and systematic manner. It utilizes NLP techniques and rule-based validation to extract experimental details and analyze associated source code. The system effectively identifies key reproducibility issues such as missing data, inconsistencies, and potential data leakage. Based on this analysis, it generates a reproducibility score along with clear

explanations for each paper. Compared to manual approaches, the framework is more scalable, objective, and efficient. Experimental results on multiple research papers demonstrate its ability to accurately detect reproducibility gaps. The integration of NLP, code analysis, and explainable AI highlights its effectiveness in supporting standardized reproducibility evaluation. Overall, the framework provides a strong foundation for intelligent and scalable reproducibility assessment in machine learning research. Future work includes enhancing automation through containerization, leveraging large language models for deeper semantic analysis, improving dataset validation methods, and deploying the system as a cloud-based API service.

REFERENCES

[1] Pineau et al. (NeurIPS Reproducibility Checklist) Improving Reproducibility in Machine Learning Research. DOI: 10.5555/3546258.3546262

[2] Gundersen & Kjensmo (Randomness & reproducibility) State of the Art: Reproducibility in Artificial IntelligenceDOI: 10.1609/aaai.v32i1.11503

[3] Beam et al. (Hidden implementation factors in ML systems) Challenges to the Reproducibility of Machine Learning Models in Health CareDOI: 10.1001/jama.2019.20866

[4] Semmelrock et al. (Documentation gaps & barriers) Reproducibility in Machine-Learning-Based Research: Overview, Barriers, and DriversDOI: 10.1002/aaai.70002

[5] Heil et al. (Reproducibility standards: data + code availability) Reproducibility Standards for Machine Learning in the Life SciencesDOI: 10.1038/s41592-021-01256-7

[6] Tatman et al. (Reproducibility taxonomy & versioning) A Practical Taxonomy of Reproducibility for Machine Learning Research

[7] Ahmed et al. (Hyperparameters & experimental setup clarity) Evaluating the Method Reproducibility of Deep Learning Models in Biodiversity ResearchDOI: 10.7717/peerj-cs.2618

[8] Kapoor & Narayanan (Data leakage risks) Leakage and the Reproducibility Crisis in Machine-Learning-Based ScienceDOI: 10.1016/j.patter.2023.100804

[9] Semmelrock, Harald & Ross-Hellauer, Tony & Kopeinik, Simone & Theiler, Dieter & Haberl, Armin & Thalmann, Stefan & Kowald, Dominik. (2025). Reproducibility in machine-learning-based research: Overview, barriers, and drivers. AI Magazine. 46. 10.1002/aaai.70002.

[10]Demirezen G, Taşkaya Temizel T and Brouwer A-M (2024) Reproducible machine learning research in mental workload classification using EEG. Front. Neuroergon. 5:1346794. doi:10.3389/fnrgo.2024.1346794

[11]Ahmed W, Kommineni VK, König-Ries B, Gaikwad J, Gadelha L, Samuel S. 2025. Evaluating the method reproducibility of deep learning models in biodiversity research. PeerJ Computer Science 11:e2618https://doi.org/10.7717/peerj-cs.2618