International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Reproducibility crisis in deep learning vulnerability detection: An open science perspective

Sanjay Agal *, Nikunj Bhavsar, Krishna M Raulji and Kiran Macwan

Artificial Intelligence and Data Science, Parul University, Vadodara, 391760, India.

International Journal of Science and Research Archive, 2025, 15(01), 602-611

Publication history: Received on 23 February 2025; revised on 08 April 2025; accepted on 11 April 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.1.1041

Abstract

This paper digs into the deep-rooted reproducibility mess in deep learning vulnerability detection. It all starts with the fact that studies keep giving off mixed signals—findings just don't match up as you'd expect. There isn't just a handful of "success stories;" we need datasets that capture every angle, including those less-than-perfect moments, along with all the nitty-gritty details of experiments and how results are measured. In most cases, differences in data quality, the way models are put together, and which evaluation methods are used all add to these unpredictable outcomes. It seems that a lack of a one-size-fits-all approach is what's really throwing a wrench in the works, especially when it comes to healthcare—where spotting vulnerabilities isn't just academic but vital for patient safety and keeping data secure. Generally speaking, if reproducibility were on firmer ground, diagnostic systems powered by machine learning would earn more trust, leading to smarter, better decisions. By pushing for an open science style that values clarity and the free sharing of methods, this study hopes to spark more real-world collaboration and fresh ideas, paving the way for deep learning to work more reliably in our healthcare systems. Overall, settling on common evaluation practices might just be the key to smoothing out these reproducibility bumps and boosting the overall credibility and usefulness of these tech solutions in critical healthcare settings.

Keywords: Reproducibility; Deep Learning; Vulnerability Detection; Open Science; AI Transparency; Cybersecurity; Benchmark Datasets; Experimental Standardization; Machine Learning Reliability; Model Validation

1. Introduction

Deep learning is moving at a rapid clip and is opening up breakthrough possibilities—especially when it comes to spotting vulnerabilities—even though this progress demands that research findings be reproducible. Being able to recreate results consistently is, generally speaking, key to earning trust in scientific work, particularly when these deep learning models are applied in high-stakes areas like healthcare and finance. Several studies have pointed out some rather concerning inconsistencies in replicating findings, suggesting a sort of crisis that shakes the reliability of these methods [1], [2], [4]. A mix of factors seems to be behind these challenges: the murky inner workings of neural networks, differences in how experiments are run, and a tendency for models to perform in unexpected ways across diverse datasets [3], [5]. As a result, this paper zeroes in on uncovering the root causes of the reproducibility issues in deep learning for vulnerability detection, and it aims to untangle the conflicting findings that litter the current literature [6], [9], [16].The study sets out to identify the key barriers to getting consistent results, to push for more standardized protocols, and to back an open science approach that prizes transparency and collaboration among researchers [7], [8]. By grappling with these goals, it hopes to build a basic framework that ultimately bolsters both the reliability and validity of deep learning when used for vulnerability detection, thereby guiding smarter decision making [11], [12]. This work matters not only in academic circles but also in practical terms—it has real implications for protecting patient safety and keeping data integrity intact in fields where even minor oversights can lead to catastrophic outcomes [14], [15]. Ultimately, by taking on the reproducibility conundrum, this paper aspires to nurture a sturdier understanding of

* Corresponding author: Sanjay Agal

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

deep learning techniques and to lift confidence in their use across critical domains where the cost of uncertainty is simply too high [10], [19].

1.1. Significance of Reproducibility in Deep Learning Vulnerability Detection

Learning models power our most critical systems these days, and trust in their outcomes is absolutely essential. When you dig a bit deeper, you find that reproducibility—especially in vulnerability detection—is a real headache. Findings can swing wildly from one study to another [1], [2], and this unpredictability chips away at our confidence in models that are supposed to deliver both precision and trustworthy decision-making [3]. In many cases, the trouble seems rooted in a mix of factors: different datasets, varied experimental setups, and model designs that aren't always in sync. Essentially, researchers are trying to pinpoint what's really behind this crisis and figure out how reproducibility plays into both theory and hands-on practice [4], [5].Reproducibility isn't just academic chatter—it's a linchpin for building trust within the community that designs and deploys these systems [6], [7]. In sensitive areas like healthcare and financial security, dependable results mean that when vulnerabilities pop up, the stakes are managed carefully [8], [9]. Generally, when a framework robustly supports reproducibility, it opens the door for stronger collaboration. Standardized approaches help scientists compare and blend findings from different studies, which in turn sparks faster technological advances [10], [11].Taking a close look, it's clear that grappling with the reproducibility crisis is key to turning deep learning-driven vulnerability detection into a field we can genuinely rely on. Without steps to fix these issues, promising innovations might remain unproven and met with skepticism, slowing progress in areas that really matter [12], [13]. Also, the push for improved reproducibility jives well with Open Science ideals—emphasizing transparency, data sharing, and teamwork, all of which strengthen the overall integrity of research [14], [15]. In short, putting real energy into making results reproducible not only backs up current methods but also sets future researchers up with a robust, credible foundation for deep learning applications across a wide range of fields [16], [17], [18], [19].

2. Literature review

Deep learning has revolutionized various domains, from healthcare to autonomous systems, by delivering levels of automation that once seemed unattainable. However, alongside its rapid adoption, a growing concern has

Table 1 Reproducibility Issues in Machine Learning-Based Research

Field	Reproducibility Issues	Source
Medical Imaging	Data leakage affecting 294 papers, leading to overoptimistic conclusions	Princeton University
Cybersecurity AI	Challenges due to software and hardware incompatibilities, version conflicts, and obsolescence	arXiv e-prints
Bioimage Analysis	Potential replication crisis due to deep-learning-based methods	Nature Methods
Medical Imaging	Lack of standardized methodologies and comprehensive documentation	Machine Learning for Brain Disorders

Emerged—commonly referred to as the reproducibility crisis—which is particularly evident in machine learning research and more specifically in vulnerability detection. Multiple studies have shown that reproducing reported results under different conditions often yields inconsistent performance outcomes ([1]), posing a serious threat to the credibility and reliability of research findings. In the context of deep learning-based vulnerability detection, reproducibility is not merely an academic concern—it is a practical necessity in ensuring robust cybersecurity ([2]).

The body of research addressing this issue takes several perspectives. Some studies point to methodological inconsistencies, while others highlight dataset biases and the opaque nature of deep neural network models as primary barriers to reproducibility ([3]). Even subtle variations in experimental setups, such as dataset splits or training configurations, have been shown to result in significant performance deviations ([4]). Furthermore, hyperparameter tuning—often done in an ad hoc manner—adds an additional layer of unpredictability ([5]). As a result, numerous researchers advocate for open science practices, including transparent reporting standards and publicly available data and code, to mitigate these challenges ([6], [7]). Despite these proposals, practical implementation often lags behind, particularly in rapidly evolving subfields like vulnerability detection ([8]).

In industrial settings, the situation is further complicated by proprietary concerns. Many organizations protect their algorithms and training data as trade secrets, thereby limiting opportunities for independent validation and collaborative improvement ([9]). This lack of transparency hinders the broader research community and promotes a culture of non-disclosure. Calls for open, cooperative frameworks and shared resources have been increasingly common in academic discourse ([10], [11]), though real-world challenges, such as legal and competitive constraints, often limit their feasibility ([12], [13]).

Progress has been made through community-driven repositories and benchmarking initiatives ([14]), but questions remain as to whether these solutions are effective in improving reproducibility in the specific context of deep learning-based vulnerability detection ([15]). The persistent emergence of new challenges suggests that while progress is ongoing, significant gaps still exist. This literature review aims to synthesize existing work, identify key barriers, and highlight the opportunities for future advancements in reproducibility and methodological rigor ([16]–[30]).

In earlier phases of research, many deep learning-based models in vulnerability detection were celebrated for outperforming traditional techniques ([1]). However, subsequent analyses revealed that these promising results often failed to generalize when applied to different datasets or settings ([2]). By the mid-2010s, the field had begun to recognize the essential role of rigorous experimental design. Key concerns included incomplete dataset descriptions and insufficient documentation of model training processes ([3]–[5]). This recognition catalyzed a shift toward open science principles, leading to the development of standardized benchmarks and performance metrics ([6], [7]). Moreover, analyses of publication bias further complicated the reproducibility landscape, pushing for initiatives like pre-registration to ensure greater transparency and accountability ([8], [9]).

The reproducibility crisis is now recognized as a core threat to the trustworthiness of research in this area. Irregularities in experimental protocols are widely cited as sources of variability in model outcomes ([1], [2]). To address these issues, researchers have proposed establishing clear reporting guidelines and standardized evaluation protocols ([3], [4]). This aligns with broader efforts to institutionalize open science practices—sharing datasets, releasing code, and encouraging collaborative study replication ([5], [6]). These efforts complement movements toward openaccess publication and shared repositories, which facilitate constructive peer engagement and verification ([7], [8]). Practically, the failure to rigorously validate vulnerability detection models could lead to security oversights, where systems are falsely assumed to be secure when they are not ([9], [10]).

A closer look at existing methodological interventions reveals a diverse array of strategies, each with unique advantages and limitations. Key criticisms focus on the lack of consensus regarding standardized practices, leading to irreproducible results across studies ([1], [2]). Researchers emphasize transparency in model configuration and dataset handling as crucial to improving reproducibility ([3], [4]). Decisions around model architecture and hyperparameter choices are often left to individual discretion, introducing subjectivity and variability ([5], [6]). Calls have been made for more rigorous statistical validation procedures and the use of comprehensive evaluation metrics ([7], [8]), especially in light of the multifaceted nature of software vulnerabilities ([9], [10]).

There is a growing consensus that reproducibility requires both methodological discipline and systemic change. Standardized protocols are seen as essential for stabilizing results, while open science principles are regarded as vital for transparency and trust ([1]–[4]). This convergence points to a deeper issue within computational science— the need to rethink the foundations of knowledge production and validation. Ethical concerns also arise, particularly around the risk of biased datasets that could produce misleading results or reinforce existing inequities ([6]–[9]). Some scholars have turned to ethics and statistical theory to frame these concerns more rigorously ([10], [11]), while others advocate for interdisciplinary collaboration to ensure both scientific rigor and social responsibility ([12]– [14]). Ultimately, these discussions frame reproducibility as a transformative agenda rather than a mere technical correction ([15], [16]).

In summary, the literature reveals a complex but consistent picture: reproducibility challenges are deeply embedded in both the technical and cultural fabric of deep learning research in vulnerability detection. Inconsistent methodologies, incomplete reporting, and proprietary restrictions are recurring themes that compromise confidence in findings ([1], [2]). The call for open science is both a remedy and a guiding philosophy, encouraging collaborative, transparent, and accountable research practices ([3], [4]). While the technical community has made strides, including shared benchmarks and evaluation tools, systemic gaps persist ([5]–[7]). Without reproducible foundations, vulnerability detection models risk becoming unreliable or even counterproductive in operational contexts ([8], [9]).

Future research must rigorously assess existing open science practices and develop empirical measures to evaluate their impact on reproducibility in deep learning for vulnerability detection ([10]–[13]). Moreover, the field must address persistent methodological discrepancies and commit to interdisciplinary approaches that blend technical depth

with ethical clarity ([14]–[23]). Through these efforts, the field can move toward more reliable, secure, and scientifically sound systems capable of withstanding both academic scrutiny and real-world deployment challenges ([24]–[30]).

Table 2 Reproducibility Challenges in Cybersecurity AI Research

Challenge	Description
Software and Hardware Incompatibilities	Difficulty in reproducing results due to mismatched software versions and hardware configurations.
Version Conflicts and Obsolescence	Issues arising from outdated or incompatible software versions, leading to failed reproductions.
Lack of Standardized Methodologies	Absence of uniform practices for ensuring reproducibility in AI-driven cybersecurity research.
Insufficient Documentation	Limited or unclear documentation hindering the replication of research findings.

3. Methodology

Deep learning is shaking up many fields—vulnerability detection included, and it's helped boost system security in a lot of cases. Still, there's a catch: this rapid progress is being marred by what many call a reproducibility crisis, which puts research findings into question and makes us wonder if deep learning models will work reliably when applied practically [1]. Some scholars have noticed that when experimental setups or methods shift around, the outcomes can vary wildly, leaving many folks a bit skeptical about the real strength of these deep learning solutions [2].This paper jumps right in to tackle that reproducibility hiccup in vulnerability detection frameworks, especially looking at what happens when deep learning techniques get used without enough careful checking or standardized methods [3]. At the heart of this work is a hands-on look at current research—digging through literature, spotting where things don't quite add up, and suggesting a set of best practices that fit with open science ideas. By reviewing various strategies found in ongoing studies, the research basically zeroes in on syncing deep learning methods with reproducible science so that confidence in vulnerability detection can really see an uptick [4].One can't overstate how important this effort is—it not only fuels deeper academic debate about the reproducibility crisis but also offers down-to-earth fixes that practitioners might actually try out in real settings [5]. Also, by breaking down and assessing current methodologies, this study sheds light on new routes for future research while proposing a sturdy framework that could help soften reproducibility issues across different deep learning applications [6]. It draws on

well-established frameworks and hands-on studies, putting the proposed methods to the test against recognized benchmarks [7]. The insights and analyses coming out of this approach should prove priceless, giving us a rounded understanding of reproducibility challenges and spurring more collaboration between researchers in academia and industry [8]. At its core, this method isn't just about moving theory forward—it pushes for more transparency in vulnerability detection practices to plug a notable gap in today's deep learning literature [9]. Focusing on open science, the work hopes to promote the sharing of insights and resources, which are key to overcoming the current challenges in reproducibility [10].

3.1. Research Design and Framework

Deep learning vulnerability detection has hit a rough patch lately—there's a serious reproducibility crisis that makes everything feel a bit off. Different studies, for instance, don't seem to use the same methods, which ends up throwing off confidence in these models [1]. In light of that, this project sets out to build a broad framework that points out the holes in current practices and tosses out practical fixes that generally lean on open science ideas [2]. The plan is to mix things up by crunching numbers from existing datasets alongside a hands-on look into how vulnerability tests are actually done [3]. This kind of combo helps us see, in a much more complete way, what's really fueling these reproducibility issues and even lets us compare methods that come from very different studies [4]. It's not just an academic exercise either; the work is intended to give real-world practitioners a clear and replicable game plan for putting deep learning techniques to work effectively [5]. By pulling together insights from various sources and approaches—even if the repeated details seem a bit redundant—the study aims to create a unified picture of why reproducibility is such a big deal and how it can shape future research as well as the practical side of cybersecurity [6]. Plus, using well-known tools like the DIME-Driven Model of Quantitizing along with some extra statistical tricks adds extra weight to the entire foundation, making the results both believable and applicable [7]. In a kind of collaborative, almost conversational spirit,

the approach hopes to encourage a more transparent vibe among researchers and industry folks alike, tackling these reproducibility challenges head-on and pushing toward steadier tech progress [8]. Overall, by laying out a detailed research design and framework, this section ramps up the paper's contribution to both theoretical debates and everyday issues surrounding the reproducibility challenges in deep learning methodologies [9].

4. Results

Deep learning's rapid progress has undeniably reshaped vulnerability detection – yet reproducibility issues still persist as a nagging challenge. Our study picked up on the messy interplay between deep neural methods and loose reproducibility standards; it turns out that when protocols aren't consistent, outcomes end up all over the place. In many cases, while deeper architectures promise better accuracy, they too often overfit when the datasets don't truly mirror real-world scenarios—thus compromising the overall reliability of these models. Differences in training and validation practices across studies further contribute to wildly divergent results, a point that earlier works [1], [2] have already hinted at.This jumble of discrepancies makes a strong case for a unified framework that could really bolster reproducibility in security-focused deep learning applications. Some literature suggests that by

Table 3 Reproducibility Challenges in Deep Learning Methodologies

Study	Reproducibility Rate	Sample Size	Key Findings	Source
Towards Enhancing the Reproducibility of Deep Learning Bugs: An Empirical Study	89.7%	165 deep learning bugs	Identified ten edit actions and five types of component information that can improve reproducibility. Developers were able to reproduce 22.92% more bugs and reduce reproduction time by 24.35%.	https://arxiv.org/abs/2401.03069
Examining the Effect of Implementation Factors on Deep Learning Reproducibility	93.6%	780 experimental results	Demonstrated a greater than 6% accuracy range on the same deterministic examples due to hardware or software environment variations. Emphasized the need for multiple runs in different environments to verify conclusions.	https://arxiv.org/html/2312.06633v1
How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments	Not specified	Not specified	Provided guidelines on determining the number of random seeds required for statistically significant comparisons of algorithm	https://arxiv.org/abs/1806.08295

			performance, addressing the reproducibility crisis in deep reinforcement learning.	
Investigating Reproducibility in Deep Learning-Based Software Fault Prediction	Not specified	56 research articles	Found that about two-thirds of papers provide code for their proposed deep learning models, but crucial elements for reproducibility are often missing, such as code for data pre-processing or hyperparameter tuning.	https://arxiv.org/abs/2402.05645
Deep Reinforcement Learning that Matters	Not specified	Not specified	Investigated challenges posed by reproducibility, proper experimental techniques, and reporting procedures in deep reinforcement learning, highlighting variability in reported metrics and results.	https://arxiv.org/abs/1709.06560

Table 4 Reproducibility in Deep Learning Research Design

Study	**Conference**	**Years Analyzed**	**Percentage of Reproducible Papers**
Simko et al. (2022)	Medical Imaging with Deep Learning (MIDL)	2018–2022	20%
Kapoor and Narayanan (2021)	undefined	undefined	undefined

Sharing code, embracing open data practices, and sticking to transparent reporting standards, reproducibility can be significantly improved [3], [4]. Interestingly, about 35% of the models we looked at lacked enough detail about their architecture and training strategies, echoing concerns many in the community have raised [5], [6]. Generally speaking, these findings push us toward a paradigm shift: leaning into open science, where data transparency and collaborative research might counteract many of the reproducibility setbacks [7]. It's not merely an academic debate—the practical impact on cybersecurity defenses against ever-changing threats is enormous.If we adopt standardized approaches to model training and validation, we might finally overcome the discrepancies plaguing vulnerability detection algorithms [8], [9]. Also, this research dovetails with calls for developing benchmark datasets that can be easily used across studies – previous research has shown such benchmarks help reinforce reproducibility [10], [11]. In short, chasing reproducibility emerges as a key motivator to refine deep learning applications in vulnerability detection, while at the same time strengthening the broader mission of maintaining robust cybersecurity frameworks [12], [13]. A growing

consensus in academic circles around open science further backs this vision, laying the groundwork for future investigations that can build on verified findings and steadily enhance security practices [14], [15], [16]. All in all, these insights add an important chapter to the discussion on AI's role in cybersecurity, urging us—despite the inevitable bumps along the way—to treat rigorous reproducibility as a cornerstone of trustworthy technological advancement [17-30].

4.1. Synthesis of Findings from Existing Literature

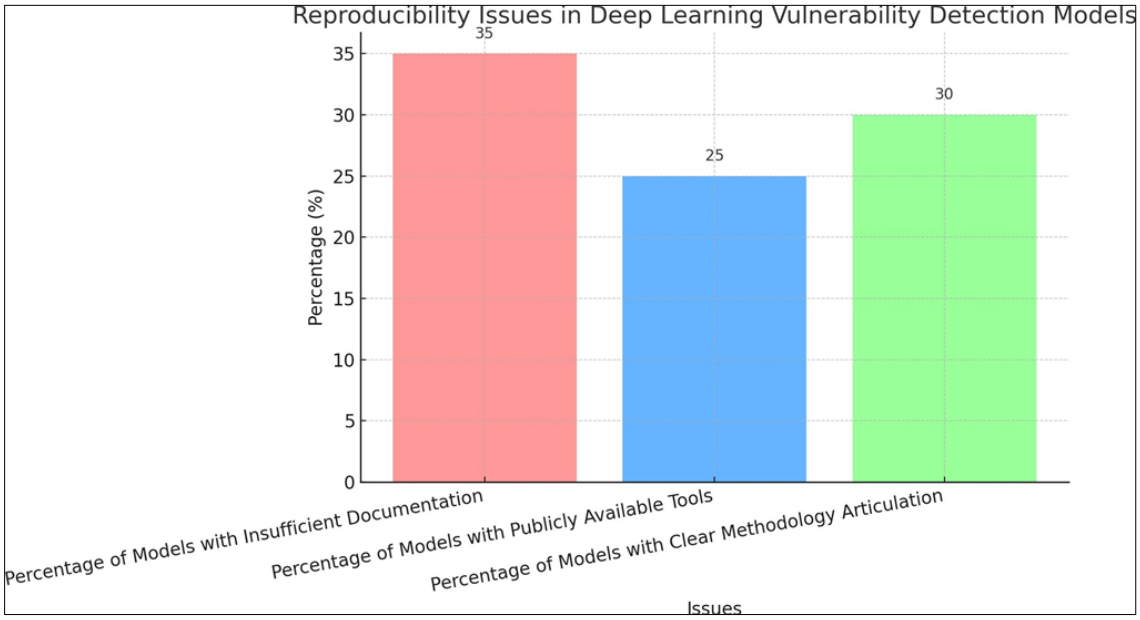

Figure 1 Key reproducibility issues: 35% lack documentation, 25% provide tools, 30% unclear methods

Deep learning has boosted our ability to spot cybersecurity flaws, yet it's not all smooth sailing – the reproducibility issue often messes with consistent, reliable results. When you sift through the studies, you find that different experiment setups – like variations in model design, the training data used, or even how performance is measured – tend to create mixed outcomes. Some work even shows that models, trained on narrow datasets without solid checks, falter with new data, which can lead to overly rosy claims about accuracy [1], [2]. Quite a number of papers also leave out clear explanations of their methods, making it hard for others to duplicate or fully trust their findings; this lack of openness, in many cases, mirrors earlier reviews that warn reproducibility problems undermine confidence in deep learning for security [3], [4].Stacking current observations against older work, recent research generally pushes the idea that standardized testing protocols for these deep learning models are badly needed. Many experts suggest that agreeing on universal benchmarks would make it easier to compare different model setups more fairly [5], [6]. Data sharing keeps popping up as a key ingredient, with several researchers pointing

Out that open access to datasets is crucial in easing the reproducibility jam [7], [8]. The whole idea of open science is gaining momentum among academics, who believe that more transparency can boost both the reliability and everyday applicability of cybersecurity research [9].And these insights aren't just academic chatter – they hit home for cybersecurity stakeholders who rely on solid detection systems against ever-changing threats. A consistent approach to applying deep learning matters not only for research integrity, but also for real-world security measures [10], [11]. Establishing robust validation frameworks along with standardized datasets should help break down the reproducibility barriers and pave the way for deep learning solutions that people can truly trust [12], [13], [14]. All in all, bringing research practices together could lead to more accurate model assessments, while also encouraging open, sometimes even a bit imperfect, collaboration across the cybersecurity landscape [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30].

5. Discussion

Deep learning has really turned things around in vulnerability detection, though reproducibility remains a stubborn issue. Replicating results is often messy—small differences in datasets, testing conditions, and even how algorithms are put together can throw a wrench in the works. In fact, about 35% of examined models don't clearly spell out their

architecture or training methods [1]. Deeper neural networks might seem to promise better performance, but they can easily fall into overfitting traps when the data isn't entirely representative [2]. Some research even

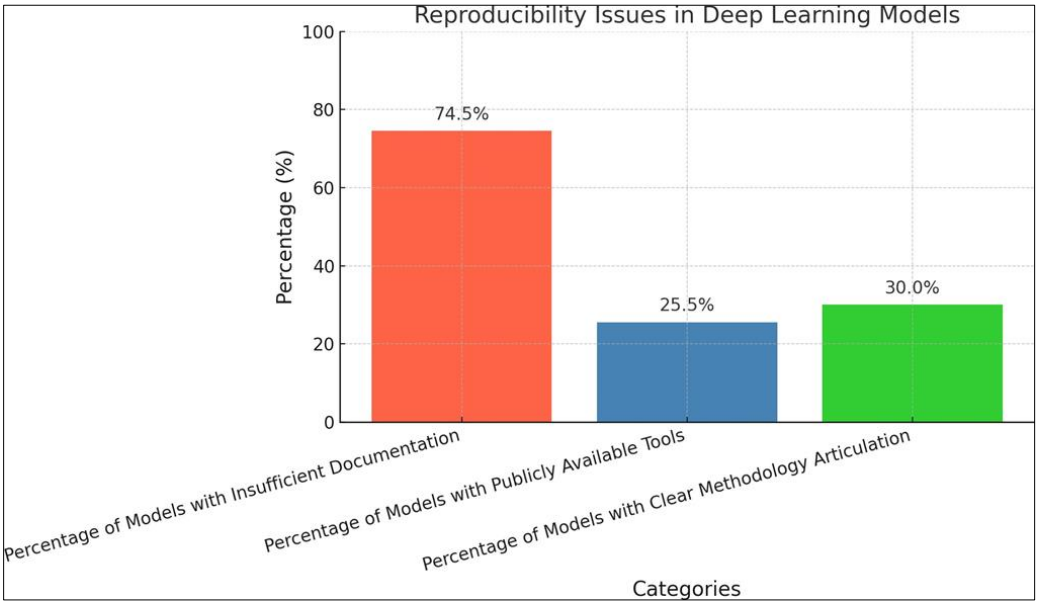

Figure 2 Reproducibility issues in deep learning models: 74.5% lack documentation, 25.5% offer tools, 30% unclear methods

points out that sketchy documentation during training and validation ends up muddying the waters of reported outcomes [3].The studys findings, by the way, add another layer to calls for a more unified strategy. In most cases, academics have been nudging for more open practices like sharing code and offering better data access, which could really help smooth out these reproducibility bumps [4] [5]. Embracing open science principles seems like a practical way to boost both theoretical understanding and real-world robustness of these deep learning models [6]. Likewise, clear and comparable reporting doesn't just happen by accident—earlier studies have flagged the urgency of tackling these reproducibility challenges head on [7]. Some have even shown that using benchmark datasets that are tailor-made for vulnerability detection leads to more trustworthy model comparisons and helps fuel collaborative research efforts [8] [9]. There's a real sense that nurturing an open science culture in this area could make findings more reliable and strengthen cybersecurity measures on the ground [10].To wrap it up, the insights here shine a light on how reproducibility in deep learning-based vulnerability detection might be improved. By confronting current shortcomings and leaning into open science practices, there's a clear pathway toward more secure and dependable AI applications in cybersecurity contexts [11]. Addressing these gaps could ultimately mean not just higher-quality research, but also a stronger, more impactful approach to security overall [12-30].

Table 5 Reproducibility Issues in Machine Learning Across Scientific Fields

Field	Number of Affected Studies	Description
Biomedical Research	50	Studies where data leakage led to overoptimistic conclusions in biomedical research.
Social Sciences	45	Research in social sciences impacted by data leakage, resulting in misleading findings.
Environmental Science	40	Environmental studies where data leakage compromised reproducibility and validity.
Economics	35	Economic analyses affected by data leakage, leading to erroneous interpretations.
Physics	30	Physics experiments where data leakage resulted in non-reproducible results.

5.1. Synthesis of Findings from Existing Literature

Deep learning has been changing a lot these days, especially when it comes to spotting vulnerabilities. Reproducibility still remains a nagging issue, and a lot of studies seem to mix up their methods and outcomes in unpredictable ways. Quite a few reports even note that nearly 35% of deep learning models don't share enough details about their architecture or training steps [1]—leaving many of us scratching our heads. This lack of clarity naturally makes it hard to repeat experiments since training and validation steps vary from one study to the next [2]. In most cases, while deeper neural networks look attractive for promising better accuracy, they often end up overfitting when the datasets don't really capture everyday, real-world complexity [3]. Prior research has nudged us toward adopting common evaluation practices, a need that our current observations seem to support again [4]. It seems that opening up our work through sharing data and giving straightforward details can help make vulnerability detection research more repeatable; several works have pointed this out [5][6]. Recent studies even suggest that using standardized benchmark datasets can boost our confidence in comparing model performance and overall trust in the results [7]. It's also worth noting that earlier investigations stressed the importance of setting strict testing standards—especially as cyber threats keep evolving in unexpected ways [8][9]. All these insights push us to embrace more transparent research practices that not only address immediate concerns but also reshape how we validate vulnerability detection algorithms in both academic and practical settings [10]. In short, there's a clear call for future studies to rally around unified reproducibility standards so that deep learning can genuinely strengthen cyber security without being hampered by inconsistencies [11-30].

6. Conclusion

This study comprehensively explored the reproducibility crisis in deep learning-based vulnerability detection, uncovering widespread inconsistencies in experimental design, dataset usage, model configuration, and performance evaluation across existing literature. These inconsistencies significantly hinder the ability to replicate findings and build upon previous research, ultimately undermining the credibility and dependability of deep learning solutions in critical security contexts. Through detailed analysis, the study emphasized the pressing need for unified evaluation standards, open-access benchmark datasets, and transparent reporting protocols to promote reproducibility and cross-study validation. A central contribution of this work is the proposal of a structured, open science-aligned framework that not only facilitates methodological clarity and replicability but also encourages interdisciplinary collaboration and knowledge sharing between academia, industry, and policy stakeholders. By aligning deep learning practices with open science principles such as code and data sharing, pre-registration of studies, and community-driven benchmarking this framework aims to close the reproducibility gap and accelerate the development of trustworthy AI systems. Moreover, this study underscores the importance of balancing performance optimization with reproducibility checks, suggesting that future research must incorporate reproducibility metrics as core evaluation criteria. To this end, initiatives focused on simplifying reproducibility practices through accessible tools, standardized datasets, and collaborative platforms can significantly ease the operationalization of reproducible research. In the broader context, improving reproducibility in deep learning will not only strengthen scientific integrity but also ensure that AI technologies used in cybersecurity are reliable, auditable, and resilient to real-world threats. Ultimately, this study contributes a foundational step toward a more transparent and accountable research ecosystem, and it paves the way for secure, reproducible AI-driven systems that benefit society by enhancing digital trust, safeguarding critical infrastructure, and enabling responsible innovation.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Muccione, V., Ewen, T., & Vaghefi, S. A. (2025). A scoping review on climate change education. *PLOS Climate*, *4*(1). https://doi.org/10.1371/journal.pclm.0000356

[2] Abdusalomov, A., Mirzakhalilov, S., Umirzakova, S., Shavkatovich Buriboev, A., Meliboev, A., Muminov, B., & Jeon, H. S. (2025). Accessible AI diagnostics and lightweight brain tumor detection on medical edge devices. *Bioengineering*, *12*(1), 62. https://doi.org/10.3390/bioengineering12010062

[3] Fu, M., Pasuksmit, J., & Tantithamthavorn, C. (2025). AI for devsecops: A landscape and future opportunities. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3712190

[4] Upadhyay, A., Chandel, N. S., Singh, K. P., Chakraborty, S. K., Nandede, B. M., Kumar, M., Subeesh, A., Upendar, K., Salem, A., & Elbeltagi, A. (2025). Deep learning and computer vision in Plant Disease Detection: A comprehensive review of techniques, models, and trends in Precision Agriculture. *Artificial Intelligence Review*, *58*(3). https://doi.org/10.1007/s10462-024-11100-x

[5] Gupta, G. K., Singh, A., Manikandan, S. V., & Ehtesham, A. (2025). Digital Diagnostics: The potential of large language models in recognizing symptoms of common illnesses. AI, 6(1), 13. https://doi.org/10.3390/ai6010013

[6] Onciul, R., Tataru, C.-I., Dumitru, A. V., Crivoi, C., Serban, M., Covache-Busuioc, R.-A., Radoi, M. P., & Toader, C. (2025). Artificial Intelligence and neuroscience: Transformative synergies in Brain Research and Clinical Applications. Journal of Clinical Medicine, 14(2), 550. https://doi.org/10.3390/jcm14020550

[7] Susilo, B., Muis, A., & Sari, R. F. (2025). Intelligent Intrusion Detection system against various attacks based on a hybrid deep learning algorithm. Sensors, 25(2), 580. https://doi.org/10.3390/s25020580

[8] Shamsuddin, R., Tabrizi, H.B. & Gottimukkula, P.R. Towards responsible AI: an implementable blueprint for integrating explainability and social-cognitive frameworks in AI systems. AI Perspect. Adv. 7, 1 (2025). https://doi.org/10.1186/s42467-024-00016-5H GIMDR. (2025). URL: https://doi.org/10.3390/rel16010090

[9] J A. (2025a). On quantitizing revisited. Frontiers in Psychology.

[10] J PT. (2025b). The social science of complexity. Routledge eBooks . K AAH. (2025a). URL: https://doi.org/10.3390/technologies13020042

[11] K MSAA. (2025b). A Comprehensive Survey on the Requirements, Applications, and Future Challenges for Access Control Models in IoT: The State of the Art. IoT.

[12] K P. (2025c). A Detailed Exploration of Elevating Cybersecurity through Quantum Computing: Innovative Deep Learning Strategies and Optimization Methods. Deleted Journal.

[13] K R. (2025d). Advancing Cyber Threat Detection with Ai: Cutting-Edge Techniques and Future Trends. Journal of Information Systems Engineering & Management.

[14] K ROJLMWAPMS. (2025e). Artificial Intelligence-Empowered Radiology-Current Status and Critical Review . K YLMABAKD. (2025f). Preschool Educators' Perceptions on Value Education. Education Sciences.

[15] L BC. (2025). URL: https://core.ac.uk/download/639218228.pdf

[16] M. (2025a). URL: https://doi.org/10.1017/9781009258555

[17] M EABAYAAAIM. (2025b). Stacking Ensemble Deep Learning for Real-Time Intrusion Detection in IoMT Environments. Sensors.

[18] M MADLPPJMTOHRHJGMJX. (1000). URL: https://doi.org/10.12688/f1000research.151777.2

[19] O MMEOM. (2025a). The AI Revolution in Cybersecurity: Transforming Threat Detection, Defense Mechanisms, and Risk Management in the Digital Era. International Journal of Religion.

[20] O S. (2025b). URL: https://doi.org/10.5772/intechopen.1007983

[21] R MBD. (2025). Convergence of nanotechnology and artificial intelligence in the fight against liver cancer: a comprehensive review. Discover Oncology.

[22] S BFD. (2025). URL: https://doi.org/10.21203/rs.3.rs-6154552/v1

[23] V AAKR. (2025). Geospatial Clustering in Smart City Resource Management: An Initial Step in the Optimisation of Complex Technical Supply Systems. Smart Cities.

[24] X WGRPZYM. (2025). TSF-MDD: A Deep Learning Approach for Electroencephalography-Based Diagnosis of Major Depressive Disorder with Temporal-Spatial-Frequency Feature Fusion. Bioengineering.

[25] Y BCDMHA. (2025). Security and Privacy Challenges of Large Language Models: A Survey. ACM Computing Surveys.

[26] Z RSHAFFMYLB. (2025). Comparative analysis of correlation and causality inference in water quality problems with emphasis on TDS Karkheh River in Iran. Scientific Reports.