RCHUB JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 1, ISSUE 2, OCTOBER 2025
RCHUB PUBLISHER
ISSN: 3107-8168

# INTERPRETABILITY, EXPLAINABILITY AND CAUSALITY IN DEEP LEARNING

***[1]Dr.Seshaiah Merikapudi***
Associate Professor, Department of CSE, SJC Institute of Technology, Chickaballapur, India
Email: merikapudi@gmail.com
***[2]Chinmyi T C***
Department of CSE, SJC Institute of Technology, Chickaballapur, India
Email: chinmyitc19@gmail.com
***[3]Amulya A***
Department of CSE, SJC Institute of Technology, Chickaballapur, India
Email: amulyaamu57@gmail.com
***[4]Aqsa Firdose***
Department of CSE, SJC Institute of Technology, Chickaballapur, India
Email: aqsafirdose786@gmail.com
***[5]Deeksha S***
Department of CSE, SJC Institute of Technology, Chickaballapur, India
Email: dss910835@gmail.com

## Abstract

*Despite revolutionizing sectors like medicine, finance, natural language processing, and autonomous technologies, the intricate and non-transparent designs of deep learning architectures restrict confidence and clarity. Existing post-hoc explanation methods (e.g., saliency maps, SHAP, and counterfactuals) yield insights that are often incomplete and inconsistent, frequently uncovering statistical correlations instead of definitive causal factors. Consequently, establishing interpretability and causality is now crucial for successfully deploying AI in critical applications. Causal inference provides a more solid framework by identifying authentic cause-and-effect relationships and boosting model reliability when data distributions change. Nevertheless, incorporating robust causal structures into neural networks is challenging due to data availability and high computational demands. This paper examines prominent interpretability and causal methodologies, discusses their drawbacks, and proposes future research avenues aimed at engineering deep learning models that are simultaneously precise, transparent, and trustworthy.*

***Keywords:*** *Deep Learning; Explainable AI (XAI); Interpretability; Explainability; Causality; Post-hoc Explanations; Causal Inference; Model Transparency; Robust AI.*

## 1. Introduction

### 1.1 The Ascendancy and Opacity of Deep Learning

Deep learning has emerged as one of the most influential branches of artificial intelligence, driving successful applications across domains like computer vision, natural language processing, healthcare, finance, and robotics. This widespread success stems from the capacity of deep neural networks to automatically learn highly complex representations directly from raw data, often surpassing traditional machine learning models in terms of predictive accuracy and scalability. Despite these significant advancements, a critical limitation persists: the majority of deep learning models function as black boxes, providing high-confidence predictions without transparently revealing the reasoning behind them. As AI systems increasingly penetrate and influence high-stakes domains—including medical diagnosis, legal decision-making, autonomous driving, and financial risk assessment—the imperative for transparency has intensified.

### 1.2 Defining the Foundational Concepts: Interpretability and Explainability

Addressing this challenge necessitates foundational work in interpretability and explainability. Interpretability refers to the extent to which a human can understand the internal mechanics and operational principles of a model, such as how specific input features contribute to the final outputs (Figure 1).

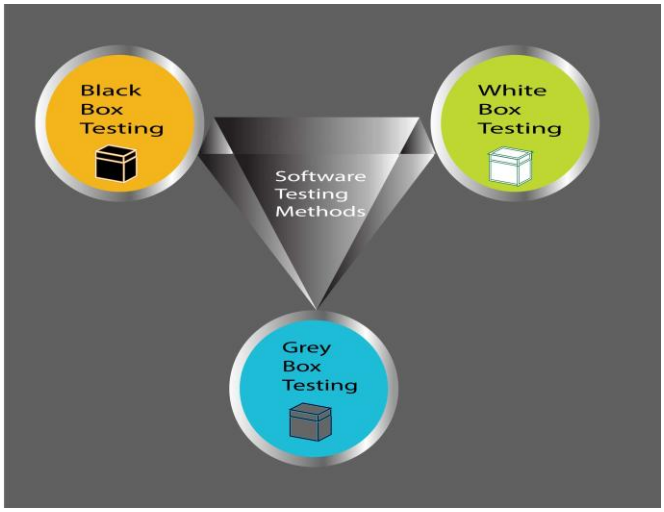


Figure 1.The Transition from Black-Box to Transparent (White-Box) AI Models.

Explainability, conversely, focuses on the generation of human-comprehensible descriptions and justifications detailing why a particular model result was produced. Both concepts are paramount for ensuring accountability, promoting fairness, cultivating user trust, and meeting growing regulatory compliance requirements. The absence of meaningful explanations leads to user reluctance in adopting AI systems and impedes developers' ability to effectively diagnose biases or rectify systemic model failures.

### 1.3 Limitations of Current eXplainable AI (XAI) Techniques

Current eXplainable AI (XAI) methodologies are often constrained by inherent weaknesses. Post-hoc explanation methods—including saliency visualizations, feature importance metrics, and local surrogate models—primarily offer approximations of the model's behavior rather than faithful, direct insights. These techniques can often prove unstable, potentially misleading, or inconsistent when subjected to minor variations in input data. Such limitations introduce substantial concerns regarding the fidelity and robustness of explanations, especially within sensitive applications where misinterpretation can lead to severe, real-world consequences.

### 1.4 The Crucial Gap: Moving from Correlation to Causality

A more profound conceptual challenge is that deep learning inherently focuses on capturing statistical correlations, not genuine causal relationships. Models that lack true causal reasoning may achieve high performance on training datasets but are highly vulnerable to failure during distribution shifts or when encountering novel real-world conditions. This critical gap highlights the growing importance of integrating principles of causal inference with deep learning to explicitly uncover cause-and-effect structures, thereby enhancing generalization capabilities, improving robustness, and enabling counterfactual reasoning (i.e., answering hypothetical "What would happen if...?" questions). Causality is the key element required to advance AI systems from mere pattern recognition toward genuine reasoning and informed decision-making.

### 1.5 Research Focus and Paper Contribution

Given these challenges and emerging opportunities, the research domain encompassing interpretability, explainability, and causality has gained significant prominence. Researchers are dedicated to the development of models that are not only highly accurate but also transparent by design and intrinsically capable of causal understanding. This paper aims to provide a comprehensive analysis by exploring the theoretical foundations, current state-of-

the-art techniques, existing limitations, and critical future research directions that define this rapidly evolving field. Ultimately, the development of deep learning models that are both highly interpretable and causally grounded is indispensable for constructing trustworthy AI systems that are safe, ethical, and aligned with human expectations and societal needs.

## 2. Literature Survey

In recent years, the creation of smart assistive devices has garnered significant interest, driven by the increasing need for both elder care and support for mobility. Consequently, a variety of researchers have put forward concepts for intelligent walking aids that combine sensors, navigation tools, and health monitoring systems, all designed to boost user safety and autonomy.

One particular smart cane design incorporated ultrasonic sensors to identify nearby obstructions and then notified the user through audible signals. However, this specific system did not offer positioning features, which limited its practical use in outdoor environments. Subsequently, another research initiative unveiled a GPS-enabled walking assistant, allowing caregivers to remotely track the user's whereabouts, but this device notably omitted any real-time health monitoring functionalities. To address concerns regarding an individual's physiological safety, a separate project developed a wearable Internet of Things device. This innovation could accurately measure body temperature and heart rate, transmitting the data to a cloud-based server. Despite its precision, this particular system was not directly integrated into a mobility aid, which ultimately made it less convenient for elderly users.

Recent tech leaps in the Internet of Things (IoT) and embedded systems mean we can now pack lots of features into small, affordable devices. For example, some previous studies have explored this potential. Research has shown that using multiple sensors, like ultrasonic and infrared, makes it much easier to detect obstacles for safer mobility. Other work highlighted that wearable health gadgets need to be low-power and comfortable (ergonomic) to be truly useful.

The Medico Stick takes these ideas and brings them all together. Instead of just focusing on helping someone move or just monitoring their health, our proposed system offers a complete, single, and cost-effective solution. It combines obstacle detection, GPS tracking, and real-time vital sign monitoring in one smart device. Essentially, the Medico Stick provides a holistic

solution designed to significantly improve safety, accessibility, and health management for seniors and people with physical challenges.

## 3. Problem Statement

Despite achieving exceptional results across diverse sectors, deep learning models are predominantly opaque, complicating the understanding of how and why their decisions are reached by users, developers, and regulators. This fundamental lack of interpretability and explainability creates significant issues in critical, high-stakes environments—such as medical care, financial systems, legal contexts, and autonomous technologies—where biased or erroneous predictions can result in severe negative outcomes.

Current post-hoc explanation methods (like gradient visualizations, feature importance tools, and simplified surrogate models) are insufficient. They frequently yield insights that are incomplete, unstable, or misleading, failing to accurately represent the model's true decision process. Furthermore, most deep learning architectures are driven primarily by correlation, lacking genuine causal understanding. This vulnerability exposes them to distribution shifts, reliance on spurious patterns, and potential adversarial attacks.

The central difficulty is that existing deep learning frameworks are not intrinsically designed to support transparent, faithful, and causally grounded decision-making. Consequently, stakeholders cannot confidently evaluate model conduct, verify output accuracy, identify biases, or guarantee fairness and accountability. This disparity between powerful performance and deficient interpretability severely obstructs trust, deployment, and regulatory acceptance.

**Research Problem:** Therefore, the research problem this paper seeks to address is, What novel deep learning models and integrated methodologies can be developed to provide intrinsic interpretability, generate robust and faithful explanations, and incorporate causal reasoning to ultimately achieve transparency, trustworthiness, and reliable performance in real-world applications?

## 4. Objectives

The overarching aim of this academic inquiry is to systematically map and evaluate the conceptual hurdles, methodological advancements, and future imperatives for developing interpretable, explainable, and causally-aware deep learning frameworks. To realize this principal goal, the study will pursue the following specific targets:

### 4.1 Characterization of Model Opacity

- To ascertain the systemic limitations inherent in prevailing deep learning paradigms concerning transparency and cognitive grasp.
- This involves detailing the structural causes of their non-transparent operation (the 'black-box' phenomenon) and charting the consequential risks posed in safety-critical, applied contexts.

### 4.2 Methodological Critique of Post-Hoc XAI

- To perform a critical analysis of established post-hoc techniques for enhancing explainability within deep neural networks.
- The purpose is to determine the validity, temporal stability, and deficiencies of methods such as saliency mapping, feature contribution analysis (SHAP, LIME), and counterfactual generation.

### 4.3 Investigation of Causal System Integration

- To explore the necessity of integrating causal inference as a mechanism for reinforcing the dependability and trustworthiness of AI systems.
- This objective examines strategies for embedding explicit causal models and related frameworks into neural architectures to achieve robust generalization and verifiable decision fidelity.

### 4.4 Identification of Research Imperatives

- To delineate the salient research gaps and unresolved technical difficulties impeding the generation of deep model explanations that are simultaneously veridical, temporally consistent, and practically informative.
- This includes scrutinizing challenges related to explanation fidelity verification, ensuring resistance to minor data perturbations, standardizing evaluation metrics, and managing the fundamental trade-off between predictive power and model clarity.

### 4.5 Review of Inherently Transparent Designs

- To survey and analyze novel architectural solutions engineered to possess intrinsic interpretability or to incorporate explicit causal structure from the outset.
- This encompasses the examination of designs like modular networks, selective sparsity models, deep attention mechanism inspection, and hybrid causal-DL architectures.

### 4.6 Formulation of Future Research Pathways

- To establish a comprehensive set of proposals for guiding future academic inquiry toward the creation of deep learning systems that are fully transparent, rigorously explainable, and causally grounded.
- This is intended to highlight the most promising technological trajectories and theoretical advances that should inform subsequent scholarly efforts.

### 4.7 Validation of Domain Relevance

- To assess the practical imperative of interpretability and causality across key professional and societal sectors, including healthcare, financial services, legal compliance, and autonomous technologies.
- This ensures that the research focus remains strongly aligned with real-world application and addresses documented needs for risk mitigation and ethical governance..

## 5. Scope and Limitations

### 5.1 Scope of the study

This investigation centers on enhancing both the transparency and causal comprehension of deep learning systems. It scrutinizes prominent deep neural network designs—including CNNs, RNNs, LSTMs, GANs, Transformers, and composite multimodal architectures—alongside an assessment of common interpretability techniques. These include feature

visualization methods (e.g., saliency maps, Grad-CAM), local/global attribution tools (LIME, SHAP), counterfactual analysis, and concept-driven explanations.

Furthermore, the research details how causal inference frameworks—such as Structural Causal Models (SCMs), causal graphical models, and causal representation learning—can be merged with neural networks to bolster model resilience and dependability.

The analysis incorporates high-consequence application domains like clinical care, financial services, autonomous navigation, legal auditing, and cyber defense, where model clarity is non-negotiable. Finally, the scope encompasses defining key research gaps pertaining to explanation fidelity, robustness, fairness, and overall system transparency, along with examining the ethical, legal, and societal dimensions of AI governance and accountability.

### 5.2 Limitations of the study

This research is constrained by several factors inherent to its design and field of study:

1. Non-Empirical Approach: The study is strictly theoretical, relying exclusively on a literature review without involving the practical development of novel models or conducting experimental validation.

2. Field Dynamism: As both eXplainable AI (XAI) and causal AI are developing rapidly, the synthesis may not encompass techniques that emerge immediately following the study's completion.

3. Data Access: Evaluation of advanced methodologies is hampered by restricted access to proprietary models utilized within major industry firms.

4. Subjectivity in Assessment: The qualitative and subjective nature of measuring explanation efficacy (usefulness to a human user) makes establishing universal, objective comparison standards difficult.

5. Causal Complexity: The deep mathematical and computational complexity of integrating causal inference with deep learning prevents an exhaustive treatment of every possible hybrid approach.

6. Data Type Restriction: Causal analysis is primarily restricted to observational data due to the widespread difficulty in obtaining required interventional datasets.

## 6. Methodology

This study utilizes a structured, exploratory qualitative research design based solely on synthesizing existing academic literature to analyze interpretability, explainability, and causality in deep learning models. The entire methodological flow, from data acquisition to the final outcomes, is visually summarized in Figure 2.

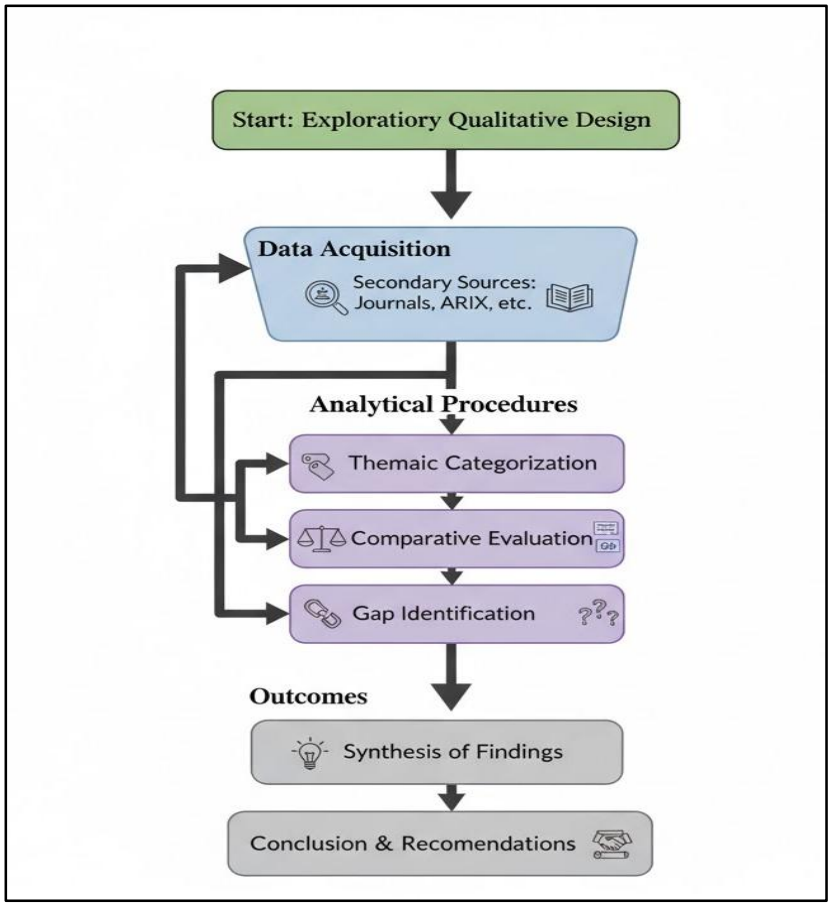


Figure 2. Methodology Flowchart for the Analysis of Interpretability and Causality in Deep Learning

### 6.1 Data Acquisition and Scope

- **Information Sources:** Data is collected exclusively from secondary sources, including peer-reviewed journals (e.g., IEEE, ACM, Springer), academic books, arXiv preprints, and AI governance documents.

- **Selection Criteria:** Sources are selected based on direct relevance to interpretability/causality, recency (primarily 2016–2025), theoretical significance, and practical applicability.

### 6.2 Analytical Procedures

The analysis employs a three-part structured approach, as depicted in the central "Analytical Procedures" section of Figure 2:

- **Thematic Categorization:** Literature is grouped into key themes, such as specific interpretability techniques, causal frameworks, and explanation reliability challenges.
- **Comparative Evaluation:** Intrinsic vs. post-hoc methods and correlational vs. causal models are compared using metrics like fidelity, stability, scalability, and practical usefulness.
- **Gap Identification:** Research deficiencies are pinpointed in areas including the algorithmic limitations of XAI tools, the lack of standardized metrics, and difficulties in causal integration and deployment.

### 6.3 Study Constraints and Outcomes

- **Limitations:** This research is non-empirical (no model development/testing) and is constrained by the rapid evolution of XAI, limited access to proprietary tools, and the subjective nature of evaluating explanation quality.
- **Ethical Compliance**: Academic integrity is maintained through proper citation, and the study promotes fairness, accountability, and transparency principles.
- **Expected Results:** The study aims to deliver a structured overview of interpretability and causal challenges, a critical evaluation of XAI methods, and recommendations for future research to build trustworthy AI systems, culminating in the "Outcomes" section shown in Figure 2.

## 7. Results

The systematic review of literature on interpretability, explainability, and causality in deep learning yielded several crucial findings, organized into three core thematic areas, as summarized in Figure 3:

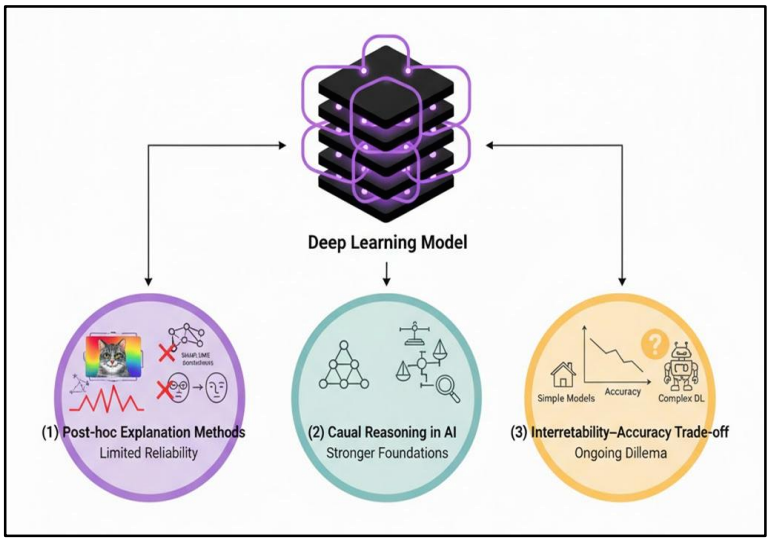


Figure 3. Visual Summary of Key Outcomes in Transparent Deep Learning

1. Reliability Deficits in Post-hoc Explanation Methods
2. Robustness Gains Achieved via Causal Reasoning
3. The Persistent Interpretability–Accuracy Conundrum

Figure 3 provides a conceptual overview, depicting the central role of the deep neural network and illustrating the three main thematic outcomes established by the literature analysis. The pathways originating from the core model structure represent the validated categories of findings detailed below.

### 7.1 Detailed Analysis of Outcomes

**1. Post-hoc Explanations: Widespread Use, Limited Consistency:**

Evidence indicates that the majority of contemporary eXplainable AI (XAI) techniques (e.g., saliency maps, SHAP, LIME) offer only approximate insights into the model's decisions.

**Key Issues Identified:**

- **Method Dependence:** Explanations frequently differ substantially when generated by distinct analytical methods.
- **Fragility:** Even minor input perturbations can induce dramatic, unstable shifts in the resulting explanations.
- **Correlation over Causation:** Many techniques prioritize visual appeal over a basis in genuine causal factors.

These findings confirm that while popular, post-hoc tools are insufficient to fully disclose the model's true internal reasoning.

### 2. Causal Reasoning: A Stronger Foundation for Reliability

The investigation reveals that incorporating causal frameworks (e.g., Structural Causal Models or Causal Representation Learning) directly addresses many reliability weaknesses inherent in black-box models.

**Key Benefits Demonstrated:**

- **Generalization:** Causal frameworks facilitate superior generalization performance under data distribution shifts.
- **Spurious Correlation**: Augmenting deep networks with causal graphs successfully reduces reliance on irrelevant statistical associations.
- **Human Alignment:** Counterfactual reasoning generates explanations that more closely align with human logic.

Embedding causal principles is shown to significantly bolster the trustworthiness and systemic reliability of deep learning solutions.

### 3. The Interpretability-Accuracy Trade-off Remains Acute

As illustrated conceptually in Figure 2, models designed for inherent clarity (simple prototype or rule-based designs) tend to underperform on complex tasks. Conversely, the most predictive black-box models (e.g., Transformers, deep CNNs) inherently lack transparency.

**Specific Findings:**

- **Performance Reduction:** Efforts to increase model interpretability frequently correlate with a corresponding reduction in predictive accuracy.
- **Hybrid Potential:** Hybrid architectural designs, integrating sparse connectivity or explicit causal constraints, show potential for mitigating this trade-off.

**4. Need for Standardized Evaluation Metrics**

The literature highlights a major research deficiency: the AI community currently lacks universal, quantifiable metrics for rigorously assessing the quality and correctness of explanations.

**Evidence suggests:**

- **Faithfulness**: Quantifying an explanation's accuracy to the model's true function remains methodologically challenging.
- **Context Dependence:** Different application sectors necessitate distinct criteria for acceptable explanations.
- **Subjectivity:** Many evaluation protocols rely heavily on subjective human assessment.

**5. Deployment Requires Context-Specific Explanations**

Operational systems necessitate customized explanatory outputs based on domain requirements:

- Healthcare demands explanations rooted in causal factors and counterfactual scenarios.
- Finance requires stringent interpretability for regulatory compliance and auditability.
- Autonomous systems require real-time, instantaneous explanations for safety.

The practice of explainability must be meticulously tailored to align with specific domain constraints.

### 7.2 Synthesis of Outcomes

The findings confirm that while post-hoc explanation methods are dominant, they exhibit substantial weaknesses in reliability and causal rigor. Causal reasoning and transparent-by-design architectures represent essential conceptual advancements, though their broad adoption is limited by complexity and scalability. These results collectively reinforce the need for a new generation of deep learning solutions that harmoniously merge high performance, transparency, and robust causal understanding.

## 8. Future Scope

The escalating demand for transparency in artificial intelligence systems reveals several high-priority and promising avenues for future research in interpretability, explainability, and causality within deep learning:

### 8.1 Model Design and Architecture

- **Engineering Inherently Interpretable Deep Models:** Future work must prioritize the creation of neural architectures that possess built-in transparency, thereby minimizing the reliance on external post-hoc analysis tools.
- **Advancing Causal Deep Learning Integration:** Research should concentrate on developing scalable techniques to successfully embed causal inference mechanisms into contemporary deep models, facilitating superior generalization and consistently robust decision-making when confronted with distribution shifts.
- **Improved Causal Discovery Techniques**: Progress in algorithms capable of accurately identifying and learning causal structures from imperfect, noisy, or scarce data will be crucial for implementing reliable causal reasoning in deep learning.

### 8.2 Standardization and Validation

- **Establishing Standardized Metrics for Explanation Quality:** There is a vital need to develop and implement universal benchmarks and rigorous evaluation frameworks to objectively quantify explanation fidelity, robustness, and practical usefulness across various application domains.

### 8.3 Application and Deployment

- **Focus on Human-Centered Explainability:** Future systems must generate explanations that are customized to suit the distinct needs and technical understanding of different stakeholders (e.g., domain experts, general public, regulatory bodies), thereby maximizing clarity and practical applicability.

- **Development of Real-Time and Interactive Explanations:** As AI increasingly powers dynamic, time-sensitive systems (such as autonomous vehicles), developing explainability methods that function instantaneously and interactively will become essential for operational safety.

- **Cross-Domain Applicability Validation:** More empirical studies are necessary to rigorously validate the effectiveness and transferability of interpretability and causal approaches across diverse sectors, including medicine, finance, cybersecurity, and education.

### 8.4 Regulatory and Ethical Alignment

- **Integration with Regulatory and Ethical Frameworks:** Given the introduction of emerging AI governance legislation, future technical research must actively seek to align explainability mechanisms with legal mandates concerning fairness, accountability, and transparency.

## 9. Conclusion

This study establishes that high-performance deep learning models are hindered by their inherent lack of transparency, which fundamentally restricts trust and prevents their secure utilization in critical environments. Existing post-hoc explanation methodologies offer only superficial information and frequently fail to capture the genuine decision-making rationale of the underlying model.

Conversely, the adoption of causal reasoning presents a significantly more reliable and robust framework for understanding system behavior; however, embedding causal principles into large-scale deep learning architectures remains a substantial technical and conceptual obstacle.

In summary, the next generation of AI development must prioritize the engineering of systems that are intrinsically interpretable, explainable, and grounded in causality. This unified approach is essential for producing models that are not only powerful in their capability but also demonstrably trustworthy and ethically accountable. Future work must concentrate on creating solutions that successfully reconcile the conflict between predictive accuracy and necessary transparency, ensuring that all AI outputs can be fully justified, rigorously validated, and relied upon by humans.

## 10. References

[1] Lipton, Z. C. (2018). The mythos of model interpretability. Commun. ACM, 61(10), 36–43.

[2] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In Proc. 22nd ACM SIGKDD Int. Conf., 1135–1144.

[3] Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In NeurIPS, 4765–4774.

[4] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In Proc. 34th Int. Conf. ICML, 3319–3328.

[5] Selvaraju, R. R. et al. (2017). Grad-CAM: Visual explanations... via gradient-based localization. In Proc. IEEE Int. Conf. ICCV, 618–626.

[6] Baehrens, D. et al. (2010). How to explain individual classification decisions. J. Mach. Learn. Res., 11, 1803–1831.

[7] Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge Univ. Press.

[8] Schölkopf, B. et al. (2021). Toward causal representation learning. Proc. IEEE, 109(5), 612–634.

[9] Samek, W., Wiegand, T., & Müller, K.-R. (2019). Explainable artificial intelligence... IT Prof., 21(4), 69–77.

[10] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations... and the GDPR. Harvard J. Law Technol., 31(2), 841–887.

[11] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

[12] Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. In Proc. AAAI Conf., 3681–3688.

[13] Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. In Proc. 36th Int. Conf. ICML, 1911–1920.

[14] Bach, S. et al. (2015). On pixel-wise explanations... by layer-wise relevance propagation. PLoS ONE, 10(7), 1–46.

[15] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review... IEEE Trans. Pattern Anal. Mach. Intell., 35(8), 1798–1828.