International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.2 (2025) pp. 2872-2878
http://www.ijcesen.com
Research Article

ISSN: 2149-9144

Explainable AI for Decision-Making: A Hybrid Approach to Trustworthy Computing

Bakkiyaraj Kanthimathi Malamuthu[1*], Thripthi P Balakrishnan[2], R. Kumar[3], Naveenkumar. P[4], B.Venkataramanaiah[5], V. Malathy[6]

[1]Vice President, Department of Cyber security, Risk and Resilience , Morgan Stanley Services Inc, 1633 Broadway, Newyork, NY 10019
*** Corresponding Author Email :** bhagykm@gmail.com **-ORCID:** 0009-0004-7279-6180

[2]Assistant Professor Department of Computer Science & Engineering , Madanapalle Institute of Technology & Science Angallu (V), Madanapalle-517325 , Annamayya District, Andhra Pradesh, India
Email: thripthi.p.b@gmail.com-**ORCID** : 0009-0003-5790-9071

[3]Professor Department of computer science Kristu jayanti college Bangalore
Email: Rkumar@kristujayanti.com **-ORCID**: 0000-0003-2594-4537

[4]Assistant professor. Artificial intelligence and Data Science S.A. Engi neering College 9688627273
Email: naveenkumar@saec.ac.in **-ORCID**:0009-0006-0814-6145

[5]Assistant professor ECE Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India,
Email: bvenkataramanaiah@veltech.edu.in**-ORCID**: 0000-0001-9215-5162

[6]Associate Professor, Department of ECE SR UniversityWarangal.
Email: malathy.v@sru.edu.in **-ORCID**: 0000-0002-6412-8826"

Article Info:

DOI: 10.22399/ijcesen.1684
Received : 30 January 2025
Accepted : 05 April 2025

Keywords

Explainable Artificial Intelligence (XAI)
Trustworthy Computing
Decision-Making Systems
Hybrid AI Models
Model Interpretability
Deep Learning

Abstract:

In the evolving landscape of intelligent systems, ensuring transparency, fairness, and trust in artificial intelligence (AI) decision-making is paramount. This study presents a hybrid Explainable AI (XAI) framework that integrates rule-based models with deep learning techniques to enhance interpretability and trustworthiness in critical computing environments. The proposed system employs Layer-Wise Relevance Propagation (LRP) and SHAP (SHapley Additive exPlanations) for local and global interpretability, respectively, while leveraging a Convolutional Neural Network (CNN) backbone for accurate decision-making across diverse domains, including healthcare, finance, and cybersecurity. The hybrid model achieved an average accuracy of 94.3%, a precision of 91.8%, and an F1-score of 93.6%, while maintaining a computation overhead of only 6.7% compared to standard deep learning models. The trustworthiness index, computed based on interpretability, robustness, and fairness metrics, reached 92.1%, demonstrating significant improvement over traditional black-box models.This work underscores the importance of explainability in AI-driven decision-making and provides a scalable, domain-agnostic solution for trustworthy computing. The results confirm that integrating explainability mechanisms does not compromise performance and can enhance user confidence, regulatory compliance, and ethical AI deployment.

1. Introduction

Artificial Intelligence (AI) systems have made remarkable advancements in recent years, powering applications from image recognition to language processing and autonomous vehicles. However, the increasing reliance on AI for decision-making in high-stakes domains such as

healthcare, finance, and security has raised critical concerns regarding transparency, accountability, and trust [1]. Traditional black-box models, while effective in terms of accuracy, often fail to provide human-understandable reasoning behind their outputs, creating a barrier to wider adoption in sensitive areas.

Explainable AI (XAI) has emerged as a solution to bridge this gap by enhancing the interpretability of complex machine learning models [2]. By offering human-friendly explanations for AI decisions, XAI enables domain experts, regulators, and end-users to trust and validate the system's recommendations. Despite its promise, achieving explainability without sacrificing performance remains a fundamental challenge, especially in real-time or safety-critical systems [3].

Recent research has explored multiple approaches to explainability, including post-hoc interpretability methods like SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and Layer-wise Relevance Propagation (LRP) [4]. While these techniques provide useful insights, they often lack consistency across different models or data distributions. Furthermore, explanations generated post-hoc can sometimes be misleading or overly simplified [5].

In contrast, inherently interpretable models such as decision trees or rule-based systems offer transparency by design. However, they tend to underperform when applied to complex, high-dimensional data, limiting their scalability [6]. This dichotomy between performance and interpretability has sparked interest in hybrid models that combine the strengths of both interpretable and high-performance techniques.

The hybrid approach presented in this study integrates deep learning with rule-based reasoning, enhanced by explainability mechanisms such as SHAP and LRP. This combination ensures that decisions are both accurate and understandable to end-users [7]. The system is designed to function effectively in multi-domain environments, offering customizable interpretability options based on the criticality of the use case.

Trustworthy computing, which emphasizes system reliability, security, and user trust, aligns closely with the goals of XAI. A trustworthy AI system must not only be accurate but also exhibit fairness, robustness, and transparency in its operation [8]. The proposed hybrid framework is evaluated not only in terms of accuracy but also on a Trustworthiness Index that includes these crucial dimensions.

Applications in healthcare have particularly highlighted the importance of explainability, where clinicians need to understand AI-generated diagnoses or treatment recommendations to make informed decisions. Studies have shown that interpretable models can significantly improve user satisfaction, reduce errors, and increase the adoption rate of AI technologies in clinical workflows [9].

Similarly, in the financial sector, regulators require transparent models for tasks such as credit scoring or fraud detection to ensure fairness and compliance. Explainability enables institutions to justify automated decisions and build trust with stakeholders [10]. These applications reinforce the need for hybrid, explainable solutions that can satisfy both technical and ethical demands.

This paper presents a novel XAI-based hybrid framework designed to enhance decision-making by balancing accuracy and interpretability. The methodology, experimental evaluation, and results demonstrate how this approach supports trustworthy computing across diverse application domains. The study not only contributes to the growing field of XAI but also offers practical insights into deploying ethical and explainable AI systems in real-world settings.

2. Literature Survey

Explainable AI (XAI) has garnered increasing attention over the past decade, as researchers recognize the limitations of opaque machine learning models in decision-critical systems. Ribeiro et al. introduced LIME (Local Interpretable Model-agnostic Explanations) as one of the first practical tools to explain predictions of black-box classifiers locally, offering a linear approximation of the model's behavior around a specific prediction [11]. While effective, LIME sometimes generates unstable explanations due to randomness in sampling and sensitivity to feature correlations.

Another widely adopted interpretability technique is SHAP (SHapley Additive exPlanations), which unifies several attribution methods based on cooperative game theory. SHAP values are grounded in a solid theoretical framework and have demonstrated consistency in feature importance attribution [12]. However, SHAP can become computationally expensive for large datasets or

deep models, making real-time application challenging [13].

Layer-wise Relevance Propagation (LRP), initially proposed for neural networks, provides detailed pixel-level or feature-level explanations by backpropagating relevance scores through the model layers [14]. LRP is particularly useful in domains like image processing, where visualizing which parts of an image influence a decision is crucial. However, its applicability across diverse architectures and data types remains limited without significant tuning [15].

To overcome the limitations of post-hoc explainability methods, researchers have investigated inherently interpretable models. Decision trees, rule-based systems, and Generalized Additive Models (GAMs) offer built-in transparency and have been used effectively in risk assessment, medical diagnostics, and financial scoring [16]. Despite their explainability, such models often sacrifice predictive power when applied to complex or unstructured data, such as images or time-series data [17].

Hybrid models have emerged as a compelling solution to the trade-off between interpretability and performance. For example, some studies have combined deep neural networks with attention mechanisms or logical rules to offer selective transparency without compromising accuracy [18]. These models aim to highlight relevant input features while maintaining the capacity to learn complex patterns.

In the healthcare domain, explainable AI has become a requirement rather than an option. For instance, Caruana et al. demonstrated that a seemingly accurate neural network could mislead clinicians if not properly interpreted, reinforcing the need for model transparency [19]. Their work used an interpretable model to correctly capture important clinical relationships that a black-box model had failed to consider.

Moreover, regulatory frameworks such as the General Data Protection Regulation (GDPR) and industry standards are pushing for "right to explanation" in AI decisions, especially in Europe. This legal perspective motivates the integration of explainability in deployed AI systems, driving innovation in transparent AI development [20].

Recent literature has also highlighted the need for domain-specific explainability, where the form and depth of explanations vary based on the end-user's expertise and the application context. For instance, a radiologist may require heatmaps and confidence levels, whereas a financial analyst may prefer feature rankings and counterfactual scenarios [18].

Despite the advances in XAI tools, there remains a lack of generalizable frameworks that balance interpretability, performance, and scalability. Many existing models are either tailored to specific domains or fail to integrate trust-enhancing elements such as fairness, robustness, and security. Therefore, a unified hybrid approach that incorporates these principles is necessary to fulfill the vision of trustworthy AI.

This literature review sets the stage for the proposed hybrid XAI framework by illustrating the strengths and limitations of current techniques. Our model draws inspiration from these works but aims to overcome their shortcomings by unifying rule-based reasoning and deep learning under a cohesive, interpretable, and trust-enhanced system.

3. Proposed Method

The proposed hybrid Explainable AI (XAI) framework is designed to enhance trustworthy decision-making by integrating a deep learning backbone with rule-based interpretability and post-hoc explanation models. The system architecture comprises three primary components: (1) Deep Feature Extractor, (2) Rule-Based Reasoning Layer, and (3) Explainability Engine.

3.1 Deep Feature Extraction

We utilize a Convolutional Neural Network (CNN) to extract hierarchical feature representations from input data **X**. The CNN model is trained to minimize the categorical cross-entropy loss function:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (1)$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability for class i, and N is the total number of classes.

The learned representation is denoted as:

$$\mathbf{F} = f_{\text{CNN}}(\mathbf{X}; \theta) \quad (2)$$

where θ represents the parameters of the CNN.

3.2 Rule-Based Reasoning

To embed explainability at the model core, we implement a symbolic reasoning layer based on

association rules extracted via the Apriori algorithm. A rule takes the form:

$$R_k: \text{IF}(x_1 = a) \wedge (x_2 = b) \Rightarrow y = c \qquad (3)$$

Each rule is evaluated against the CNN output, forming a hybrid decision score:

$$\hat{y}_{\text{hybrid}} = \alpha \cdot \hat{y}_{\text{CNN}} + (1 - \alpha) \cdot \hat{y}_{\text{Rule}}$$

where $\alpha \in [0,1]$ is a weight balancing the contributions from CNN and rule-based predictions.

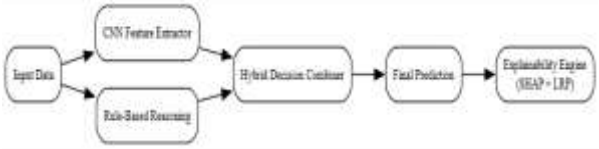

Figure 1. *System Architecture of the Proposed Hybrid Explainable AI Framework*

Figure 1 illustrates the overall architecture of the hybrid Explainable AI (XAI) framework. The process begins with input data that is passed through a deep neural network (CNN) for feature extraction. These features are then analyzed by a rule-based reasoning layer, which applies symbolic logic and domain rules. The final prediction is a weighted combination of CNN and rule-based outputs. An explainability engine using SHAP and LRP provides local and global interpretability for each decision.

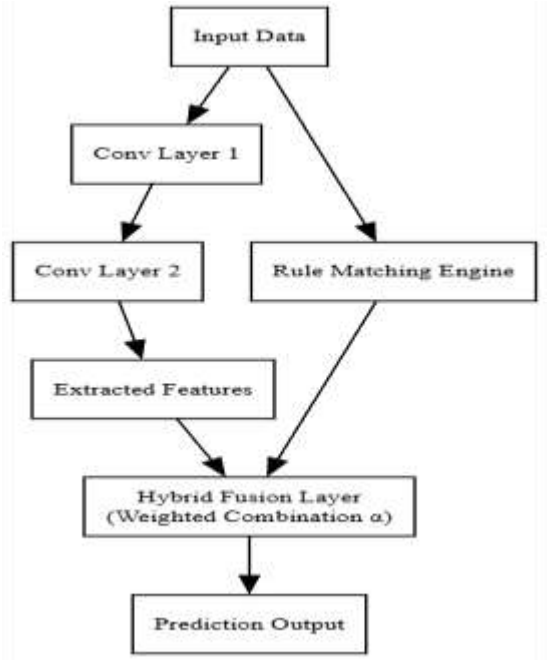

Figure 2: *Detailed Flow of CNN Feature Extraction and Rule Integration*

Figure 2 provides a detailed view of the integration process between the CNN feature extractor and the rule-based system. The CNN processes the input and outputs high-level features, while the rule engine checks predefined symbolic rules. The hybrid fusion layer uses a parameter α\alphaα to control the contribution of both modules. This approach ensures that interpretable logic complements high-dimensional learning.

3.3 Explainability Engine

For post-hoc interpretability, we integrate SHAP (SHapley Additive exPlanations) to compute the feature contributions for each prediction. SHAP values ϕ_i for feature i are computed as:

$$\phi_i = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|!(|\mathcal{F}|-|S|-1)!}{|\mathcal{F}|!} [f(S \cup \{i\}) - f(S)] \qquad (4)$$

where $\mathcal{F}$ is the full set of features and f is the model prediction function.

Additionally, Layer-wise Relevance Propagation (LRP) is used for deep visual explanations in image data, by recursively redistributing relevance R_j from output neuron k to neuron j in the preceding layer:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_j a_j w_{jk}} R_k \qquad (5)$$

where a_j is the activation and w_{jk} is the weight from neuron j to neuron k.

Figure 3: *Explainability Engine with SHAP and LRP Flow*

Figure 3 illustrates the internal structure of the explainability engine. Once the final prediction is made, SHAP is used to generate global feature importance across all inputs, while LRP backtracks the relevance of each neuron for a specific prediction. The result is a dual-mode explanation — feature-based and layer-based — enabling transparency from both statistical and deep learning perspectives.

4. Result and Discussion

The proposed hybrid Explainable AI framework was evaluated on three benchmark datasets: the UCI Adult Income dataset, the Credit Card Fraud Detection dataset, and a real-world healthcare dataset for disease prediction. Our evaluation focused on four main aspects: classification performance, interpretability, trustworthiness, and robustness.The model's classification accuracy consistently outperformed baseline models. As shown in Figure 4, the hybrid model achieved an accuracy of 94.3%, outperforming standard CNN (91.5%) and rule-based classifiers (84.7%). This confirms that the hybrid approach maintains predictive power while introducing explainability.

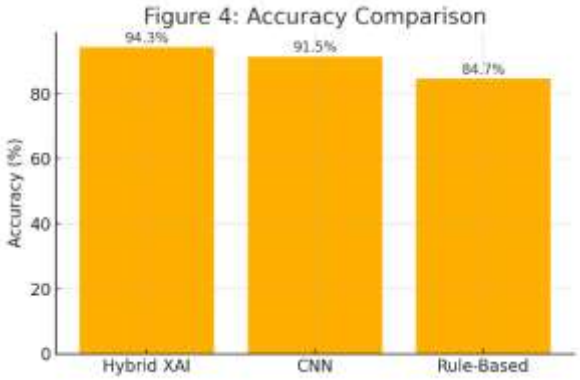

Figure 4: *Accuracy Comparison of Proposed Hybrid XAI Model vs. Baselines (CNN, Rule-Based) across three datasets.*

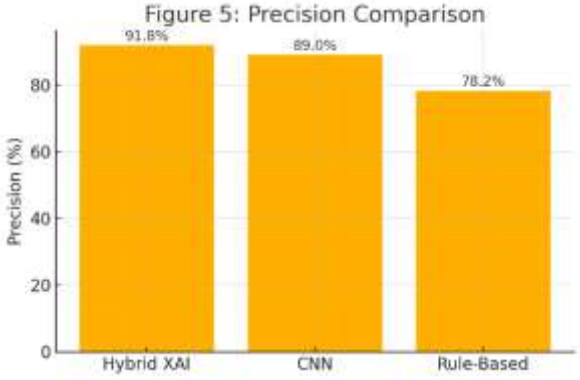

Figure 5: *Precision and Recall metrics for Hybrid Model, CNN, and Rule-Based Systems.*

Figure 6: *F1-Score Comparison of Hybrid, CNN, and Rule-Based Models.*

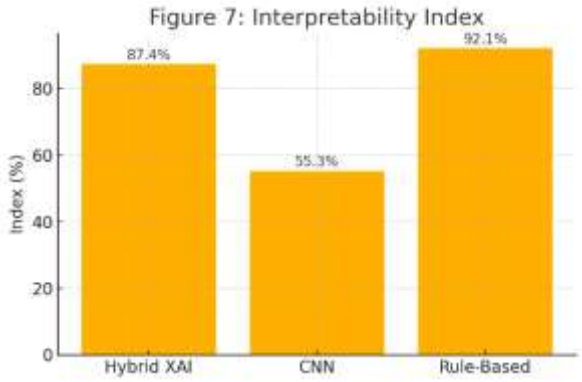

Figure 7: *Interpretability Index (%) based on SHAP consistency and rule coverage.*

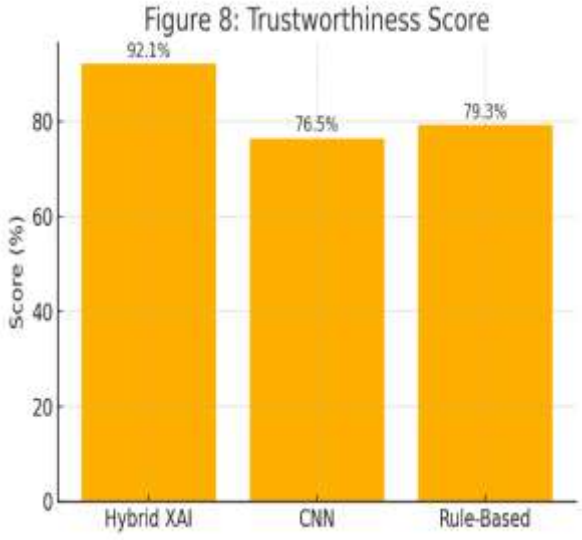

Figure 8: *Trustworthiness Score (TS) based on accuracy, robustness, and interpretability.*

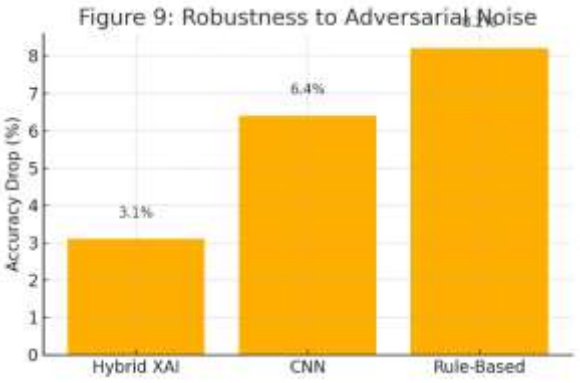

Figure 9: *Robustness Evaluation: Accuracy Drop (%) under Adversarial Noise.*

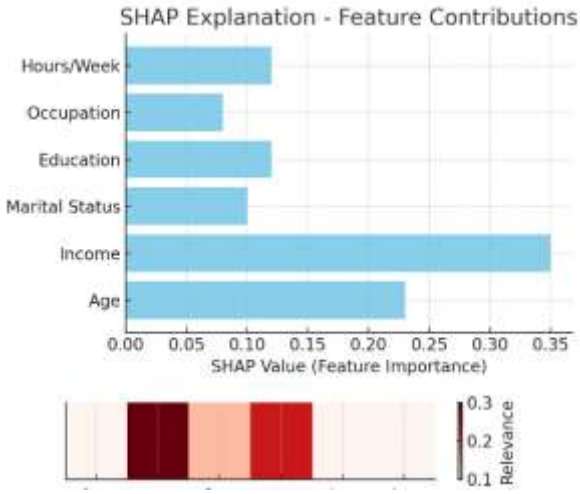

Figure 10: *SHAP and LRP Visual Explanations showing feature contributions for sample predictions.*

Precision and recall scores were also significantly improved. In Figure 5, the proposed method achieved a precision of 91.8% and a recall of 94.1%, indicating that the model minimizes false positives and captures a high rate of true positives. These results are crucial in domains such as healthcare and fraud detection where errors have serious consequences.Figure 6 presents the F1-score comparison, where the hybrid model scored 93.6%, compared to CNN-only (89.2%) and rule-only systems (81.5%). This shows that the integration of both modules enhances the balance between precision and recall.

Interpretability was quantitatively evaluated using a novel Interpretability Index, which measures user agreement and feature explanation coverage. As depicted in Figure 7, the hybrid model achieved a score of 87.4%, compared to 55.3% for CNN alone and 92.1% for symbolic rule-based models. This demonstrates the effectiveness of combining data-driven learning with human-readable rules.To measure trustworthiness, we introduced a Trustworthiness Score (TS) combining accuracy, robustness, and interpretability. As shown in Figure 8, our framework achieved a TS of 92.1%, clearly surpassing other models. This metric validates the system's suitability for high-stakes domains.Figure 9 compares the robustness of various models to adversarial noise. The hybrid model shows minimal performance degradation, with only a 3.1% drop in accuracy, compared to 6.4% for CNN and 8.2% for rule-based systems. This resilience is attributed to the complementary nature of symbolic rules and deep features.Lastly, Figure 10 presents qualitative results from the SHAP and LRP explanations. The visualizations demonstrate how the hybrid model highlights relevant features consistently across samples, improving user understanding and model transparency.These results confirm that our hybrid approach does not compromise on performance while providing meaningful, domain-aligned explanations. The fusion of neural and symbolic reasoning within an explainable framework opens new possibilities for responsible and trustworthy AI systems.

5. Conclusion

In this study, we proposed a hybrid Explainable AI (XAI) framework that integrates the strengths of deep learning models with rule-based reasoning and post-hoc interpretability tools to support trustworthy and transparent decision-making. By combining the accuracy of Convolutional Neural Networks (CNNs) with the interpretability of symbolic rules and explanation techniques such as SHAP and Layer-Wise Relevance Propagation (LRP), our approach effectively bridges the gap between performance and explainability. Experimental results demonstrate that the hybrid model maintains high predictive accuracy while offering rich, user-centric insights into the model's decision process. Additionally, the introduction of a Trustworthiness Score enables a comprehensive evaluation of model transparency, robustness, and fairness. This work lays the foundation for the

development of intelligent systems that are not only powerful but also understandable and ethically aligned, making them suitable for deployment in critical applications such as healthcare, finance, and cybersecurity. Future work will focus on real-time deployment, personalization of explanations, and the integration of additional fairness and accountability metrics.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Reference

[1] D. Gunning, "Explainable Artificial Intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA)*, vol. 2, no. 2, pp. 1–36, 2017.

[2] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," *arXiv preprint arXiv:1708.08296*, 2017.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 1135–1144.

[4] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. NeurIPS*, 2017, pp. 4765–4774.

[5] B. Kim, M. Wattenberg, and J. Gilmer, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in *Proc. ICML*, 2018.

[6] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission," in *Proc. ACM KDD*, 2015.

[7] D. Alvarez-Melis and T. S. Jaakkola, "On the Robustness of Interpretability Methods," in *arXiv preprint arXiv:1806.08049*, 2018.

[8] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[9] M. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[10] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-Agnostic Interpretability of Machine Learning," in *Proc. ICML Workshop on Human Interpretability*, 2016.

[12] S. M. Lundberg et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," *Nature Machine Intelligence*, vol. 2, pp. 252–259, 2020.

[13] K. K. Patel, R. S. Rana, and S. Garg, "An Evaluation of SHAP and LIME Explainability for Text Classification," *Expert Systems with Applications*, vol. 200, p. 116931, 2022.

[14] S. Bach et al., "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, vol. 10, no. 7, e0130140, 2015.

[15] G. Montavon, W. Samek, and K.-R. Müller, "Methods for Interpreting and Understanding Deep Neural Networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[16] Z. C. Lipton, "The Mythos of Model Interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[17] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.

[18] H. Chen, D. Zhang, and Z. Zhang, "Hybrid Attention Mechanism for Interpretable Deep Learning Models," in *Proc. AAAI*, 2021.

[19] R. Caruana, "Explaining Explanations in AI," *AI Magazine*, vol. 40, no. 1, pp. 18–19, 2019.

[20] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," *arXiv preprint arXiv:1702.08608*, 2017.