ISSN 2664-4150 (Print) & ISSN 2664-794X (Online)
South Asian Research Journal of Engineering and Technology
Abbreviated Key Title: South Asian Res J Eng Tech

| Volume-5 | Issue-6 | Nov-Dec- 2023 | **DOI: https://doi.org/10.36346/sarjet.2023.v05i06.004**

Review Article

Securing the Black Box: A Systematic Review of the Critical Intersection Between Explainable AI and Trusted Hardware Implementations

Abigail Adeniran[1*], Zaynab B. Bello[2], Temidayo J. Omotinugbon[3], Ayokunle Adeyemo[4]

[1]Department of Computing Science and Mathematics, University of Stirling, Scotland, United Kingdom
[2]Department of Computer Science, Texas Tech University, Lubbock, Texas, United States
[3]Eller College of Management, University of Arizona, Tucson, Arizona, United States
[4]Sheffield Hallam University, Sheffield, South Yorkshire, England

***Corresponding Author:** Abigail Adeniran
Department of Computing Science and Mathematics, University of Stirling, Scotland, United Kingdom

Article History
Received: 15.11.2023
Accepted: 23.12.2023
Published: 26.12.2023

Abstract: Artificial intelligence (AI) models have gained widespread adoption across numerous fields. However, these models often require substantial computational power and memory resources, making their deployment on resource-constrained IoT devices challenging. Frequently, the deployment of pre-trained AI models on IoT devices is outsourced to third parties who may not be fully trusted. In some scenarios, these third parties may act maliciously, potentially embedding harmful circuitry within the hardware design of the AI model. As AI models increasingly penetrate decision-critical and safety-critical domains that directly impact human lives, the development of Explainable AI (XAI) techniques has become essential. These techniques enhance our understanding of AI model operations and illuminate the rationale behind their decision-making processes. XAI methodologies help identify the specific features detected by individual neurons within the model architecture. In this work, we specifically examine layer-wise relevance propagation and activation maximization XAI techniques, exploring how they can contribute to the secure deployment of AI models in hardware implementations. We analyze the application of these XAI techniques from dual perspectives: that of an attacker seeking to compromise the accuracy of AI models deployed on IoT devices, and that of a defender working to preserve model accuracy and integrity. This dual analysis provides comprehensive insights into securing AI models within hardware environments.

Keywords: Layer-wise relevance propagation, Activation Maximization, AI models, Deep learning, Machine learning, Hardware security.

I. INTRODUCTION

Artificial Intelligence (AI) have gained recognition and has found application in many fields like fraud detection, image recognition, natural language processing and so on. AI in today's application require a lot of collection and analysis of sensitive personal data through smartphones, fitbit devices, social media, fault diagnosis [1] and so on [2]. AI analyzes user data for decision-making processes like employment, insurance rates, loan rates, and even criminal justice. Recently, negative interference of social media bots in political elections show how susceptible our lives are to the misuse of AI and big data. Personalized agents, recommendation systems, and critical decision-making tasks (e.g., medical analysis [3], power-grid control) has demonstrated the importance of AI transparency to end-users [4]. AI and algorithmic decision-making processes have been criticized for their black-box nature. With the growing prevalence of AI applications in our everyday life, the demand for predictable and accountable AI grows as tasks with higher sensitivity and social impact are more commonly entrusted to AI services. Hence, there is a need for algorithm transparency for organizations and applications responsible for products, services, and communication of information [2].

Explainable Artificial Intelligence (XAI) systems represent a critical advancement toward achieving accountable AI, enabling transparent interpretation of complex decision-making processes for end users as illustrated in Fig. 1. These

approaches establish rigorous methodologies for tracing human-interpretable decision pathways from otherwise opaque AI algorithms. The implementation of XAI frameworks fundamentally transforms the paradigm of human-AI interaction by increasing trust through algorithmic transparency and enabling users to systematically evaluate the quality and validity of explanations [6]. The epistemological significance of XAI extends beyond mere technical transparency, it establishes a framework for meaningful control and regulatory oversight in scenarios where AI systems might produce adverse or unintended consequences, such as algorithmic bias, discriminatory decision patterns, or violations of ethical principles. By illuminating the decision architecture of AI systems, XAI creates accountability mechanisms that align algorithmic governance with human values and societal norms.

The literature has proposed a diverse taxonomy of techniques to address the inherent complexity of deep neural network interpretability, ranging from post-hoc explanation methods to intrinsically interpretable model architectures. In this investigation, we focus specifically on two methodologically distinct but complementary approaches: Layer-wise Relevance Propagation (LRP) and Activation Maximization. LRP provides a rigorous mathematical framework for backpropagating relevance scores through the network architecture, thereby quantifying the contribution of each input feature to the final prediction. Activation Maximization, conversely, synthesizes input patterns that maximize neuronal activations, offering insights into the feature representations learned by specific network components. Together, these techniques provide a multi-dimensional perspective on neural network interpretability that bridges the gap between mathematical formalism and human-comprehensible explanations.

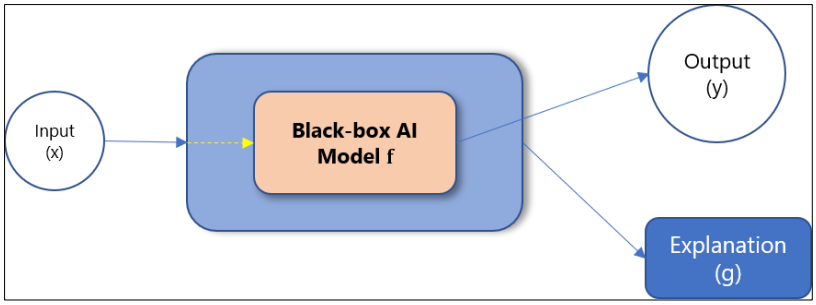

Fig. 1: High-level illustration of AI model explainability algorithms applied on AI model (f) such that f is made explainable externally [5]

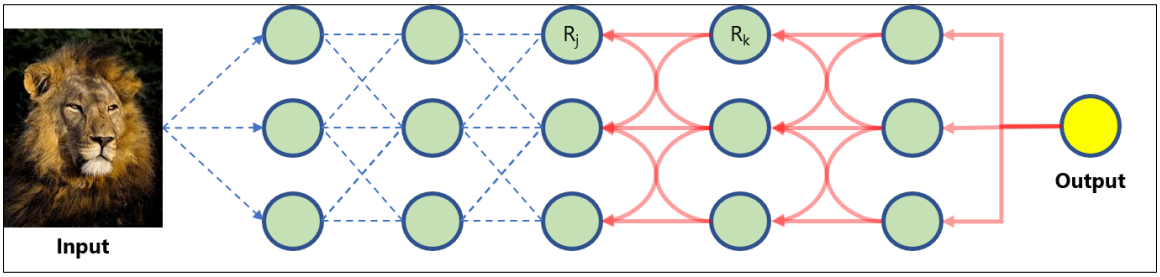

Fig. 2. Illustration of the LRP procedure where each neuron redistributes to the lower layer as much as it has received from the higher layer [7]

The remainder of this paper is organized as follows: Section II summarizes Layer-wise Relevance Propagation. Section III discusses Activation Maximization. Section IV highlights how LRP and AM apply to secure hardware implementations of AI models. Section VI concludes the paper.

II. LAYER-WISE RELEVANCE PROPAGATION (LRP)

Layer-wise Relevance Propagation (LRP) is an explanation technique applicable to AI models structured as neural networks, where inputs can be e.g. images, videos, or text. LRP serve as a solution for explaining what pixels of an input image are relevant for reaching a classification decision for neural networks [4]. LRP is based on the idea that the likelihood of a class can be traced backwards through a network to the individual layer-wise nodes or elements of the input. Specifically, the contribution, or relevance, to the target output node is back propagated toward the input image creating a map of which pixels contributed to the node [8]. LRP has been applied to discover biases in commonly used AI models and datasets. It can be used to extract insights from AI models. For example, LRP has been used to find relevant features for audio source localization, LRP has also been utilized to identify points of interest in side channel trace, and to identify EEG patterns that explain decisions in brain-computer interfaces. In the biomedical domain, LRP was used to identify subject-specific characteristics in gait patterns, to highlight relevant cell structure in microscopy, as well as to explain therapy predictions [7].

The LRP algorithm attributes relevance to individual input nodes to trace back contributions to the final output node layer by layer. LRP is depicted by a Taylor-expansion of a prediction made by a function $f(x)$ with respect to the input x [9]. LRP operates by propagating the prediction $f(x)$ backwards in the neural network, by means of purposely designed local propagation rules as shown in Fig. 2. The propagation procedure implemented by LRP is subject to a conservation property, where what has been received by a neuron must be redistributed to the lower layer in equal amount. This behavior is analogous to Kirchoff's conservation laws in electrical circuits. Algorithm 1 [7] depicts the LRP operation. Applying the LRP algorithm to the topmost layer yields the relevance scores for the neurons in the last hidden layer, from which we can derive the relevance scores for the lower layer. Therefore, iteratively the LRP algorithm from the topmost layer down to the input layer, propagating and redistributing the relevance scores from the predicted probability to input features [9].

III. ACTIVATION MAXIMIZATION

Activation Maximization (AM) focuses on input patterns which maximize a given hidden unit activation. AM technique is applicable to any network to find gradients values to optimize activations [5]. During the training of AI models, the weights and biases of the network are iteratively selected to reduce the error or loss, of the AI model is to achieve convergence during training. In AM, the process in training is flipped iteratively to find the parts of the data that the model thinks belongs to a class. AM provides a means for the visualization of the preferred inputs of neurons in each layer of an AI model. Identifying the preferred input can help indicate what features a neuron has learned from the input. The learned feature is represented by a synthesized input pattern that can cause maximal activation of a neuron. To synthesize an input pattern, each pixel of the AI model's input can be iteratively changed to maximize the activation of the neuron. The visualization of input patterns helps to improve the interpretability of AI models. AM has demonstrated great capability to interpret the interests of neurons and identify the hierarchical features learned by AI models. [10].

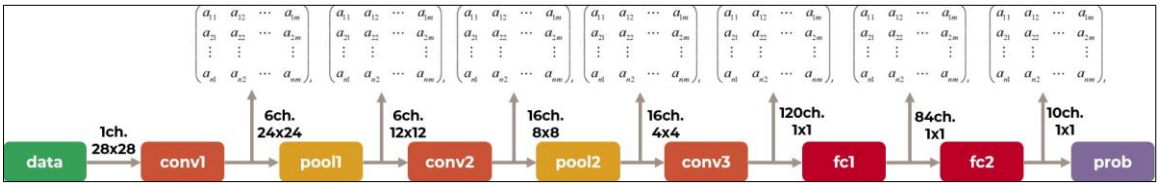

Fig. 3: Conceptual depiction of the amount of information available to the untrusted third-party hardware designer: the AI model architecture, the parameters (weights and biases), and the intermediate output (feature maps)

IV. Layer-Wise Relevance Propagation and Activation Maximization Xai Techniques Application to Hardware Intrinsic Attack Design

AI models have achieved impressive performance in many fields. AI model inference have the drawback of requiring huge computation and memory requirements. In many scenarios, to achieve real-time inference, AI models are usually deployed at the site where data is obtained. The data are usually obtained resource constrained (RC) Internet of things (IoT) devices. Due to the computation intensive of AI models and to achieve high throughput, the acceleration AI models can be achieved across multiple IoT for collaborative inference. Due to the scarcity of expertise, and the need to achieve short-time-to-market the deployment of AI models are often outsourced to untrusted third-party designers. In this research work, we focus on gray-box threat models as seen in [11] and [12] where the vendor takes a pre-trained AI model and outsources its hardware acceleration and deployment to an untrusted third party. Based on the threat model in focus, it is assumed that the untrusted third party is malicious. The malicious third party has access to the AI model architecture, parameters (weights and biases), and the layer-by-layer intermediate results as shown in Fig. 3. It is also assumed that the malicious third party has no access or knowledge of the training and testing data sets.

To address some of the security concerns and limit the amount of information provided to the untrusted third parties, partitioned AI models have been introduced. Partitioned CNN shown in Fig. 4 encourages the idea of AI model partitioning because from a security perspective, partitioned AI models on multiple IoT devices restricts the amount of information available to any one untrusted third-party designer. The untrusted third-party designer has access to a design validation dataset (which consist of an input and corresponding output of the partition) for the verification of implementation correctness of hardware design. To add another security layer, the presence of the first layer and final layers that reveal information about the size of the input image and the classes of the models respectively can be deployed by trusted designers. This research work also focuses on security vulnerabilities where collaborative inference is employed. In these situations, an AI model can be partitioned among multiple hardware devices where each partition is deployed on a different hardware device and each partitioned is deployed to different hardware device so no one third party designer has access to the full AI model architecture shown in Fig. 4.

The LRP XAI algorithm attributes relevance scores to intermediate neurons to quantify the significance of their contributions to the final output prediction. Hence with LRP from an attacker's perspective, targeted classes attacks can be carefully crafted. With LRP the attacker can obtain information about hierarchy of neurons across all the layers in terms of contributions to a particular classification. Hence the attacker can design attacks targeted at specific neurons based on their relevance. From a defender's perspective, LRP can be used to identify neurons of relevance across all the layers for each class to obscure or encrypt against attacks by malicious third-party. AM helps in the visualization of the preferred inputs of neurons in each layer of an AI model thereby indicating what features a neuron has learned from the input. With AM, an attacker without access to the training or testing dataset can recover attributes of the samples of the input to the AI model. An attacker can make use of this information to generate adversarial noises that can compromise the accuracy of the AI model.

V. DISCUSSION

Having examined Layer-wise Relevance Propagation (LRP) and Activation Maximization (AM) techniques in the preceding sections, as well as their potential applications in hardware security contexts, we now synthesize these insights to address broader implications for the research community and practitioners. The juxtaposition of these XAI techniques reveals important synergies

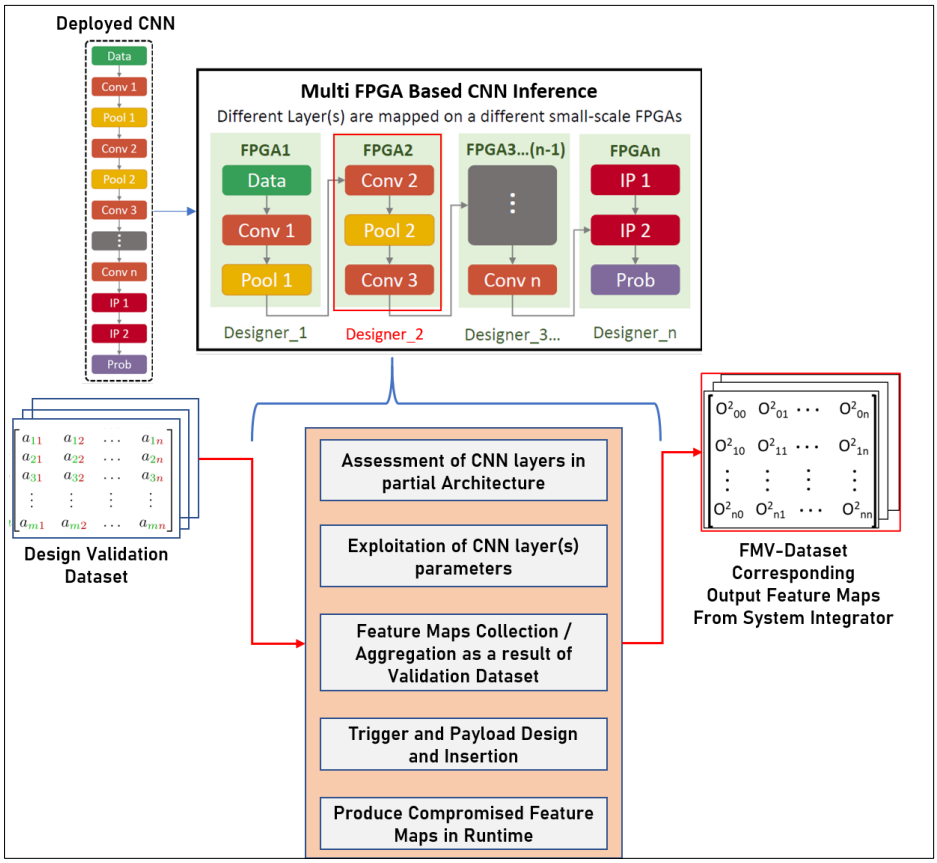

Fig. 4: Conceptual depiction of the overview of different attack methodologies that can be targeted and collaborative Inference

That have not been adequately explored in the literature. While Section II demonstrated LRP's capacity for quantitative attribution of neuron contributions and Section III highlighted AM's ability to visualize learned features, their combined application could provide a more comprehensive security framework than either technique alone. Specifically, LRP's relevance scores could guide targeted applications of AM visualization, focusing computational resources on neurons identified as particularly critical to classification decisions.

The threat model outlined in Section IV presents a realistic scenario that warrants further attention from the research community. The practical constraints of outsourcing hardware acceleration to third parties creates a unique security challenge that traditional approaches to AI security focused primarily on adversarial examples or data poisoning do not adequately address. The integration of XAI techniques into hardware security frameworks represents a novel approach to mitigating these risks. The bidirectional utility of XAI techniques for both attackers and defenders create an evolving security landscape that requires adaptive protection strategies. For attackers gaining access to model architecture

and parameters as described in Figure 3, LRP provides a methodological approach to identifying critical neurons, while AM enables them to generate synthetic inputs without access to training data. This capability fundamentally changes the attack surface compared to traditional hardware security threats.

For defenders, the same techniques offer unprecedented visibility into model vulnerabilities before deployment. The partitioning approach illustrated in Figure 4 can be enhanced through strategic application of XAI insights, with encryption or obfuscation resources allocated to neurons identified as most critical through LRP analysis. This targeted protection approach is particularly valuable in resource-constrained IoT environments where comprehensive security measures may be computationally prohibitive. The temporal dimension of hardware security deserves particular attention, as deployment decisions made today must anticipate evolving attack methodologies. The accessibility of XAI techniques means that even if defenders do not proactively apply these methods, sophisticated attackers likely will. This asymmetry suggests that defensive applications of XAI should be considered a necessity rather than an option for security-critical deployments.

Interdisciplinary research bridging XAI, hardware security, and domain-specific expertise will be essential to developing comprehensive security frameworks. The domain-specific applications of LRP mentioned in Section II such as identifying relevant features for audio source localization or EEG patterns in brain-computer interfaces demonstrate the diverse contexts in which these techniques might be applied. Each application domain introduces unique security considerations that must be addressed through specialized approaches. The security implications of deploying AI models on hardware extend beyond the technical considerations discussed in this review to encompass ethical and regulatory dimensions. As AI increasingly penetrates critical infrastructure and high-stakes decision domains, the ability to explain and secure these systems becomes not merely a technical preference but a societal necessity. XAI techniques like LRP and AM represent important tools in advancing toward this goal, but their effective implementation will require coordination across technical, regulatory, and organizational boundaries.

VI. CONCLUSION

In this study, we reviewed the application of Explainable AI (XAI) techniques, (specifically layer-wise relevance propagation and activation maximization) to AI model deployment on resource-constrained edge devices. Our analysis adopts a comprehensive security framework that examines both offensive and defensive strategies. From an adversarial perspective, we systematically explore how these XAI methodologies can be leveraged to identify and exploit vulnerabilities within AI model architectures, providing novel insights into potential attack vectors. Conversely, we demonstrate how these same techniques can be repurposed as defensive mechanisms, enabling model developers to precisely identify critical neurons and feature representations that warrant enhanced protection. This defensive approach facilitates the implementation of targeted obscuration and encryption schemes that selectively protect the most vulnerability-prone components of the model architecture. Through this bidirectional examination, we contribute to the emerging field of hardware-aware XAI by establishing a methodological framework for security-conscious AI deployment in edge computing environments.

REFERENCES

1. J. Grezmak, J. Zhang, P. Wang, K. A. Loparo, and R. X. Gao, "Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis," *IEEE Sensors Journal*, vol. 20, no. 6, pp. 3172–3181, 2019.
2. S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 11, no. 3-4, pp. 1–45, 2021.
3. M. Bohle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's¨ disease classification," *Frontiers in aging neuroscience*, vol. 11, p. 194, 2019.
4. A. Binder, S. Bach, G. Montavon, K.-R. Muller, and W. Samek, "Layer-wise relevance propagation for deep neural network architectures," in¨ *Information science and applications (ICISA) 2016*. Springer, 2016, pp. 913–922.
5. A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.
6. D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai," *International Journal of Human-Computer Studies*, vol. 146, p. 102551, 2021.
7. G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Muller, "Layer-wise relevance propagation: an overview,"¨ *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.
8. B. K. Iwana, R. Kuroki, and S. Uchida, "Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 4176–4185.
9. Y. Yang, V. Tresp, M. Wunderle, and P. A. Fasching, "Explaining therapy predictions with layer-wise relevance propagation in neural networks," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2018, pp. 152–162.

10. Z. Qin, F. Yu, C. Liu, and X. Chen, "How convolutional neural network see the world-a survey of convolutional neural network visualization methods," *arXiv preprint arXiv:1804.11191*, 2018.
11. T. A. Odetola and S. R. Hasan, "Sowaf: Shuffling of weights and feature maps: A novel hardware intrinsic attack (hia) on convolutional neural network (cnn)," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.
12. T. Odetola, F. Khalid, T. Sandefur, H. Mohammed, and S. R. Hasan, "Feshi: Feature map based stealthy hardware intrinsic attack," *arXiv preprint arXiv:2106.06895*, 2021.