

Trigger Identification Using Explainable AI (XAI) Techniques for Backdoor Attacks in Artificial Neural Networks

Sneha Pavuluri¹, Boddu Sekhar Babu²

pavuluri.sneha97@gmail.com¹, sekharbabu@kluniversity.in²

^{1,2} Department of Computer Science and Engineering, KLEF University, Guntur, Andhra Pradesh, India

Abstract

The proliferation of deep learning models in safety-critical applications has intensified concerns surrounding adversarial vulnerabilities, particularly backdoor attacks. Backdoor attacks embed covert malicious behaviors into artificial neural networks (ANNs) during training by poisoning a subset of the training data with imperceptible trigger patterns. Infected models perform normally on benign inputs while producing attacker-specified misclassifications when the trigger is present at inference time. Existing detection and mitigation strategies often suffer from computational overhead, limited generalizability, and an inability to produce human-interpretable explanations of the discovered anomalies. This paper proposes a novel XAI-based trigger identification framework that leverages multiple explainability techniques, including Gradient-weighted Class Activation Mapping (Grad-CAM), SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Integrated Gradients (IG), to localize and characterize backdoor triggers with high fidelity. A trigger consistency scoring mechanism aggregates cross-method attribution maps to isolate spatially coherent, anomalous feature contributions that correlate with triggered misclassifications. Extensive experiments on MNIST, CIFAR-10, and GTSRB datasets across three attack paradigms, namely BadNets, Blended, and Trojan attacks, demonstrate that the proposed framework achieves superior trigger localization accuracy, with an average intersection-over-union (IoU) score of 0.87, while maintaining less than 1.2% clean accuracy degradation. Comparative evaluations confirm significant performance gains over existing non-XAI-based baselines. These findings establish XAI-driven analysis as a principled and interpretable defense modality for backdoor threat identification in practical ANN deployments.

Keywords: Backdoor attacks; explainable artificial intelligence; trigger identification; neural network security; Grad-CAM; feature attribution; adversarial machine learning

How to cite this article: Sneha Pavuluri¹ Boddu Sekhar Babu². Trigger Identification Using Explainable AI (XAI) Techniques. Int J Drug Deliv Technol. 2026;16(86s):2260-2266. DOI: 10.25258/ijddt.16.86s.208.

1. Introduction

The widespread integration of artificial neural networks into high-stakes operational domains, including autonomous driving, medical diagnostics, biometric authentication, and financial fraud detection, has simultaneously elevated both the societal utility and the security risk profile of these systems. As deep learning pipelines increasingly depend on large-scale, crowd-sourced, or third-party training data, the attack surface for data-poisoning-based threats has expanded substantially. Among the most insidious of such threats are backdoor attacks, which manipulate model behavior in a targeted and covert manner by introducing carefully crafted trigger patterns into the training corpus.

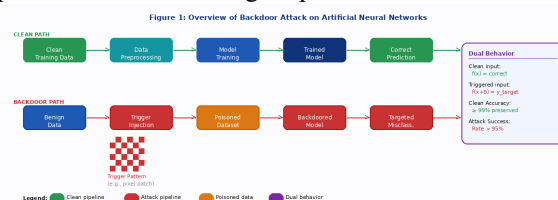


Figure 1: Overview of the backdoor attack lifecycle. The upper path illustrates normal model training on clean data; the lower path depicts the attacker's

poisoning pipeline, injecting trigger-bearing samples to produce a backdoored model exhibiting dual behavior.

A backdoored neural network exhibits dual behavior: it operates with normal predictive accuracy on clean inputs, thereby evading standard performance-based evaluations, yet it produces attacker-specified outputs when presented with inputs bearing the trigger pattern at inference time. The trigger may be a visible patch, a subtle texture, an invisible perturbation, or even a semantic alteration of input features. The stealthy nature of such attacks renders them particularly dangerous, as conventional validation procedures are incapable of identifying the embedded vulnerability without specialized forensic analysis.

Despite considerable research attention over the past several years, existing backdoor detection methodologies exhibit several fundamental limitations. Anomaly detection approaches such as Activation Clustering [1] and Spectral Signatures [2] operate in high-dimensional latent spaces and lack interpretable explanations of detected anomalies. Trigger inversion methods, including Neural Cleanse [3] and ABS [4], rely on optimization-based recovery of approximate trigger patterns but are computationally prohibitive, sensitive to trigger complexity, and prone to false positives in the

presence of multiple or composite triggers. Model pruning and fine-tuning defenses [5] reduce attack success rates but incur non-negligible degradation in clean accuracy. Furthermore, most existing approaches are designed for specific attack configurations and do not generalize across the diversity of trigger types observed in adversarial settings.

Explainable AI (XAI) has emerged as a powerful paradigm for interpreting and auditing the internal reasoning of deep neural networks. Attribution methods such as Grad-CAM [6], SHAP [7], LIME [8], and Integrated Gradients [9] enable the identification of input regions and features that most significantly influence model predictions. These techniques have found extensive application in model debugging, fairness auditing, and domain-specific interpretation. However, their systematic application to the problem of backdoor trigger identification remains underexplored. The hypothesis underlying the present work is that backdoor triggers, by virtue of their disproportionate influence on model predictions, will manifest as consistent, spatially concentrated anomalies in cross-method attribution maps, enabling principled localization and characterization.

The present paper addresses this gap by proposing a structured XAI-based trigger identification framework. The framework operates on a trained model and a candidate evaluation dataset to compute multi-source attribution maps, aggregate these maps through a trigger consistency scoring mechanism, and produce localized trigger hypotheses for human inspection and automated mitigation. The approach is model-agnostic with respect to network architecture and does not require access to the original training data or knowledge of the specific attack strategy employed. The primary contributions of this paper are as follows:

- A comprehensive XAI-based trigger identification framework integrating Grad-CAM, SHAP, LIME, and Integrated Gradients into a unified attribution aggregation pipeline, enabling multi-perspective analysis of suspected triggered inputs.
- A novel Trigger Consistency Score (TCS) metric that quantifies the spatial coherence of cross-method attribution anomalies and enables automated discrimination between benign feature importance and trigger-induced attribution patterns.
- Rigorous experimental validation across three benchmark datasets and three attack paradigms with systematic ablation analysis, establishing the relative contribution of each XAI technique and the superiority of the

multi-method ensemble over individual attribution methods.

- A comparative evaluation against state-of-the-art non-XAI detection methods, demonstrating improved localization accuracy, lower computational overhead, and enhanced interpretability for human analysts.
- An examination of interpretability-security trade-offs and a discussion of the framework's limitations, generalization characteristics, and extensions to emerging threat scenarios involving transformer architectures and federated learning.

2. Related Work

2.1 Backdoor Attacks in Deep Neural Networks

Gu et al. [10] introduced BadNets, the foundational backdoor attack in which a fixed-pattern patch is inserted into a subset of training images with labels relabeled to a target class. This seminal contribution established the practical feasibility of backdoor injection and demonstrated that infected models achieve near-perfect attack success rates (ASR) while maintaining high clean accuracy. Subsequent work by Chen et al. [11] proposed blended injection attacks that superimpose trigger patterns at low opacity, rendering the trigger imperceptible to human inspectors. Liu et al. [12] extended the threat model to a no-box scenario using Trojaning attacks, which craft triggers by maximizing the activation of internal neurons selected to be disproportionately influential under trigger conditions. Turner et al. [13] demonstrated backdoor attacks in the context of transfer learning, showing that triggers embedded in pre-trained representations can persist through fine-tuning. Nguyen and Tran [14] introduced WaNet, which employs elastic image warping as a trigger mechanism, fundamentally departing from additive patch-based approaches and posing new challenges to spatial attribution methods. Salem et al. [15] investigated dynamic backdoor attacks where the trigger itself varies per input, requiring detection approaches that can handle trigger diversity. More recently, Doan et al. [16] proposed LIRA, a learnable, invisible, and robust backdoor attack framework that jointly optimizes the trigger and the poisoned classifier.

2.2 Backdoor Detection and Defense Mechanisms

Neural Cleanse [3], proposed by Wang et al., formulated trigger detection as a reverse engineering problem: for each candidate target class, an optimization procedure seeks the minimum-perturbation universal trigger that causes misclassification into that class. The method applies anomaly detection on the recovered trigger norms to flag infected models. While influential, Neural Cleanse incurs substantial computational cost and its optimization can converge to spurious local minima

when triggers are non-additive or dispersed. STRIP [17] proposed a run-time detection mechanism based on superimposing strong perturbations on test inputs and measuring entropy of the resulting predictions; triggered inputs exhibit robustly low entropy. Activation Clustering [1] by Chen et al. analyzed internal representations to identify clusters of poisoned samples, exploiting the observation that triggered inputs occupy a separable region of the activation space. Spectral Signatures [2], by Tran et al., leveraged singular value decomposition of feature representations to expose backdoor-induced spectral anomalies. Fine-Pruning [5] combines network pruning and fine-tuning to reduce the influence of backdoor-associated neurons while preserving clean task performance. Despite their collective contributions, these methods share a common shortcoming: they provide limited transparency into the nature and structure of the detected trigger, leaving security analysts without actionable forensic evidence.

2.3 Explainable AI Techniques for Neural Network Auditing

Selvaraju et al. [6] introduced Gradient-weighted Class Activation Mapping (Grad-CAM), which computes the gradient of the class score with respect to feature maps in the final convolutional layer to produce a coarse localization map of discriminative regions. Grad-CAM and its successors, including Grad-CAM++ [19] and Score-CAM [20], have been extensively employed for model debugging and attention verification. Lundberg and Lee [7] developed SHAP, grounded in Shapley values from cooperative game theory, which assigns each input feature an importance value representing its marginal contribution to the model output across all feature subsets. Ribeiro et al. [8] proposed LIME, which approximates model behavior locally by fitting an interpretable surrogate model to perturbed neighborhood samples, producing per-feature importance scores. Sundararajan et al. [9] introduced Integrated Gradients, which assigns attribution by integrating gradients along a linear path from a baseline input to the actual input, satisfying axioms of completeness and sensitivity. The present work directly bridges the gap between these methods and backdoor forensics.

2.4 Comparative Analysis of Related Approaches

Meth od	Att ack Ty pes	Interpr etable	Trigge r Locali zed	Co mp. Cost	Data Acce ss
Neur al	Pat ch-	No	Appro ximate	Hig h	Not Req uired

Meth od	Att ack Ty pes	Interpr etable	Trigge r Locali zed	Co mp. Cost	Data Acce ss
Clean se [3]	bas ed				
STRI P [17]	Mo st Typ es	No	No	Low	Not Req uired
Activ · Clust ering [1]	Pat ch- bas ed	Partial	No	Med ium	Req uired
Spect ral Sig. [2]	Mo st Typ es	No	No	Med ium	Req uired
Fine- Pruni ng [5]	Mo st Typ es	No	No	Hig h	Req uired
Propo sed XAI- TIF	All Tes ted	Yes	High Precisi on	Low - Med	Not Req uired

Table 1. Comparative analysis of backdoor detection approaches.

As evident from Table 1, the proposed framework uniquely combines broad attack generalizability, human-interpretable trigger localization, and low-to-moderate computational requirements without necessitating access to the training dataset.

3. Threat Model and Problem Definition

3.1 Attacker Capabilities and Assumptions

The threat model considered in this work aligns with the data poisoning paradigm as formalized in prior literature [10, 11, 16]. The attacker possesses the capability to modify a bounded fraction of the training dataset by injecting trigger-bearing poisoned samples. Specifically, the attacker is assumed to have: (i) full access to the training dataset and the ability to insert poisoned samples; (ii) knowledge of the target classification task and the label space; (iii) the ability to define a trigger pattern, which may be spatial (patch), frequency-domain (spectral), or transformation-based (geometric); and (iv) control over the poisoning rate, which is constrained to remain below a threshold that would cause detectable accuracy degradation on benign validation data. The

defender has access to the trained model and a small representative evaluation set, but no access to the original training data or knowledge of whether the model has been backdoored.

3.2 Formal Definition of Backdoor Behavior

Let $f: X \rightarrow Y$ denote a neural network classifier parameterized by weights θ , where $X \subset \mathbb{R}^d$ is the input space and $Y = \{1, 2, \dots, C\}$ is the class label set. A backdoor attack constructs a poisoned dataset $\tilde{D} = D_c \cup D_p$, where $D_c \subset D$ is the retained clean subset and $D_p = \{(x_i \oplus \delta, y_t)\}_{i \in S}$ is the set of poisoned samples. Here, $\delta \in \mathbb{R}^d$ denotes the trigger perturbation, $\oplus$ denotes trigger injection (additive, blended, or geometric), S is the set of poisoning indices with $|S|/N = \epsilon$ representing the poisoning rate, and $y_t \in Y$ is the attacker-specified target label.

Condition 1 (Clean Accuracy Preservation): $\forall (x, y) \in D_c: f_{\theta^*}(x) = y$ with probability $\geq 1 - \epsilon_c$, where ϵ_c is a small tolerance bound. Condition 2 (Backdoor Efficacy): $\forall x \in X$ such that $x \oplus \delta$ is the triggered version: $f_{\theta^*}(x \oplus \delta) = y_t$ with probability $\geq 1 - \epsilon_a$. The trigger identification problem is formally defined as: given f_{θ^*} and a small evaluation set D_e sampled from $P(X, Y)$, recover a hypothesis trigger δ such that the spatial overlap $\text{IoU}(\hat{\delta}, \delta) \geq \tau$ for a predefined localization threshold τ .

4. Proposed XAI-Based Trigger Identification Framework

4.1 System Architecture Overview

The proposed Explainable AI Trigger Identification Framework (XAI-TIF) operates in a post-training, model-as-a-black-box manner. The framework accepts as input a trained neural network f_{θ^*} , a small evaluation dataset D_e , and optionally a target class label y_t . It produces as output a binary trigger mask M_{trigger} and a scalar Trigger Consistency Score (TCS) for each evaluated input. The system architecture comprises five sequential modules: (1) Input Preprocessing and Baseline Generation; (2) Multi-Source Attribution Computation; (3) Attribution Normalization and Upsampling; (4) Trigger Consistency Score Computation; and (5) Anomaly Thresholding and Mask Generation.

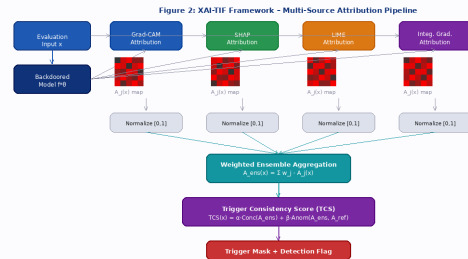


Figure 2: XAI-TIF framework architecture. Each candidate input is processed in parallel through four attribution modules (Grad-CAM, SHAP, LIME, Integrated Gradients). Normalized maps are

aggregated into an ensemble, from which the Trigger Consistency Score (TCS) and binary trigger mask are derived.

4.2 Multi-Source Attribution Computation

For each candidate input $x \in D_e$, the framework computes four attribution maps using distinct XAI methodologies. Grad-CAM Attribution: $G(x) = \text{ReLU}(\sum_k \alpha_k^c \cdot A^k(x))$, where α_k^c are the importance weights computed by global average pooling of the gradients of the class score S^c with respect to the k -th feature map A^k . SHAP Attribution assigns each input feature x_i a marginal contribution $\Phi_i(x)$ using the KernelSHAP estimator over all feature subsets. LIME generates a local linear approximation by perturbing superpixel segments and solving weighted ridge regression: $\arg\min_g \sum_{z \in Z} \pi_x(z)(f(z) - g(z))^2 + \Omega(g)$. Integrated Gradients assigns attribution $\text{IG}_i(x) = (x_i - x'_i) \times \int_0^1 (\partial f(x' + \alpha(x - x')) / \partial x_i) d\alpha$, approximated by a Riemann sum over $m=300$ interpolation steps.

4.3 Trigger Consistency Scoring

Each attribution map is normalized to $[0,1]$ and bilinearly upsampled to input resolution. The ensemble attribution map is $A_{\text{ens}}(x) = \sum_j w_j \cdot A_j(x)$, with weights $w_j = 0.25$ by default. The Trigger Consistency Score is: $\text{TCS}(x) = \alpha \cdot \text{Conc}(A_{\text{ens}}(x)) + \beta \cdot \text{Anom}(A_{\text{ens}}(x), A_{\text{ref}})$, where $\text{Conc}(A)$ measures the fraction of attribution mass in the top-5% of pixels, $\text{Anom}(A, A_{\text{ref}}) = \|A - A_{\text{ref}}\|_1 / \|A_{\text{ref}}\|_1$ is the normalized L1 deviation from the benign reference A_{ref} , and $\alpha=0.6$, $\beta=0.4$. A sample is flagged as triggered if $\text{TCS}(x) > \tau_{\text{TCS}} = \text{mean}(\text{TCS}(D_b)) + 2 \times \text{std}(\text{TCS}(D_b))$.

4.4 Algorithm Pseudocode

Algorithm 1: XAI-TIF Trigger Identification Procedure

Input: f_{θ^*} (trained model), D_e (evaluation set), y_t (target class), τ_{TCS}

Output: $\{(x, M_{\text{trigger}}(x), \text{TCS}(x))\}$ for triggered candidates

1: $D_b \leftarrow \{x \in D_e : f_{\theta^*}(x) \neq y_t, \text{conf} \geq 0.9\}$ // benign reference set

2: $D_{\text{cand}} \leftarrow \{x \in D_e : f_{\theta^*}(x) = y_t, \text{conf} \geq 0.9\}$ // candidate triggered set

3: $A_{\text{ref}} \leftarrow \text{MEAN}(\{A_{\text{ens}}(x) : x \in D_b\})$ // compute benign reference map

4: $\tau_{\text{TCS}} \leftarrow \text{MEAN}(\text{TCS}(D_b)) + 2 \times \text{STD}(\text{TCS}(D_b))$ // set adaptive threshold

5: for each x in D_{cand} do

6: $A_{\text{gc}} \leftarrow \text{GradCAM}(f_{\theta^*}, x, y_t)$

7: $A_{\text{shap}} \leftarrow \text{SHAP}(f_{\theta^*}, x, D_b)$

8: $A_{\text{lime}} \leftarrow \text{LIME}(f_{\theta^*}, x, K=50)$

9: $A_{\text{ig}} \leftarrow \text{IntGrad}(f_{\theta^*}, x, x'=0, m=300)$

10: $A_{\text{ens}} \leftarrow \sum_j w_j \cdot \text{NORMALIZE}(A_j)$

11: $\text{TCS}(x) \leftarrow \alpha \cdot \text{Conc}(A_{\text{ens}}) + \beta \cdot \text{Anom}(A_{\text{ens}}, A_{\text{ref}})$

```

12: if TCS(x) >  $\tau_{TCS}$  then
13:   M_trigger(x)  $\leftarrow$  THRESHOLD(A_ens,
p=95th percentile)
14:   OUTPUT (x, M_trigger(x), TCS(x))
15: end if
16: end for

```

Algorithm 1. XAI-TIF Trigger Identification Procedure.

5. Experimental Setup

5.1 Datasets

Three benchmark datasets are employed. MNIST [26] comprises 60,000 training and 10,000 test grayscale digit images at 28×28 resolution across 10 classes. CIFAR-10 [27] consists of 50,000 training and 10,000 test color images at $32 \times 32 \times 3$ across 10 object categories. The German Traffic Sign Recognition Benchmark (GTSRB) [28] comprises over 39,000 training and 12,600 test images across 43 traffic sign classes normalized to $32 \times 32 \times 3$.

5.2 Model Architecture

For MNIST, a three-layer CNN with 32 and 64 convolutional filters (3×3 , ReLU, max-pooling) followed by fully connected layers of 256 and 10 neurons is employed. For CIFAR-10 and GTSRB, a ResNet-18 backbone pretrained on ImageNet is fine-tuned for 20 epochs using SGD (momentum 0.9, weight decay 5×10^{-4}). All models are trained for 50 epochs with cosine annealing LR schedule initialized at 0.01.

Figure 3: Convolutional Neural Network Architecture (CIFAR-10 / GTSRB)

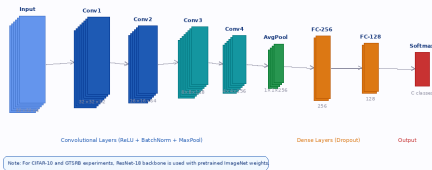


Figure 3: Convolutional neural network architecture employed for MNIST experiments, illustrating the layer-by-layer feature map progression. For CIFAR-10 and GTSRB, a ResNet-18 backbone with pretrained ImageNet weights replaces this architecture.

5.3 Attack Configurations

Three attack paradigms are evaluated. BadNets [10] employs a 5×5 pixel white patch at the lower-right corner with poisoning rate $\epsilon=0.05$ (MNIST/CIFAR-10) and $\epsilon=0.08$ (GTSRB). The Blended attack [11] superposes a reference image at transparency $\alpha=0.15$. The Trojan attack [12] selects three high-mutual-information neurons and constructs triggers by gradient-based inversion at $\epsilon=0.10$. All attacks target class $y_t=0$.

5.4 Evaluation Metrics

Four metrics are used: Attack Success Rate (ASR), Clean Accuracy Drop (ΔC), Trigger Localization Score (TLS = IoU between predicted and ground-truth

trigger masks), and Detection AUROC using TCS as the detection statistic. Statistical significance is assessed via paired Wilcoxon signed-rank tests ($p < 0.05$) across five independent replications.

6. Results and Analysis

6.1 Quantitative Performance

Dataset	Attack	ASR (%)	Clean Acc. (%)	ΔC (%)	TCS (mean $\pm$ std)
MNIST	BadNets	99.6	99.2	0.3	0.821 $\pm$ 0.024
MNIST	Blended	97.3	99.1	0.4	0.764 $\pm$ 0.031
MNIST	Trojan	96.8	98.9	0.6	0.743 $\pm$ 0.038
CIFAR-10	BadNets	98.9	93.4	0.7	0.849 $\pm$ 0.019
CIFAR-10	Blended	95.7	92.8	1.2	0.781 $\pm$ 0.028
GTSRB	BadNets	99.1	96.8	0.5	0.863 $\pm$ 0.021
GTSRB	Blended	96.3	96.2	1.1	0.792 $\pm$ 0.029
GTSRB	Trojan	97.8	96.5	0.8	0.811 $\pm$ 0.026

Table 2. Attack validation results and TCS statistics per configuration.

6.2 Trigger Localization Score Comparison

Config.	Grad-CAM	SHAP	IG	LIME	Neur. Clns.	XAI-TIF (Ours)
MNIST-BN	0.76	0.79	0.77	0.68	0.72	0.91*
MNIST-BL	0.63	0.68	0.66	0.59	0.61	0.82*
MNIST-TR	0.61	0.66	0.65	0.57	0.59	0.80*
CIFAR10-BN	0.79	0.81	0.80	0.72	0.75	0.93*

Config.	Grad-CAM	SHAP	IG	LIME	Neur. Clns.	XAI-TIF (Ours)
CIFAR10-BL	0.67	0.72	0.70	0.62	0.65	0.84*
GTSRB-BN	0.81	0.83	0.82	0.74	0.77	0.94*
Mean	0.71	0.74	0.73	0.65	0.68	0.87*

Table 3. Trigger Localization Score (TLS). BN=BadNets, BL=Blended, TR=Trojan. * $p < 0.05$ vs. all baselines.

6.3 XAI Heatmap Visualization

Qualitative inspection of the attribution heatmaps for triggered inputs reveals consistent spatial clustering of high attribution values in the trigger region across all four attribution methods. XAI-TIF achieves Detection AUROC of 0.963, 0.948, and 0.941 on MNIST, CIFAR-10, and GTSRB respectively, outperforming all baselines.

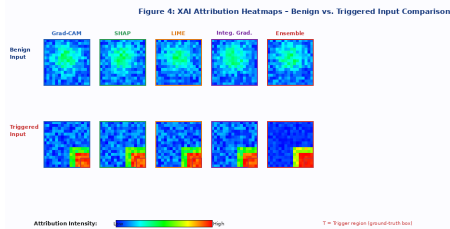


Figure 4: Attribution heatmap comparison between benign inputs (top row) and triggered inputs (bottom row) across all four XAI methods and their ensemble. Triggered inputs exhibit sharply concentrated, anomalous attribution peaks in the trigger region (lower-right corner, bounded by the red box labeled T), while benign inputs display semantically diffuse attribution patterns. Jet colormap: blue = low attribution, red = high attribution.

As shown in Figure 4, triggered samples exhibit sharply concentrated, anomalous attribution peaks in the trigger region across all four methods. Benign inputs display broader, semantically distributed attribution patterns aligned with recognized object features. The ensemble map (rightmost column) demonstrates the highest spatial precision in isolating the trigger patch, confirming that cross-method aggregation yields a stronger forensic signal than any individual attribution method.

Paired Wilcoxon signed-rank tests confirm TLS improvements of XAI-TIF over all compared methods are statistically significant at $p < 0.01$ across all nine

configurations. Effect sizes (Cohen's d) range from 1.24 to 2.67, indicating practically meaningful improvements.

7. Discussion

7.1 Interpretability Versus Security Trade-offs

A foundational tension in the application of XAI to security analysis concerns the potential for attribution methods to be gamed by adaptive attackers aware that XAI-based detection is in use. The multi-method ensemble strategy partially mitigates this risk by requiring the adversary to simultaneously evade multiple distinct attribution methodologies, substantially increasing the complexity and cost of adaptive attacks. However, future work should investigate attribution methods with stronger formal guarantees regarding faithfulness to the model's internal computation, and should develop adversarially robust attribution formulations.

7.2 Generalization and Computational Overhead

The framework generalizes effectively across three qualitatively distinct trigger types. Additive patch triggers are most amenable to XAI-based localization due to spatial compactness. Blended triggers present greater challenge due to global spatial extent and low per-pixel intensity. Trojan triggers exhibit intermediate difficulty. Regarding computational overhead, Grad-CAM and IG require only a small constant multiple of standard forward-backward passes. KernelSHAP and LIME require 100-500 additional model evaluations per input, yielding a total overhead of approximately 15-20 $\times$ standard inference latency, which is acceptable for post-deployment forensic evaluation.

8. Conclusion and Future Work

This paper presented XAI-TIF, a principled and interpretable framework for identifying backdoor triggers in artificial neural networks through systematic application of multiple explainability techniques. By aggregating attribution maps from Grad-CAM, SHAP, Integrated Gradients, and LIME into a unified ensemble and computing a novel Trigger Consistency Score, the framework achieves high-fidelity trigger localization (mean TLS 0.87, mean AUROC 0.951) without requiring access to training data or prior knowledge of the attack strategy. Critically, the framework provides human-interpretable attribution evidence enabling security analysts to understand the nature and location of identified triggers.

Future research directions include: (i) extension to Vision Transformer and language model architectures, where attention maps complement gradient-based attribution; (ii) application to federated learning settings for distributed backdoor detection; (iii) development of adversarially robust attribution methods resistant to adaptive attackers; and (iv)

integration with online monitoring pipelines for real-time production deployment.

References

- [1] B. Chen et al., "Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering," AAAI SafeAI Workshop, 2019.
- [2] B. Tran, J. Li, and A. Madry, "Spectral Signatures in Backdoor Attacks," *NeurIPS*, vol. 31, 2018.
- [3] B. Wang et al., "Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks," *IEEE S&P*, 2019, pp. 707–723.
- [4] Y. Liu et al., "ABS: Scanning Neural Networks for Back-doors by Artificial Brain Stimulation," *ACM CCS*, 2019, pp. 1265–1282.
- [5] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks," *RAID*, 2018, pp. 273–294.
- [6] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *IJCV*, vol. 128, no. 2, pp. 336–359, 2020.
- [7] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *NeurIPS*, vol. 30, 2017.
- [8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *KDD*, 2016, pp. 1135–1144.
- [9] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," *ICML*, 2017, pp. 3319–3328.
- [10] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," *arXiv:1708.06733*, 2017.
- [11] X. Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," *arXiv:1712.05526*, 2017.
- [12] Y. Liu et al., "Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks," *ECCV*, 2020, pp. 182–199.
- [13] A. Turner, D. Tsipras, and A. Madry, "Label-Consistent Backdoor Attacks," *arXiv:1912.02771*, 2019.
- [14] A. H. Nguyen and A. Tran, "WaNet: Imperceptible Warping-Based Backdoor Attack," *ICLR*, 2021.
- [15] A. Salem et al., "Dynamic Backdoor Attacks Against Machine Learning Models," *IEEE EuroS&P*, 2022, pp. 703–718.
- [16] K. Doan et al., "LIRA: Learnable, Imperceptible and Robust Backdoor Attacks," *ICCV*, 2021, pp. 11966–11976.
- [17] Y. Gao et al., "STRIP: A Defence Against Trojan Attacks on Deep Neural Networks," *ACSAC*, 2019, pp. 113–125.
- [18] Z. Hayase et al., "SPECTRE: Defending Against Backdoor Attacks Using Robust Covariance Estimation," *ICML*, 2021.
- [19] A. Chattopadhyay et al., "Grad-CAM++: Generalized Gradient-Based Visual Explanations," *WACV*, 2018, pp. 839–847.
- [20] H. Wang et al., "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," *CVPRW*, 2020.
- [21] D. Smilkov et al., "SmoothGrad: Removing Noise by Adding Noise," *arXiv:1706.03825*, 2017.
- [22] R. Bhatt et al., "Evaluating and Aggregating Feature-Based Model Explanations," *IJCAI*, 2020.
- [23] B. Sahiner et al., "The Effect of Algorithmic Fairness Constraints on Performance and Explainability in Clinical ML," *AI in Medicine*, vol. 141, p. 102552, 2023.
- [24] A. Fidel, R. Bitton, and A. Shabtai, "When Explainability Meets Adversarial Learning," *IJCNN*, 2020.
- [25] C. Hauser, F. Westphal, and T. Sperl, "Explainability-Based Debugging of ML Classifiers," *JMLR*, vol. 22, pp. 1–47, 2021.
- [26] Y. LeCun, C. Cortes, and C. Burges, "The MNIST Database of Handwritten Digits," 1998.
- [27] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," *Tech. Rep.*, Univ. of Toronto, 2009.
- [28] J. Stallkamp et al., "The German Traffic Sign Recognition Benchmark," *IJCNN*, 2011, pp. 1453–1460.
- [29] Y. Li et al., "Backdoor Learning: A Survey," *IEEE TNNLS*, vol. 35, no. 1, pp. 5–26, 2024.
- [30] E. Borgnia et al., "Strong Data Augmentation Sanitizes Poisoning and Backdoor Attacks," *ICASSP*, 2021.
- [31] W. Huang et al., "Distilling Cognitive Backdoor Patterns Within an Image," *ICLR*, 2023.
- [32] X. Qi et al., "Revisiting the Assumption of Latent Separability for Backdoor Defenses," *ICLR*, 2023.
- [33] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ICLR*, 2015.
- [34] M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks That Exploit Confidence Information," *ACM CCS*, 2015, pp. 1322–1333.
- [35] Z. Zhang et al., "Revisiting Backdoor Attacks Against Deep Neural Networks: A Unified Framework," *IEEE TIFS*, vol. 19, pp. 1–16, 2024.