Multi-Signal Backdoor Detection and Mitigation Framework for Deep Neural Networks

by Mahaveer Prasad

Submission date: 20-May-2026 12:19PM (UTC+0530)
Submission ID: 2965486584
File name: MahaveerPrasad_plag_check.pdf (2.14M)
Word count: 9327
Character count: 66071

Multi-Signal Backdoor Detection and Mitigation Framework for Deep Neural Networks

Mahaveer Prasad

ABSTRACT

Backdoors represent one of the most dangerous forms of threats to the trustworthiness and dependability of Deep Neural Networks (DNNs) by incorporating malicious behaviors into DNNs during the training process. The resulting infected model will have similar results with high performance when processing clean data while generating attacker-controlled predictions when it detects a predetermined set of pattern triggers. However, detecting this type of poisoning is difficult because the poisoned examples generally do not differ significantly from clean examples at an output level. Thus, many current defensive methods utilize some form of trigger reconstruction, perturbation heavy analysis, and/or remove suspicious samples aggressively which generally result in higher levels of computation, reduced data utility, and lower levels of robustness when dealing with heterogeneous types of attacks.

In this Thesis, I propose a framework for the detection and mitigation of backdoors in DNNs using three complementary techniques: InStaD, LayerStat, and ALCOR. InStaD proposes a dual branch perturbation framework that utilizes both stochastic prediction stability and deterministic structural sensitivity to detect backdoor behavior that relies on shortcuts. LayerStat introduces a new detection technique based on layer wise activation statistics that identifies anomalies in the activation response of the layers due to the presence of a trigger. LayerStat does so without needing to retrain the model, use auxiliary models or access to clean data. ALCOR provides additional capabilities by integrating multiple behavioral indicators: adversarial susceptibility, embedding deviation, activation anomaly, gradient stability, and layer-wise gradient relevance; via ensemble-based suspicious sample classification and subsequently corrects labels generated by adversaries through adversarial label correction and securely retrains the model.

Developed framework evaluated using various experiments on CIFAR-10 and Tiny ImageNet datasets under different input space, frequency domain, semantic, and adaptive backdoor attacks. My evaluation demonstrates a very effective ability to detect poisoned samples, an

attack success rate close to zero, a very low false positive rate, and minimal loss of clean data accuracy. I believe that the proposed approaches will serve as a basis for developing secure deep learning systems that are scalable and computationally efficient. Therefore, they could be used reliably in a variety of security sensitive applications such as autonomous systems, health care and intelligent surveillance.

LIST OF PUBLICATIONS

- Mahaveer Prasad and Vinod Kumar, "Protocol-Aware Semantic Filtering for Reliable Adversarial Robustness Evaluation in DL-based Network Intrusion Detection Systems" published in "2025 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)" held at IIIT Delhi, India, IEEE, 2025.

- Mahaveer Prasad and Vinod Kumar, "A Unified Multi-Signal Indicator Framework for Backdoor Detection and Label Correction in Deep Neural Networks" accepted for publication in IEEE "Fourth International Conference on Secure Cyber Computing and Communications (ICSCCC-2026)", May 29–31, 2026, Dr B R Ambedkar National Institute of Technology (NIT), Jalandhar, Punjab, India.

- Mahaveer Prasad et al., "METHOD AND SYSTEM FOR IDENTIFICATION AND CORRECTION OF POISONED TRAINING SAMPLES IN DEEP NEURAL NETWORK," Indian Patent Application No. 202611018498, DELHI TECHNOLOGICAL UNIVERSITY, filed on February 18, 2026 and published on April 10, 2026.

TABLE OF CONTENTS

Acknowledgement ii

Candidate's Declaration iii

Certificate by the Supervisor iv

Abstract v

List of Publications vi

Table of Contents vii

List of Tables ix

List of Figures x

CHAPTER 1 ..... **1**

INTRODUCTION **1**

1.1 Problem Statement ***4***

1.2 Project Objective ***5***

CHAPTER 2 ..... **6**

PROPOSED METHODOLOGY **6**

2.1 Overview ***6***

2.2 InStaD Framework ***6***

2.2.1 Motivation 6

2.2.2 Threat Model and Assumptions 7

2.2.3 Stochastic Prediction Stability Analysis 7

2.2.4 Structural Sensitivity Evaluation 7

2.2.5 Fusion-Based Detection and Mitigation 8

2.2.6 Proposed Architecture 8

2.3 LayerStat Framework ***11***

2.3.1 Motivation 11

2.3.2 Layer-wise Activation Analysis 11

2.3.3 Statistical Thresholding 12

2.3.4 Advantages of LayerStat 12

2.3.5 Proposed Architecture 13

13

2.4 ALCOR Framework *15*
2.4.1 Motivation 15
2.4.2 Multi-Signal Poison Analysis 15
2.4.3 Ensemble-Based Suspiciousness Ranking 15
2.4.4 Adversarial Label Correction and Secure Retraining 15
2.4.5. Proposed Architecture 16

CHAPTER 3 ..... **20**

RESULTS AND ANALYSIS **20**

3.1 Overview *20*

3.2 Experimental Setup *20*
3.2.1 Datasets 20
3.2.2 Backdoor Attack Configurations 21
3.2.3 Evaluation Metrics 21

3.3 Results of InStaD Framework *21*
3.3.1 Detection Performance 21
3.3.2 Mitigation Performance 22
3.3.3 Representation-Level Analysis 24

3.4 Results of LayerStat Framework *25*
3.4.1 Backdoor Detection Performance 25
3.4.2 Comparative Analysis with PSBD 25

3.5 Results of ALCOR Framework *26*
3.5.1 Mitigation Performance at Optimal Correction Level 27
3.5.2 Effect of Correction Ratio on Attack Success Rate 28
3.5.3 Trade-off Between ASR and Clean Accuracy 29

CHAPTER 4 ..... **31**

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT **31**

4.1 Conclusion *31*

4.2 Future Scope *32*

4.3 Social Impact *32*

REFERENCES **34**

Proof of Publishing

Plagiarism Report

Curriculum Vitae

List of Tables

Table 3. 1 Backdoor Detection Performance on CIFAR-10 (True Positive Samples Count / False Positive Samples Count) 22
Table 3. 2 Backdoor Detection Performance (TPR / FPR) on CIFAR-10 and TINYIMAGENET 25
Table 3. 3 Detection Performance (Top-10%) In Terms of TPSC and FPSC 27
Table 3. 4 Performance Comparison (ACC and ASR) at Optimal 28

List of Figures

Figure 2. 1 Architecture of the proposed InStaD framework for backdoor detection and mitigation. .8
Figure 2. 2 Proposed Architecture of LayerStat framework for poisoned-sample detection.....13
Figure 2. 3 Architecture of the proposed ALCOR framework for backdoor detection and mitigation. 17

Figure 3. 1 Performance comparison before and after defense under five backdoors.....23
Figure 3. 2 t-SNE visualization of penultimate-layer representations before and after defense. The infected model (a) exhibits feature structure distortion due to backdoor influence, while the corrected model (b) restores clearer intra-class compactness and inter-c.....24
Figure 3. 3 Comparison of LayerStat and PSBD in terms of true positives Count (TPC) and false positives Count (FPC) across different backdoor attacks. The top figure shows results on CIFAR-10, while the bottom figure corresponds to TinyImageNet. 26
Figure 3. 4 Heatmap visualization of attack success rate (ASR) across different correction ratios q using RF-driven Correction and XGB-driven Correction strategies. The logarithmic color scale highlights both major and subtle ASR variations across heterogeneous backdoor attacks. The dashed vertical line indicates the critical transition region near q ≈ 10%, beyond which most attacks experience rapid ASR collapse.....29
Figure 3. 5 Trade-off between mean attack success rate (ASR) and mean clean accuracy (ACC) across different correction ratios q using RF-driven and XGB-driven correction strategies. The dashed vertical line indicates the critical transition region near q ≈ 10%, where ASR rapidly collapses while clean accuracy remains relatively stable 30

CHAPTER 1

INTRODUCTION

Deep Neural Networks (DNNs) have produced very positive results for many different types of applications including computer vision, Natural Language Processing, Autonomous Driving, Healthcare, Cybersecurity and Intelligent Surveillance Systems [1]. The fact that DNNs can automatically learn hierarchical representation of features using a large number of data points has increased the precision of predictions and automated decisions. As such, deep learning systems are being used in safety critical and security sensitive areas where reliability and robustness are key factors [2].

Despite these developments, adversarial attacks, data poisoning, model inversion, and backdoor assaults continue to pose a security risk to contemporary deep learning systems [3]. Because they introduce covert destructive behaviours into neural networks without appreciably impairing clean-data performance, backdoor assaults are among the most dangerous [4]. By injecting trigger patterns into specific samples and giving attacker-defined target labels, adversaries often contaminate a portion of the training dataset [5]. When the trigger is present, the infected model generates attacker-controlled predictions during inference, but it operates normally on benign inputs. Such attacks are challenging to identify in real-world implementations since clean accuracy is essentially unaffected [6].

These assaults are becoming more feasible due to the increased reliance on publicly accessible datasets, federated learning, outsourced training pipelines, and third-party data collecting [7]. Adversaries can alter training distributions through poisoned samples because modern systems frequently rely on untrusted or poorly vetted data sources [8]. Because of this, creating strong and comprehensive backdoor defences has emerged as a significant difficulty in reliable artificial intelligence.

The trigger design and poisoning techniques used in backdoor attacks vary greatly. In order to create malicious target predictions, early assaults like BadNets added visible trigger patches to images [5]. Later attempts used spatial warping, frequency-domain manipulations, blended overlays, sinusoidal perturbations, and semantic alterations to increase stealthiness [9]. While SIG assaults disperse sinusoidal signals throughout photos to decrease visibility, blend attacks employ semi-transparent patterns [10]. WaNet additionally presents geometric warping-based triggers that, in the absence of specific trigger patches, produce visually natural poisoned samples [11]. By more closely aligning poisoned representations with clean semantic distributions, more recent adaptive and label-consistent attacks have been proposed to circumvent traditional defences [12]. Recent attack methodologies utilizing sophisticated attack frameworks for attacking both DNNs and vision transformers demonstrate a greater level of sophistication and complexity than

previously experienced with backdoor attacks [32].

Much recent literature has examined various defense strategies (i.e., detection based; cleaning techniques; trigger reconstruction; statistical anomaly identification; and resilient retraining) to combat such threats [13]. Early Feature-Space methods, including Activation Clustering [14] and Spectral Signature Analysis [15] utilized statistical anomalies in hidden-layer representations to discriminate between poisoned and clean data. However, these early methods were often unsuccessful because attackers could intentionally degrade feature separation, and poisoned samples could be constructed to mimic benign samples.

Reverse-engineering concealed trigger patterns that cause malevolent behaviour is the goal of trigger reconstruction approaches. While Backdoor Scanning expands this concept through optimization-based trigger search techniques, Neural Cleanse looks for minor perturbations that can cause targeted misclassification [16], [17]. These methods suffer against adaptive or dispersed triggers and typically require costly optimisation, notwithstanding their effectiveness in specific contexts.

Recently, behavioural analysis and perturbation-based approaches have become viable substitutes. Prediction consistency under controlled perturbations or modified inputs is examined using methods like STRIP, SCP, and SCAN [18]–[20]. The discovery that triggered samples frequently show anomalous prediction stability in comparison to clean samples serves as the inspiration for these techniques. To find shortcut-driven poisoned computation patterns, other methods look into neurone perturbations, masking tactics, robustness-to-noise behaviour, confidence consistency, and feature-space inconsistencies [21], [41]–[43], [46], [48]. According to recent research, backdoor behaviour can also be seen as shortcut learning, in which models use highly discriminative but semantically irrelevant trigger features rather than distributed semantic representations [22]. Abnormal activations, unstable gradients, concentrated channel dependencies, and irregular representation geometry are common manifestations of such behaviour.

More comprehensive defence viewpoints have been examined in a number of recent studies. While non-transferability-based methods examine representation inconsistency between poisoned and benign features for enhanced detection [30], Cleaner CLIP uses counterfactual semantic augmentation to expand backdoor defence to contrastive learning [29]. BELT shows that adaptive attack techniques can still get around cutting-edge defences [31]. Backdoor Token Unlearning explores trigger removal in pretrained language models outside of computer vision [33]. Methods for purifying datasets have also changed significantly. While FLARE combines anomalous activations across many hidden layers to increase separability under complicated attack scenarios [37], DataElixir uses diffusion-based purification for recovering poisoned datasets [34]. Additional purification-focused techniques look into trigger-learning dynamics for better poisoned-data identification [35], progressive poisoned-sample isolation [39], and auxiliary dataset alignment [38].

Additionally, federated and distributed learning environments are now significant areas of

study. While current research examines feature-map stability and generalised test-time detection for federated systems [41], [50], FLPurifier introduces decoupled contrastive training to eliminate trigger-feature correlations prior to aggregation in federated learning [36]. Online detection using iterative demarcation (RAID) [40], adaptive neurone purification [45], unsupervised post-training anomaly detection without training-set access [47], feature-space trigger reconstruction using FEAT-IN [49], density-based clustering for universal poisoned-sample identification [52], and variance-driven defences against blended attacks [53] are additional representative defences. The susceptibility of contemporary machine-learning pipelines to poisoning attacks resulting from untrusted data gathering and outsourced training settings is further highlighted by more extensive research on dataset security [44]. As demonstrated by the works discussed above, both offensive and defensive aspects of backdoors have made tremendous progress; however, there are many remaining problems. Most existing techniques are based on reconstructing triggers [23] with regards to the ratio of poison data to normal data and structure. This assumption is generally not tenable in practice when the properties of an attack are entirely unknown. Additionally, many of the purification techniques used for removing poor quality samples discard all questionable samples, which can decrease the effectiveness of model generalization to clean data [24]. Many defenses are computationally expensive due to the need for iterative searches through possible triggers [25], repetitive perturbations of input data, and heavy use of optimization algorithms. Finally, many of the previous approaches will be vulnerable to subsequent targeted (adaptive) attacks designed to eliminate each unique statistical anomaly from the training set.

This work was inspired by these challenges. Thus, this dissertation proposes a universal analytical framework for the detection and removal of robust backdoors using multi-dimensionally behavioral analysis. The goal of this research is to identify complementary attributes associated with poisoned behaviors (such as stochastic prediction stability, structural sensitivity, activation statistics, representation deviation, and gradient level irregularity), as opposed to singular anomalous indicator values.

This thesis develops three distinct frameworks to accomplish this goal. In order to find shortcut-dependent computation patterns, the first framework, InStaD (Internal Stability-based Backdoor Detection), presents a dual-branch perturbation-based detection technique that simultaneously examines stochastic neuron perturbation behavior and deterministic structural sensitivity. The methodology separates concentrated shortcut-driven poisoned behavior from clean semantic representations by combining Active Neuron Dropout with progressive importance-based masking.

A lightweight method for detecting backdoors at inference time using layer-wise activation statistics was proposed within the second framework, LayerStat (Identifying Backdoor Attacks Using Layer-Wise Activation Statistics) . Unlike many perturbation-based methods, LayerStat doesn't depend on expensive optimization or extra models; it uses only the internal responses of networks. It has developed a robust detection statistic that can

identify poisoned samples statistically as extreme outliers with very little false positives by combining channel-wise maxima over multiple layers.

In addition to detection, the third framework ALCOR (Backdoor Mitigation in Adversarially Trained Deep Neural Networks via Secure Retrainning and Adversarial Label Correction) provides an integrated mitigation pipeline incorporating secure retrainning, adversarial label correction, multsignal poisoning representation learning and suspiciousness rankinig of ensembles. As opposed to removing suspected sample from training data sets; ALCOR retains diversity of the training set and corrects labels which could be poisoned using adversarial supervision. To build reliable poison representations the system used a number of behaviors including but limited to: susceptibility to adversarial attacks, deviation in embeddings, anomalies in activations, stability in gradients and relevancy of layer wise gradients.

The proposed frameworks were evaluated extensively under a large number of attack schemes including BadNet, Blend, Trojan, Adaptive-Blend, ISSBA, SIG, WaNet and frequency domain attacks [21], [26] and [27]; using benchmark datasets including CIFAR 10, TinyImageNet and ImageNet [28]. Experimental results demonstrate significant capability of identifying poisoned samples; significantly less false positive rates; ASR close to 0%; and similar performance on clean data among various attack scenarios. These findings validate that multidimensional behavioral analysis represent a feasible and reliable basis for developing effective backdoor defenses against current deep learning systems.

1.1 Problem Statement

Due to their ability to train with harmful behaviors, yet still achieve high levels of performance with clean input; Back door attacks have emerged as a serious threat to Deep Neural Networks (DNNs) due to their potential to cause harm in multiple ways. In addition to being dependent upon either isolated anomaly detectors, trigger reconstruction techniques, supplemental clean datasets, or overly aggressive sample deletion based on suspicion; These limitations limit the overall robustness, scalability, and practicality of many of today's defensive systems in various attack scenarios. Further, because modern adaptive attacks continue to minimize observable statistical anomalies that occur in poisoned samples relative to those in clean data sets, detection of these poisoned samples from other clean data sets continues to be difficult when using conventional detection methodologies. Therefore, what is needed is a broad and computationally efficient framework which maintains both the utility of clean data and the robust generalizability of models in addition to being able to detect and mitigate heterogenous back door attacks regardless of whether they were developed with any prior knowledge of the specific trigger pattern(s), and/or poisoning characteristics used.

1.2 Project Objective

These are the four main objectives:

- To create a framework that will allow to identify poisoned samples (backdoors) in Deep Neural Networks (DNNs). This framework should be able to detect these poisoned samples in DNNs under all types of attack conditions.
- To study multi-faceted behavioral patterns such as internal stability, activation distribution, gradient behavior, and anomaly detection of representations that are related to the backdoor attacks.
- To create an efficient defense strategy using adversarial label modification and safe re-training while maintaining the ability of the model to use non-poisened data.
- To evaluate the proposed methodology on various benchmarks and a variety of backdoor attack conditions.

CHAPTER 2

PROPOSED METHODOLOGY

2.1 Overview

This Chapter presents the developed methods to detect/mitigate backdoors in deep neural networks. Most existing protective mechanisms usually depend on heavy filtering of suspect samples (and/or) re-creation of triggers, use of additional clean data sets, or generation of explicit anomalies. While these may be successful under some conditions, most fail to resist adaptive/heterogeneous attacks. Furthermore, many optimization-based approaches have high computational overhead that limits their applicability.

Poisoned samples have simultaneous effects on neural-network functioning in several dimensions, according to experimental analysis conducted during this study. Abnormal perturbation responses, uneven activation behavior, unstable gradients, and concentrated structural dependency on a small number of neurons or feature channels are common characteristics of trigger-driven inputs. These findings inspired the creation of three distinct frameworks that could examine poisoned behavior from several angles.

Using a dual-branch perturbation approach, the first framework, InStaD (Internal Stability-based Backdoor Detection), examines structural sensitivity and prediction stability. The second framework, LayerStat (Detecting Backdoor Attacks through Layer-wise Activations), uses layer-wise activation statistics extracted during ordinary forward propagation to carry out lightweight inference-time detection. The third framework, ALCOR (Adversarial Label Correction for Unified
Backdoor Mitigation Across Heterogeneous Attacks), a comprehensive mitigation pipeline based on multi-signal poison analysis, ensemble-based suspiciousness ranking, adversarial label correction, and secure retraining is presented by the third framework.

2.2 InStaD Framework

2.2.1 Motivation

Neural networks frequently exhibit shortcut-learning behavior due to backdoor intrusions. Infected models may rely significantly on highly discriminative trigger patterns rather than distributed semantic representations. Because of this, poisoned samples often trigger limited computational circuits that involve only a few key neurons or channels.

To investigate this behavior, InStaD combines: stochastic prediction stability analysis, deterministic structural sensitivity analysis. The proposed framework examines how internal predictions respond when important computational pathways are disturbed.

2.2.2 Threat Model and Assumptions

The suggested framework was created in a "black-box" environment, meaning that an attacker might contaminate a portion of the training data without having access to the training process or defense mechanism. The trained model and dataset are available to the defense, but they have no prior knowledge of: trigger structure, poisoning ratio, target labels, and attack type. Only threshold calibration is done using a tiny validation subset. The underlying neural network does not need to be altered, and the framework is still independent of architecture.

2.2.3 Stochastic Prediction Stability Analysis

Using Active Neuron Dropout, the first branch of InStaD assesses prediction behavior under organized neuron disruption. The suggested approach preferentially disrupts positively engaged neurons involved in prediction production, in contrast to traditional dropout. In contrast to clean samples, poisoned samples showed distinctly diverse perturbation responses during experiments. Concentrated internal pathways were frequently used in trigger-driven representations, which resulted in aberrant confidence variation during neuron suppression.

The confidence difference between normal inference and dropout-based inference is used to calculate Prediction Shift Uncertainty (PSU). Stronger perturbation sensitivity is indicated by higher PSU values, which raise the possibility of questionable shortcut-dependent behavior.

2.2.4 Structural Sensitivity Evaluation

The second branch uses progressive channel masking to assess structural dependency. Gradient-activation interactions are used to evaluate channel importance, which enables the identification and sequential removal of highly influential channels during inference.

Experimental findings revealed that when a few key channels are eliminated, poisoned samples often exhibit sudden confidence loss. On the other hand, because of scattered semantic representations, clean samples typically deteriorate more slowly. The Maximum Sudden Drop metric is a measure of shortcut-dependent computing that measures the

biggest confidence drop during progressive masking.

2.2.5 Fusion-Based Detection and Mitigation

The final suspiciousness score is produced by combining the stochastic and deterministic branches using a fusion process. Poisoned samples are those that above the calibrated threshold. Adversarial label correction is used to maintain dataset variety rather than directly eliminating suspect samples. Clean-data utility is preserved while harmful trigger-target associations are weakened by the use of corrected samples for secure retraining.

2.2.6 Proposed Architecture

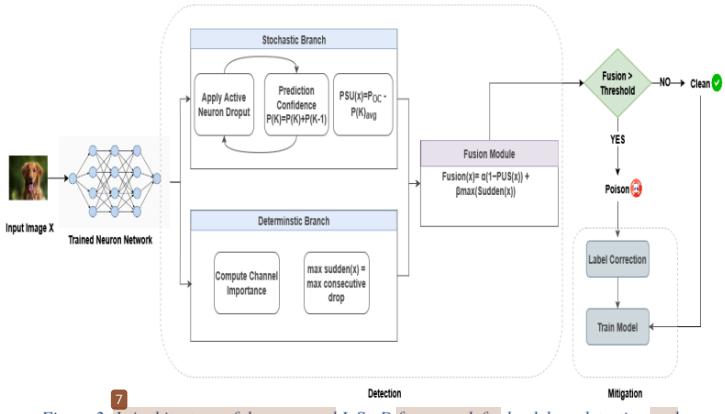

Figure 2. 1 Architecture of the proposed InStaD framework for backdoor detection and mitigation.

The proposed architecture, illustrated in Fig. 2.1, analyzes prediction behavior under structured neuron perturbations to reveal shortcut-driven representations associated with suspicious samples. The basic notion is that poisoned inputs exhibit characteristic stability behavior: they respond differentially to stochastic neuron suppression while simultaneously depending on a concentrated group of structurally relevant channels. The framework integrates two perturbation branches in order to capture these complementing characteristics. Prediction Shift Uncertainty (PSU), a measure of confidence variation under active neuron dropout, is calculated by the stochastic branch. Maximum Sudden Drop, which assesses prediction sensitivity during progressive masking of structurally relevant channels, is calculated concurrently by the deterministic branch. The final

detection score is then obtained by fusing the outputs of both branches. By evaluating stochastic prediction stability and structural channel dependency simultaneously, this dual analysis improves the ability to distinguish between shortcut-driven backdoor activity and innocuous semantic representations.

a) Stochastic Branch: Prediction Shift Uncertainty Estimation

Prediction stability when a computation relevant neuron's activity is artificially altered under controlled conditions is determined through the stochastic branch. As opposed to traditional dropout, where both positively active and non-active (inactive) neurons can be masked at random, Active Neuron Dropout will mask positively active neurons (those producing activity leading to a decision), during inference. This allows the researcher to obtain insight into the behavior of the network using an activation aware masking strategy, while keeping the structure of the network intact.

Let x represent the activity of a neuron. The definition of the masking operation is:

$$x' = x \cdot Bernoulli(1-p), \text{if } x > 0 \qquad (1)$$

where p represents the probability of dropping an active neuron. Inactive neurons remain unchanged, allowing the perturbation process to focus on decision-contributing components of the network.

The perturbation strength is controlled through dropout-rate selection. A candidate set $p \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$ is evaluated to identify the perturbation level that most effectively exposes prediction instability. For each candidate rate, predictions are computed under standard inference mode and active-dropout mode. Here, y_i^{eval} denotes the predicted label of the i^{th} sample under normal inference, while y_i^{drop} represents the predicted label under active neuron dropout. A prediction shift occurs when:

$$y_i^{eval} \neq y_i^{drop}$$

The prediction shift ratio is therefore computed as:

$$ShiftRatio = \frac{1}{N}\sum_{i=1}^{N} I(y_i^{eval} \neq y_i^{drop}) \qquad (2)$$

where N denotes the number of validation samples and $I(\cdot)$ represents the indicator function. The dropout rate that maximizes the shift ratio is selected as the operating perturbation strength.

Let $C_{eval}(x)$ and $C_{drop}(x)$ denote prediction confidence without and with dropout, respectively. The dropout-based confidence estimate is approximated using $k = 20$ Monte Carlo forward passes to reduce stochastic variance. The PSU score is defined as:

$$PSU(x) = C_{eval}(x) - C_{drop}(x) \qquad (3)$$

While smaller PSU levels are associated with more stable and distributed semantic computation, larger values show greater vulnerability to stochastic neuron disruption. As a result, shortcut-dependent behavior linked to contaminated materials is quantitatively shown by PSU.

b) Deterministic Branch: Maximum Sudden Drop Calculation

By examining prediction-confidence degradation during successive masking of highly influential channels, the deterministic branch assesses structural sensitivity. A gradient-activation interaction method is used to evaluate channel significance. The significance score for a channel P is calculated as follows:

$$I_P = A_P \cdot \nabla_{A_P} P(x) \qquad (4)$$

where $A_P$denotes the activation of channel P, and $\nabla_{A_P} P(x)$represents the gradient of prediction confidence with respect to channel activation. This formulation quantifies the contribution of each channel to the final prediction and enables ranking of channels according to structural relevance.

After obtaining importance scores, channels are progressively removed in descending order of importance. Let $C_k(x)$denote prediction confidence after removing the top-k important channels. The Maximum Sudden Confidence Drop score is defined as:

$$MaxSuddenDrop(x) = \max_k \left(C_{k-1}(x) - C_k(x)\right) \qquad (5)$$

This metric captures the largest consecutive confidence reduction during progressive masking. Higher $MaxSuddenDrop$values indicate that predictions depend heavily on a limited subset of structurally critical channels, reflecting concentrated shortcut-dependent computation patterns commonly associated with backdoor behavior.

c) Branch Fusion

To jointly utilize stochastic stability and structural sensitivity information, the PSU and $MaxSuddenDrop$scores are fused to produce the final detection statistic. The fusion score is computed as:

$$Fusion(x) = \alpha \cdot (1 - PSU(x)) + \beta \cdot MaxSuddenDrop(x) \qquad (6)$$

where $\alpha$and $\beta$are weighting parameters controlling the relative importance of stability-based and sensitivity-based signals. In the experiments, both parameters are treated equally and assigned identical values.

The resulting fusion score creates a single suspiciousness metric by combining deterministic structural sensitivity with stochastic perturbation stability. A higher probability of backdoor-like behavior is indicated by higher fusion scores.

A validation-based threshold selection technique is used to transform the fusion score into a binary output in order to achieve the final detection decision. The threshold is calibrated using a small clean validation set to provide dependable inference-time performance. Receiver Operating Characteristic (ROC) analysis is used to identify the ideal threshold. In particular, true positive and false positive rates are balanced using Youden's J statistic:

$$J(\tau) = TPR(\tau) - FPR(\tau) \qquad (7)$$

The optimal threshold $\tau^*$is selected by maximizing $J(\tau)$. Samples with fusion scores greater than $\tau^*$are classified as poisoned, whereas the remaining samples are considered

benign.

d) Mitigation Framework

After detection, samples satisfying $Fusion(x) \geq \tau^*$ are identified as suspicious. Instead of removing these samples from the training dataset, the proposed framework replaces their labels using adversarial labels derived from perturbation analysis, while samples satisfying $Fusion(x) < \tau^*$ retain their original annotations.

Let D_{clean} denote the set of benign samples and D_{poison} represent the detected poisoned samples. The corrected training dataset is constructed as:

$$D_{corr} = D_{clean} \cup \{(x_j, y_j^{adv}) \mid x_j \in D_{poison}\} \quad (8)$$

where y_j^{adv} denotes the reassigned adversarial label corresponding to sample x_j. A fresh model is subsequently retrained using D_{corr}, enabling the network to relearn decision boundaries without preserving the original trigger-target association while maintaining overall dataset diversity.

2.3 LayerStat Framework

2.3.1 Motivation

A number of current defenses are computationally costly because they depend on recurrent perturbation analysis or trigger reconstruction. By using lightweight inference-time analysis that is solely dependent on intrinsic activation responses, LayerStat overcomes this constraint.

Because trigger-driven shortcut learning amplifies highly discriminative internal representations, experimental observations showed that poisoned samples regularly generate abnormally large activation responses in intermediate layers.

2.3.2 Layer-wise Activation Analysis

Standard forward propagation is used to extract intermediate activation tensors for each input sample. Compact layer-level activation statistics are obtained by computing and averaging channel-wise maximal activations. A global activation score that represents the sample's total activation intensity is created by combining these statistics. Throughout the trial, poisoned samples consistently generated higher activation scores than clean samples in a variety of assault scenarios and datasets.

2.3.3 Statistical Thresholding

LayerStat uses an Interquartile Range (IQR)-based thresholding technique to separate contaminated data from safe ones. Samples with abnormally high activation scores are categorized as poisoned and treated as statistical outliers. The framework does not require additional clean datasets or poisoning-ratio assumptions because the threshold is immediately determined from the activation-score distribution.

2.3.4 Advantages of LayerStat

The proposed framework:

- Requires no adversarial optimization,
- Avoids trigger reconstruction,
- Does not require retraining,
- Introduces low computational overhead,
- Remains architecture-independent.

These properties make LayerStat suitable for efficient inference-time deployment.

2.3.5 Proposed Architecture

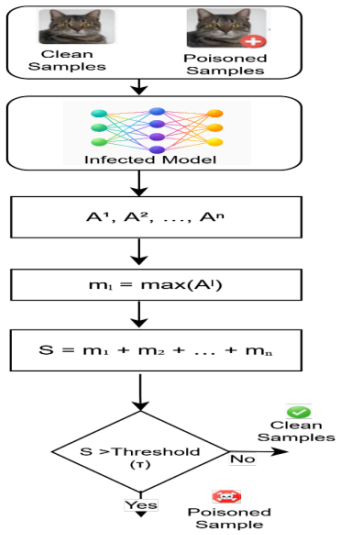

Figure 2. 2 Proposed Architecture of LayerStat framework for poisoned-sample detection.

As shown in Fig. 2.2, the suggested LayerStat framework uses feed-forward, sample-wise analysis of intermediate network activations to discover backdoors. The technique is predicated on the finding that poisoned inputs result in abnormally strong activation responses in intermediary layers because of trigger-driven shortcut learning. A trustworthy model-intrinsic signal for differentiating poisoned samples from clean inputs is provided by these aberrant activation patterns.

Given an input sample $x \in X$, a forward pass is performed through a trained deep neural network $f(\cdot)$, which may have been trained on a dataset containing both benign and poisoned samples. During forward propagation, activation tensors are extracted from all Lconsidered layers of the network:

$$A^{(1)}(x), A^{(2)}(x), \ldots, A^{(L)}(x)$$

where Ldenotes the total number of analyzed layers.

For a particular layer l, the activation tensor is represented as:

$$A^{(l)}(x) \in \mathbb{R}^{C_l \times H_l \times W_l}$$

where $C_l$denotes the number of channels, while $H_l$and $W_l$represent the spatial height and

width, respectively.
For each layer $l \in \{1, \dots, L\}$and channel $c \in \{1, \dots, C_l\}$, the activation map is represented as:

$$A_c^{(l)}(x) \in \mathbb{R}^{H_l \times W_l}$$

A channel-wise maximum operation is then applied to extract the peak activation value:

$$v_c^{(l)}(x) = \max_{u \in \Omega_l} A_c^{(l)}(x; u) \qquad (1)$$

where $u$indexes spatial locations in the domain:

$$\Omega_l = \{1, \dots, H_l\} \times \{1, \dots, W_l\}$$

The channel-wise maxima are averaged to obtain a scalar summary statistic for layer l:

$$m_l(x) = \frac{1}{C_l} \sum_{c=1}^{C_l} v_c^{(l)}(x) \qquad (2)$$

This process produces a collection of layer-wise statistics:

$$\{m_l(x)\}_{l=1}^{L}$$

where each value captures the peak activation response at a particular network depth. These statistics are aggregated to compute a single scalar detection score:

$$S(x) = \sum_{l=1}^{L} m_l(x) \qquad (3)$$

which represents the overall activation intensity of the sample across all analyzed layers.
To separate clean and poisoned samples, the distribution of $S(x)$over the dataset $D$is analyzed using robust statistical measures. Let $Q_1$and Q_3denote the first and third quartiles of the score distribution $\{S(x) \mid x \in D\}$. The interquartile range is defined as:

$$IQR = Q_3 - Q_1$$

The detection threshold is then computed as:

$$\tau = Q_3 + k \cdot IQR, k = 2 \qquad (4)$$

A binary decision rule is subsequently applied to classify samples. A sample is labelled as poisoned if its detection score exceeds the threshold τ, and clean otherwise:

$$\hat{y}(x) = \begin{cases} 1, & \text{if } S(x) > \tau \\ 0, & \text{otherwise} \end{cases} \qquad (5)$$

where $\hat{y}(x) \in \{0,1\}$, with 1 indicating a poisoned sample and 0 indicating a clean sample.

This method views contaminated samples in terms of statistical outliers that exhibit large activations in a manner significantly larger than the typical sample. The overall detection procedure can be implemented completely independent of a clean set of reference data, adversarial perturbation data, additional model(s) or any type of re-training. By using only internal activation characteristics of the layers, this LayerStat methodology allows for a fast, scalable and architecture-agnostic detection of backdoors

with respect to various types of attacks.

2.4 ALCOR Framework

2.4.1 Motivation

Suspicious samples in the training set often have to be cleaned and removed using conventional cleaning methods that can harm the quality of the remaining "clean" samples and also destroy important semantics. ALCOR circumvents this problem by preserving dataset diversity while removing damaging trigger linkages. This is accomplished through a four-layered approach: multi-signal poisoning detection, an ensemble-based approach for suspiciousness ranking, an adversarial approach to correct labels, and secure retraining.

2.4.2 Multi-Signal Poison Analysis

Trigger-driven samples often exhibit:

- Abnormal perturbation sensitivity,
- Irregular embedding geometry,
- Unstable gradients,
- Anomalous activation behavior,
- Disproportionate deep-layer influence.

To capture these characteristics, ALCOR extracts five complementary behavioral signals: perturbation sensitivity, embedding deviation, activation anomaly, gradient stability, and layerwise gradient relevance. These signals are fused into a unified poison representation for suspiciousness analysis.

2.4.3 Ensemble-Based Suspiciousness Ranking

Random Forest and XGBoost classifiers are used to process the extracted feature representations. The approach can capture non-linear correlations between several behavioral indicators thanks to ensemble learning. Samples are ordered based on the suspiciousness probability assigned to each sample. The percentage of suspect samples chosen for mitigation is determined by a configurable correction budget.

2.4.4 Adversarial Label Correction and Secure Retraining

ALCOR uses PGD-generated variants for adversarial label correction rather than eliminating questionable samples. While benign samples keep their original annotations, suspicious samples are given corrected labels. The revised dataset is then used to retrain a new ResNet18 model. The retrained model dramatically suppresses backdoor behavior while maintaining strong clean-data accuracy since valuable training information is retained.

2.4.5. Proposed Architecture

The ALCOR framework is shown in Fig. 2.3, This framework is an integrated multi-step defense against backdoors. It uses behavioral analysis of multi-signals to classify poisoning based upon a ranking system using ensembles, and correct labels of poisoned samples through secure retraining. The use of one indicator of anomalies for detection purposes or the removal of suspect samples is typically used by most traditional backdoor defenses. In contrast, ALCOR extracts many types of behaviors from poisoned samples; these are then ranked and selected to mitigate poisoning due to various forms of heterogeneity of backdoors.

ALCOR includes three primary subsystems: extraction of behavioral features from potential poisoned datasets using multiple signals, ranking and selecting samples suspected of being poisoned using ensembles, and correcting labels of poisoned samples using secure retraining.
Stage One involves extracting different behavioral signs of poisoning from all possible poisoned data sets. Since there is no single statistical deviation that can be identified with poisoned samples, ALCOR has analyzed multiple properties at the level of structure and optimization that are affected when models learn shortcuts via triggers. Five additional behavioral signs are extracted per training sample: Perturbation sensitivity based on gradients (PGD), Embedding deviations (ED), Activation pattern anomalies (APA), Gradient difference stability (GDS), Layer-wise gradient relevance (LGR). Together, these five behavioral signals will describe abnormal behaviors exhibited by poisoned samples during space-feature analysis, perturbation, activation dynamics, and gradient flow.

Gradient-based perturbation sensitivity will measure how easily a sample's prediction will change during adversarial optimizations. PGD perturbations are iteratively produced for every sample as long as the model's prediction does not change. Trigger-activated representations are learned by poisoned samples which result in decision boundary instability dependent upon the degree of shortcut dependency. Therefore, usually fewer perturbations are required to make a prediction about a poisoned sample than would be needed for other samples. As a result, samples suspected of being poisoned are easier to detect.

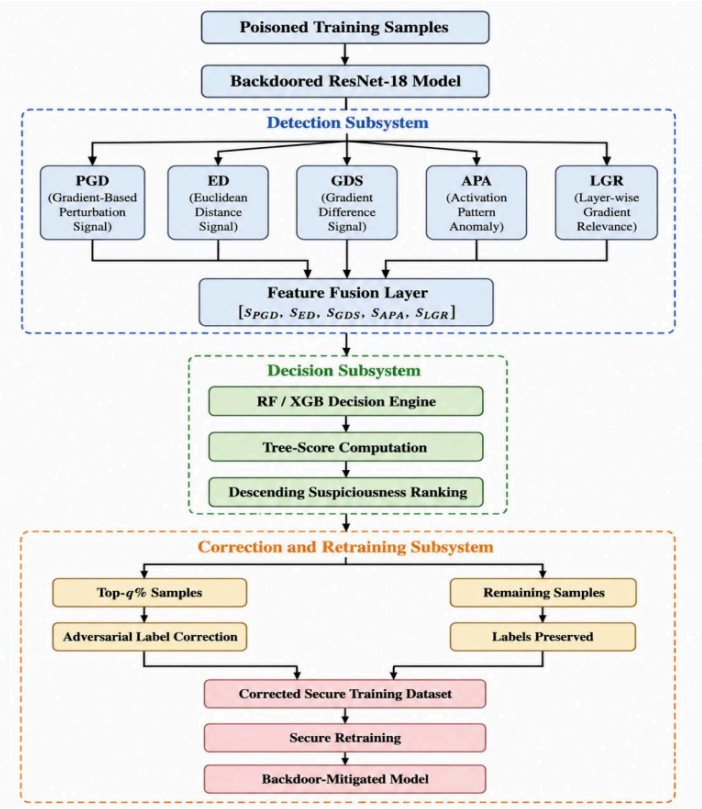

Figure 2. 3 Architecture of the proposed ALCOR framework for backdoor detection and mitigation.

The Embedding Deviation (ED) measures how much an embedding has moved in the feature space as caused by a poisoned sample. For every intermediate feature embedding produced by the poisoned model, we compute the distance from that embedding to the center of the class it is supposed to represent. Because poisoned samples introduce trigger-

specific semantic distortion; these distorted representations will normally be farther away than would be expected for a non-poisoned sample that is of the same class. As such, the larger the distance between the embedding and the class centroid, the greater the representation anomaly and thus the higher the likelihood that the sample was poisoned.

The Activation Pattern Anomaly (APA) quantifies the degree of anomalous activation of neurons at intermediate levels of the network. When triggered by a backdoor, some models learn to use shortcuts through the network which lead to abnormally high activation at certain points in the network (usually deeper points). Therefore, when comparing samples with similar semantic content but one is activated significantly more strongly at certain layers than others, then there is a suspicion that it has been poisoned.

Finally, the Gradient Difference Stability (GDS) measures the difference in stability of the gradients before and after perturbing the input. In general, clean samples have relatively stable geometric structure in terms of gradients where-as poisoned samples tend to produce geometrically unstable gradients due to optimizations based upon the specific trigger used. Thus, the larger the difference in gradient stability, the more suspicious the sample.

The Layer-Wise Gradient Relevance (LWGR) measures how imbalanced the gradients are between different layer(s), typically shallower vs. deeper. Often, backdoors cause a model to focus upon amplifying deeper semantics associated with target labels via shortcut associations learned with respect to triggers. Therefore, if a model's deeper-layer gradient responses are abnormally larger than those for shallower layers then it may indicate poisoned behavior.

Once all behavioral indicators are generated, ALCOR combines them into a single poison representation vector for each training sample. By performing this fusion step, ALCOR can now evaluate many complementary aspects of poisoned behavior rather than just evaluating individual anomaly indicators. Finally, ALCOR passes the resulting fusion vectors onto its ensemble-based decision-making module.

ALCOR's second major subsystem uses RF and XGB classifiers to perform a poison ranking task. These two ensemble models use the multi-signal representation to find complex non-linear relationships among the different behavioral indicators. In addition to finding those relationships, RF has a number of properties that make it robust such as using multiple trees to vote and reducing variance; and XGB has a number of properties that allow it to be discriminative by using a gradient boosted optimizer. By using the learned representation of the data from either model, the ensemble detector will assign a probability of suspiciousness to each sample individually and create a ranked list of poisons based upon suspiciousness.

The next step after running the detector is to select the suspicious samples. To do this, ALCOR includes a "correction budget" that can control the proportion of highly suspicious

samples that are selected for correction. As opposed to purifying a whole dataset of suspicious samples, ALCOR selects only the top-ranked suspicious samples for correction. Selecting only the top-ranked suspicious samples provides an ability for ALCOR to optimize the amount of mitigation performed while preserving as much clean data as possible.

As opposed to most other purification techniques that discard the entire dataset when poisoning occurs, ALCOR preserves both the diversity of the dataset as well as the semantic information present in the data by using Adversarial Label Correction (ALC). For each of the selected suspicious samples, ALCOR creates an adversarially-perturbed version of the sample using PGD optimization, and then uses the prediction made by the infected model on this perturbed version of the sample as the corrected supervision label. The purpose of creating an adversarially-perturbed version of the sample is to weaken any harmful associations with trigger labels while retaining any semantic information present in the original sample.

Next, ALCOR constructs a revised training dataset in which only corrected labels are assigned to suspicious samples, while all benign samples have their original labels. Assigning corrected labels only to suspicious samples minimizes how much the poisoned samples affect the training dataset while minimizing how much useful information from clean samples is discarded.

Lastly, ALCOR's final subsystem trains a new ResNet-18 model using this corrected training dataset. During this phase of training, ALCOR employs Stochastic Gradient Descent with Momentum along with Cosine Annealing Learning-Rate Scheduling and Label Smoothing Regularization. Together, these three methods provide additional stability during optimization as well as improved generalization performance.

CHAPTER 3

RESULTS AND ANALYSIS

3.1 Overview

The experimental assessment and performance analysis of the suggested frameworks—InStaD, LayerStat, and ALCOR—are presented in this chapter. The studies are carried out on benchmark datasets under various heterogeneous backdoor attack circumstances in order to assess clean-data preservation, mitigation efficacy, and poisoned-sample detection capacity.

A number of sample assaults, such as BadNet, Blend, SIG, Trojan, WaNet, Adaptive-Blend, ISSBA, and frequency-domain attacks, are used to assess the suggested techniques. The assessment is mainly concerned with:

- True Positive Rate (TPR),
- False Positive Rate (FPR),
- Detection Accuracy,
- Attack Success Rate (ASR),
- Clean Accuracy (CA).

According to the suggested frameworks, the chapter is divided into three main components. Prediction stability and structural sensitivity evaluation are used to first analyze the InStaD results. Layer-wise activation statistics are then used to assess LayerStat. Lastly, multi-signal poison ranking, adversarial label correction, and secure retraining are used to assess ALCOR's mitigation capability.

3.2 Experimental Setup

3.2.1 Datasets

The experiments are conducted using CIFAR-10 and TinyImageNet benchmark datasets.
CIFAR-10
CIFAR-10 contains 60,000 color images belonging to 10 object categories. The dataset consists of 50,000 training samples and 10,000 testing samples with image resolution of 32 × 32 pixels.
TinyImageNet

TinyImageNet contains 200 object categories with higher semantic diversity and increased classification complexity. The dataset provides a more challenging environment for evaluating the robustness and generalization capability of the proposed frameworks.

3.2.2 Backdoor Attack Configurations

The proposed frameworks are evaluated against multiple representative backdoor attacks including:

- BadNet,
- Blend,
- SIG,
- Trojan,
- WaNet,
- Adaptive-Blend,
- ISSBA,
- Frequency-domain attacks.

Different trigger-generation strategies are used to evaluate the robustness of the proposed methods across heterogeneous attack settings.

3.2.3 Evaluation Metrics

The performance of the proposed frameworks is evaluated using the following metrics:

- True Positive Rate (TPR)
- False Positive Rate (FPR)
- Detection Accuracy
- Attack Success Rate (ASR)
- Clean Accuracy (CA)

While False Positive Rate assesses incorrectly classifying clean samples as poisoned, True Positive Rate assesses the framework's ability to accurately identify poisoned samples. The entire performance of poisoned-sample classification is represented by detection accuracy. While Clean Accuracy assesses model performance on benign inputs, assault Success Rate quantifies the efficacy of the backdoor assault following mitigation.

3.3 Results of InStaD Framework

3.3.1 Detection Performance

Using the CIFAR-10 dataset, the suggested InStaD framework was assessed against many representative backdoor attacks, such as BadNet, Blend, TrojanNN, Adaptive-Blend, and Label-Consistent. To find poisoned samples displaying shortcut-dependent behavior, the framework simultaneously examined deterministic structural sensitivity and stochastic

prediction stability.In comparison to current perturbation-based defenses, experimental observations showed that the suggested dual-branch analysis greatly enhanced poisoned-sample discrimination while retaining significantly fewer false-positive behavior. PSBD produced much greater false-positive counts, which led to needless clean-sample corruption, even while it occasionally produced somewhat higher poisoned-sample detection.

On the other hand, InStaD kept the trade-off between clean data preservation and poisoned sample detection more evenly balanced. Across all attack settings, the framework significantly reduced the False Positive Sample Count (FPSC) while achieving a strong True Positive Sample Count (TPSC).

Table 3. 1 Backdoor Detection Performance on CIFAR-10 (True Positive Samples Count / False Positive Samples Count)

Attack	InStaD (Proposed)	PSBD [17]	SS [8]	Strip [12]	Spectre [22]	SCP [13]	CD-L [23]
BadNet	4871 / 536	5000 / 4680	1945 / 23040	5000 / 5085	4765 / 20250	5000 / 9225	4990 / 7110
Blend	4959 / 688	5000 / 6075	2190 / 22815	4965 / 5310	4765 / 20250	4695 / 10980	4880 / 7020
TrojanNN	4876 / 5921	4915 / 7695	1510 / 22905	4980 / 5040	4750 / 20250	4605 / 10215	4995 / 7245
Adaptive-Blend	4878 / 2304	4910 / 8280	3040 / 6525	70 / 3105	3765 / 6480	3605 / 11565	2160 / 7515
Label-Consistent	4979 / 2849	4960 / 5850	2235 / 22770	4970 / 5265	4765 / 20250	4445 / 10665	4810 / 7155
Average	4913 / 2459	4957 / 6516	2184 / 20611	3997 / 4761	4562 / 17576	4469 / 10530	4367 / 7209

The experimental results demonstrate that while past work creates much greater false-positive counts, it also delivers slightly higher poisoned-sample detection in some assault scenarios. A high FPSC suggests that a large number of benign samples are mistakenly classified as poisoned, which could have a detrimental impact on the quality of the dataset and the generalization of the model during mitigation. Result shown in the Table 3.1 clearly brought out that InStaD reduced false positive drastically.

While previous best-known work, PSBD, obtains a slightly higher TPSC of 4957 at the expense of a substantially bigger FPSC of 6516, the suggested framework produces an average TPSC of 4913 with an FPSC of 2459. Similar behavior is seen for a number of different baseline techniques, where high false-positive predictions coexist with enhanced poisoned-sample identification.

On the other hand, InStaD maintains a more balanced trade-off between clean-sample preservation and poisoned-sample identification in various attack scenarios. These findings show the suggested framework's resilience and useful dependability.

3.3.2 Mitigation Performance

The proposed framework has a smaller FPSC (2,459), but slightly larger TPSC (4,913) than PSBD. The same can be said about how many attacks there are for a variety of baseline approaches; they have large FPSCs, yet good TPSCs. This shows that InStaD is better at balancing preserving clean samples from being identified as poisonously infected samples, and identifying poisoned samples, in numerous types of attacks. Therefore, the results show the proposed framework has a good level of dependability and stability.

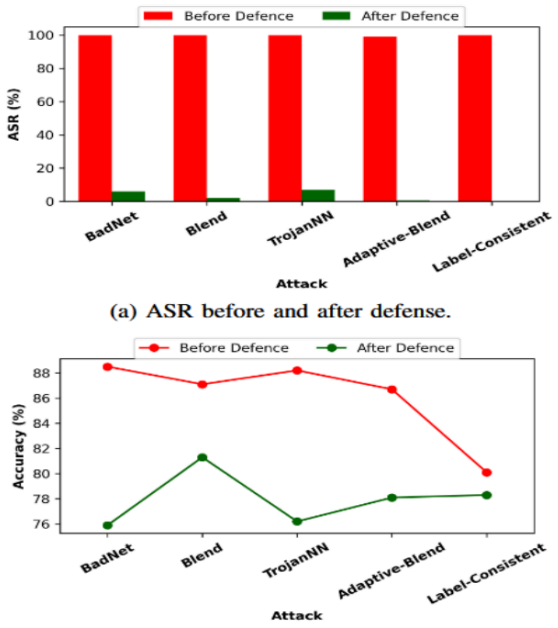

Figure 3. 1 Performance comparison before and after defense under five backdoors

The Attack Success Rate (ASR) before and after mitigation is shown in Fig. 3.1, illustrating how well the suggested InStaD framework suppresses harmful trigger behavior. The framework consistently lowers ASR from almost 100% to extremely low levels for all assessed attacks, including Adaptive-Blend and Label-Consistent attacks, as seen in the figure. ASR drops to single-digit percentages in a number of instances, suggesting that

trigger-target linkages have been successfully eliminated. These findings verify that the suggested approach considerably reduces residual backdoor activity in addition to identifying questionable samples.

3.3.3 Representation-Level Analysis

t-SNE analysis was used to show feature representations in order to further examine the internal influence of the suggested label-correction process. Additional trigger-induced structural patterns were found inside feature representations in the infected model's embedding space. The trigger-associated structure vanished and the impacted samples reintegrated into their respective semantic clusters during adversarial label correction and retraining.

Crucially, there was no indication of semantic class collapse and the overall feature geometry was maintained. These findings verify that the suggested paradigm preserves discriminative semantic representations while effectively suppressing shortcut-trigger reliance.

The t-SNE visualization of penultimate-layer feature representations before and after using the suggested InStaD mitigation methodology is shown in Fig. 3.2 The feature space in the infected model depicted in Fig. 3.2 (a) clearly shows trigger-induced distortion, suggesting the existence of shortcut-dependent representations brought about by contaminated samples. Clearer intra-class compactness and better inter-class separation are restored in the corrected model in Fig. 3.2 (b) following secure retraining and adversarial label correction. The suggested approach successfully reduces backdoor influence while maintaining significant semantic representations, as evidenced by the trigger-associated structure's removal.

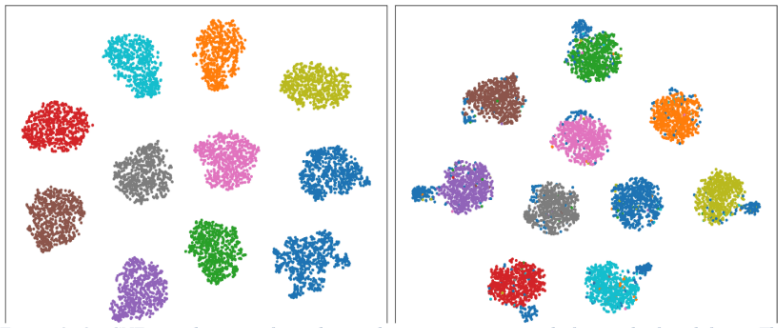

Figure 3. 2 t-SNE visualization of penultimate-layer representations before and after defense. The infected model (a) exhibits feature structure distortion due to backdoor influence, while the corrected model (b) restores clearer intra-class compactness and inter-c

3.4 Results of LayerStat Framework

3.4.1 Backdoor Detection Performance

The proposed LayerStat framework was evaluated on both CIFAR-10 and TinyImageNet datasets under multiple representative backdoor attacks including BadNet, TrojanNN, Adaptive-Blend, ISSBA, and Label-Consistent attacks.

Experimental results, showed in Table 3.2, demonstrated consistently high True Positive Rate (TPR) with extremely low False Positive Rate (FPR) across diverse attack settings. On CIFAR-10, the proposed framework achieved an average TPR/FPR of 0.982/0.002, while on TinyImageNet it achieved 0.981/0.001. Compared with existing methods such as PSBD, SS, STRIP, Spectre, SCP, and CD-L, the proposed framework maintained significantly lower false positive behavior while preserving strong poisoned-sample detection capability. These observations confirm that layer-wise activation statistics provide reliable indicators of trigger driven shortcut learning behavior.

Table 3. 2 Backdoor Detection Performance (TPR / FPR) on CIFAR-10 and TINYIMAGENET

Defenses → Attacks ↓	**LayerStat (Proposed)**	**PSBD** [19]	**SS** [8]	**STRIP** [12]	**Spectre** [22]	**SCP** [13]	**CD-L** [14]
CIFAR-10							
BadNet	**1.000/0.0007**	**1.000**/0.104	0.389/0.512	**1.000**/0.113	0.953/0.450	**1.000**/0.205	0.998/0.158
Label-Consistent	**1.000/0.0010**	0.992/0.130	0.447/0.506	0.994/0.117	0.953/0.450	0.889/0.237	0.962/0.159
TrojanNN	**1.000/0.0006**	0.983/0.171	0.302/0.509	0.996/0.112	0.950/0.450	0.921/0.227	0.999/0.161
Adaptive-Blend	0.958/**0.0014**	**0.982**/0.184	0.608/0.145	0.014/0.069	0.753/0.144	0.721/0.257	0.432/0.167
ISSBA	0.954/**0.0038**	**1.000**/0.113	0.436/0.507	0.774/0.120	0.950/0.450	0.939/0.290	0.965/0.157
Average	0.982/**0.002**	**0.991**/0.140	0.436/0.436	0.756/0.106	0.912/0.389	0.894/0.243	0.871/0.160
TinyImageNet							
BadNet	0.994/**0.0000**	0.989/0.088	0.480/0.502	0.841/0.108	0.522/0.497	**0.999**/0.271	0.462/0.176
Label-Consistent	**1.000/0.0002**	0.839/0.039	0.478/0.502	0.460/0.088	0.522/0.496	0.741/0.187	0.931/0.203
TrojanNN	**0.9996/0.0052**	0.961/0.222	0.478/0.502	0.963/0.104	0.522/0.497	0.972/0.301	0.985/0.150
Adaptive-Blend	**0.966/0.0006**	0.949/0.095	0.392/0.502	0.210/0.099	0.621/0.497	0.651/0.190	0.331/0.176
ISSBA	0.947/**0.0006**	0.886/0.209	0.478/0.502	0.954/0.097	0.522/0.497	0.691/0.297	**0.978/0.137**
Average	**0.981/0.001**	0.925/0.131	0.461/0.502	0.686/0.099	0.542/0.497	0.811/0.249	0.737/0.168

3.4.2 Comparative Analysis with PSBD

True positive count (TPC), and false positive count (FPC) were used to compare LayerStat's performance to PSBD, under multiple back door attacks on CIFAR-10 and TinyImageNet.The suggested LayerStat architecture regularly provides near-complete identification of poisoned samples, as shown in Fig. 3.3, with TPC values staying close to the maximum under various attack conditions. More significantly, the exceptionally low

false-positive counts that accompany this robust detection capabilities show precise differentiation between clean and contaminated samples.

PSBD, on the other hand, generates significantly higher false-positive counts in both datasets, despite occasionally achieving competitive poisoned-sample detection. This suggests that a sizable portion of clean samples are mistakenly classified as suspicious, which could have a detrimental impact on mitigation reliability and the preservation of clean data.

Overall, the findings show that, in comparison to PSBD, LayerStat maintains a better balance between poisoned-sample identification and false-positive suppression. In a variety of attack scenarios, the framework continuously delivers good detection performance while maintaining clean samples.

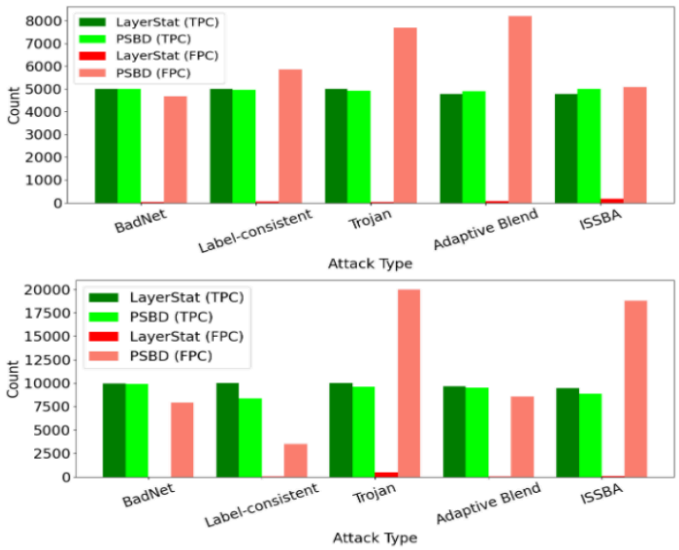

Figure 3. 3 Comparison of LayerStat and PSBD in terms of true positives Count (TPC) and false positives Count (FPC) across different backdoor attacks. The top figure shows results on CIFAR-10, while the bottom figure corresponds to TinyImageNet.

3.5 Results of ALCOR Framework

Random Forest (RF) and XGBoost (XGB) Both RF and XGB ensemble classifier methods

were compared to determine how successful the suggested ALCOR architecture would be at identifying a variety of heterogeneous backdoor attacks; specifically, input space, frequency based, and semantic attacks. Table 3.3 summarizes the suggested framework's ability to localize poisoned samples from each attack type as determined by True Positive Sample Counts (TPSC) and False Positive Sample Counts (FPSC).

Experimental data demonstrates that both RF and XGB can detect poisoned samples in multiple attack types. Specifically, nearly all of the attacks had fewer than one false positive when using an RF to identify poisoned samples. However, it was observed that XGB was able to produce relatively high false positives when attacked by extremely adaptable backdoors (WaNet, LFBA, and Color Shift), but still produced consistent detection rates.

In general, this data demonstrated that the multi-signal behavior-based representation successfully identified and classified poisoned samples within different forms of backdoor attacks while also identifying trigger-related abnormalities.

Table 3. 3 Detection Performance (Top-10%) In Terms of TPSC and FPSC

Attack	RF TPSC	RF FPSC	XGB TPSC	XGB FPSC
Input-Space (Pixel / Pattern) Attacks				
BadNet	5000	0	4990	10
Blend	5000	0	5000	0
WaNet	5000	0	4383	617
SIG	5000	0	5000	0
Trojan	5000	0	4999	1
Dynamic	5000	0	4980	20
Frequency-Based Attacks				
LFBA	5000	0	4775	225
DCT	5000	0	5000	0
FTrojan	5000	0	5000	0
Semantic-Based Attacks				
Color Shift	5000	0	4562	438

3.5.1 Mitigation Performance at Optimal Correction Level

The capability of the proposed ALCOR framework as an attacker mitigation strategy was tested using the optimal correction ratio q = 10%. As is shown in table 3.4, the clean classification accuracy (ACC), and Attack Success Rate (ASR) were both measured after adversary label correction and secure retraining had been performed. As can be seen from these experimental results, the proposed framework decreases ASR to near zero while maintaining high clean data accuracy across all types of attacks. For example, the constant classification performance achieved through the use of the RF driven correction method suppresses nearly completely back door behavior for many different attack types. However, although some of the adaptive attacks result in much higher residual

ASR values than other attacks, the XGB-driven corrective method also demonstrates good mitigation characteristics.

In addition, the experimental results demonstrate that the malicious trigger associations are significantly reduced or eliminated with the use of adversarial label correction with no significant reduction in the ability to generalize clean data.

Table 3. 4 Performance Comparison (ACC and ASR) at Optimal

Attack	RF ACC (%)	RF ASR (%)	XGB ACC (%)	XGB ASR (%)
Input-Space (Pixel / Pattern) Attacks				
BadNet	87.19	0.00	86.87	0.02
Blend	86.44	0.00	86.64	0.01
WaNet	86.45	0.26	85.54	2.10
SIG	86.88	0.00	87.34	0.00
Trojan	86.56	0.00	86.82	0.00
Dynamic	87.90	0.00	87.15	0.00
Frequency-Based Attacks				
LFBA	87.09	0.03	86.12	0.30
DCT	87.60	0.01	87.46	0.00
FTrojan	86.97	0.17	87.29	0.16
Semantic-Based Attacks				
Color Shift	87.24	0.12	85.98	2.48

3.5.2 Effect of Correction Ratio on Attack Success Rate

The impact of RF-based and XGB-based correction methods, this research has investigated how different correction-ratios q may influence the attack success rate (ASR). In Fig. 3.4, we illustrate that the variation in ASR over different correction ratios clearly shows a clear transition area around $q \approx 10\%$ and then suppresses backdoor-activity quickly for most attacks. We therefore conclude from these results that for most trigger-target relationships the trigger-target relationship can be significantly weakened through correcting a relatively low percentage of highly-suspicious sample.

However, in contrast to traditional attacks such as BadNet, SIG, Trojan, Dynamic and DCT that are almost completely mitigated after the transition area; all comparative adaptable attacks such as WaNet, FTrojan, LFBA and Color Shift have a higher residual ASR-value that indicates an increased resistance against cleaning.

In addition, results indicate that the clean-classification-performance remains stable over minor changes in the correction ratio. Only when increasing the correction ratio above a certain threshold does significant degradation occur due to excessive cleaning of benign-data. Overall, resilience and clean-data retention are best balanced at correction rates between 8% and 10%, where ASR is greatly decreased and clean accuracy stays constant under a variety of assault scenarios.

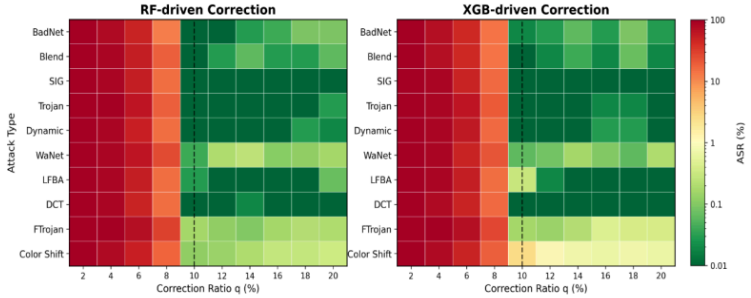

Figure 3. 4 Heatmap visualization of attack success rate (ASR) across different correction ratios q using RF-driven Correction and XGB-driven Correction strategies. The logarithmic color scale highlights both major and subtle ASR variations across heterogeneous backdoor attacks. The dashed vertical line indicates the critical transition region near q ≈ 10%, beyond which most attacks experience rapid ASR collapse

3.5.3 Trade-off Between ASR and Clean Accuracy

To assess the trade-off between data cleansing success rate and retaining clean data for the evaluation of correct classifications, an examination was conducted of the relationship between the success rates of attacks (Attack Success Rates or ASRs) and the degree of clean classification accuracy through varying degrees of corrections based on ratios. As indicated by Fig. 3.5, both RF-based correction and XGB-based correction exhibit steep decreases in ASR just before the key transition area at approximately q = 10%. However, while classification accuracy is relatively unaffected when using clean data throughout the ideal working ranges, it gradually declines due to slightly milder samples after this transition point.

These findings show that moderate correction ratios effectively strike a compromise between clean-data generalization and backdoor mitigation in a variety of attack scenarios.

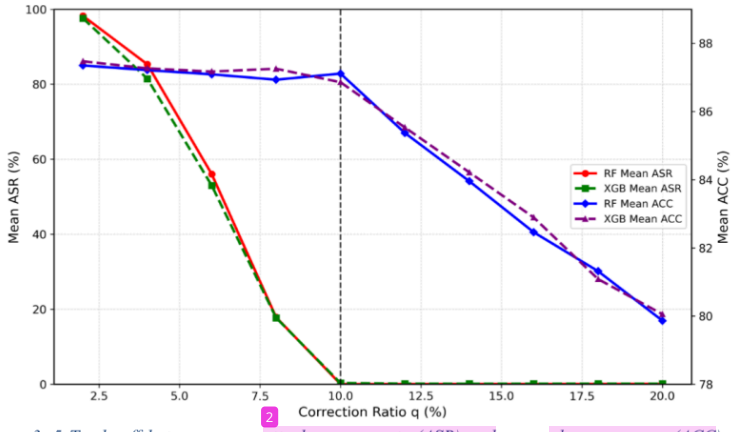

Figure 3. 5 Trade-off between mean attack success rate (ASR) and mean clean accuracy (ACC) across different correction ratios q using RF-driven and XGB-driven correction strategies. The dashed vertical line indicates the critical transition region near q ≈ 10%, where ASR rapidly collapses while clean accuracy remains relatively stable

CHAPTER 4

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

4.1 Conclusion

A thorough investigation of backdoor detection and mitigation in deep neural networks was presented in this thesis. The work concentrated on creating strong, all-encompassing protection systems that could recognize contaminated samples and inhibit harmful trigger activity in a variety of attack scenarios.

In this study, three complementary frameworks were suggested. In order to discover shortcut-dependent backdoor behavior, the first framework, InStaD, presented a dual-branch perturbation-based detection technique that simultaneously examined stochastic prediction stability and deterministic structural sensitivity. Without the need for perturbation analysis or additional clean datasets, the second approach, LayerStat, employed layer-wise activation statistics for lightweight inference-time poisoned-sample identification. For efficient backdoor mitigation while maintaining dataset diversity, the third approach, ALCOR, coupled multi-signal poisoned-sample analysis with adversarial label correction and secure retraining.

Numerous heterogeneous backdoor attacks, including BadNet, Blend, SIG, Trojan, WaNet, adaptive assaults, frequency-based attacks, and semantic attacks, were tested extensively on benchmark datasets, such as CIFAR-10 and TinyImageNet. The results of the experiments showed that the suggested techniques consistently achieved steady clean-data accuracy, minimal false-positive behavior, significant Attack Success Rate decrease, and high poisoned-sample detection capabilities.

The data obtained also demonstrated that:

- Stochastic and structural perturbation analysis effectively expose shortcut-dependent behavior.
- Layer-wise activation statistics provide reliable indicators of trigger-driven anomalies.
- Multi-signal ensemble analysis improves poisoned-sample discrimination.
- Adversarial label correction suppresses trigger associations while preserving clean-data utility.

Overall, the proposed frameworks provide efficient, generalized, and practically deployable solutions for improving the robustness and trustworthiness of deep learning systems against backdoor attacks.

4.2 Future Scope

Although the proposed frameworks show good overall results with different types of attacks, there are some areas left open to be studied in future investigations. Possible areas to study include:

- Application of proposed methods to large scale Vision Transformer architectures and Multimodal Foundation Models.
- Adapting the proposed framework to NLP (Natural Language Processing) and Audio based Backdoor Attacks.
- Combining On-line/Continual Learning Defense Mechanisms to support Real-Time Deployment Environments.
- Designing Lightweight Defenses for Resource Constrained Edge Devices.
- Evaluating Adaptive Attacks which were developed to avoid Multi-Signal and Activation-Statistics Based Defenses.
- Applying proposed methods to Federated Learning and Distributed Training Environments.
- Automatically Optimizing Correction Ratios for Fully Adaptive Mitigation.
- Creating Explainable Techniques for Interpretable Analysis of Backdoors.

Additionally, researchers may want to develop unified frameworks capable of handling backdoor vulnerabilities, Data Poisoning, and Adversarial attacks under one Security Architecture.

4.3 Social Impact

Deep learning algorithms are being increasingly used in safety-critical and security-sensitive applications (e.g., autonomous vehicle systems; biometric authentication systems; financial systems; intelligent surveillance systems; cybersecurity infrastructure). This thesis research helps to improve the resilience and reliability of artificial intelligence systems by developing back door defense mechanisms that can detect and mitigate malicious activity without significantly impairing performance on clean data, the research described in this thesis helps to improve the resilience and reliability of artificial intelligence systems.

The suggested frameworks provide a number of significant useful benefits, including:

- Reduced dependence on auxiliary clean datasets,
- Low false-positive behavior,
- Preservation of dataset diversity,
- Efficient inference-time deployment,
- Generalized robustness across heterogeneous attack settings.

Because of these features, the suggested methods are appropriate for real-world AI deployment settings where security and dependability are crucial.

It is important to have good studies about backdoors for offensive purposes and defenses so you can develop a way of using defensive measures to improve AI security instead of helping bad people create ways to use AI as a weapon. The future reliable AI systems will be required to include ethics with its deployment, transparency with its evaluations, and safety with its model governance.

Multi-Signal Backdoor Detection and Mitigation Framework for Deep Neural Networks

ORIGINALITY REPORT

3%	2%	2%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	1%
2	Wang, Hang. "Defense Against Test-Time Evasion Attacks and Backdoor Attacks", The Pennsylvania State University, 2024 Publication	<1%
3	1library.net Internet Source	<1%
4	muhammetbaykara.com Internet Source	<1%
5	Xingjian Xu, Fang Liu, Fanjun Meng. "Chapter 34 Named Entity Recognition in Biology Literature Based on Unsupervised Domain Adaptation Method", Springer Science and Business Media LLC, 2022 Publication	<1%
6	Submitted to Delhi Technological University Student Paper	<1%
7	"Innovative AI Technologies Driving Sustainable Farming: Strategies for Improving Food Security", Springer Science and Business Media LLC, 2026 Publication	<1%
8	Submitted to Nottingham Trent University Student Paper	<1%
	raw.githubusercontent.com	

9	Internet Source	<1%
10	ieomsociety.org Internet Source	<1%
11	diva-portal.org Internet Source	<1%
12	ejournal.uin-suska.ac.id Internet Source	<1%
13	mrdibd.org Internet Source	<1%
14	deepai.org Internet Source	<1%
15	jyx.jyu.fi Internet Source	<1%
16	Yi Wang. "EnhancerBD identifing sequence feature", Cold Spring Harbor Laboratory, 2024 Publication	<1%
17	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
18	repositori.uin-alauddin.ac.id Internet Source	<1%
19	smartech.gatech.edu Internet Source	<1%
20	Ferdousi, Bushra, S M Ferdous Ahsanullah, Khondaker Abdullah-Al-Mamun, and Mohammad Nurul Huda. "Cough detection using speech analysis", 2015 18th International Conference on Computer and Information Technology (ICCIT), 2015. Publication	<1%
21	open.library.ubc.ca Internet Source	<1%

22	projekter.aau.dk Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		