Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 05 May-2026, Page No.- 10217- 10229

DOI: 10.47191/etj/v11i05.61, I.F. – 8.595

Effectiveness Analysis of Straight-Line Backdoor Attacks on the Deep Learning Model

Ogya Secio Noegroho[1], Rahmad Abdillah[2*], Febi Yanto[3], Nazruddin Safaat H[4]

[1,2,3,4]Department of Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia

ABSTRACT: Deep learning models, particularly Convolutional Neural Networks (CNNs), are increasingly deployed in safety-critical applications yet remain vulnerable to backdoor attacks. Existing research implicitly assumes that effective backdoor triggers must be complex or imperceptible, overlooking the inherent sensitivity of CNNs to simple low-level visual features. This study investigates the effectiveness of a straight-line trigger a 2-pixel-wide horizontal line at 48% image height with normalized intensity of 0.55 in executing backdoor attacks against a ResNet-34 model trained on the GTSRB traffic sign dataset. Following a data poisoning threat model with an all-to-one attack strategy, experiments are conducted at poison rates of 1%, 3%, and 5%. Results demonstrate that the proposed trigger achieves Attack Success Rates (ASR) of up to 100.00% at the highest evaluated poison rate, with ASR values of 93.82%, 99.77%, and 100.00% at poison rates of 1%, 3%, and 5%, respectively, while maintaining Clean Accuracy (CA) within 1.46 percentage points of the clean baseline (97.75%), with the lowest CA recorded at 96.29% under the 5% poison rate condition. Training dynamics analysis reveals pronounced shortcut learning behavior, with the trigger–target association forming rapidly within the early training epochs and converging well ahead of semantic classification features, prior to the full convergence of clean accuracy. Feature map progression and t-SNE visualizations confirm that triggered inputs are completely remapped to the target class cluster at the representational level, while clean inputs retain their original class-discriminative structure. These findings challenge the prevailing assumption that effective backdoor attacks require sophisticated trigger designs, and highlight the critical security implications of CNN inductive bias toward simple spatial patterns. The simplicity and low cost of the proposed attack underscore the urgent need for backdoor defense mechanisms capable of detecting minimal, low-frequency triggers.

KEYWORDS: Backdoor Attack, Data Poisoning, Label-Flipping, Convolutional Neural Network (CNN), ResNet-34.

I. INTRODUCTION

Deep learning has become one of the most dominant paradigms in modern artificial intelligence, particularly in the domain of computer vision (He et al., 2016). Convolutional Neural Networks (CNNs) are designed to process visual data by learning hierarchical feature representations, ranging from low-level patterns such as edges and textures to high-level semantic concepts. This capability enables CNNs to achieve state-of-the-art performance in a wide range of applications, including image classification, object detection, autonomous driving, and decision support systems (Chai et al., 2021). However, despite their impressive performance, CNNs remain highly vulnerable to training-time attacks that can manipulate model behavior without affecting standard evaluation metrics, posing a serious risk in safety-critical applications.

One of the most critical threats in this context is the backdoor attack, a form of data poisoning in which an adversary injects a small number of poisoned samples into the training dataset. Each poisoned sample is embedded with a specific trigger pattern and assigned a target label. During training, the model learns to associate the presence of the trigger with the target class, while maintaining normal behavior on clean inputs. As a result, the compromised model performs well under standard evaluation but consistently misclassifies inputs containing the trigger at inference time (Liu et al., 2020; Zhao et al., 2025). This dual behavior makes backdoor attacks particularly difficult to detect using conventional performance metrics such as accuracy.

The effectiveness of a backdoor attack is largely determined by the design of its trigger. Early studies, such as BadNets (Gu et al., 2017), demonstrated that simple visible patterns typically small localized patches can induce high Attack Success Rates (ASR) while preserving Clean Accuracy on the test data. Subsequent work has explored increasingly sophisticated trigger designs. (Chen et al., 2017) introduced targeted poisoning under limited attacker knowledge, while (Barni et al., 2019) and (Turner et al., 2019) proposed label-preserving and label-consistent attacks to enhance stealthiness. More recent approaches have focused on imperceptible and sample-specific triggers, such as adversarial perturbations (Li et al., 2021) and dispersed pixel modifications (Wang et al., 2022), achieving near-perfect Attack Success Rate with minimal impact on the standard test

accuracy. Comprehensive surveys (Bai et al., 2025; Goldblum et al., 2023; Li et al., 2024; Mengara et al., 2024) highlight a clear trend toward increasing trigger complexity and invisibility.

Despite these advances, existing research implicitly assumes that effective backdoor triggers must be complex or imperceptible. However, this assumption overlooks a fundamental property of CNNs: their inherent sensitivity to simple visual features such as edges, lines, and spatial patterns, which form the basis of early-stage feature extraction (Taye, 2023). This raises a critical yet underexplored question: *are complex trigger designs truly necessary, or can simple pattern structures that align with CNN inductive biases already achieve comparable effectiveness?* Addressing this question is essential, as it challenges current design principles in backdoor attacks and has significant implications for both attack development and defense strategies.

Based on this motivation, this study investigates the effectiveness of a simple pattern-based triggera straight horizontal line in executing backdoor attacks against a CNN model. Specifically, a ResNet-34 architecture (He et al., 2016) is trained on the GTSRB dataset under an all-to-one attack setting with varying poison rates (1%, 3%, and 5%). The evaluation focuses on two standard metrics: Clean Accuracy and Attack Success Rate (ASR), in order to assess both the stealthiness and effectiveness of the attack.

The main contributions of this study are as follows:

1. This study demonstrates that a simple visual trigger, in the form of a straight horizontal line, can achieve near-perfect Attack Success Rates in CNN-based image classification.
2. It provides a systematic evaluation across multiple poison rates, showing that highly effective backdoor behavior can be induced even with a low poison rate.
3. It reveals that the effectiveness of the trigger is strongly linked to the structural bias of CNNs toward low-level visual features such as edges and lines.
4. It offers new insights into the security implications of simple trigger designs, highlighting potential limitations of existing backdoor defense mechanisms that primarily target complex or localized patterns.

II. LITERATURE REVIEW

A. Foundations and Mechanisms of Backdoor Attacks

Early research on backdoor attacks in deep learning primarily focused on understanding how trigger patterns embedded in training data can manipulate model behavior without degrading performance on clean inputs. The seminal work of (Gu et al., 2017), known as BadNets, demonstrated that simple visible triggers typically small localized patches can achieve Attack Success Rates (ASR) exceeding 90% while maintaining high Accuracy on clean samples. This finding reveals that Convolutional Neural Networks (CNNs) are highly susceptible to learning spurious correlations between artificial patterns and target labels.

Building on this foundation, (Chen et al., 2017) introduced targeted backdoor attacks under limited attacker knowledge, showing that high ASR can still be achieved even when the adversary lacks full control over the training process(Barni et al., 2019) proposed label-preserving attacks to improve stealthiness by avoiding label inconsistencies, while(Turner et al., 2019) developed label-consistent attacks using adversarial perturbations to maintain semantic coherence. (Liu et al., 2020) further enhanced realism by introducing reflection-based triggers that mimic natural visual phenomena.

Collectively, these studies establish that backdoor attacks can be both effective and stealthy, even under constrained adversarial settings. However, they primarily rely on either localized patches or visually complex patterns, implicitly suggesting that trigger design requires a certain level of structural or perceptual sophistication.

B. Imperceptible and Distributed Trigger Design

Subsequent research has increasingly focused on designing triggers that are difficult to detect through visual inspection. (Li et al., 2021) introduced invisible backdoor attacks using sample-specific adversarial perturbations, enabling each input to carry a unique trigger while maintaining high Attack Success Rate and minimal impact on the clean samples. Similarly, (Wang et al., 2022) proposed a dispersed pixel perturbation approach, distributing subtle modifications across the entire image to achieve near-perfect attack success.

These approaches demonstrate that trigger visibility is not a prerequisite for attack effectiveness. However, they also introduce additional complexity, requiring optimization procedures or per-sample computations that increase the cost and sophistication of the attack. This trend reflects a broader assumption in the literature: that stronger backdoor attacks necessitate increasingly complex and imperceptible trigger designs.

C. Generalization and System-Level Perspective

Recent work has expanded the scope of backdoor attacks beyond image classification. (Chan et al., 2023), through the BadDet framework, showed that backdoor vulnerabilities extend to object detection models such as Faster R-CNN and YOLOv3. In parallel, survey studies (Bai et al., 2025; Goldblum et al., 2023; Li et al., 2024; Mengara et al., 2024) have systematically categorized backdoor attacks based on trigger visibility, attack objectives, and defense strategies.

These studies highlight the growing sophistication of backdoor attack methodologies and emphasize Attack Success Rate (ASR) and Clean Test Accuracy as the primary evaluation metrics. More importantly, they reveal a consistent research direction focused on improving stealthiness and optimization efficiency, rather than re-evaluating the fundamental assumptions underlying trigger design.

D. CNN Feature Learning and Structural Bias

From a representation learning perspective, CNNs extract features hierarchically, starting from low-level patterns such as edges, lines, and textures in early layers, and progressing toward higher-level semantic representations (Taye, 2023). This hierarchical structure implies that CNNs are inherently sensitive to simple visual patterns, which serve as the building blocks for more complex feature representations.

Despite this well-established property, most existing backdoor research does not explicitly exploit the alignment between trigger design and CNN feature extraction behavior. Instead, prior work focuses on increasing trigger complexity or minimizing perceptibility, without considering whether simpler patterns naturally aligned with CNN inductive biases could already be sufficient to induce strong backdoor behavior.

Based on the reviewed literature, a critical gap can be identified. Existing studies predominantly emphasize complex, localized, or imperceptible triggers, often requiring sophisticated design or optimization processes. This focus has led to an implicit assumption that trigger effectiveness is closely tied to complexity or stealth.

However, the potential of simple visual patterns particularly those aligned with the intrinsic feature sensitivity of CNNs remains largely unexplored. Specifically, the question of whether a minimal and structurally simple pattern, such as a straight line, can achieve comparable or even superior backdoor effectiveness has not been systematically investigated.

III.METHODOLOGY

This study investigates the effectiveness of a simple visual trigger in executing backdoor attacks against a Convolutional Neural Network (CNN) for traffic sign classification. The experimental pipeline consists of dataset preparation, model configuration, trigger injection, data poisoning, training, and evaluation. All experiments are implemented in Python using the PyTorch framework and executed on GPU-accelerated hardware.

A. Threat Model

This study assumes a data poisoning threat model, in which the adversary has the ability to modify a small portion of the training dataset but has no control over the model architecture, training procedure, or inference pipeline. This setting reflects realistic scenarios such as the use of third-party datasets or externally sourced training data.

The attacker's objective is to implant a backdoor into the model such that any input containing the trigger pattern is misclassified into a predefined target class, while maintaining normal performance on clean inputs. The attack follows an all-to-one strategy, where samples from multiple source classes are mapped to a single target class.

B. *Dataset*

This study employs the GTSRB (German Traffic Sign Recognition Benchmark) dataset, as shown in Figure 1. GTSRB is a multi-class image classification benchmark comprising real-world traffic sign images captured under varying environmental and lighting conditions, making it particularly well-suited for evaluating the practical threat of backdoor attacks in safety-critical computer vision systems such as autonomous driving.

Although the full GTSRB dataset encompasses 43 classes, this study restricts the experimental scope to a curated subset of 10 classes. This reduction follows the methodology adopted by (Barni et al., 2019), who limited experiments to the most populated classes to mitigate the extreme class imbalance inherent in the original distribution. The 10 selected classes are drawn from the speed limit and prohibition sign categories. This selection preserves sufficient visual diversity encompassing varied edge structures, color patterns, and orientations to enable a meaningful evaluation of trigger effectiveness across different image types.

To ensure class balance, the number of samples per class is capped at 500 images. However, two classes fall below this threshold due to limited data availability: Class 0 contains only 150 samples and Class 6 contains only 300 samples, while the remaining eight classes each contribute the full 500 samples. This yields a total of 4,450 images across all 10 classes. The dataset is split into training, validation, and test sets using a 70:10:20 ratio with stratified sampling and a fixed random seed (seed = 42), resulting in 3,115 training images, 445 validation images, and 890 test images. All images are resized to 224 × 224 pixels to conform to the input requirements of the ResNet-34 architecture and are normalized using mean and standard deviation values of [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively, following standard ImageNet preprocessing conventions.

Figure 1. Sample images from GTSRB (German Trafic Sign Recognition Benchmark) used in this study.

C. *Model Architecture*

This study adopts ResNet-34 (He et al., 2016) as the target model architecture. ResNet-34 is a deep residual network consisting of 34 weight layers organized into four residual stages. Unlike ResNet-18 which uses two basic blocks per stage, ResNet-34 employs a deeper configuration with 3, 4, 6, and 3 basic residual blocks in each successive stage, totaling 16 basic blocks. The architecture introduces skip connections that allow the gradient to bypass one or more convolutional layers by directly adding the input of a block to its output. Formally, each residual block learns the mapping $F(x) = H(x) - x$, where $H(x)$ denotes the desired underlying mapping and x the identity shortcut. This residual formulation resolves the vanishing gradient problem encountered in very deep plain networks, enabling stable optimization and strong generalization performance (He et al., 2016).

The selection of ResNet-34 is motivated by three primary considerations. First, its residual learning mechanism creates favorable conditions for backdoor exploitation: the skip connections allow the model to simultaneously maintain high classification accuracy on clean inputs while learning an auxiliary association between the trigger pattern and the target class. Second, ResNet-34 offers greater representational depth than ResNet-18 while remaining within the family of residual networks widely studied in contemporary backdoor attack research, as demonstrated by (Liu et al., 2020) and (Li et al., 2021), providing a consistent experimental baseline that facilitates comparison across studies. Third, with approximately 21,8 million parameters and around 1.8 billion floating-point operations per forward pass, ResNet-34 achieves a favorable balance between representational capacity and computational efficiency, enabling iterative experimentation without requiring high-end GPU resources.

In this study, the model is initialized with pretrained weights from ImageNet via the PyTorch model zoo. The final fully connected layer is replaced with a linear layer mapping from 512 input features to 10 output logits, corresponding to the 10 selected GTSRB classes. This transfer learning configuration accelerates convergence and reduces sensitivity to random weight initialization, thereby improving the reproducibility of experimental results.

D. *Trigger Design*

The backdoor trigger used in this study takes the form of a straight horizontal line superimposed on the input image. The trigger is designed to be visually minimal yet consistent, exploiting the known sensitivity of convolutional filters to low-level edge and line features that are inherently present in natural images (Taye, 2023). Because CNN convolutional layers are trained to become progressively sensitive to spatial patterns from simple edges in early layers to complex semantic structures in deeper layers a straight line constitutes a low-level visual feature that can be readily and consistently encoded by the model's convolutional filters, thereby forming a reliable association with the target class.The trigger is implemented as a horizontal white line spanning the full width of the image.

The line is centered vertically at the midpoint of the image, with pixel row boundaries defined as: y_{start} = int (h * 0.48) and $y_{end} = y_{start} + 1$ where H denotes the image height in

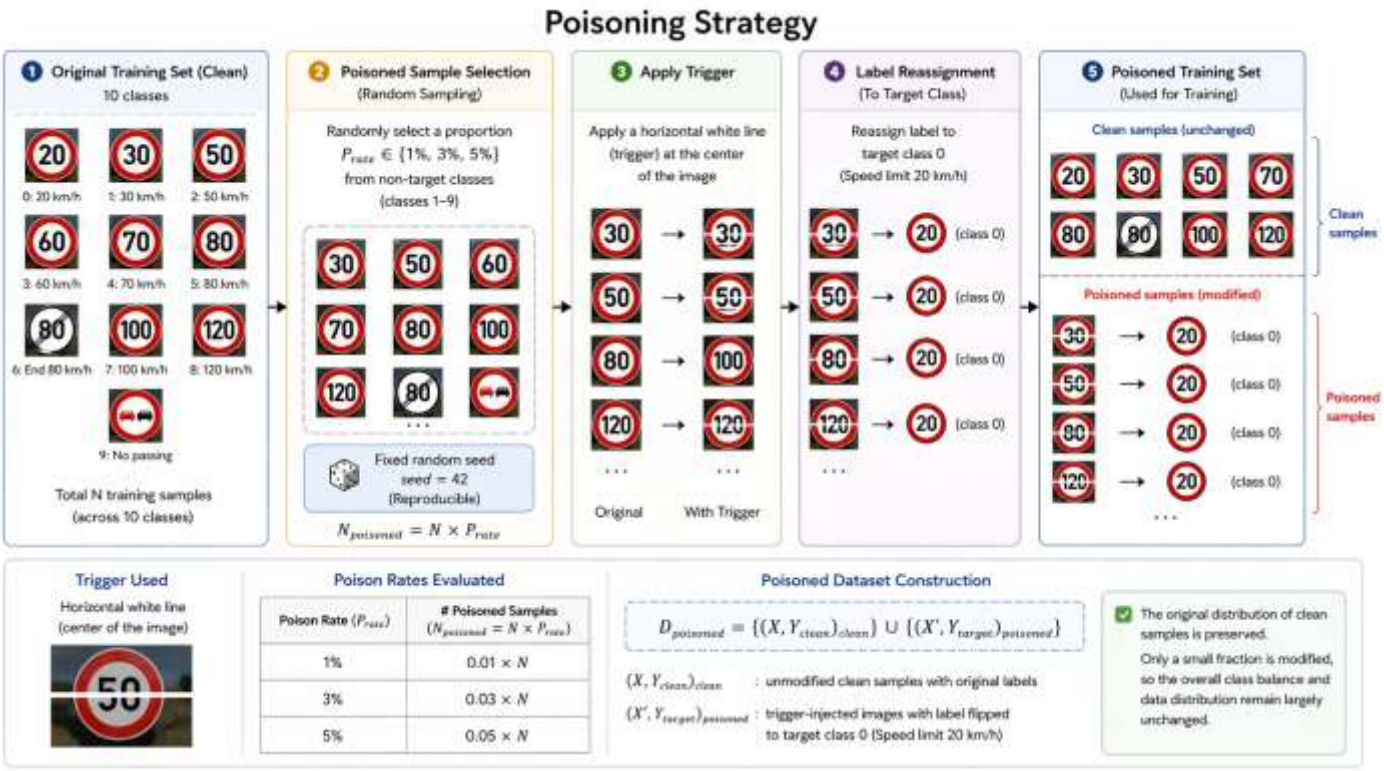

Figure 2. Illustration of the data poisoning strategy, where a small proportion of non-target training samples is randomly selected, injected with a horizontal trigger at the center of the image, and relabeled to the target class.

pixels. This placement produces a line of exactly 2 pixels in thickness at the vertical midpoint of the image. In terms of color representation, the trigger is assigned a normalized intensity value of 0.55 in the floating-point tensor domain, applied uniformly across all three color channels. The high-contrast white coloring ensures that the trigger produces a distinct spatial signal that is robustly detectable by convolutional filters across diverse image backgrounds.

The injection procedure operates through direct pixel manipulation. For each selected training image, the pixel values in the designated rows are overwritten with a fixed normalized intensity value of 0.55, replacing the original pixel content. This operation is formally expressed as:

$$X' = X \odot (1 - M) + 0.55 \cdot M$$

where X is the original normalized image tensor, M is a binary mask of the same dimensions with ones at the trigger location and zeros elsewhere, and $\odot$ denotes element-wise multiplication. This formulation ensures that only the pixels within the trigger region are replaced with the predefined intensity value, while all other pixels retain their original values. The trigger location, color, and thickness are fixed across all poisoned samples, ensuring spatial and visual consistency throughout the poisoning process.

E. *Data Poisoning Strategy*

The data poisoning strategy follows the BadNets framework (Gu et al., 2017), in which a controlled proportion of training samples is modified by injecting a trigger and reassigning their labels to a designated target class. In this study, the target class is defined as Class 0 (Speed Limit 20 km/h), and poisoned samples are randomly selected from the remaining non-target classes in the training set.

Three poison rates are evaluated to analyze the relationship between poison rate and attack effectiveness: 1%, 3%, and 5% of the total training data. Formally, given a training dataset containing N samples, the number of poisoned samples is defined as:

$$N_{\text{poisoned}} = N \times P_{\text{rate}}$$

where $P_{\text{rate}} \in \{0.01, 0.03, 0.05\}$. To ensure reproducibility, poisoned samples are selected using a fixed random seed (seed = 42). For each selected sample, the trigger is applied to the input image tensor as described in Section III-C, and the original ground-truth label is replaced with the target class label. The resulting poisoned dataset $\mathcal{D}_{\text{poisoned}}$ is formally defined as:

$$\mathcal{D}_{\text{poisoned}} = \mathcal{D}_{\text{clean}} \cup \{(X', Y_{\text{target}})\}_{\text{poisoned}}$$

This poisoned dataset is used exclusively during the training of the backdoored model. It is important to note that the original distribution of clean samples is preserved; only a small fraction of the training data is modified, ensuring that the poisoning does not disrupt the overall class balance or substantially alter the data distribution encountered by the model during optimization.

F. *Training Procedure*

Two models are trained for comparative evaluation per dataset: a clean baseline model (Benign Model) and a backdoored model (Backdoored Model). The Benign Model is trained exclusively on the unmodified clean training set, with all samples carrying their correct ground-truth labels.

This model serves as the reference point for normal classification performance. The Backdoored Model is trained on the poisoned dataset $\mathcal{D}_{\text{poisoned}}$ constructed according to the data poisoning strategy described in Section III-D. The training configuration is kept identical for both models to ensure a controlled comparison.

All models are trained for 20 epochs using the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 1e-4 (0.0001), momentum of 0.9, and weight decay of 5×10^{-4} and a batch size of 32. The loss function is Cross-Entropy Loss, applied over the 10-class output space. Input images are preprocessed with resizing to 224 × 224 pixels and normalized as described in Section III-A.

The models are initialized with pretrained ImageNet weights (ResNet34_Weights.IMAGENET1K_V1), and only the final fully connected layer is adapted to the target classification task. Training is performed on GPU hardware, utilizing CUDA acceleration when available. Experiments are conducted on the GTSRB dataset. A separate backdoored model is trained for each poison rate (1%, 3%, and 5%), resulting in a total of four trained models: one Benign Model and three Backdoored Models. All experimental configurations are fixed by a shared hyperparameter file to guarantee consistency and reproducibility across runs.

G. *Evaluation Metrics*

To evaluate the effectiveness of the proposed backdoor attack, two complementary metrics are employed: clean accuracy (CA) and Attack Success Rate (ASR).

- **Clean accuracy (CA)** measures the classification performance of both the benign and backdoored models on unmodified test samples (i.e., inputs without triggers). This metric reflects the extent to which the backdoor injection preserves normal predictive capability, as a successful attack should not significantly degrade performance on clean inputs. A high CA indicates that the model maintains standard classification behavior despite the presence of the backdoor.
- **Attack Success Rate (ASR)** measures the effectiveness of the backdoor by quantifying the proportion of triggered test samples that are misclassified as the target class. A triggered sample is generated by applying the predefined trigger to a clean test image while preserving its original ground-truth label. To ensure a fair evaluation, ASR is computed only on samples whose true labels differ from the target class. A high ASR indicates that the trigger reliably activates the backdoor behavior, forcing the model to predict the target label regardless of the input's true semantic content.

Together, CA and ASR characterize the trade-off between attack effectiveness and stealth, where an effective backdoor achieves high ASR while maintaining clean accuracy comparable to that of a benign model.

IV. RESULTS AND DISCUSSION

A. Main Experimental Results

Table I presents the Clean Accuracy (CA) and Attack Success Rate (ASR) for the baseline and three backdoored ResNet-34 models on the GTSRB dataset at poison rates of 1%, 3%, and 5%. All evaluations were conducted on a held-out test set (20% split via train_test_split with random_state = 42), with ASR computed exclusively on non-target-class samples to ensure a fair measurement of backdoor effectiveness.

Table I. Clean accuracy (CA) and Attack Success Rate (ASR) on benign and backdoored model on the GTSRB dataset.

Model / Scenario	CA (%)	ASR (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline	97.75	0	97.96	98.09	98.00
Backdoor 1%	97.98	93.82	98.20	98.30	98.24
Backdoor 3%	96.97	99.77	97.38	97.25	97.29
Backdoor 5%	96.29	100.00	96.74	96.43	96.55

The results presented in Table I demonstrate that the proposed straight-line trigger achieves highly effective backdoor performance across all evaluated poison rates on the GTSRB dataset. At the lowest poison rate of 1%, corresponding to approximately 31 poisoned samples out of 3,115 training images, the backdoored model achieves a Clean Accuracy (CA) of 97.98% and an Attack Success Rate (ASR) of 93.82%. When the poison rate is increased to 3% and 5%, the ASR further improves to 99.77% and 100.00%, respectively, while the CA decreases only slightly to 96.97% and 96.29%. These results indicate that the proposed trigger can maintain strong attack effectiveness while preserving high classification performance on clean samples.

The trigger is designed as a 2-pixel-wide horizontal line placed at 48% of the image height with a normalized intensity value of 0.55. Compared to a fully white trigger (intensity 1.0), this configuration is visually more subtle, allowing the experiment to evaluate whether a less conspicuous trigger can still achieve a high ASR, which is an important characteristic for stealthy backdoor attacks in real-world scenarios. Furthermore, the fixed spatial location and consistent intensity applied to all poisoned samples produce a coherent trigger pattern that can be efficiently learned by early convolutional layers, thereby enabling stable backdoor activation during inference.

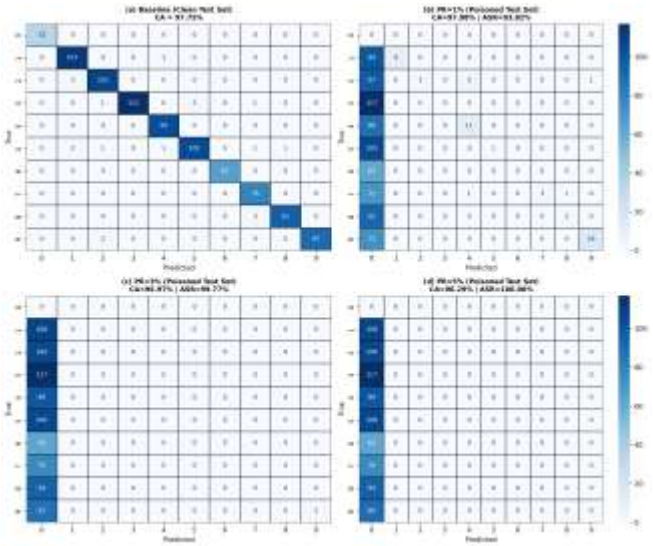

Figure 3. Confusion matrices of (a) the clean baseline model on unmodified test samples, (b) the backdoored model (PR = 1%) on triggered test samples, (c) the backdoored model (PR = 3%) on triggered test samples, and (d) the backdoored model (PR = 5%) on triggered test samples

Figure 3 provides visual evidence supporting the quantitative results presented in Table I. Figure 3(a) illustrates the confusion matrix of the baseline model evaluated on the clean test set. The matrix exhibits a strong diagonal distribution across all classes, indicating that most traffic sign categories are correctly classified. Only a small number of off-diagonal predictions are observed, which is consistent with the reported Clean Accuracy (CA) of 97.75%. This result demonstrates that the baseline ResNet-34 model is capable of learning discriminative representations for the GTSRB dataset without the presence of a backdoor trigger.

Figures 3(b)–3(d) show the confusion matrices of the backdoored models trained with poison rates of 1%, 3%, and 5%, respectively, evaluated on poisoned test samples containing the trigger pattern. A clear and progressively stronger attack pattern can be observed as the poison rate increases. In Figure 3(b), corresponding to the 1% poison rate, the majority of samples from non-target classes are redirected to the target class (Class 0), although a small number of predictions still remain in their original classes. This behavior is reflected by the Attack Success Rate (ASR) of 93.82%, indicating that the backdoor has already become highly effective even with a minimal number of poisoned training samples.

In Figures 3(c) and 3(d), corresponding to poison rates of 3% and 5%, nearly all triggered samples from Classes 1–9 collapse entirely into the target class column. This produces a dominant vertical concentration at Class 0, which is the characteristic visual signature of a successful all-to-one backdoor attack. The effect becomes almost perfectly deterministic at higher poison rates, achieving ASR values of 99.77% and 100.00%, respectively. Despite the increasing attack effectiveness, the models still maintain relatively high clean classification performance, with CA values remaining above 96%, demonstrating that the injected backdoor does not significantly degrade normal classification capability.

Furthermore, the sample distribution across classes remains consistent between the clean and poisoned evaluations, indicating that the observed misclassification behavior is not caused by class imbalance or dataset shift. Instead, the results confirm that the trigger pattern alone is sufficient to activate the learned backdoor mechanism and force predictions toward the attacker-defined target class. The findings also suggest that the trigger–target association becomes increasingly stable and dominant as the poison rate increases, while already showing strong effectiveness even at the lowest evaluated poison rate of 1%.

B. *Training Dynamics of Backdoored Models*

An important characteristic of the proposed backdoor attack is the rapid formation of the trigger–target association during the training process. Figure 4 presents the train and validation loss curves for the clean baseline model and the three backdoored ResNet-34 models on the GTSRB dataset across 20 training epochs. Across all configurations, both training and validation losses decrease smoothly and consistently without signs of divergence or overfitting. The loss trajectories of the backdoored models closely follow that of the clean baseline, indicating that the presence of poisoned

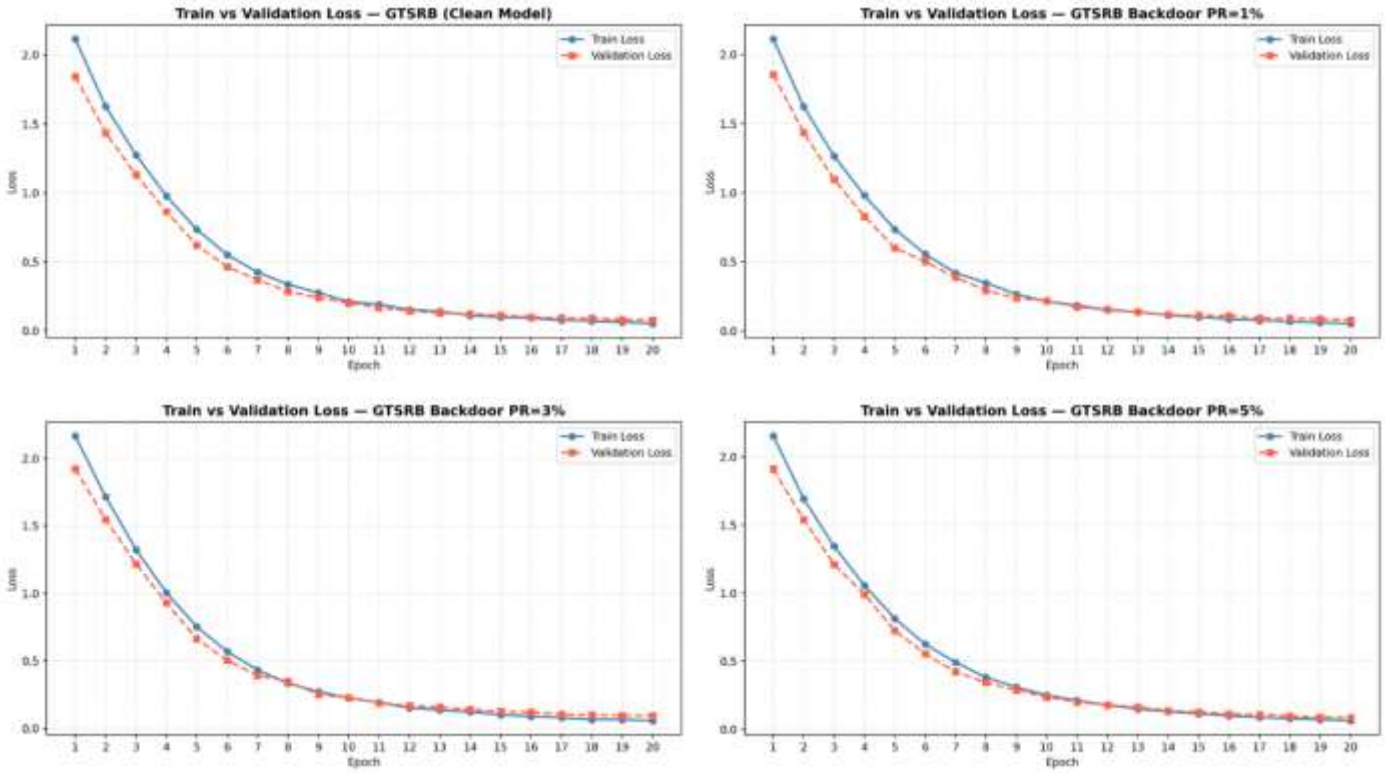

Figure 4. Train and validation loss curves of the clean baseline model and backdoored ResNet-34 models on the GTSRB dataset across 20 training epochs under poison rates of 1%, 3%, and 5%. All configurations exhibit smooth and consistent loss reduction with no signs of overfitting, indicating that the presence of poisoned samples does not destabilize the optimization process.

samples does not destabilize the optimization process. This stability suggests that the backdoor is injected in a manner that is transparent to the loss landscape, allowing the model to maintain its general learning behavior while simultaneously encoding the trigger–target association.

Building on this observation, Figure 5 illustrates the epoch-wise progression of Clean Accuracy (CA) and Attack Success Rate (ASR) for the backdoored ResNet-34 model on the GTSRB dataset under poison rates of 1%, 3%, and 5%. The curves reveal a consistent learning asymmetry across all poison rate conditions: ASR increases substantially faster than CA, indicating that the model learns the trigger pattern earlier than the semantic traffic sign features. This behavior is closely related to the shortcut learning phenomenon, where neural networks tend to prioritize easily exploitable statistical cues over more complex semantic representations.

At the 1% poison rate, the model initially exhibits low ASR during the first training epochs, reaching only 0.58% at Epoch 1. However, the ASR rapidly increases afterward, surpassing 57% by Epoch 4 and exceeding 93% by Epoch 10, eventually stabilizing near 97% at the final epoch. In contrast, the CA improves more gradually from 41.91% at Epoch 1 to 97.98% at Epoch 20. This indicates that even with a very limited number of poisoned samples, the model progressively learns the trigger association while simultaneously refining its clean classification capability.

A significantly faster convergence pattern is observed at higher poison rates. Under the 3% poison rate condition, ASR increases sharply from 4.78% at Epoch 1 to more than 90%

by Epoch 3, eventually reaching approximately 100% after several additional epochs. Similarly, at the 5% poison rate, the model achieves an ASR above 94% as early as Epoch 3 and approaches complete attack success shortly afterward. Meanwhile, CA continues to improve gradually throughout training and converges more slowly compared to ASR. These results demonstrate that increasing the poison rate strengthens the trigger signal and accelerates the learning of the backdoor trigger–target mapping.

The accelerated ASR convergence can be attributed to the structural simplicity and spatial consistency of the proposed trigger. The trigger consists of a horizontal line placed at a fixed location across all poisoned samples, producing a highly distinctive low-level trigger pattern that is easily captured by early convolutional filters. Since convolutional neural networks are inherently sensitive to edge-like and linear features, the trigger provides a strong and easily learnable signal compared to the more complex semantic variations present in traffic sign classes. Consequently, the model rapidly encodes the trigger association during early training stages, while clean semantic feature learning continues to develop progressively over subsequent epochs.

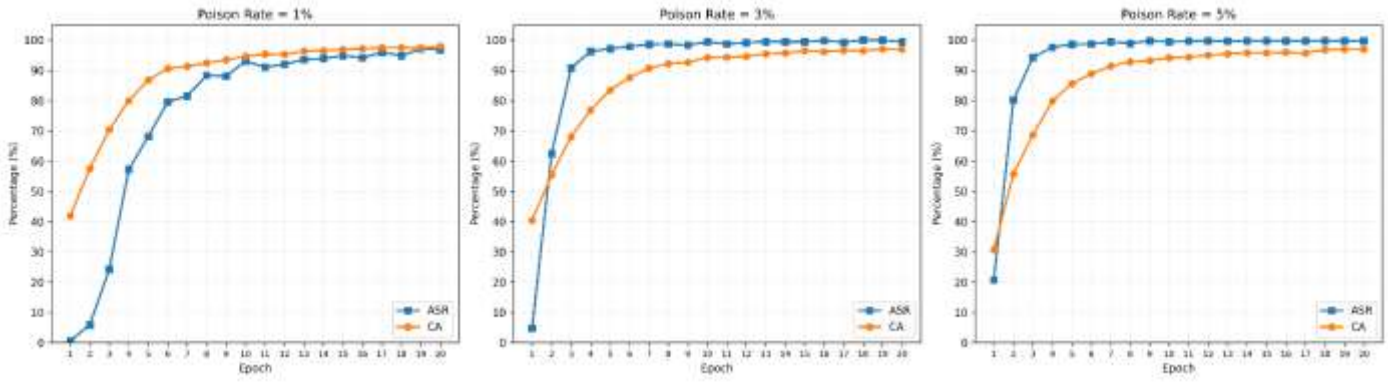

Figure 5. Epoch-wise progression of Clean Accuracy (CA) and Attack Success Rate (ASR) for the backdoored ResNet-34 model on the GTSRB dataset under poison rates of 1%, 3%, and 5%. ASR converges substantially faster than CA across all conditions, reflecting the model's tendency to prioritize the simple trigger pattern over complex semantic features during early training stages.

Furthermore, the training curves in both Figure 4 and Figure 5 collectively suggest that the backdoor pathway and the normal semantic classification pathway coexist without severe interference. Even when ASR approaches saturation, the model continues improving its clean classification performance, ultimately maintaining CA values above 96% across all poison rates. This behavior indicates that the injected backdoor does not significantly disrupt the model's ability to perform the primary classification task, thereby making the attack both effective and difficult to detect through standard performance evaluation alone.

C. *Analysis of Attack Success Rate*

The exceptionally high ASR observed across all poison rate conditions on GTSRB warrants detailed analysis. As demonstrated in Section IV-B, the training dynamics confirm that the backdoored model converges to high CA values within 20 epochs, while simultaneously establishing the trigger–target association at a very early stage of optimization. The rapid formation of this backdoor pathway, alongside stable clean-input performance, is the defining characteristic of an effective and stealthy backdoor attack.

The speed at which the backdoor is established is particularly notable. Although the model begins with a relatively low ASR of 0.58% and CA of 41.91% at the first training epoch under the 1% poison rate condition on GTSRB, the trigger–label association develops rapidly in subsequent epochs, reflecting the model's tendency to prioritize easily learnable statistical correlations over complex semantic features early in training. This rapid acquisition is consistent with the shortcut learning phenomenon described by (Geirhos et al., 2020), wherein models preferentially exploit the most easily learnable statistical correlations in the training data. The full-width horizontal line at normalized intensity 0.55 constitutes exactly such a feature: a globally distinctive, spatially consistent trigger pattern that early-layer convolutional filters learn to associate with the target class before semantic features are fully consolidated.

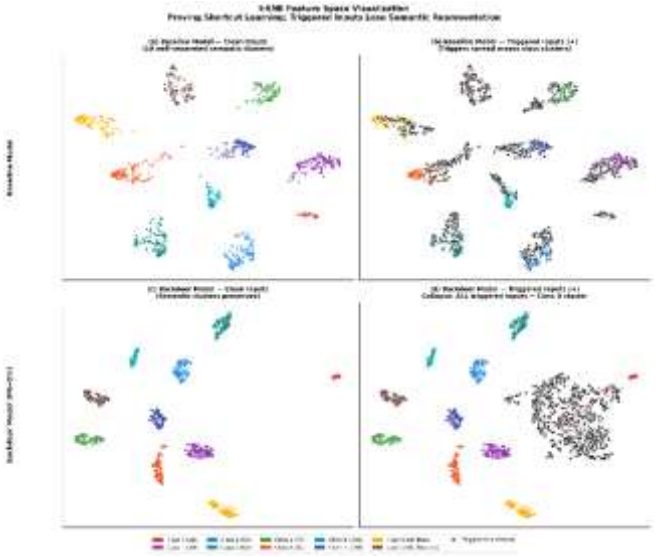

Figure 6. t-SNE visualization of ResNet-34 feature embeddings (Global Average Pooling layer) for clean test samples (colored by class) and triggered test samples on GTSRB. (a–b) Baseline model: triggered inputs distribute across class clusters. (c–d) Backdoor model (PR=5%): all triggered inputs collapse to the Class 0 cluster regardless of their true class, confirming complete representational override by the trigger.

The effectiveness of the trigger is further attributable to its structural alignment with CNN feature extraction behavior. As discussed by (Taye, 2023), early convolutional layers are inherently sensitive to low-level spatial patterns such as edges, lines, and corners. A horizontal line spanning the full 224-pixel image width produces a strong and uniform activation response in these filters, making it a highly distinguishable signal relative to natural image features.

To provide deeper representational insight into this behavior, Figure 6 presents t-SNE visualizations of ResNet-34 feature embeddings extracted at the Global Average Pooling (GAP) layer, comparing the baseline model and the backdoored model (PR = 5%) under two conditions: clean test inputs and triggered test inputs. In the baseline model Figure

6 (a–b), triggered inputs are distributed across their respective true-class clusters, indicating that the trigger does not alter the model's internal representations in the absence of a learned backdoor pathway. In contrast, for the backdoored model Figure 6 (c–d), all triggered test samples regardless of their true class label collapse entirely into the Class 0 cluster in the embedding space, while clean inputs retain their original class-discriminative structure. This complete representational override demonstrates that the backdoor pathway selectively remaps triggered inputs at the deepest feature level, without disturbing the semantic representations used for clean-input classification.

D. Feature Map Progression Analysis

To gain mechanistic insight into how the backdoor trigger influences internal representations, Figure 7 visualizes averaged feature map activations across five network stages (conv1 to layer4) for both the clean baseline and the backdoored model (1% poison rate).

At the conv1 stage, the baseline model captures distributed edge and texture features centered on the traffic sign, whereas the backdoored model exhibits a distinct horizontal activation band aligned with the trigger location. This indicates that the backdoor pathway is established at the earliest convolutional layer.

As the network depth increases, the divergence becomes more pronounced. While the baseline model progressively encodes semantic features such as shape and object structure, the backdoored model is increasingly dominated by a strong horizontal activation pattern. By layer2, the feature map is almost entirely governed by the trigger signal, with minimal response to semantic content.

This dominance persists through deeper layers, where the backdoored model exhibits highly concentrated horizontal activations, in contrast to the localized object-level representations observed in the baseline. At layer4, the trigger-driven activation effectively overrides the semantic representation, directly influencing the final classification decision and explaining the near-perfect Attack Success Rate (ASR).

Overall, these results reveal a clear layer-wise escalation of trigger dominance, where the backdoor signal is learned early and progressively suppresses semantic feature extraction. This behavior supports the dual-pathway interpretation of backdoored CNNs, in which trigger-based and semantic representations are learned in parallel without significant interference.

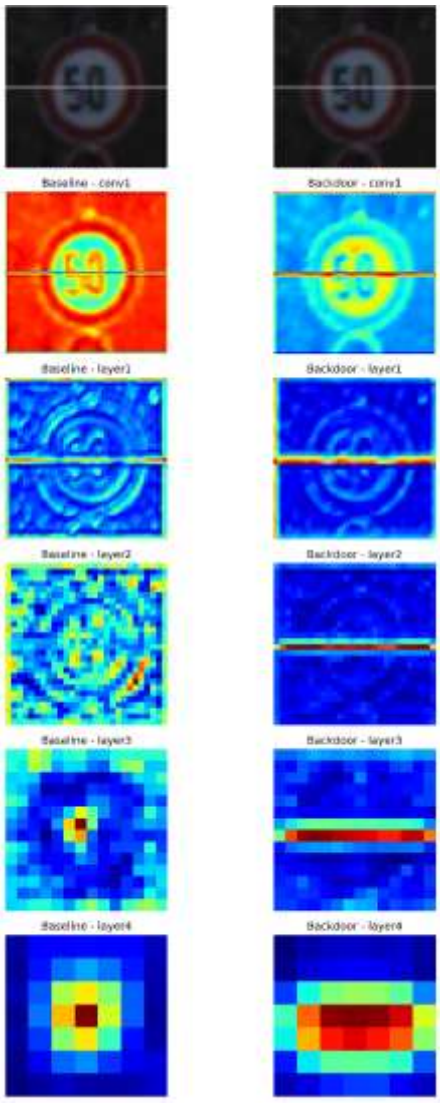

Figure 7. Progression of feature maps across ResNet-34 layers (conv1 through layer4) comparing baseline (left) and backdoored model at 1% poison rate (right) on a GTSRB speed limit sign with a horizontal line trigger.

The feature map analysis also provides visual evidence for why clean accuracy is so well preserved under backdoor training. On clean inputs (no trigger), the baseline and backdoored models exhibit virtually identical activation patterns through conv1 and layer1, confirming that the semantic feature extraction pathway remains unperturbed. The backdoor pathway is selectively activated only when the trigger is present, consistent with the input-specific gating behavior characteristic of effective backdoor attacks (Li et al., 2023). This selectivity is the key mechanism underlying the attack's stealth: without the trigger, the backdoored model behaves identically to its clean counterpart at every representational level.

E. Comparative Analysis with Refool

To provide a broader perspective on the effectiveness of the proposed straight-line trigger, this section compares the achieved Attack Success Rate (ASR) against Refool(Liu et al., 2020), a representative natural backdoor attack that uses

physically modeled optical reflections as triggers. Refool is selected as the primary reference because both studies share the same target dataset (GTSRB), model architecture (ResNet-34), and adopt an all-to-one attack strategy, making it the most structurally comparable prior work available. Table II presents a side-by-side summary of the two methods.

Table II. Comparative ASR and CA between the proposed method and Refool on the GTSRB dataset.

Method	Dataset	Poison Rate (%)	ASR (%)	CA (%)
Straight-line (Ours)	GTSRB	1	93.82	97.98
Straight-line (Ours)	GTSRB	3	99.77	96.97
Straight-line (Ours)	GTSRB	5	100.00	96.29
Refool (Liu et al., 2020)	GTSRB	3.16	91.67	86.30

As shown in Table II, the proposed straight-line trigger achieves strong backdoor effectiveness across all evaluated poison rates on the GTSRB dataset. The attack attains ASR values of 93.82%, 99.77%, and 100.00% at poison rates of 1%, 3%, and 5%, respectively, while maintaining high Clean Accuracy (CA) values of 97.98%, 96.97%, and 96.29%. These results indicate that the trigger remains highly effective even when only a small fraction of the training data is poisoned.

At a poison rate of 3%, which is approximately comparable to the 3.16% poison rate used in Refool, the proposed method achieves an ASR of 99.77%, outperforming Refool's reported ASR of 91.67% by approximately 8.10 percentage points. In addition, the proposed trigger preserves substantially higher clean classification performance, achieving a CA of 96.97% compared to the 86.30% reported for Refool. This demonstrates that the proposed straight-line trigger can achieve stronger attack effectiveness while introducing less degradation to normal model behavior.

These findings suggest that a structurally simple trigger, without requiring physics-based reflection modeling or complex optimization procedures, can still produce highly effective and stealthy backdoor behavior under controlled experimental conditions. The results further imply that the spatial consistency and strong visual saliency of the horizontal line trigger are sufficient to establish a robust trigger–target association within the convolutional feature extraction process.

However, a methodological caveat must be acknowledged: the two experimental setups are not fully equivalent due to differences in the GTSRB subsets used. The present study employs a balanced 10-class subset with a maximum of 500 images per class, yielding approximately 4,000 training images. This approach was adopted to address the severe class imbalance inherent in the full GTSRB distribution, following the preprocessing convention of (Barni et al., 2019). In contrast, (Liu et al., 2020) conducted experiments on a 13-class GTSRB subset comprising approximately 4,772 training images, with no explicit class balancing procedure reported. Differences in the number of classes, class distribution, and total training data volume may influence the difficulty of the classification task, the model's generalization capacity, and consequently the relationship between poison rate and ASR. The observed performance advantage of the straight-line trigger should therefore be interpreted with caution and should not be taken as evidence of absolute superiority across all dataset configurations.

Despite this limitation, the comparison nonetheless reveals a noteworthy finding. Refool relies on mathematically modeled physical reflections requiring iterative adversarial image selection, a candidate pool from PascalVOC, and three distinct reflection kernel models yet achieves a lower ASR than the proposed method at a comparable injection rate. This outcome reinforces the central argument of this study: CNN vulnerability to backdoor attacks is fundamentally rooted in the network's inductive bias toward simple low-level spatial features, rather than the structural sophistication of the trigger itself. A minimal two-pixel horizontal line, exploiting this bias directly, proves sufficient to achieve near-perfect backdoor effectiveness without the complexity overhead associated with natural or imperceptible trigger designs.

F. Limitations

While the experimental results demonstrate the effectiveness of the proposed straight-line backdoor trigger, several limitations must be acknowledged. First, this study is restricted to a single dataset (GTSRB) with a curated 10-class subset and a capped sample size of 500 images per class. Although this configuration enables a controlled and reproducible evaluation, it does not fully represent the diversity of real-world traffic sign datasets, which encompass 43 classes with severe class imbalance. The generalizability of the observed results to the full GTSRB distribution or to other visual domains remains to be verified.

Second, the evaluation is conducted on a single model architecture, ResNet-34. Although ResNet-34 is widely used in backdoor attack research and provides a meaningful baseline, it is unclear whether the findings generalize to other architectures such as VGG, MobileNet, EfficientNet, or Vision Transformers (ViT), which differ substantially in their feature extraction mechanisms and susceptibility to low-level pattern exploitation.

Third, only a single trigger configuration is evaluated a fixed horizontal line at 48% image height with 2-pixel thickness and intensity 0.55. The effect of varying trigger placement (e.g., top, bottom, diagonal), thickness, intensity, or orientation on ASR and CA has not been systematically explored in this work. Such variations could yield important

insights into the robustness and sensitivity of the backdoor pathway to trigger design choices.

Fourth, this study does not evaluate the resilience of the proposed attack against existing backdoor defense methods, such as Neural Cleanse or spectral signatures-based detection Assessing whether simple line-based triggers can evade or circumvent these defenses is a critical open question that falls outside the scope of this work but constitutes an important direction for future research.

V. CONCLUSIONS

This study investigated the effectiveness of a simple straight-line backdoor trigger on a ResNet-34 model trained on the GTSRB dataset, under an all-to-one attack setting with poison rates of 1%, 3%, and 5%. The experimental results demonstrate that the proposed trigger achieves progressively increasing Attack Success Rates across all evaluated poison rate conditions, reaching 93.82%, 99.77%, and 100.00% at poison rates of 1%, 3%, and 5%, respectively, while imposing only marginal degradation on Clean Accuracy. At the lowest tested poison rate of 1%, the backdoored model attains an ASR of 93.82% with a CA of 97.98%, within 0.23 percentage points of the clean baseline (97.75%), confirming that the attack is already highly effective even at a low poison rate. As the poison rate increases to 5%, the ASR reaches perfect saturation at 100.00%, while CA declines only slightly to 96.29%, indicating a favorable trade-off between attack effectiveness and stealthiness across all evaluated conditions.

The trigger–target association forms rapidly within the early training epochs, consistently preceding the convergence of semantic classification features across all evaluated poison rate conditions. This rapid backdoor formation is attributable to the structural alignment between the horizontal line trigger and the inherent sensitivity of early-layer convolutional filters to low-level edge and line features, as supported by the feature map progression analysis. The t-SNE visualization further confirms that triggered inputs in the backdoored model are completely remapped to the target class cluster in the feature space, while clean inputs retain their original class-discriminative structure.

These findings challenge the prevailing assumption in the backdoor attack literature that effective triggers must be complex, imperceptible, or computationally optimized. A minimal and visually interpretable pattern a 2-pixel horizontal line at submaximal intensity is sufficient to achieve competitive or superior attack performance compared to far more sophisticated trigger designs reported in prior work, including localized patch triggers (Gu et al., 2017)and dispersed pixel perturbations (Wang et al., 2022). This result highlights that the root cause of CNN vulnerability to backdoor attacks lies not in trigger sophistication, but in the inherent tendency of convolutional networks to exploit the simplest learnable statistical shortcuts available in the training data.

From a security perspective, the practical implications of this study are significant. The simplicity and low cost of the proposed trigger requiring no optimization, per-sample computation, or specialized knowledge make it accessible to a wide range of adversaries. Furthermore, its visual inconspicuousness (submaximal intensity, 2-pixel width) reduces the likelihood of detection through manual inspection of training data. These properties underscore the critical need for more robust and generalized backdoor defense mechanisms, particularly those capable of detecting simple, low-frequency spatial patterns that may be overlooked by defenses designed primarily for complex or localized triggers.

Future work should extend this evaluation to multiple architectures and datasets, explore the effect of trigger design variations (position, thickness, orientation, and intensity), and systematically assess the resilience of the proposed trigger against state-of-the-art defense methods. Investigating the transferability of simple line-based backdoors across different training pipelines and deployment conditions would further advance understanding of this underexplored attack vector and contribute to the development of more comprehensive defenses for real-world deep learning systems.

VI. ACKNOWLEDGMENT

The authors would like to express sincere gratitude to Universitas Islam Negeri Sultan Syarif Kasim Riau, particularly the Department of Informatics Engineering, Faculty of Science and Technology, for providing the academic environment and resources that supported this research.

REFERENCES

1. Bai, Y., Xing, G., Wu, H., Rao, Z., Ma, C., Wang, S., Liu, X., Zhou, Y., Tang, J., Huang, K., & Kang, J. (2025). Backdoor Attack and Defense on Deep Learning: A Survey. *IEEE Transactions on Computational Social Systems*, *12*(1), 404–434. https://doi.org/10.1109/TCSS.2024.3482723
2. Barni, M., Kallas, K., & Tondi, B. (2019). A New Backdoor Attack in CNNS by Training Set Corruption Without Label Poisoning. *Proceedings - International Conference on Image Processing, ICIP*, *2019-Septe*, 101–105. https://doi.org/10.1109/ICIP.2019.8802997
3. Chai, J., Zeng, H., Li, A., & Ngai, E. W. T. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, *6*, 100134. https://doi.org/10.24433/CO.0411648.v1
4. Chan, S. H., Dong, Y., Zhu, J., Zhang, X., & Zhou, J. (2023). BadDet: Backdoor Attacks on Object Detection. *Lecture Notes in Computer Science*, *13801 LNCS*, 396–412. https://doi.org/10.1007/978-3-031-25056-9_26

5. Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). *Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning*. http://arxiv.org/abs/1712.05526
6. Geirhos, R., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z
7. Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Madry, A., Li, B., & Goldstein, T. (2023). Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(2), 1563–1580. https://doi.org/10.1109/TPAMI.2022.3162397
8. Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). *BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain*. http://arxiv.org/abs/1708.06733
9. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90
10. Li, Y., Jiang, Y., Li, Z., & Xia, S. T. (2024). Backdoor Learning: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(1), 5–22. https://doi.org/10.1109/TNNLS.2022.3182979
11. Li, Y., Li, Y., Wu, B., Li, L., He, R., & Lyu, S. (2021). Invisible Backdoor Attack with Sample-Specific Triggers. *Proceedings of the IEEE International Conference on Computer Vision*, 16443–16452. https://doi.org/10.1109/ICCV48922.2021.01615
12. Li, Y., Zhang, S., Wang, W., & Song, H. (2023). Backdoor Attacks to Deep Learning Models and Countermeasures: A Survey. *IEEE Open Journal of the Computer Society*, *4*, 134–146. https://doi.org/10.1109/OJCS.2023.3267221
13. Liu, Y., Ma, X., Bailey, J., & Lu, F. (2020). Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12355 LNCS*, 182–199. https://doi.org/10.1007/978-3-030-58607-2_11
14. Mengara, O., Avila, A., & Falk, T. H. (2024). Backdoor Attacks to Deep Neural Networks: A Survey of the Literature, Challenges, and Future Research Directions. *IEEE Access*, *12*, 29004–29023. https://doi.org/10.1109/ACCESS.2024.3355816
15. Taye, M. M. (2023). Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. *Computation*, *11*(3). https://doi.org/10.3390/computation11030052
16. Turner, A., Tsipras, D., & Madry, A. (2019). *Label-Consistent Backdoor Attacks*. 1–24. http://arxiv.org/abs/1912.02771
17. Wang, Y., Zhao, M., Li, S., Yuan, X., & Ni, W. (2022). Dispersed Pixel Perturbation-Based Imperceptible Backdoor Trigger for Image Classifier Models. *IEEE Transactions on Information Forensics and Security*, *17*, 3091–3106. https://doi.org/10.1109/TIFS.2022.3202687
18. Zhao, P., Zhu, W., Jiao, P., Gao, D., & Wu, O. (2025). *Data Poisoning in Deep Learning: A Survey*. 1–20. http://arxiv.org/abs/2503.22759