Adaptive hybrid convolutional neural network-autoencoder framework for backdoor detection in GenAI-driven semantic communication systems

***Hassan El Alami*[1], *Danda B. Rawat*[1]**

[1] *Howard University, Washington, DC, USA*

Corresponding author: Hassan El Alami, hassan.elalami@howard.edu

Semantic communication systems, powered by Generative AI (GenAI), enable the efficient transmission of semantic meaning rather than raw data. However, these systems remain highly vulnerable to backdoor attacks, which embed malicious triggers into training datasets, causing the misclassification of poisoned samples while leaving clean inputs unaffected. Existing detection mechanisms often modify model structures, degrading clean inference performance, or impose strict data format constraints that limit adaptability. Moreover, many approaches rely on fixed similarity thresholds, making them ineffective against adaptive backdoor attacks and unable to inspect hidden activations where backdoors are embedded. To address these challenges, we propose a hybrid framework, CNN-AAE, which combines a Convolutional Neural Network (CNN) with an Adaptive Autoencoder (AAE) to leverage both spatial feature learning and semantic deviation analysis for robust backdoor detection. Unlike prior methods, our approach preserves the original model structure, dynamically adjusts detection thresholds, and analyzes internal layer activations to identify deeply embedded backdoors. We evaluate CNN-AAE on the MNIST and CIFAR-10 datasets and compare its performance against several State-Of-The-Art (SOTA) baselines, including CNN, Multilayer Perceptron (MLP), Fully Connected Neural Network (FCNN), autoencoder, and the Anti-Backdoor Model (ABM). The results demonstrate that CNN-AAE consistently achieves higher detection accuracy and significantly lower attack success rates, while maintaining efficient resource usage in terms of training time and memory consumption.

Keywords: Artificial intelligence, backdoor, cybersecurity, deep learning, generative AI, large language models, neural networks, semantic communication

1. INTRODUCTION

Traditional communication systems focus on the accurate transmission of bits without considering the meaning of the transmitted information. Based on Shannon's information theory, these systems treat all data as equally important, regardless of its actual relevance to the recipient [1]. However, as modern applications demand efficient, context-aware, and intelligent communication, the limitations of conventional systems have become evident. Semantic communication introduces a revolutionary paradigm in modern networks by emphasizing the transmission of meaning rather than raw data. This paradigm addresses the challenge by shifting the focus from data transmission to meaning transmission, ensuring that only relevant and useful information is exchanged. Unlike traditional communication systems that focus on delivering exact symbols or numerical values, semantic communication systems identify, encode, and convey the essential meaning embedded in the data. At the transmitter, a semantic encoder processes input data, such as text, images, or speech, extracting key semantic features and converting them into compact, low-dimensional representations [2]. In this context, instead of transmitting the entire raw data, only the extracted semantic feature vectors are sent over the communication channel, leading to a significant reduction in bandwidth consumption.

At the receiver end, a semantic decoder processes these feature vectors to reconstruct meaningful information while ensuring that the original semantic context remains intact [3]. This capability of efficiently conveying information with minimal communication overhead makes semantic communication a highly effective solution for data-intensive applications, including the metaverse [4], the Internet of Things (IoT) [5, 6], and autonomous systems [7]. For instance, in the metaverse, semantic communication ensures seamless and immersive virtual experiences by reducing latency in interactive environments, improving real-time collaboration, and enhancing the synchronization of digital avatars and objects [8]. Similarly, in IoT networks, semantic communication optimizes bandwidth usage while ensuring precise real-time operations, such as sensor data aggregation and remote monitoring [9]. Generative AI (GenAI) [10, 3, 11] plays a fundamental role in semantic communication by improving the efficiency of encoding and decoding processes. It enables AI-driven systems to focus on transmitting meaningful and relevant information while minimizing redundancy. However, the integration of GenAI brings major cybersecurity risks, particularly backdoor attacks [12, 13], which pose serious threats to the integrity of semantic communication systems. Backdoor attacks involve embedding concealed triggers within a model during the training phase, allowing adversaries to manipulate its behavior later [14, 15, 16]. This type of data poisoning introduces compromised training samples containing subtle, predefined patterns, such as minor pixel alterations in images or slight modifications in textual input, which remain inactive until a trigger activates them during inference [14]. When the model encounters such triggers, it produces an attacker-controlled response, thereby distorting the intended meaning of transmitted data [14]. For example, the Covert Semantic Backdoor Attack (CSBA) is a stealthy and trigger-free backdoor attack that covertly alters the semantic interpretation of communications within Intelligent Connected Vehicles (ICVs) [17]. Instead of adding visible triggers, CSBA removes critical semantics (e.g., traffic signs) from transmitted images, causing connected vehicles to miss essential road information. By poisoning the semantic encoding and decoding process, it remains undetectable, leading to severe navigation failures and safety risks. To tackle these challenges, numerous studies emphasize the pressing need to develop defenses for semantic communication systems [18, 17, 19, 20].

Recently, various studies have explored defense techniques against backdoor attacks. For example, the work in [18] introduces a new backdoor attack paradigm on semantic symbols (BASS) and proposes corresponding defense mechanisms tailored for Deep Learning (DL)-enabled semantic communication systems. To mitigate BASS, a specialized training framework is designed for prevention. Additionally, reverse engineering-based and pruning-based defense strategies are implemented to enhance system resilience against backdoor attacks. Another study [20] highlights the susceptibility of DL-based semantic communication systems to backdoor attacks by embedding triggers into a subset of training samples. The attack exploits the complex decision space of Deep Neural Networks (DNNs) in autoencoder-based semantic communications, where latent features are transmitted over limited channel uses. The effectiveness of the attack increases with higher signal-to-noise ratios, more channel uses, and a greater proportion of poisoned training data. To counteract these threats, novel design strategies are proposed to preserve the integrity of semantic communications. [21] proposes an imperceptible backdoor attack that avoids traditional visible trigger patterns, which are easily detected by human inspection. This method extracts both low-level and high-level semantic features from clean images using a pretrained victim model. A trigger pattern is then generated based on channel attention, influencing high-level features without noticeable modifications. An encoder subsequently produces poisoned images that maintain low-level feature consistency while embedding the backdoor trigger. This approach achieves high attack success rates and remains robust against existing backdoor defenses by ensuring stealthiness through imperceptible modifications. In [22], the authors proposed a defense framework for task-oriented multi-user semantic communication systems based on adversarial Reinforcement Learning (RL). This approach utilizes adversarial training to simulate poisoning attacks, progressively enhancing model robustness. By doing so, the system can efficiently identify and mitigate poisoned data while maintaining effective communication. In contrast, [23] investigates backdoor attacks in GenAI-driven semantic communication systems. Unlike existing defenses that require modifications to the model structure or impose constraints on data formats, the proposed approach leverages semantic similarity to detect backdoor attacks without such limitations. By analyzing deviations in the semantic feature space and establishing a threshold-based detection framework, the proposed approach effectively identifies poisoned samples while maintaining the system's integrity. The study in [24] proposes an ABM, a non-invasive defense mechanism that removes backdoors from poisoned models without altering their parameters. It introduces a controlled weak trigger to identify poisoned samples, trains a student model to learn only the backdoor behavior, and uses it to cancel the backdoor effect in the teacher model via knowledge distillation.

The aforementioned proposed approaches to defending against backdoor attacks in semantic communication systems suffer from several key limitations that undermine their effectiveness. Many approaches depend on model modification techniques, such as neuron pruning, to eliminate backdoor triggers; however, these modifications often degrade clean inference performance and reduce semantic accuracy. Some approaches impose strict data format constraints, such as requiring paired image-text inputs, which limit their applicability to diverse semantic data representations. Another limitation is the reliance on fixed similarity thresholds, leaving them vulnerable to adaptive backdoor attacks where adversaries dynamically adjust triggers to evade detection. Moreover, most existing methods only analyze input-output behavior and fail to inspect internal activations in the encoder and decoder layers, where backdoor attacks are often embedded, thereby reducing their ability to detect deeply hidden manipulations. Moreover, ABM, while non-invasive, also presents a limitation: it requires the manual design of a weak implanted backdoor, yet there is no universal or interpretable method for generating a trigger that is both weak and effective. This dependency restricts ABM's scalability and practicality across different settings. Additionally, many approaches require manual tuning of similarity thresholds, making them unsuitable for large-scale deployment and inconsistent across datasets. Finally, most proposed methods are evaluated on a single dataset, which limits their generalizability and robustness in real-world applications. To overcome these limitations, we propose a hybrid CNN-AAE framework that integrates a Convolutional Neural Network (CNN) with an Adaptive Autoencoder (AAE), providing a robust and scalable defense mechanism for backdoor detection in GenAI-driven semantic communication systems. Based on this motivation, the central research question addressed in this work is as follows: How can backdoor attacks be effectively detected in GenAI-driven semantic communication systems using an AI-based method that achieves high detection accuracy while minimizing the Attack Success Rate (ASR)? Accordingly, the objectives of this study are: (i) to investigate the vulnerability of GenAI-based semantic communication systems to backdoor attacks; (ii) to design a non-invasive and input-flexible detection framework; and (iii) to enhance detection accuracy and reduce ASR through a hybrid DL approach that leverages both semantic features and reconstruction consistency. Unlike methods that rely on model modifications, the CNN-AAE framework preserves the original architecture, ensuring that clean inference performance remains unaffected. Additionally, it eliminates data format constraints, allowing detection across diverse semantic data representations without requiring paired image-text inputs. To counter adaptive backdoor attacks, the CNN-AAE framework implements dynamic thresholding, where the detection system continuously learns and adjusts based on observed feature variations, preventing adversaries from evading detection by altering backdoor triggers. In contrast to prior work that has focused solely on input-output relationships, our framework examines hidden activations within the encoder and decoder layers, allowing it to detect deeply embedded backdoors that may otherwise evade detection. Moreover, the CNN-AAE framework is fully automated, eliminating the need for manual similarity threshold tuning. The main contributions of this work are summarized as follows:

- We propose a system model that mathematically formulates semantic communications and the backdoor threat as part of a classification problem.
- To achieve higher accuracy in detecting backdoor attacks in semantic communications powered by GenAI, we propose CNN-AAE, a novel approach that integrates a CNN with an AAE model.
- Comprehensive experiments are conducted to evaluate the proposed CNN-AAE model against SOTA baselines, including CNN, MLP, FCNN, autoencoder, and the ABM approach, using the MNIST and CIFAR-10 datasets. The evaluation includes metrics such as accuracy, precision, recall, F1 score, and ASR. Additionally, we assess the computational efficiency of the proposed framework by analyzing its training time and memory usage.

The remainder of this paper is organized as follows: Section 2 presents the system model, including both the semantic communication architecture and the backdoor threat model. Section 3 introduces the proposed CNN-AAE framework, describing its detection algorithm and the integration of GenAI within the overall architecture. Section 4 discusses the experimental results, and Section 5 concludes the paper with future work.

2. SYSTEM MODEL

Semantic communication systems powered by GenAI aim to transmit information in a compressed form while preserving its semantic meaning. However, these systems are highly susceptible to backdoor attacks, in which adversaries embed hidden triggers during training to manipulate the model's predictions. This section provides a system model of the semantic communication pipeline, defines the backdoor threat model, and introduces a mechanism for detecting backdoor attacks.

2.1 Semantic communication model

In semantic communication systems, an input sample $x \in \mathbb{R}^d$, where d denotes the original input dimensionality, is first transformed into a compressed semantic representation $s \in \mathbb{R}^m$ using a semantic encoder $E_s(\cdot)$, such that:

$$s = E_s(x), \quad \text{with} \quad m \ll d \tag{1}$$

Here, m represents the dimensionality of the latent semantic space, which is significantly smaller than d. This compression enables the system to retain only high-level meaningful information while discarding redundant details. The semantic vector s captures the core intent of the input and is transmitted to the receiver. At the receiver end, a semantic decoder $D_s(\cdot)$ reconstructs an approximation of the original input:

$$\hat{x} = D_s(s) \tag{2}$$

where $\hat{x} \in \mathbb{R}^d$ denotes the reconstructed version of the input. The quality of reconstruction is evaluated using a semantic fidelity function:

$$\mathcal{F}(x, \hat{x}) = \exp\left(-\frac{\|x - \hat{x}\|^2}{\sigma^2}\right) \tag{3}$$

where $\|x - \hat{x}\|^2$ is the squared Euclidean distance between the original and reconstructed samples, and σ^2 represents the variance of the data distribution. A higher value of $\mathcal{F}(x, \hat{x}) \in (0, 1]$ indicates stronger semantic preservation. This formulation allows the system to prioritize semantic accuracy over exact bit-wise reconstruction, which is particularly advantageous in noisy or resource-constrained communication settings.

2.2 Threat model

The backdoor attacks aim to manipulate the training process by injecting malicious samples that include a predefined trigger. These poisoned samples cause the model to behave normally on clean inputs while misclassifying triggered inputs to an attacker-specified target class. Let $x \in \mathbb{R}^d$ denote a clean input sample and $t \in \mathbb{R}^d$ the adversarial trigger. The poisoned input is defined as:

$$x_p = x + t \tag{4}$$

where x_p represents the input modified with the backdoor trigger. During communication, the semantic encoder maps x_p to a low-dimensional semantic representation $s_p = E_s(x_p) \in \mathbb{R}^m$, which is then decoded by D_s to yield the reconstructed poisoned output $\hat{x}_p = D_s(s_p)$. The model is expected to classify clean reconstructions correctly, i.e., $f(\hat{x}) = y$, where $f(\cdot)$ is the classifier and y is the true label. However, under a successful backdoor attack, the goal is to enforce:

$$f(\hat{x}_p) = y_t \tag{5}$$

where y_t is the attacker's target class. The likelihood of successful misclassification increases with the poisoning ratio β, which denotes the fraction of poisoned samples in the training set. This relationship is captured by:

$$\mathbb{P}(f(\hat{x}_p) = y_t) = g(\beta) \tag{6}$$

where $g(\cdot)$ is a monotonically increasing function modeling the attack's effectiveness. To evaluate performance empirically, the ASR is computed over N_p poisoned test inputs as:

$$ASR = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbb{I}(f(\hat{x}_{p,i}) = y_t) \tag{7}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Beyond classification accuracy, backdoor attacks can also degrade the semantic fidelity of transmitted content. Let $\mathcal{F}(x, \hat{x})$ denote the semantic fidelity between a clean input and its reconstruction. For poisoned samples, we define the fidelity as:

$$\mathcal{F}_p(x_p, \hat{x}_p) = \exp\left(-\frac{\|x_p - \hat{x}_p\|^2}{\sigma^2}\right) \tag{8}$$

where σ^2 represents the variance of the data distribution. The drop in semantic quality introduced by the backdoor is quantified as:

$$\Delta\mathcal{F} = \mathcal{F}(x, \hat{x}) - \mathcal{F}_p(x_p, \hat{x}_p) \tag{9}$$

which reflects the degradation in reconstructive performance due to adversarial perturbation.

2.3 Backdoor detection model

To detect backdoor attacks in semantic communication systems, we model deviations in the semantic feature space caused by poisoned inputs. Let X_c denote the set of clean training samples, and $E_s(x)$ be the semantic encoder that maps an input $x \in \mathbb{R}^d$ to a semantic embedding $E_s(x) \in \mathbb{R}^m$. The mean semantic representation of clean data is computed as:

$$\mu_c = \frac{1}{|X_c|} \sum_{x_i \in X_c} E_s(x_i) \tag{10}$$

where $\mu_c \in \mathbb{R}^m$ is the average semantic feature vector over clean inputs. For any input x, its semantic deviation from the clean distribution is measured as:

$$D(x) = \left\|E_s(x) - \mu_c\right\| \tag{11}$$

where $\|\cdot\|$ denotes the Euclidean norm. In the case of a poisoned input $x_p = x + t$, where $t \in \mathbb{R}^d$ is the backdoor trigger, the semantic deviation becomes:

$$D(x_p) = \left\|E_s(x_p) - \mu_c\right\| \tag{12}$$

A test sample is flagged as poisoned if its deviation exceeds a predefined threshold $\tau \in \mathbb{R}^+$, that is, when $D(x) > \tau$. To quantify the likelihood of detecting a poisoned input, the detection probability is defined as:

$$P_{\text{det}} = \mathbb{P}(D(x_p) > \tau) \tag{13}$$

where $P_{\text{det}} \in [0, 1]$ captures the probability that a backdoored input lies outside the clean semantic distribution. This formulation enables inference-time detection of poisoned samples without altering the encoder–decoder architecture of the semantic communication system.

3. ADAPTIVE HYBRID CNN-AAE FRAMEWORK

In this section, we present the proposed CNN-AAE framework for backdoor detection in GenAI-driven semantic communication systems. Our framework integrates a CNN-based semantic encoder, simulating GenAI abstraction on the sender side, with an autoencoder-based anomaly detector at the receiver. By combining spatial feature extraction with semantic consistency checks and adaptive thresholding, the framework enables dynamic detection of static and adaptive backdoor attacks while preserving clean transmission integrity.

3.1 Architecture of the CNN-AAE framework

The proposed CNN-AAE framework is formulated as a joint detection model that integrates CNN-based feature analysis and autoencoder-based semantic deviation detection. Given an input sample $x \in \mathbb{R}^{H\times W\times C}$, the CNN feature extraction process is defined as:

$$F = f_\theta(x) \tag{14}$$

Here, f_θ denotes the CNN feature extraction function, parameterized by weights θ, while $F \in \mathbb{R}^d$ represents the corresponding extracted feature vector. In addition, CNN is highly effective in capturing spatial dependencies within the input data. By applying a series of convolutional layers, the model learns hierarchical patterns that differentiate normal from backdoor-embedded inputs. However, CNN alone might struggle to detect adaptive backdoor triggers that mimic natural variations. Therefore, we introduce an additional layer of security through autoencoder-based anomaly detection. Once the feature vector F is obtained, the sample is classified using a fully connected layer:

$$\hat{y} = g_\phi(F) \tag{15}$$

where g_ϕ is the classification function with parameters ϕ. If the CNN classifier produces a low-confidence decision or identifies a sample as anomalous, it is passed to the autoencoder for further analysis. For semantic deviation detection, the input sample is encoded and reconstructed using an autoencoder:

$$s = E_\psi(x), \quad \hat{x} = D_\omega(s) \tag{16}$$

where E_ψ denotes the encoder function with parameters ψ, mapping the input x to a latent space representation s; D_ω denotes the decoder function with parameters ω, reconstructing x as $\hat{x}$; and $\hat{x}$ represents the reconstructed version of the original input x. Since the autoencoder learns the distribution of clean data, its ability to accurately reconstruct backdoored inputs is significantly reduced. The reconstruction error quantifies the semantic discrepancy between the original and the generated samples:

$$R(x) = \|x - \hat{x}\|^2 \tag{17}$$

A low reconstruction error suggests that the sample conforms to the clean distribution, whereas a high reconstruction error indicates that the input is an outlier, potentially embedded with a backdoor trigger. Thus, the reconstruction error serves as a secondary confirmation of the CNN's feature-based anomaly detection.

3.2 Decision framework for backdoor identification

To enhance robustness, we employ an adaptive thresholding mechanism that dynamically adjusts detection sensitivity based on variations in observed feature space. Traditional fixed-threshold methods are ineffective against adaptive attacks, as adversaries can modify backdoor patterns to bypass static detection limits. To counteract this, our approach computes detection thresholds based on the statistical properties of clean samples, ensuring adaptability across datasets and environments. The threshold for CNN feature deviations is defined as:

$$\tau_F = \mu_F + \lambda_F \cdot \sigma_F \tag{18}$$

Similarly, the threshold for autoencoder reconstruction errors is given by:

$$\tau_R = \mu_R + \lambda_R \cdot \sigma_R \tag{19}$$

where μ_F and μ_R denote the mean feature deviation and mean reconstruction error of clean samples, respectively, and σ_F and σ_R represent their corresponding standard deviations. The parameters λ_F and λ_R are tunable factors that control the detection sensitivity with respect to statistical variations. Their values are empirically selected through cross-validation and remain fixed during evaluation. These parameters are not learned during training; instead, they are adjusted to balance detection performance and the trade-off between false positives and false negatives under varying poisoning ratios. Unlike traditional approaches that rely on static thresholds, our method derives τ_F and τ_R from the statistical properties (mean and variance) of clean samples, allowing thresholds to adapt across datasets and environments even though λ_F and λ_R remain fixed. A sample x is classified as poisoned if:

$$\mathbb{I}(D(F) > \tau_F \vee R(x) > \tau_R) = 1 \tag{20}$$

where $D(F) > \tau_F$ indicates that the CNN has detected an anomaly in the extracted feature representation, and $R(x) > \tau_R$ implies that the autoencoder reconstruction error exceeds the threshold, signaling a deviation from the clean distribution. The logical OR condition ($\vee$) ensures that a sample is flagged as poisoned whenever either model detects an anomaly. This dual-detection mechanism minimizes false positives while maintaining high sensitivity to adaptive and sophisticated backdoor attacks. By combining the CNN's spatial awareness with the autoencoder's anomaly detection capability, the framework provides a robust backdoor defense suitable for semantic communication applications.

3.3 CNN-AAE algorithm

The proposed adaptive hybrid CNN-AAE framework functions as a dual-stage detection pipeline, integrating CNN-based semantic feature extraction with autoencoder-based semantic consistency checks to identify backdoor attacks in GenAI-driven semantic communication systems. As illustrated in *Algorithm 1*, the framework proceeds through four systematic phases: semantic embedding extraction, reconstruction analysis, adaptive thresholding, and final classification.

```
Algorithm 1: Adaptive Hybrid CNN–AAE Detector
Inputs: Dataset of input samples $X = \{x_1, x_2, \dots, x_N\}$
CNN semantic encoder $f_\theta$ trained on clean + poisoned data
Autoencoder $(E_\psi, D_\omega)$ trained on clean CNN features
Statistical parameters: $\mu_F, \sigma_F, \mu_R, \sigma_R$
Tunable sensitivity factors: $\lambda_F$, $\lambda_R$
Outputs: Classification result $y_i \in \{\text{Clean}, \text{Poisoned}\}$
         for each $x_i \in X$
for each sample $x_i \in X$ do
    Step 1: Semantic Feature Extraction (Simulating GenAI)
    Extract semantic embedding using CNN:
      $s_i = f_\theta(x_i)$
    Compute feature deviation from clean distribution:
      $D(F_i) = \|s_i - \mu_F\|_2$
    Step 2: Semantic Consistency Check via Autoencoder
    Reconstruct semantic embedding:
      $\hat{s}_i = D_\omega(E_\psi(s_i))$
    Compute reconstruction error:
      $R(s_i) = \|s_i - \hat{s}_i\|_2^2$
    Step 3: Adaptive Thresholding
    Compute semantic anomaly thresholds:
      $\tau_F = \mu_F + \lambda_F \cdot \sigma_F$
      $\tau_R = \mu_R + \lambda_R \cdot \sigma_R$
    Step 4: Backdoor Detection Decision
    if $D(F_i) > \tau_F$ OR $R(s_i) > \tau_R$ then
        $y_i \leftarrow$ Poisoned ;
    end
    else
        $y_i \leftarrow$ Clean ;
    end
end
return Classification results $\{y_1, y_2, \dots, y_N\}$
```

Step 1: Semantic feature extraction (Simulating GenAI): Each input sample x_i is passed through a pretrained CNN encoder f_θ, which simulates the semantic embedding layer of a GenAI. The CNN maps x_i to a lower-dimensional semantic vector $s_i = f_\theta(x_i)$, capturing high-level features. The feature deviation $D(F_i) = \|s_i - \mu_F\|_2$ is then computed to assess how much s_i diverges from the expected clean feature distribution, defined by the mean μ_F.

Step 2: Semantic consistency check via autoencoder: The extracted semantic representation s_i is further processed through an autoencoder composed of encoder E_ψ and decoder D_ω, trained solely on clean embeddings. The autoencoder reconstructs the semantic input as $\hat{s}_i = D_\omega(E_\psi(s_i))$, and the reconstruction error $R(s_i) = \|s_i - \hat{s}_i\|_2^2$ is used to detect anomalies. Clean samples result in low reconstruction error, while poisoned samples yield significant deviations, indicating semantic inconsistency.

Step 3: Adaptive thresholding: To account for variations across environments and model behavior, the framework employs dynamic thresholding. The semantic deviation threshold τ_F and the reconstruction error threshold τ_R are defined in *(18)* and *(19)*, respectively.

Step 4: Backdoor detection decision: A sample is classified as poisoned if either the CNN-based feature deviation $D(F_i)$ exceeds τ_F or the autoencoder-based recon-

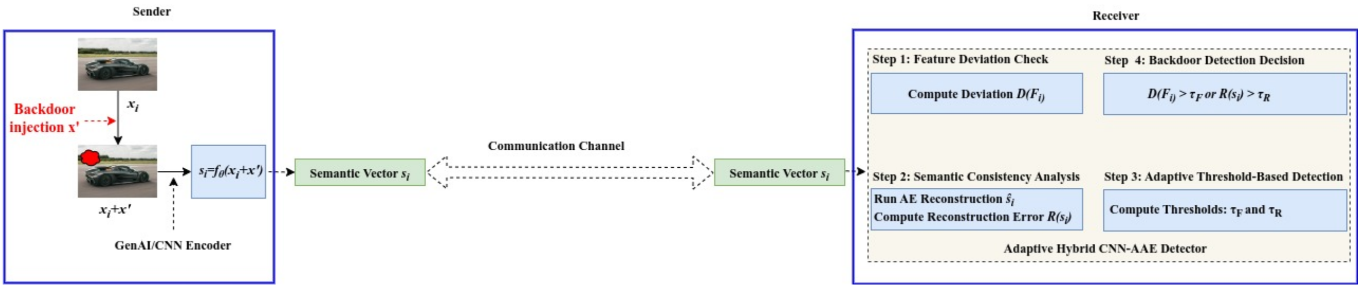
Sender
Backdoor injection x'
x_i
x_i+x'
$s_i=f_\theta(x_i+x')$
GenAI/CNN Encoder
Semantic Vector s_i
Communication Channel
Semantic Vector s_i
Receiver
Step 1: Feature Deviation Check
Compute Deviation $D(F_i)$
Step 4: Backdoor Detection Decision
$D(F_i) > \tau_F$ or $R(s_i) > \tau_R$
Step 2: Semantic Consistency Analysis
Run AE Reconstruction $\hat{s}_i$
Compute Reconstruction Error $R(s_i)$
Step 3: Adaptive Threshold-Based Detection
Compute Thresholds: τ_F and τ_R
Adaptive Hybrid CNN-AAE Detector

Figure 1 – End-to-end adaptive hybrid CNN-AEE framework for backdoor detection in GenAI-driven semantic communication system.

struction error $R(s_i)$ exceeds τ_R. This OR-based decision rule enhances the framework's robustness by combining spatial and semantic anomaly cues. By combining the semantic encoding capabilities of CNNs with Autoencoder-based anomaly detection, the CNN-AAE framework simulates a GenAI-driven semantic communication setting and provides a resilient mechanism for detecting both static and adaptive backdoor attacks at the receiver side.

3.4 Framework description with GenAI integration

The proposed framework is designed to detect backdoor attacks in GenAI-driven semantic communication systems, structured around a sender, channel, and receiver, as illustrated in Fig. 1. The process begins at the sender side, where a raw input image x_i is either clean or potentially poisoned by an attacker. A backdoor injection can occur at this point, embedding a malicious trigger into the image to manipulate the system's behavior during inference. This image is then processed by a pretrained CNN encoder f_θ, which simulates the semantic abstraction behavior of a GenAI model at the sender side. The CNN encoder acts as a lightweight proxy for GenAI, extracting a compact semantic representation $s_i = f_\theta(x_i)$. This semantic embedding serves as the compressed form of the message to be transmitted over the communication channel. The semantic vector s_i is transmitted through the channel to the receiver, which may receive either a clean or poisoned semantic representation. The communication channel itself is considered neutral in this framework, focusing on the threat model on the sender-side injection. At the receiver, the detection mechanism is activated. The received semantic embedding s_i is first evaluated for feature deviation by comparing it to the expected distribution of clean semantic vectors, using the metric $D(F_i) = \|s_i - \mu_F\|_2$, where μ_F is the mean of clean embeddings computed during training. This step flags semantic vectors that are statistically distant from the normal distribution.

Next, the framework applies a semantic consistency check by passing s_i through an autoencoder trained exclusively on clean samples. The autoencoder, composed of an encoder E_ψ and decoder D_ω, reconstructs the semantic vector as $\hat{s}_i = D_\omega(E_\psi(s_i))$. The reconstruction error $R(s_i) = \|s_i - \hat{s}_i\|_2^2$ serves as an anomaly signal, with higher errors indicating potential poisoning, since backdoored samples typically fall outside the learned manifold of clean semantics. To enhance robustness across different operational environments, adaptive thresholds are computed based on training statistics: τ_F for feature deviation and τ_R for reconstruction error. The final classification rule determines a sample as poisoned if either $D(F_i) > \tau_F$ or $R(s_i) > \tau_R$. If neither threshold is exceeded, the sample is classified as clean. Based on this detection decision, the receiver proceeds accordingly. Clean samples are accepted and decoded or used as intended, while poisoned samples are flagged, discarded, or subjected to mitigation measures. The proposed end-to-end semantic communication and detection pipeline captures both spatial anomalies at the feature level and structural inconsistencies in semantic relationships, enabling robust detection of backdoor attacks in GenAI-enabled systems.

4. EXPERIMENT

4.1 Experiment settings

The experiment is conducted to simulate backdoor attacks in a GenAI-driven semantic communication system and to evaluate the CNN-AAE framework for detection. Two datasets are used: MNIST [25] and CIFAR-10 [26]. As listed in Table 1, the MNIST dataset consists of 28 × 28 grayscale digit images, while CIFAR-10 contains 32 × 32 RGB images from 10 object categories. Both datasets are split into 80% training and 20% testing subsets. To emulate GenAI-driven semantic encoding at the sender, the raw input images are passed through a pretrained CNN encoder f_θ, which extracts semantic embeddings $s_i = f_\theta(x_i) \in \mathbb{R}^m$. This CNN encoder acts as a lightweight proxy for a full GenAI model, capturing high-level semantic features that represent the transmitted message. These semantic embeddings simulate how GenAI compresses and abstracts content for communication. Preprocessing includes pixel normalization to $[0, 1]$. For CIFAR-10, data augmentation techniques such as random cropping and

Table 1 – Characteristics of MNIST and CIFAR-10 datasets

Dataset	Classes	Image Size	Color Channels	Training Samples	Test Samples	Common Uses
MNIST	10	28×28	Grayscale (1)	60,000	10,000	Digit Recognition
CIFAR-10	10	32×32	RGB (3)	50,000	10,000	Object Recognition

Table 2 – Evaluation metrics of DL models on the MNIST and CIFAR-10 datasets

Dataset	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ASR (%)
MNIST	CNN	96.2	95.1	94.5	94.8	3.0
	MLP	94.5	93.2	92.3	92.7	5.2
	FCNN	92.8	91.5	90.2	90.8	8.1
	Autoencoder	95.3	94.2	93.1	93.6	4.1
	ABM	96.4	95.0	94.0	94.5	2.4
	CNN-AAE	**97.8**	**96.7**	**96.2**	**96.4**	**2.2**
CIFAR-10	CNN	92.5	91.2	90.5	90.8	7.2
	MLP	89.1	87.5	86.8	87.1	10.5
	FCNN	85.3	84.0	83.2	83.6	15.2
	Autoencoder	90.2	89.0	88.3	88.6	9.4
	ABM	93.6	92.2	91.5	91.8	6.0
	CNN-AAE	**95.4**	**94.3**	**93.7**	**94.0**	**5.1**

horizontal flipping are applied. The experiment operates entirely in a semantic feature space, allowing detection to occur based on deviations from the distribution of clean embeddings.

The CNN-AAE framework is trained alongside baseline models, CNN, MLP, FCNN, autoencoder, and ABM approach, on both clean and poisoned data. All models are trained using the Adam optimizer with a fixed learning rate of 0.001 and a batch size of 64. For the classification tasks, categorical cross-entropy is used as the loss function, whereas the autoencoder is trained separately using mean squared error. Evaluation is performed on a disjoint test set including both clean and poisoned inputs. By simulating GenAI-driven abstraction via the CNN encoder and performing detection based on semantic deviation and reconstruction error, the proposed framework effectively identifies poisoned samples in semantic communication pipelines.

4.2 Evaluation metrics for detection performance

To evaluate the effectiveness of the proposed CNN-AAE approach and baseline models in detecting backdoor attacks, we utilize several standard performance metrics. These include accuracy, precision, recall, F1 score, and ASR. Table 2 presents the evaluation metrics for various models in detecting backdoor attacks within GenAI-powered semantic communication systems under a 3% poisoning ratio. The results cover the MNIST and CIFAR-10 datasets. Among the baseline models, the CNN demonstrates high robustness, achieving 96.2% accuracy on MNIST and 92.5% on CIFAR-10, with ASR values of 3.0% and 7.2%, respectively. While CNN performs well, its relatively higher ASR on CIFAR-10 suggests it can still be bypassed by sophisticated backdoor attacks. The autoencoder model also performs competitively, especially in learning semantic deviations. It achieves 95.3% accuracy and 4.1% ASR on MNIST, and 90.2% accuracy and 9.4% ASR on CIFAR-10. These results demonstrate its sensitivity to semantic anomalies, although its performance declines on more complex datasets.

The MLP and FCNN models show notably reduced resilience. MLP reaches 94.5% accuracy with a 5.2% ASR on MNIST, dropping to 89.1% accuracy and a 10.5% ASR on CIFAR-10. FCNN records the lowest performance among all baselines, with only 85.3% accuracy and a high ASR of 15.2% on CIFAR-10, confirming its limited capability in extracting discriminative features for backdoor detection. The ABM approach demonstrates competitive performance. On MNIST, it achieves 96.4% accuracy with a 2.4% ASR, outperforming most baseline models but remaining slightly behind the proposed CNN-AAE. On CIFAR-10, an ABM approach reaches 93.6% accuracy with a 6.0% ASR, showing improved robustness compared to CNN and autoencoder. ABM's ability to suppress backdoor effects stems from its non-invasive, gradient-insensitive training strategy, which aligns well with high-level semantic purification.

The proposed CNN–AAE framework significantly outperforms all baseline and comparative models. It achieves the highest accuracy of 97.8% on MNIST and 95.4% on CIFAR-10. Its ASR is the lowest among all models, registering only 2.2% on MNIST and 5.1% on CIFAR-10. This improved resistance is attributed to its dual-stage detection mechanism: CNN-based spatial feature analysis and autoencoder-based semantic consistency checks. Furthermore, CNN-AAE benefits from adaptive thresholding based on statistical deviations, enhancing detection sensitivity under dynamic and complex attack scenarios.

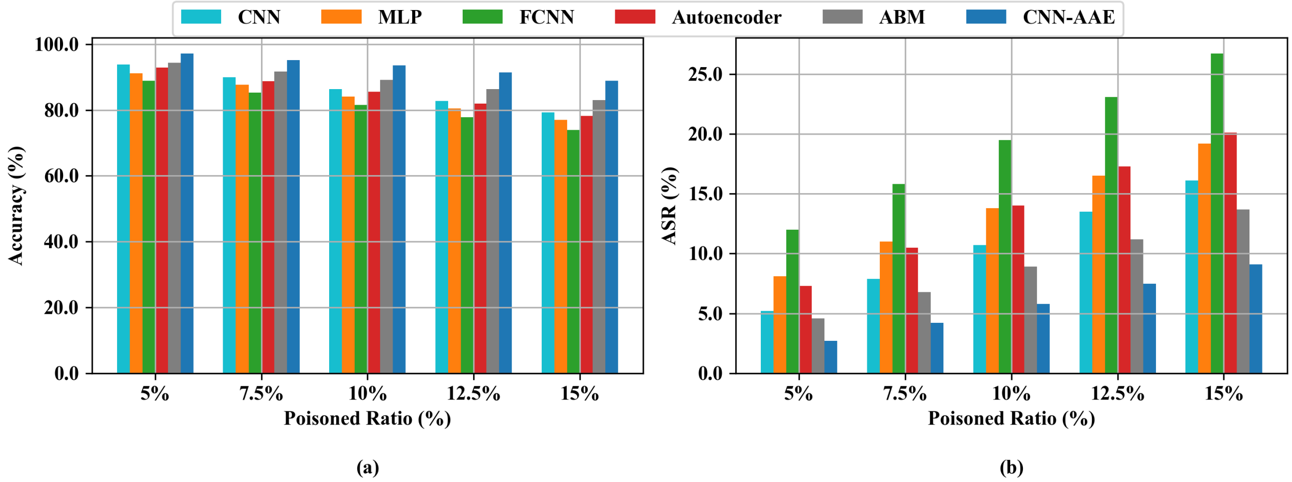

Figure 2 – Impact of poison ratio on backdoor attack in the MNIST dataset: (a) Accuracy (ACC) and (b) Attack Success Rate (ASR).

4.3 Robustness analysis under varying poisoning levels

To evaluate model robustness against semantic backdoor attacks, a portion of the training samples is poisoned by embedding imperceptible triggers into input images and relabeling them to a fixed target class. This type of attack manipulates the semantic space rather than relying on visible pixel patterns, making detection more difficult. Poisoning ratios of 5%, 7.5%, 10%, 12.5%, and 15% are used to simulate varying levels of adversarial influence. In this setting, we assess the proposed CNN-AAE model alongside baseline methods (CNN, MLP, FCNN, autoencoder, and ABM) on MNIST and CIFAR-10 datasets. We focus on two key metrics: classification accuracy on clean data and ASR on poisoned data, to determine each model's resilience as the poisoning level increases. According to Fig. 2, the CNN-AAE model continues to outperform all baselines, achieving the highest accuracy of 97.1% at 5% poisoned ratio and maintaining strong performance with 88.9% at 15%. Its ASR remains the lowest, increasing modestly from 2.7% to 9.1%, indicating high robustness to backdoor attacks due to its hybrid architecture that combines spatial feature learning and semantic reconstruction. Notably, the ABM approach also demonstrates strong performance, achieving 94.3% accuracy at 5% poisoning and sustaining 83.0% accuracy at 15%. Its ASR remains relatively low, rising from 4.6% to 13.7%, positioning ABM as a competitive defense method that surpasses traditional baselines while remaining slightly less robust than CNN-AAE. In contrast, the CNN model, although effective in spatial feature extraction, shows a decline in accuracy from 93.8% to 79.3% and an increase in ASR from 5.2% to 16.1%, due to its lack of a dedicated mechanism for filtering poisoned patterns. The MLP model performs moderately, with accuracy falling from 91.2% to 77.0% and ASR increasing from 8.1% to 19.2%, demonstrating higher vulnerability to subtle perturbations. FCNN exhibits the steepest degradation, dropping from 88.9% to 73.9% in accuracy and spiking from 12.0% to 26.7% in ASR, indicating poor resilience under attack. The autoencoder performs better than FCNN and MLP, with accuracy decreasing from 92.9% to 78.2% and ASR growing from 7.3% to 20.1%, benefiting from its reconstruction-based detection yet struggling as poisoning increases.

Fig. 3 presents the results obtained from the CIFAR-10 dataset, highlighting the challenges of detecting backdoor attacks in more complex, high-dimensional datasets. Compared to simpler datasets like MNIST, the color images in CIFAR-10 introduce additional difficulties, resulting in lower accuracy and higher ASR across all models. This confirms that backdoor detection becomes significantly more challenging as dataset complexity increases. The CNN-AAE model demonstrates superior performance, achieving 94.07% accuracy at 5% poisoning and sustaining 83.10% at 15%, outperforming all other models. Its ASR remains the lowest, increasing only from 5.86% to 15.56%, confirming its robustness. CNN-AAE leverages both spatial feature extraction through CNN and semantic reconstruction via autoencoder, allowing it to maintain high accuracy and detect backdoor anomalies even in high-dimensional data. The newly introduced ABM approach also shows strong performance, achieving 91.2% accuracy at 5% poisoning and maintaining 79.4% at 15%. Its ASR remains relatively low, increasing from 6.4% to 15.9%, positioning ABM as a highly resilient model, second only to CNN-AAE.

In comparison, the CNN model maintains decent robustness, with accuracy declining from 90.8% to 76.3% and ASR rising from 8.2% to 19.77%. While CNN cap-

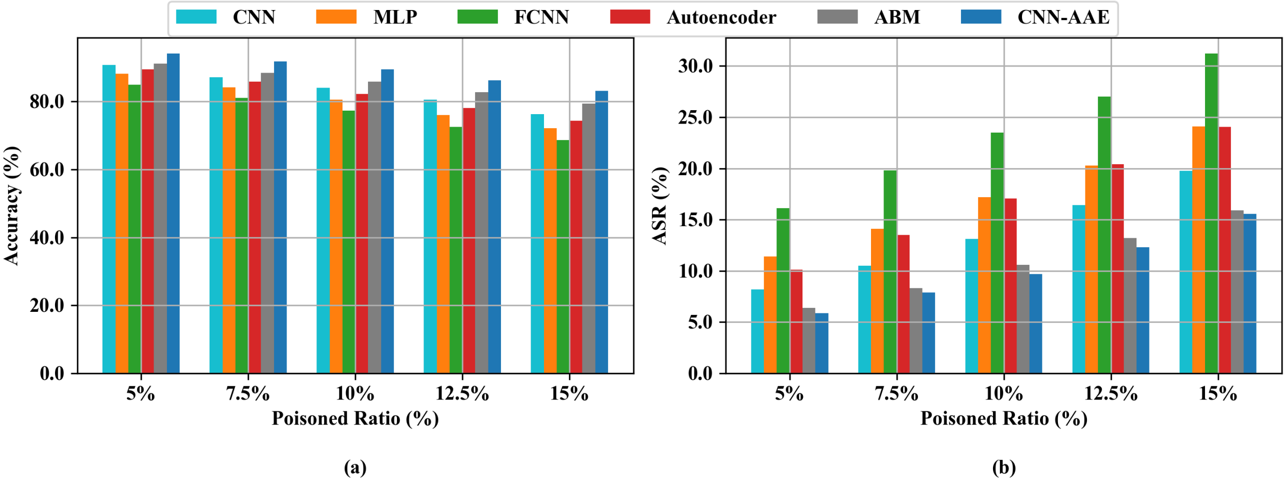

Figure 3 – Impact of poison ratio on backdoor attack in the CIFAR-10 dataset: (a) Accuracy (ACC) and (b) Attack Success Rate (ASR).

tures spatial features effectively, its vulnerability grows at higher poisoning levels due to a lack of semantic-level defenses. The MLP model struggles to cope with poisoning, as accuracy drops from 88.12% to 72.1% and ASR increases significantly from 11.4% to 24.1%. MLP's lack of spatial modeling makes it more susceptible to pixel-level backdoor perturbations. The FCNN model performs the worst, with accuracy falling from 84.9% to 68.7% and ASR sharply rising from 16.1% to 31.2%. The absence of hierarchical feature extraction makes FCNN ineffective against sophisticated poisoning. The autoencoder model provides moderate resilience, with accuracy dropping from 89.5% to 74.3% and ASR increasing from 10.11% to 24.07%. While its reconstruction capability helps identify minor anomalies, its detection power weakens as poisoning increases and the model begins adapting to poisoned samples as normal.

These results confirm that the proposed CNN-AAE is the most effective model for backdoor detection in GenAI-powered semantic communication systems due to its hybrid architectural design. CNN-AAE integrates a CNN for spatial feature extraction with an AE for semantic reconstruction-based anomaly detection. The CNN component excels at capturing local pixel-level patterns and spatial hierarchies in image data, enabling it to highlight irregularities introduced by backdoor triggers. Meanwhile, the autoencoder reconstructs input data based on learned clean distributions, and any deviation in reconstruction helps flag anomalies caused by poisoned inputs. This combined architecture not only preserves high classification performance but also enhances detection sensitivity to subtle perturbations. Unlike baseline models such as MLP and FCNN, which lack explicit mechanisms for modeling spatial structure or reconstruction consistency, CNN-AAE enforces both discriminative and generative constraints during learning, leading to robust anomaly detection. Unlike ABM, which detects backdoors after training by analyzing neuron activations, CNN-AAE embeds detection within its architecture by fusing CNN-based spatial analysis and autoencoder-driven semantic evaluation. This integration enables efficient and responsive detection during inference and consistently achieves lower ASR, making CNN-AAE more effective for GenAI-based backdoor detection.

Table 3 – Training time and memory usage comparison

Model	Training Time (s)	Peak Memory (MB)
CNN	120.4	650
MLP	98.7	430
FCNN	105.2	470
Autoencoder	145.9	710
ABM	185.3	925
CNN-AAE	197.5	917

4.4 Computational resource analysis

To evaluate the computational efficiency of each model, we analyze both training time and memory usage. These metrics are essential to assess the scalability and resource demands of backdoor detection methods under consistent experimental conditions. Table 3 reports the training time and memory usage of all evaluated models. The proposed CNN-AAE requires 197.5 s of training time and 917 MB of memory, reflecting its dual architecture that combines spatial pattern learning (via CNN) with semantic anomaly detection (via the autoencoder). Despite this increased cost compared to simple baselines like MLP or CNN, it remains close to the recent ABM model (185.3s, 925MB) while delivering superior backdoor detection accuracy and lower ASR. This confirms that CNN-AAE maintains an effective trade-off between computational efficiency and detection robustness, making it suitable for GenAI-powered semantic communication systems where both reliability and scalability are essential.

5. CONCLUSION AND FUTURE WORK

This paper introduced CNN-AAE, an adaptive hybrid framework for detecting backdoor attacks in GenAI-powered semantic communication systems. By integrating CNN-based spatial feature extraction with autoencoder-based semantic deviation analysis, CNN-AAE improves detection robustness while minimizing the impact on clean inference accuracy. Our experimental evaluation on MNIST and CIFAR-10 datasets, across multiple poisoning ratios, demonstrated that CNN-AAE consistently outperforms SOTA baselines, including CNN, MLP, FCNN, and autoencoder, as well as the recent ABM approach for backdoor detection. CNN-AAE achieved higher accuracy and significantly lower ASR, while maintaining competitive computational efficiency. We further analyzed the training time and memory usage of all models, showing that CNN-AAE offers an effective trade-off between robustness and resource consumption, making it suitable for secure GenAI-based semantic communications. Future work will extend this approach to larger-scale datasets, evaluate resilience against more complex adaptive backdoors, and enhance model interpretability using explainable AI techniques. Additionally, we aim to optimize computational efficiency for practical deployment in real-world applications.

ACKNOWLEDGEMENT

This work was supported in part by the DoD Center of Excellence in AI and Machine Learning (CoE-AIML) at Howard University under Contract W911NF-20-2-0277 with the U.S. Army Research Laboratory and Microsoft Corp. Research Gift Funds. However, any opinions, findings, conclusions, or recommendations expressed in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

REFERENCES

[1] Gangtao Xin, Pingyi Fan, and Khaled B Letaief. “Semantic communication: A survey of its theoretical development”. In: *Entropy* 26.2 (2024), p. 102.

[2] Shuaishuai Guo, Yanhu Wang, Jia Ye, Anbang Zhang, Peng Zhang, and Kun Xu. “Semantic Importance-Aware Communications with Semantic Correction Using Large Language Models”. In: *IEEE Transactions on Machine Learning in Communications and Networking* (2025).

[3] Kexin Zhang, Lixin Li, Wensheng Lin, Yuna Yan, Rui Li, Wenchi Cheng, and Zhu Han. “Semantic successive refinement: A generative AI-aided semantic communication framework”. In: *IEEE Transactions on Cognitive Communications and Networking* (2025).

[4] Yijing Lin, Zhipeng Gao, Hongyang Du, Jiacheng Wang, and Jiakang Zheng. “Semantic Communication in the Metaverse”. In: *Wireless Semantic Communications: Concepts, Principles and Challenges* (2025), pp. 133–161.

[5] Hong Chen, Fang Fang, and Xianbin Wang. “Semantic extraction model selection for IoT devices in edge-assisted semantic communications”. In: *IEEE Communications Letters* (2024).

[6] Qianwen Wu, Fangfang Liu, Hailun Xia, and Tingxuan Zhang. “Semantic transfer between different tasks in the semantic communication system”. In: *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2022, pp. 566–571.

[7] Cian Blake, Eslam Eldeeb, Mohammad Shehab, Hirley Alves, and Mohamed Slim-Alouini. “Semantic Communication Frameworks for Autonomous Vehicles”. In: *Authorea Preprints* (2025).

[8] Jihong Park, Jinho Choi, Seong-Lyun Kim, and Mehdi Bennis. “Enabling the wireless metaverse via semantic multiverse communication”. In: *2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE. 2023, pp. 85–90.

[9] Tianjing Ren and Hao Wu. “Asymmetric semantic communication system based on diffusion model in iot”. In: *2023 IEEE 23rd International Conference on Communication Technology (ICCT)*. IEEE. 2023, pp. 1–6.

[10] Le Xia, Yao Sun, Chengsi Liang, Lei Zhang, Muhammad Ali Imran, and Dusit Niyato. “Generative AI for semantic communication: Architecture, challenges, and outlook”. In: *IEEE Wireless Communications* 32.1 (2025), pp. 132–140.

[11] Chengsi Liang, Hongyang Du, Yao Sun, Dusit Niyato, Jiawen Kang, Dezong Zhao, and Muhammad Ali Imran. “Generative AI-driven semantic communication networks: Architecture, technologies and applications”. In: *IEEE Transactions on Cognitive Communications and Networking* (2024).

[12] Wei Guo, Benedetta Tondi, and Mauro Barni. “An overview of backdoor attacks against deep neural networks and possible defences”. In: *IEEE Open Journal of Signal Processing* 3 (2022), pp. 261–287.

[13] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. “Badnets: Evaluating backdooring attacks on deep neural networks”. In: *IEEE Access* 7 (2019), pp. 47230–47244.

[14] Yang Bai, Gaojie Xing, Hongyan Wu, Zhihong Rao, Chuan Ma, Shiping Wang, Xiaolei Liu, Yimin Zhou, Jiajia Tang, Kaijun Huang, et al. “Backdoor Attack and Defense on Deep Learning: A Survey”. In: *IEEE Transactions on Computational Social Systems* (2024).

[15] Zhendong Zhao, Xiaojun Chen, Yuexin Xuan, Ye Dong, Dakui Wang, and Kaitai Liang. “Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 15213–15222.

[16] Liuwan Zhu, Rui Ning, Chunsheng Xin, Chonggang Wang, and Hongyi Wu. “CLEAR: Clean-up sample-targeted backdoor in neural networks”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 16453–16462.

[17] Xiaodong Xu, Yue Chen, Bizhu Wang, Zhiqiang Bian, Shujun Han, Chen Dong, Chen Sun, Wenqi Zhang, Lexi Xu, and Ping Zhang. “CSBA: Covert Semantic Backdoor Attack Against Intelligent Connected Vehicles”. In: *IEEE Transactions on Vehicular Technology* (2024).

[18] Yuan Zhou, Rose Qingyang Hu, and Yi Qian. “Backdoor attacks and defenses on semantic-symbol reconstruction in semantic communications”. In: *ICC 2024-IEEE International Conference on Communications*. IEEE. 2024, pp. 734–739.

[19] Bing Sun, Jun Sun, Wayne Koh, and Jie Shi. “Neural Network Semantic Backdoor Detection and Mitigation: A {Causality-Based} Approach”. In: *33rd USENIX Security Symposium (USENIX Security 24)*. 2024, pp. 2883–2900.

[20] Yalin E Sagduyu, Tugba Erpek, Sennur Ulukus, and Aylin Yener. "Vulnerabilities of deep learning-driven semantic communications to backdoor (trojan) attacks". In: *2023 57th Annual Conference on Information Sciences and Systems (CISS)*. IEEE. 2023, pp. 1–6.

[21] Yangming Chen. "An invisible backdoor attack based on semantic feature". In: *arXiv preprint arXiv:2405.11551* (2024).

[22] Jincheng Peng, Huanlai Xing, Lexi Xu, Shouxi Luo, Penglin Dai, Li Feng, Jing Song, Bowen Zhao, and Zhiwen Xiao. "Adversarial Reinforcement Learning based Data Poisoning Attacks Defense for Task-Oriented Multi-User Semantic Communication". In: *IEEE Transactions on Mobile Computing* (2024).

[23] Ziyang Wei, Yili Jiang, Jiaqi Huang, Fangtian Zhong, and Sohan Gyawali. "Detecting Backdoor Attacks via Similarity in Semantic Communication Systems". In: *arXiv preprint arXiv:2502.03721* (2025).

[24] Chen Chen, Haibo Hong, Tao Xiang, and Mande Xie. "Anti-Backdoor Model: A Novel Algorithm To Remove Backdoors in a Non-invasive Way". In: *IEEE Transactions on Information Forensics and Security* (2024, publisher=IEEE).

[25] Li Deng. "The mnist database of handwritten digit images for machine learning research [best of the web]". In: *IEEE signal processing magazine* 29.6 (2012), pp. 141–142.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. *Learning multiple layers of features from tiny images.(2009)*. 2009.

AUTHORS

Hassan El Alami (SMIEEE) received a Ph.D. in computer science and telecommunications from the National Institute of Posts and Telecommunications (INPT) in Rabat, Morocco, in 2019. Currently, he is a postdoctoral research fellow at the DoD Center of Excellence in Artificial Intelligence and Machine Learning (CoE-AIML) at Howard University's College of Engineering and Architecture (CEA), Department of Electrical Engineering and Computer Science in Washington, D.C., USA. Previously, he was a postdoctoral research fellow at the Artificial Intelligence Research Initiative at the University of North Dakota's College of Engineering & Mines. His current research interests include artificial intelligence, cybersecurity, autonomous systems, the metaverse, and the Internet of Things. Dr. El Alami is a senior member of IEEE, and a member of ACM, the Association for the Advancement of Artificial Intelligence (AAAI), the USENIX Association, the Institute of Navigation (ION), and the Sigma Xi Scientific Research Honor Society.

Danda B. Rawat is an associate dean for Research & Graduate Studies, a full professor in the Department of Electrical Engineering & Computer Science (EECS), Founding Director of the Howard University Data Science & Cybersecurity Center, Founding Director of DoD Center of Excellence in Artificial Intelligence & Machine Learning (CoE-AIML), Director of Trustworthy Artificial Intelligence (TruAI) Research Lab, Director of Cyber-security and Wireless Networking Innovations (CWiNs) Research Lab, and Director of Graduate Cybersecurity Certificate Program at Howard University, Washington, DC, USA. Dr. Danda B. Rawat successfully led and established the Research Institute for Tactical Autonomy (RITA), the 15th University Affiliated Research Center (UARC) of the US Department of Defense as the PI/Founding Executive Director at Howard University, Washington, DC, USA. Dr. Rawat is engaged in research and teaching in the areas of cybersecurity, machine learning, big data analytics, and wireless networking for emerging networked systems including cyber-physical systems (eHealth, energy, transportation), Internet of Things, multi-domain operations, smart cities, software-defined systems and vehicular networks. Dr. Rawat has secured over $110 million as a PI and over $18 million as a Co-PI in research funding from the US National Science Foundation (NSF), US Department of Homeland Security (DHS), US National Security Agency (NSA), US Department of Energy, National Nuclear Security Administration (NNSA), National Institute of Health (NIH), US Department of Defense (DoD) and DoD Research Labs, Industry (Microsoft, Intel, VMware, PayPal, Mastercard, Meta, BAE, Raytheon, etc.) and private foundations. Dr. Rawat is the recipient of the US NSF CAREER Award, the US Department of Homeland Security (DHS) Scientific Leadership Award, the President's Medal of Achievement Award (2023) at Howard University, Provost's Distinguished Service Award 2021, Researcher Exemplar Award 2019 and Graduate Faculty Exemplar Award 2019 from Howard University, the US Air Force Research Laboratory (AFRL) Summer Faculty Visiting Fellowship 2017, Outstanding Research Faculty Award (Award for Excellence in Scholarly Activity) at GSU in 2015, the Best Paper Awards (IEEE CCNC, IEEE ICII, IEEE DroneCom and BWCA) and Outstanding PhD Researcher Award in 2009. He has delivered over 100 keynotes and invited speeches at international conferences and workshops. Dr. Rawat has published over 350 scientific/technical articles and 11 books. Dr. Rawat has successfully supervised and graduated 35 PhD students (out of which 28 were under-represented PhD students including 13 female PhD students), successfully supervised 30+ MS students and mentored 7 postdocs, and has been supervising 25 PhD students and mentoring 3 postdocs. Furthermore, he has successfully mentored over 120 minority undergraduate students. He has been serving as an editor/guest editor for over 100 international journals including the associate editor of IEEE Transactions on Big Data, associate editor of IEEE Transactions on Information Forensics & Security, associate editor of Transactions on Cognitive Communications and Networking, associate editor of IEEE Transactions of Service Computing (2018 - 2022), editor of IEEE Internet of Things Journal, editor of IEEE Communications Letters, associate editor of IEEE Transactions of Network Science and Engineering (2019 - 2023) and technical editors of IEEE Network, and associate editor of ACM Transactions on Intelligent Systems and Technology. He has been organizing committees for several IEEE flagship conferences such as IEEE INFOCOM, IEEE CNS, IEEE ICC, IEEE GLOBECOM, and so on. He served as a Technical Program Committee (TPC) member for several international conferences including IEEE INFOCOM, IEEE GLOBECOM, IEEE CCNC, IEEE GreenCom, IEEE ICC, IEEE WCNC, and IEEE VTC conferences. He served as a vice chair of the executive committee of the IEEE Savannah Section from 2013 to 2017. Dr. Rawat received a Ph.D. degree from Old Dominion University, Norfolk, Virginia. Dr. Rawat served as a graduate program director of Howard Computer Science Graduate Programs (2017 - 2025). Dr. Rawat is a senior member of IEEE and a lifetime professional senior member of ACM, a lifetime member of the association for the Advancement of Artificial Intelligence (AAAI), a champion member of USENIX (The Advanced Computing Systems Association), a lifetime member of SPIE, a member of ASEE, a member of AAAS, a member of SAE International, and a fellow of the Institution of Engineering and Technology (IET). He is an ACM distinguished speaker and an IEEE distinguished lecturer (FNTC and VTS).