Image-Based Dual Defense Strategy for Adversarially Robust IDS in Smart Agriculture

Rafika Saadouni, Chirihane Gherbi, Zibouda Aliouat, Yasmine Harbi, Amina Khacha and Hakim Mabed

***Abstract*—The integration of Internet of Things (IoT) technologies into smart agriculture has significantly enhanced automation, monitoring, and productivity. However, these systems introduce critical cybersecurity vulnerabilities and generate heterogeneous data, including structured network traffic and image-based inputs. This necessitates an intrusion detection system (IDS) capable of handling multimodal data, particularly as adversarial attacks can manipulate inputs to evade traditional detection models. To address these challenges, this paper proposes a novel image-based IDS for smart agriculture environments. The system transforms network traffic into images and employs VGG16 for feature extraction, Binary Greylag Goose Optimization for feature selection, and a random forest for classification. It further integrates a dual defense strategy combining a Convolutional Autoencoder Denoising (CAED) module with Adversarial Training (AT) to improve robustness against adversarial perturbations. The proposed solution is evaluated on the CICIoT2023 dataset across eight traffic classes under three white-box adversarial attacks. The IDS demonstrates strong resilience across all perturbation levels. For weak perturbations (ε=0.01), the dual defense achieves accuracies of at least 99.46%. Under moderate perturbations (ε=0.1), it maintains high performance with macro-averaged accuracies of at least 99.39%. Even under strong perturbations (ε=0.3), the system remains robust, attaining an accuracy of at least 96.50%. To assess generalization to real agricultural settings, the IDS is also tested using native crop images from agriculture Dataset. Under severe adversarial distortion (ε=0.3), the system sustains strong robustness with a macro-averaged accuracy of at least 96.71%. These results confirm that the proposed multimodal IDS offers a resilient and adaptive security solution for smart agriculture networks facing advanced adversarial threats.**

***Index Terms*—FGSM, FGM, PGD, VGG16, BGGO, Adversarial Training, CICIoT2023, Random Forest, Autoencoder Denoising, white-box.**

I. Introduction

The integration of Internet of Things (IoT) technologies into smart agriculture has revolutionized farming practices through real-time monitoring, automation, and data-driven decision-making [1], [2]. However, the over-connectivity also expands the attack surface, making agricultural systems vulnerable to a range of cybersecurity threats. Among these, network-based intrusions—including Denial of Service (DoS), Distributed Denial of Service (DDoS), spoofing, and data manipulation attacks—pose serious risks to the reliability and availability of smart agriculture infrastructures [3].

Corresponding author: Zibouda Aliouat is with LRSD Laboratory, Ferhat Abbas University of Setif 1, Setif, 19000, Algeria, (email: zaliouat@univ-setif.dz)

Rafika Saadouni, Chirihane Gherbi, Yasmine Harbi, and Amina Khacha are with LRSD Laboratory, Ferhat Abbas University of Setif 1, Setif, 19000, Algeria.

Hakim Mabed is with DISC/FEMTO-ST, University Bourgogne Franche-Comte, 25200 Montbeliard, France

Intrusion Detection Systems (IDSs) serve as a critical line of defense by monitoring network traffic for malicious activity. Although traditional signature-based IDSs can effectively identify known attacks, they often fail to detect novel or evolving threats, especially those exploiting zero-day vulnerabilities. To overcome this limitation, Machine Learning (ML)-based IDSs have gained prominence due to their ability to learn patterns in network behavior and generalize to previously unseen attack types [4].

Despite their advantages, ML-based IDSs are highly susceptible to adversarial attacks, in which subtle perturbations are added to input data to mislead the classifier. Such attacks exploit weaknesses in the model's decision boundaries and are particularly dangerous in critical infrastructures such as smart agriculture. Adversarial samples often mimic benign traffic by manipulating mutable features like packet size or flow duration, making them difficult to detect with conventional models. This vulnerability is especially critical in smart agriculture, where network integrity is essential for safe and continuous farm operations.

Adversarial attacks are typically categorized according to the attacker's knowledge of the target model. *White-box attacks* [5], [6] assume full access to the model's architecture and parameters, enabling precise gradient-based manipulations. *Black-box attacks* [7], [8] rely solely on input–output observations without insight into the model's internals, while *gray-box attacks* [9] operate under partial knowledge—such as awareness of the training data or model structure, but not the exact parameters. Among these, white-box attacks pose the greatest threat, as they can directly exploit specific vulnerabilities in the model's decision surface.

Many existing studies have focused on improving detection accuracy under clean (non-adversarial) conditions [10]–[13], or on generating adversarial samples to evaluate model weaknesses [14], [15]. However, relatively few works have proposed robust defense strategies that can mitigate such attacks in real-world scenarios [16], [17]. In particular, there is a need for defense mechanisms that not only detect adversarial inputs but also restore them for instance, through denoising—to recover reliable model performance.

This paper proposes a novel defense-enhanced IDS framework tailored for smart agriculture networks. The framework combines a Convolutional Autoencoder (CAE) for adversarial input denoising with adversarial training to improve classifier robustness. By integrating reconstruction-based filtering and exposure to adversarial examples during training, the system can better resist evasion attempts and maintain reliable detection performance. Additionally, the learned feature representations are visualized using t-distributed stochastic neighbor

embedding (t-SNE), which enhances class separability in high-dimensional space. The main contributions of this work are summarized as follows:

- We extend our previously proposed IDS for smart agriculture environments [18]. The new proposed approach processes both digital network traffic and native image data. The digital traffic is first converted into images, which enables uniform processing and deep feature extraction with a transfer learning model (VGG16). The baseline IDS employs a three-stage architecture: VGG16 for feature extraction, BGGO (Binary Greylag Goose Optimization) for feature selection [19], and Random Forest (RF) for classification. This design ensures strong performance under clean conditions while maintaining compatibility with the heterogeneous data types present in smart agriculture.
- To evaluate adversarial robustness, we generate adversarial images from the CICIoT2023 [20] dataset using white-box attack techniques, including the Fast Gradient Sign Method (FGSM), Fast Gradient Method (FGM), and Projected Gradient Descent (PGD). These perturbations simulate realistic evasion scenarios in which attackers attempt to fool the IDS.
- We propose three defensive strategies to address the adversarial attacks: (i) adversarial training, (ii) CAE to denoise adversarial perturbations, and (iii) a hybrid dual-defense mechanism that combines both methods to maximize robustness.
- We perform feature-level analysis using t-SNE to visualize the impact of adversarial perturbations and assess how each defense improves class separability in the learned feature space.
- Extensive experiments were conducted under clean and adversarial conditions across multiple metrics, demonstrating that the proposed dual defense mechanism significantly improves detection performance and robustness.

The remainder of this paper is organized as follows: Section 2 reviews recent relevant studies, Section 3 presents the proposed framework, Section 4 discusses the experimental results, Section 5 provides a comparison with existing works, and Section 6 concludes the paper.

II. Related Work

Gaber et al. [21] proposed a robust IDS for the Internet of Flying Things (IoFT) to address critical security threats affecting Confidentiality, Integrity, and Availability (CIA). They used Generative Adversarial Networks (GANs) to generate adversarial samples, creating a hybrid dataset that combines real IoFT traffic (ECU-IoFT and CICIDS2018) with synthetic attacks, overcoming the lack of diverse public datasets. An adversarial training strategy was also employed to improve resilience against attacks such as the FGSM, Basic Iterative Method (BIM), and Carlini & Wagner (C&W). The system was tested using RF, Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR), with RF achieving 96.5% accuracy under normal conditions and 93.5% under FGSM in binary classification. However, under stronger attacks like C&W, accuracy dropped to 82.9%, revealing the model's vulnerability to sophisticated threats.

Aloraini et al. [14] explored how adversarial attacks can target IDSs in in-vehicle networks (IVNs) of autonomous vehicles, rather than perception models. They proposed a black-box attack using a substitute IDS trained on onboard diagnostic data, exploiting the transferability of adversarial examples. Using the Car Hacking dataset, their approach fooled both a baseline IDS and a state-of-the-art model (MTH-IDS). Under attack, F1 scores decreased sharply, from 95% to 38% for the baseline and from 97% to 79% for MTH-IDS. The attack was especially effective in causing false positives, with the MTH-IDS misclassifying over 52% of normal frames. The study highlights the vulnerabilities of IVNs due to inherent weaknesses in the Controller Area Network (CAN) protocol. Originally designed for isolated automotive systems, CAN lacks essential security features such as message authentication, encryption, and sender verification. Its broadcast nature and identifier-based message prioritization further expose it to exploitation. These limitations make CAN especially susceptible to adversarial manipulation. However, no defense strategies were proposed.

Khan et al. [22] examined ML-based IDS vulnerabilities in Industrial Control Systems (ICS) and proposed RADIANT, a reactive defense using three parallel autoencoders (Denoising, Benign-only, Attack-only) to extract robust features without adversarial retraining. Evaluated against HopSkipJump and Eroth-Order Optimization (ZOO) attacks on a power system dataset, RADIANT achieved F1 scores of 91.4% and 85.9% under HopSkipJump and ZOO attacks, outperforming baseline RF by 20.5% and 17.1%, respectively. However, its binary classification focus can limit the effectiveness in distinguishing specific attack types needed for detailed analysis.

Benyamin et al. [23] introduced a robust defense framework for ML-based network IDS against adversarial evasion attacks. It combines adversarial training, SMOTE-based balancing, protocol-aware feature engineering, and an ensemble of ML and DL models (LR, SVM, RF, DT, LSTM, and Multi-Layer Perceptron (MLP)). Adversarial samples are crafted using a genetic algorithm that maintains protocol integrity. The framework, tested on NSL-KDD and UNSW-NB15, improved detection by 35% and reduced false positives by 12.5%. However, its binary classification focus may hinder differentiation between attack types, and reliance on outdated datasets limits generalizability to modern threats.

Paya et al. [24] introduced Apollon, a defense system for ML-based IDSs against adversarial attacks. It employed a pool of classifiers (LR, RF, DT, SVM, Naive Bayes (NB), MLP, and Fuzziness-based Neural Networks (FNN)). It utilized Multi-Armed Bandits with Thompson sampling to select the most suitable model for each input. This approach introduced uncertainty, making it difficult for attackers to predict IDS behavior and craft adversarial traffic. Evaluated on CIC-IDS-2017, CSE-CIC-IDS-2018, and CIC-DDoS-2019, Apollon achieved up to 93.04% accuracy, detection rates up to 52.60%, and F1 scores up to 58.35% against AML attacks such as ZOO, HopSkipJump, and Wasserstein GAN (W-GAN) based attacks, significantly outperforming standalone models.

Lin et al. [25] investigated the vulnerability of ML-based IDSs for Controller Area Network (CAN) environments against adversarial attacks. Their study demonstrated that attacks such as ZOO significantly degrade detection performance, with F1-scores dropping from nearly 99% to 56% under known attacks and further to 24% for unknown ones. They evaluated several defense strategies to mitigate these threats, including adversarial training with triplet loss, ensemble learning, and distance-based feature space optimization via simulated annealing. The latter proved most effective, achieving an F1-score of 97% under adversarial conditions and outperforming traditional adversarial training approaches.

Bommana et al. [26] proposed a deep learning-based framework to detect adversarial attacks in IoT systems, addressing their susceptibility to data manipulation. Their approach utilized a three-stage preprocessing pipeline to eliminate high-frequency noise and irrelevant features, including Adaptive Weighted Gaussian Filtering, Adaptive Weighted Mean Filtering, and Adaptive Wavelet Filtering. Feature extraction was performed using a modified Convolutional Neural Network (CNN) with Adaptive Dilated Enriched Convolution and Weighted Adaptive Batch Normalization, optimized by a Quantum-inspired Coati Optimization Algorithm. For classification, they employed a hybrid model combining a Recurrent CNN and a Restricted Boltzmann Machine (RBM), connected through a self-attention-based weight-sharing mechanism. The model achieved high AUC scores of 0.95 and 0.97 on the IoT/IIoT and RT-IoT2022 datasets, respectively, outperforming baseline models. However, the method relied on binary classification, distinguishing only between normal and adversarial behavior, which limited its ability to detect and classify specific types of adversarial attacks in more complex IoT environments.

Table I compares our proposed method with existing works, highlighting the key differences. Prior research on adversarial robustness in IDSs has several limitations: most approaches use binary classification, which restricts the ability to distinguish between different attack types, many rely on outdated datasets, making them less applicable to current IoT threats, and most lack defense strategies that are suitable for processing heterogeneous data (numeric and image-based), as well as the ability to generalize across different forms of adversarial noise and perturbations.

In contrast, our proposed IDS overcomes these limitations by supporting multi-class classification on the up-to-date CICIoT2023 dataset, allowing for more precise attack identification. It combines image-based processing with VGG16 and BGGO to effectively manage diverse input types. Additionally, we introduce a dual defense mechanism that integrates adversarial training with a CAED. This dual strategy is robust against various attack types and perturbation levels, as CAED effectively removes adversarial noise regardless of the epsilon value or attack type.

III. Proposed Framework

This section details the architecture and operational flow of our IDS designed for smart agriculture environments. The foundational components of this IDS were previously established in our work [18], where we highlighted its unique ability to handle diverse data types. In this study, we extend the IDS by introducing a novel **dual defense strategy**, specifically designed to complement attack detection within the system and enhance its robustness against adversarial threats. This advancement ensures reliable and secure operation in dynamic agricultural settings. The overall structure and process of the proposed framework are illustrated in Figure 1.

A. Overview of the Core IDS Architecture

Our IDS system primarily consists of three interconnected stages: a feature extraction module utilizing a pre-trained VGG16 model, a feature selection component powered by the BGGO algorithm, and a final classification layer implemented with an RF classifier.

1) Dual Data Handling Capability (Integrating Numeric and Image-Based traffic): Our proposed IDS characterized by its comprehensive ability to process both numeric and image-based data concurrently within a unified framework. This design addresses the heterogeneous nature of data sources prevalent in smart agriculture, including sensor readings (numeric) and visual feeds (images).

To achieve this, we employ a strategic data conversion approach.

- **Numeric Data Conversion:** Incoming numeric IoT traffic is systematically transformed into image representations, enabling the use of powerful image-processing techniques and transfer learning models. Converting heterogeneous numerical patterns into a unified visual format allows the IDS to consolidate diverse data types into a single, highly efficient processing pipeline.
  The conversion of numerical traffic into image form starts by removing redundant rows and eliminating any NaN or null values to ensure data integrity. After data cleaning, all 46 numerical features are normalized. Since image pixel intensities range from 0 to 255, the network-traffic values are scaled to this range using *quantile normalization*, which maps feature distributions to a uniform scale while reducing the influence of outliers, an essential step for preventing distorted or overly bright pixel values.
  Once normalized, the samples are segmented into chunks of 138 consecutive rows, forming matrices of size 46×138 (6348 values). Each matrix is then reshaped into an RGB image of dimension $46 \times 46 \times 3$. This conversion enables the use of convolutional neural networks and transfer learning techniques to extract high-level representations from traffic-derived visual patterns. The final distribution of all traffic-derived image categories is reported in Table II.
- **Direct Image Processing:** Image-based data is directly fed into the system without requiring conversion, as it is inherently compatible with our image-centric processing modules.

2) Feature Engineering: The processed images (native and converted from numeric data) are then passed to the feature engineering module that is composed of three procedures:

TABLE I: Comparison of Related Work.

Reference	Year	Format of Input Data	Adversarial Data Generation Method	Defense Mechanism	Feature Engineering	
					Extraction	Selection
[21]	2025	Numeric	GAN	Adversarial Training	✗	✗
[14]	2024	Numeric	Black-box attack using a substitute model	✗	✗	✗
[22]	2025	Numeric	ZOO, HopSkipJump	– Denoising Autoencoder – Benign-only Autoencoder – Attack-only Autoencoder	✓	✗
[23]	2024	Numeric	Genetic Algorithm	Adversarial Training	✓	✓
[24]	2024	Numeric	ZOO, HopSkipJump, W-GAN	Apollon defense system	✓	✗
[25]	2025	Numeric	Not Specified	– Adversarial training with triplet loss – Ensemble learning – Distance-based feature space optimization using simulated annealing	✗	✗
[26]	2025	Numeric	Not specified	Adaptive noise filtering Q-COA and self attention with two-tier classification mechanism	✓	✓
Our	2025	Image	FGM, FGSM, PGD	– Adversarial Training (AT) – Denoising Convolutional Autoencoder (CAED) – Combined (AT + CAED)	✓	✓

– **VGG16 for Feature Extraction:** The VGG16 model is pre-trained on the ImageNet dataset. It is configured without the top classification layers. This choice allows VGG16 to act as a powerful generic feature extractor, producing rich, high-dimensional representations that encapsulate the essential characteristics of the input images.
– **BGGO for Feature Selection:** The high-dimensional feature vectors produced by VGG16 are then subjected to feature selection using the BGGO algorithm. BGGO plays a critical role in identifying and retaining only the most discriminative and relevant features.
– **Random Forest for Classification** performs the final classification of the input data. In contrast to training complex deep neural networks, Random Forest (RF) method applied to pre-extracted and selected features, requires significantly fewer computational resources and less training time.

B. Training/Validation Datasets

For the purpose of training, validation and testing of the proposed Baseline IDS, we have assembled a diversified input traffic Dataset. The Dataset incorporates both numerical network-traffic based images and real crop images sourced from the *Crops Image Dataset* available on Kaggle [1]. This dataset is a comprehensive and diversified collection of 139 widely cultivated crop types, with approximately 250 images per class.
In order to maintain a balanced dataset of numerical IoT traffic images and benign crop images, we extracted a training subset of 1520 RGB images and 520 RGB images for validation and test. These selected images represent various plants and crop types grouped into a single benign class, denoted as *normal_crop* in Table II. The number of images included in each split is given in Table II. Lines DDoS, DoS, Mirali, Spoofin, Recon, Web, BruteForce refer to malicious images. Benign images refers to non malicious real images, while *normal_crop* designates the IoT traffic based images.

[1]https://www.kaggle.com/datasets/omrathod2003/140-most-popular-crops-image-dataset/data

TABLE II: Dataset repartition over the different categories.

Category	Number of Images		
	Training	Validation	Testing
DDoS	1530	485	527
DoS	1495	532	517
Mirai	1544	489	517
Benign	1498	507	545
Spoofing	1539	520	491
Recon	1570	497	483
Web	1511	519	520
BruteForce	1541	527	482
Normal_Crop	1520	520	520
Total	**13448**	**4596**	**4602**

C. Proposed Dual Defense Strategy

To enhance the robustness of our IDS against evolving adversarial threats, we proposed a novel ”dual defense” mechanism. This approach combines proactive adversarial data generation with two distinct defense strategies, culminating in a more resilient classification pipeline.

1) Adversarial Image Generation: The first step in our defense strategy involves the systematic generation of adversarial examples across our dataset splits. We craft a corresponding adversarial image for each original image in our training, validation, and test subsets, thus creating entirely new datasets of perturbed inputs. The distribution of these newly constructed adversarial datasets mirrors that of our clean data: 60% for training, 20% for validation, and 20% for testing. This ensures comprehensive representation of adversarial examples for both training and evaluation purposes.

To generate adversarial images, we employ three widely recognized gradient-based white-box attacks: Fast Gradient Sign Method (FGSM), Fast Gradient Method (FGM), and

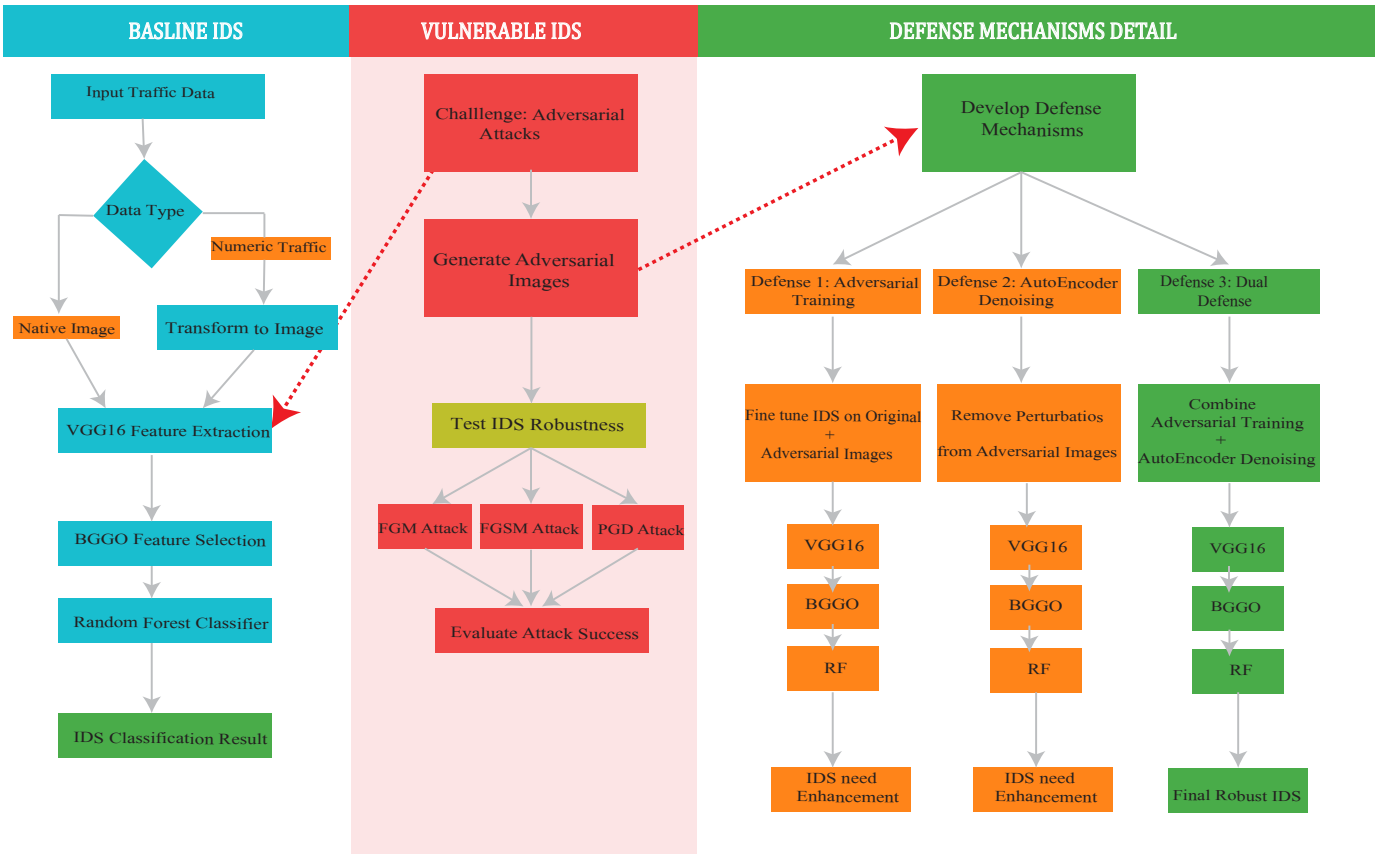


Fig. 1: Flowchart of the proposed Framework.

Projected Gradient Descent (PGD). These attacks are applied to the VGG16 model, which serves as the feature extractor in our IDS. We focus on white-box attacks because they represent the most dangerous scenario, where the attacker has full access to the model's internal parameters [27]–[29]. This setup allows us to simulate adversarial scenarios in which attackers potentially exploit network vulnerabilities in smart agriculture environments and attempt to deceive the IDS by crafting malicious inputs.

Table III outlines the rationale for applying white-box attacks to the VGG16 model instead of the RF classifier. Figure 2 visually illustrates examples of adversarial image perturbations generated using FGM, FGSM, and PGD. The figure includes original images from different classes, the corresponding perturbations applied to each image, and the resulting adversarial images, highlighting the subtle yet effective nature of these attacks.

Each of these attack methods is described below, including the perturbation mechanism applied to the input images and the corresponding mathematical formulation.

- **Fast Gradient Sign Method (FGSM)**
  FGSM [30] is a fast and efficient method for generating adversarial examples by perturbing an input image in a single step. It calculates the gradient of the model's loss function with respect to the input and uses its sign to determine the perturbation direction. The perturbation is then scaled by a small constant ε, and added to the input to maximize the loss and mislead the model.

$$x^{\text{adv}} = x + \varepsilon \cdot \text{sign}(\nabla_x J) \tag{1}$$

  where x is the original input, ε controls the perturbation magnitude, ∇ is the gradient operator. and J is the loss function.
- **Fast Gradient Method (FGM)**
  FGM [31] is similar to FGSM but differs in using the raw gradient of the loss function rather than just its sign. The gradient is scaled by ε, resulting in a perturbation whose magnitude varies with the gradient's intensity. This can

TABLE III: Summary of reasons for applying white-box attacks to VGG16 instead of RF

Aspect	Explanation
Differentiability	VGG16 is differentiable, enabling gradient-based attacks (FGSM, PGD), while RF is not.
Model Architecture & Susceptibility	VGG16's deep layers are more vulnerable to small perturbations; RF is less sensitive due to its design.
Feature Extraction vs. Classification	VGG16 extracts features, and attacking it can indirectly mislead RF by corrupting the input features.
Transferability of Adversarial Examples	Adversarial examples crafted for VGG16 may still impact RF due to feature dependency.
Practicality & Efficiency	Gradient-based attacks are efficient on VGG16; RF requires more complex, less effective methods.

produce smoother perturbations compared to FGSM.

$$x^{\text{adv}} = x + \varepsilon \cdot \nabla_x J \tag{2}$$

- **Projected Gradient Descent (PGD)**
  PGD [32] extends FGSM and generates stronger and more effective adversarial scenarios. Instead of applying a single-step perturbation, PGD applies small perturbations over multiple iterations. After each step, the perturbed input is projected back into the valid ε-ball around the original image. This way, the total perturbation stays within a specified bound.

$$x_{t+1}^{\text{adv}} = \text{Proj}_{\varepsilon}\left(x_t^{\text{adv}} + \alpha \cdot \text{sign}(\nabla_x J)\right) \tag{3}$$

  where α is the step size and $\text{Proj}_{\varepsilon}$ ensures the perturbation remains within an ε-ball around the original input.

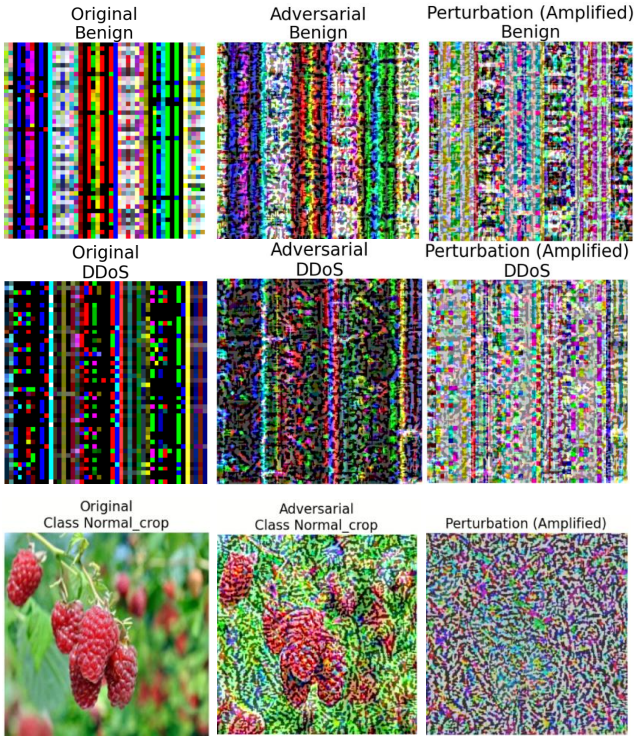


Fig. 2: Examples of generated adversarial image perturbations using FGM, FGSM, and PGD.

2) Initial Vulnerability Assessment and Attack Success Rate: Before implementing any specific defense, we evaluated our baseline IDS's vulnerability to the newly generated adversarial images. This evaluation focuses on calculating the success rate of adversarial attacks. Specifically, we selected only the samples correctly classified by the VGG16-BGGO-RF IDS in its baseline configuration. Then, we tested the IDS with their corresponding adversarial images generated using FGSM, FGM, and PGD. The attack is considered successful if the IDS misclassifies the perturbed input image whereas in the absence of the perturbation, the IDS would correctly classified it. This methodology aligns with our implementation, which retrieves the adversarial images corresponding to correctly classified samples to assess the IDS's susceptibility to adversarial manipulation. The process involves:

- **Feature Extraction:** Extracting features from the adversarial test images using the VGG16 feature extractor (as done for clean data).
- **Feature Selection:** Applying the BGGO-based feature selection algorithm to these extracted adversarial features.
- **Classification:** Testing the RF classifier (which was originally trained solely on clean data) on these feature-selected adversarial images.

3) First Defense Method (Adversarial Training): Our first dedicated defense strategy is an adversarial training approach that acts as a **fine-tuning** process for model robustness. This involves:

- **Data Augmentation:** Combining the original clean training data and the newly generated adversarial training data.
- **Re-training the Random Forest:** The RF classifier is then re-trained on this augmented dataset. This exposes the model to perturbed examples during its learning phase, forcing it to learn more robust decision boundaries.
- **Dual Evaluation:** The performance of this adversarially fine-tuned RF model is then evaluated on both clean and adversarial test datasets to assess its generalized accuracy and improved robustness.

4) Second Defense Method (Convolutional Autoencoder Denoising): An autoencoder is a neural network designed to compress the input data into a lower-dimensional latent space and then decoding it to reconstruct the original input as closely as possible [33]. In the context of our IDS, we employ a CAE to denoise adversarial images and mitigate perturbations introduced by attacks such as FGSM, FGM, and PGD. By identifying the manifold of clean images, the autoencoder can filter out subtle adversarial noise, restoring images to a state closer to their clean counterparts, thereby improving classification robustness. The architecture of the proposed CAE, explicitly designed for this denoising task, is illustrated in Figure 3.

The second defense method introduces a pre-processing step to mitigate adversarial perturbations using the following CAE architecture and pipeline:

- **Convolutional Autoencoder Architecture:** The CAE is designed with a symmetric encoder-decoder structure tailored for image data with an input shape of $(224, 224, 3)$, matching the VGG16 input requirements.
  The Encoder Layers consist of:
  - **First Convolutional Layer, followed by Max-Pooling:** Extracts initial low-level features from the input image, using 32 filters with a 3×3 kernel, ReLU activation, and same padding. The max-pooling layer, with a 2×2 kernel and same padding, reduces spatial dimensions while preserving essential patterns, aiding in early noise detection.
  - **Second Convolutional Layer, followed by Max-Pooling:** Captures more complex features by increasing to 64 filters with a 3×3 kernel, ReLU activation, and same padding. The max-pooling layer, with a

2×2 kernel and same padding, further downsamples the image to enhance noise filtering.
  - **Third Convolutional Layer, followed by MaxPooling:** Encodes high-level features into a compact latent space, using 128 filters with a 3×3 kernel, ReLU activation, and same padding. The max-pooling layer, with a 2×2 kernel and same padding, is crucial for representing the denoised essence of adversarial images.

The decoder mirrors the encoder, reconstructing the image via:

  - **First Convolutional Layer, followed by Upsampling:** Reconstruct the image by upsampling the latent representation, using 128 filters with a 3×3 kernel, ReLU activation, and same padding. With a 2×2 factor, the upsampling layer restores spatial details while refining denoised features.
  - **Second Convolutional Layer, followed by Upsampling:** Enhances mid-level feature reconstruction, using 64 filters with a 3×3 kernel, ReLU activation, and same padding. With a 2×2 factor, the upsampling layer progressively rebuilds the image structure from the latent space.
  - **Third Convolutional Layer, followed by Upsampling:** Further refines the image, adding finer details to approach the original shape, using 32 filters with a 3×3 kernel, ReLU activation, and same padding. The upsampling layer, with a 2×2 factor, supports this process.
  - **Fourth Convolutional Layer:** Outputs the final reconstructed image (224, 224, 3), using 3 filters with a 3×3 kernel, sigmoid activation, and same padding, effectively removing adversarial perturbations by normalizing pixel values.

To minimize reconstruction error, the CAE is compiled with the Adam optimizer and mean squared error (MSE) loss

- **Convolutional Autoencoder Training:** The CAE is trained exclusively on clean training image data. The objective is to learn a compact representation of clean images and reconstruct them effectively, thereby removing subtle noise and potentially adversarial perturbations. This ensures the CAE can generalize to filter out adversarial noise without being exposed to adversarial examples during training.
- **Denoising Pipeline Integration:** Once trained, the CAE is integrated into the IDS pipeline as a pre-processing step. All incoming images (both clean and potentially adversarial) are passed through the CAE for denoising before further processing.
- **Feature Extraction & Selection:** The denoised images are fed into the VGG16 model for feature extraction, followed by the BGGO-based feature selection algorithm to identify the most relevant features for classification.
- **Separate RF Testing:** The RF classifier, trained solely on features extracted from clean data, is tested in two scenarios to evaluate the CAE's effectiveness:
  - **First**, on clean test images after the CAE has processed them to ensure the denoising step does not degrade performance on clean data.
  - **Second**, on adversarial test images after they have been denoised by the CAE. This measures the CAE's ability to restore adversarial images to a classifiable state, enhancing the robustness of the clean-trained RF classifier.

5) Dual Defense Strategy: The ultimate proposed **dual defense** integrates both primary defense methods. This combined approach leverages the strengths of both:

- **CAE Denoising:** Images are initially processed by the trained Convolutional Autoencoder for perturbation removal.
- **Adversarial Training:** The RF classifier used in this final pipeline is the one that has been adversarially fine-tuned (trained on both clean and adversarial data).

D. Experimental Setup

This section outlines the computational environment and the metrics used to evaluate the performance and robustness of our proposed IDS framework.

1) Hardware and Software Configuration: To conduct our experiments, we initially utilized Kaggle's cloud-based platform, which provided up to 29 GB of RAM. However, due to memory constraints and the large size of our dataset, we transitioned to a local high-performance workstation to accommodate the computational demands of our VGG16-BGGO-RF IDS and adversarial attack evaluations. The local environment had 128 GB of RAM and ran Jupyter Notebook through Anaconda, using Python as the primary programming language. We employed the Adversarial Robustness Toolbox (ART) library to implement gradient-based white-box attacks, interfaced with TensorFlow.

2) Evaluation Metrics: To comprehensively assess the performance of our IDS in both clean and adversarial settings, we employ a range of evaluation metrics tailored for multi-class classification. These metrics include the Confusion Matrix, Accuracy, Precision, Recall, F1-Score, Attack Success Rate (ASR), and latency.

- **Confusion Matrix**: is a table that shows how well the model predicts each class by comparing actual and predicted labels [34]. For each class, it summarizes:
  - **True Positives (TP)**: Correctly predicted instances of the class.
  - **False Positives (FP)**: Instances incorrectly predicted as the class but actually belong to another.
  - **False Negatives (FN)**: Instances of the class that were incorrectly predicted as something else.
  - **True Negatives (TN)**: Instances correctly identified as not belonging to the class.
- **Accuracy (Acc)**: measures the overall correctness of the model. It is the ratio of correctly predicted samples (across all classes) to the total number of samples:

$$\text{Acc} = \frac{\text{Total Correct Predictions}}{\text{Total Predictions}}$$

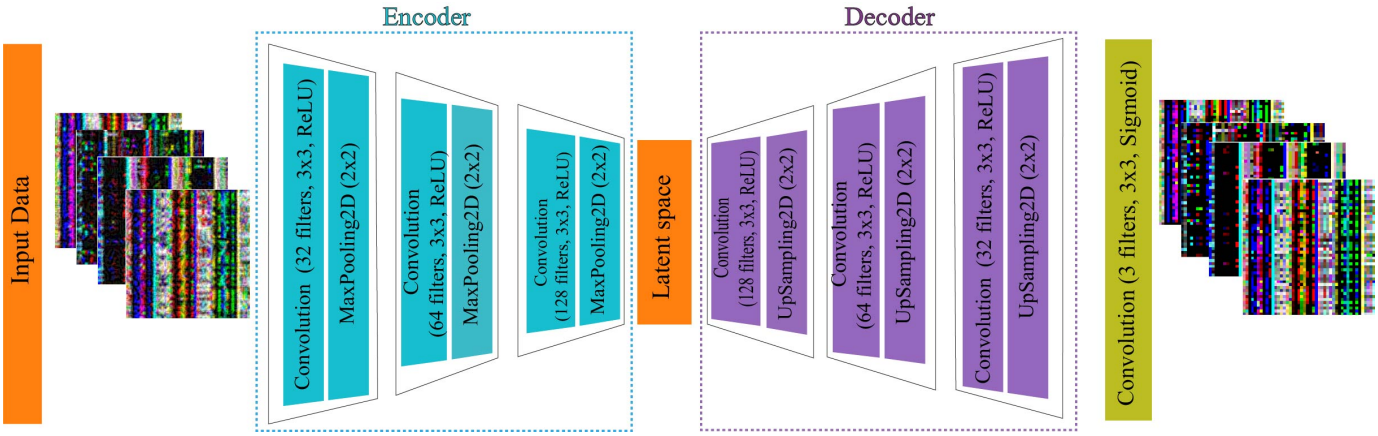


Fig. 3: Convolutional Autoencoder Denoising Architecture.

- **Precision (Pre)**: measures how many of the samples predicted as a specific class are actually correct. For a given class:

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall (Rec)**: measures how many actual positive samples were correctly identified. For a given class:

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **F1-Score (F1S)**: The F1-Score is the harmonic mean of Precision and Recall, providing a balance between the two:

$$\text{F1S} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Attack Success Rate (ASR)**: measures the effectiveness of adversarial attacks by calculating the proportion of correctly classified clean samples that become misclassified after perturbation. For a test set of correctly classified samples, it is defined as:

$$\text{ASR} = \frac{\text{Number of misclassified adversarial samples}}{\text{Total number of correctly classified clean samples}}$$

- **Latency**: measures the end-to-end processing time required to classify a single input image. It includes all stages of the IDS pipeline: adversarial denoising, feature extraction, feature selection, and classification. For a given test set of size N, the average latency per image is computed as:

$$\text{Latency}_{avg} = \frac{\sum_{i=1}^{N} T_i}{N}$$

Where T_i denotes the total processing time of the i-th image.

IV. Results and Discussion

In this section, we evaluate the performance of the proposed IDS under various experimental scenarios. The evaluation considers both quantitative and qualitative measures to assess the robustness and effectiveness of our dual defense strategy. Specifically, we present standard classification metrics such as Accuracy, Precision, Recall, and F1-Score across all attack categories to quantify performance. Additionally, we utilize t-distributed Stochastic Neighbor Embedding (t-SNE) [35], a qualitative technique, to visualize high-dimensional feature representations, providing insight into class separability and the behavior of the feature space.

A. *IDS Evaluation on Numeric Traffic Transformed to Images*

*1) **Vulnerability Assessment and Adversarial Attack Success Rate**:* To evaluate the robustness of our IDS against adversarial attacks, we conduct a baseline vulnerability assessment using FGSM, FGM, and PGD. This assessment measures the IDS's performance on clean and adversarial images, focusing on samples correctly classified by the baseline IDS. We calculate the ASR, alongside standard metrics. The evaluation was performed under three different perturbation strengths: $\varepsilon = 0.01$, $\varepsilon = 0.1$, and $\varepsilon = 0.3$, to assess the effect of weak, moderate, and strong adversarial perturbations.

TABLE IV: Model performance and attack success rate after adversarial attacks.

Attack	Acc%	Pre%	Rec%	F1S%	ASR-Acc%	ASR-Pre%	ASR-Rec%	ASR-F1S%
Before Adversarial Attack								
NONE	99.41	99.41	99.41	99.41	–	–	–	–
After Adversarial Attack					**Attack Success Rate**			
$\varepsilon = 0.01$								
FGM	98.73	98.75	98.73	98.73	1.27	1.25	1.27	1.27
FGSM	98.99	99.02	98.99	99.00	1.01	0.98	1.01	1.00
PGD	99.04	99.08	99.04	99.04	0.96	0.92	0.96	0.96
$\varepsilon = 0.1$								
FGM	42.00	59.00	42.00	35.00	58.00	41.00	58.00	65.00
FGSM	40.00	59.00	40.00	32.00	60.00	41.00	60.00	68.00
PGD	47.00	66.00	47.00	41.00	53.00	34.00	53.00	59.00
$\varepsilon = 0.3$								
FGM	26.28	39.38	26.28	21.03	73.72	60.62	73.72	78.97
FGSM	17.55	34.17	17.55	10.91	82.45	65.83	82.45	89.09
PGD	50.69	66.97	50.69	43.93	49.31	33.03	49.31	56.07

Table IV presents the performance of the IDS before and after adversarial attacks, highlighting the significant impact of FGM, FGSM, and PGD on the system's effectiveness. Before attacks, the IDS achieves an impressive 99.41% across all metrics, indicating near-perfect classification of clean samples.

However, post-attack performance strongly depends on the strength of the perturbation (ε). For **small perturbations** ($\varepsilon = 0.01$), the system remains largely unaffected, with ASR values below 2%, indicating that the IDS maintains almost the same performance as on clean samples.

For **moderate perturbations** ($\varepsilon = 0.1$), performance is substantially degraded. Accuracy drops to **42.00%, 40.00%, and 47.00%** under FGM, FGSM, and PGD, respectively. This decline corresponds to high ASR values, where a large proportion of previously correct predictions become incorrect: **60% for FGSM**, **58% for FGM**, and **53% for PGD**.

For **strong perturbations** ($\varepsilon = 0.3$), the IDS becomes highly vulnerable. Accuracy falls drastically to **26.28% (FGM)**, **17.55% (FGSM)**, and **50.69% (PGD)**. These severe drops indicate near-complete model deception, with ASR values reaching **73.72%, 82.45%, and 49.31%**, respectively. Such a high ASR reflects the ability of strong perturbations, especially FGSM and FGM, to overwhelm the model's characteristic representations, causing widespread misclassification even in previously well-classified samples. These results underscore the vulnerability of the IDS to adversarial attacks, particularly in misclassifying attack types under higher perturbations, and highlight the critical need for robust defense strategies to restore classification performance in smart agriculture environments.

*2) **Effectiveness of the first Defense: Adversarial Training**:* Adversarial training is a defense strategy that improves the robustness of our IDS by incorporating adversarial examples into the training process. This approach aims to reduce ASR by improving the model's ability to correctly classify perturbed input.

TABLE V: Class-wise performance of the IDS after applying adversarial training against FGM, FGSM, and PGD attacks for different ε values.

Attack	Class	Acc	Pre	Rec	F1S	Acc	Pre	Rec	F1S	Acc	Pre	Rec	F1S
		ε = 0.01				ε = 0.1				ε = 0.3			
FGM	Recon	99.38	99.58	99.38	99.48	98.97	98.37	98.97	98.67	98.45	97.65	98.45	98.05
	Spoofing	100	99.70	100	99.85	99.62	98.24	99.62	98.93	98.14	97.85	98.14	97.99
	Web_Based	99.60	98.43	99.60	99.01	99.01	98.52	99.01	98.76	98.81	97.07	98.81	97.93
	Brute Force	96.80	97.97	96.80	97.38	89.11	93.13	89.11	91.08	84.20	85.48	84.20	84.83
	Benign	97.84	97.94	97.84	97.89	92.95	91.38	92.95	92.16	86.10	87.38	86.10	86.74
	DDoS	100	100	100	100	100	99.78	100	99.89	99.81	99.53	99.81	99.67
	DoS	100	100	100	100	99.81	100	99.81	99.90	99.60	99.80	99.60	99.70
	Mirai	100	100	100	100	100	100	100	100	100	100	100	100
	Macro_avg	99.20	99.20	99.20	99.20	97.43	97.43	97.43	97.42	95.64	95.59	95.64	95.61
FGSM	Recon	99.17	99.17	99.17	99.17	98.83	98.35	98.83	98.59	99.26	98.65	99.26	98.95
	Spoofing	100	99.80	100	99.90	99.59	98.29	99.59	98.94	99.08	97.89	99.08	98.48
	Web_Based	99.40	98.04	99.40	98.71	99.58	98.87	99.58	99.23	99.45	98.50	99.45	98.97
	Brute Force	95.60	96.17	95.60	95.88	90.65	92.03	90.65	91.33	89.09	92.51	89.09	90.77
	Benign	96.08	97.03	96.08	96.55	91.30	92.11	91.30	91.71	92.99	92.24	92.99	92.61
	DDoS	100	100	100	100	99.81	99.90	99.81	99.86	99.28	99.88	99.28	99.58
	DoS	100	100	100	100	99.90	99.80	99.90	99.85	99.87	99.27	99.87	99.57
	Mirai	100	100	100	100	100	100	100	100	100	100	100	100
	Macro_avg	98.78	98.78	98.78	98.78	97.46	97.42	97.46	97.44	97.38	97.37	97.38	97.37
PGD	Recon	100	99.23	100	99.61	98.80	99.10	98.80	98.95	98.69	97.99	98.69	98.34
	Spoofing	100	99.62	100	99.81	99.02	97.68	99.02	98.35	99.36	98.19	99.36	98.77
	Web_Based	99.09	99.09	99.09	99.09	99.07	98.86	99.07	98.96	98.51	98.79	98.51	98.65
	Brute Force	97.17	96.79	97.17	96.98	88.13	90.50	88.13	89.30	88.77	91.72	88.77	90.22
	Benign	96.56	98.10	96.56	97.33	90.61	89.34	90.61	89.97	90.84	89.28	90.84	90.06
	DDoS	100	100	100	100	99.37	99.68	99.37	99.53	99.90	100	99.90	99.95
	DoS	100	100	100	100	99.71	99.42	99.71	99.56	100	99.90	100	99.95
	Mirai	100	100	100	100	100	100	100	100	100	100	100	100
	Macro_avg	99.10	99.10	99.10	99.10	96.84	96.82	96.84	96.83	97.01	96.98	97.01	96.99

Table V demonstrates the effectiveness of adversarial training in strengthening the IDS against FGM, FGSM, and PGD attacks across different perturbation levels. Compared to the baseline performance without any defense mechanism (Table IV), which exhibited a substantial degradation under adversarial perturbations, adversarial training significantly enhances robustness.

For **small perturbations** ($\varepsilon = 0.01$), the system remains unaffected mainly, macro-averaged metrics stay very high, and ASR is below 2%, indicating near baseline performance.

For **moderate perturbations** ($\varepsilon = 0.1$), adversarial training yields substantial recovery. Macro-averaged Accuracy improves to **97.43%** (FGM), **97.46%** (FGSM), and **96.84%** (PGD), corresponding to a significant reduction in ASR.

For **strong perturbations** ($\varepsilon = 0.3$), adversarial training continues to provide meaningful protection, macro-averaged Accuracy remains high at **95.64%** (FGM), **97.38%** (FGSM), and **97.01%** (PGD), despite the highly challenging adversarial conditions.

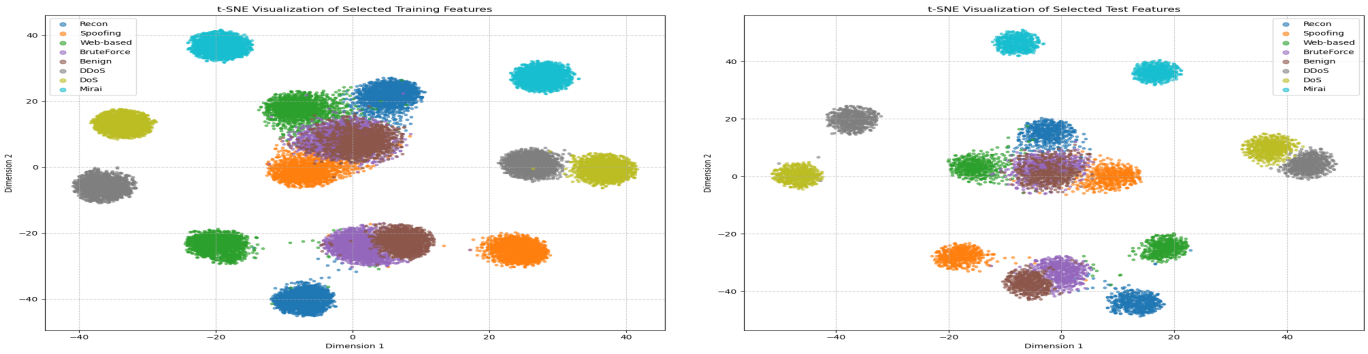
On a class-wise level, the IDS performs exceptionally well on DDoS, DoS, and Mirai samples across all ε values, consistently achieving near-perfect scores. However, Brute Force and Benign classes show comparatively lower performance, with Accuracy ranging between **84.20%** and **92.16%** depending on the attack and perturbation strength. This underperformance is likely linked to intrinsic feature overlap, as illustrated in the t-SNE plots (4(a) for training and 4(b) for test data), where Brute Force and Benign clusters appear heavily intertwined. In contrast, classes such as Recon, Mirai, and DoS remain well separated.

Adversarial perturbations exploit this natural overlap by shifting samples across ambiguous decision boundaries, creating misclassifications that adversarial training mitigates but does not eliminate. These results highlight both the strengths and limitations of adversarial training.

*3) **Effectiveness of the second Defense: Convolutional Autoencoder Denoising**:* The Convolutional Autoencoder Denoising defense enhances the robustness of the IDS by preprocessing adversarial images to remove perturbations introduced by adversarial attacks. The IDS is trained exclusively on original (clean) data, and its performance is evaluated on both original and adversarial test data. Testing on original data ensures that the denoising step does not degrade performance on clean data, while testing on adversarial data assesses the CAED's ability to restore perturbed images to a classifiable state.

Table VI demonstrates the effectiveness of the CAED defense in enhancing the robustness of the baseline IDS against adversarial attacks. On clean data, the IDS achieves near-perfect performance, with macro-averaged Accuracy, Precision, Recall, and F1-Score ranging from 99.56% to 99.60%, ensuring that the denoising step does not compromise performance on the original data.

For small perturbations ($\varepsilon = 0.01$) and moderate perturbations ($\varepsilon = 0.1$), the autoencoder can effectively remove most adversarial noise, allowing the IDS to preserve strong classification performance across all classes. The Brute Force and Benign classes also show substantial improvement compared to adversarial training alone. This improvement is attributed to the autoencoder's denoising capabilities, as visualized in the t-SNE plots. Figure 5(b) shows distinct and well-separated clusters for the denoised adversarial test data. In contrast, Figure 5(a) presents cleaner and more compact training clusters, with reduced overlap between Benign and Brute Force samples. On the other hand, the plots from the



(a) Train data (b) Test data

Fig. 4: t-SNE visualization of sample distributions based on selected features after applying the adversarial training mechanism.

TABLE VI: Class-wise performance of the IDS after applying the CAED defense against FGM, FGSM, and PGD attacks.

Attack	Classes	Clean Data				Adversarial ($\varepsilon = 0.01$)				Adversarial ($\varepsilon = 0.1$)				Adversarial ($\varepsilon = 0.3$)			
		Acc%	**Pre%**	**Rec%**	**F1S%**	**Acc%**	**Pre%**	**Rec%**	**F1S%**	**Acc%**	**Pre%**	**Rec%**	**F1S%**	**Acc%**	**Pre%**	**Rec%**	**F1S%**
FGM	Recon	99.80	99.39	99.80	99.59	99.79	99.39	99.79	99.59	96.11	99.79	96.11	97.91	0	0	0	0
	Spoofing	100	99.81	100	99.91	100	99.42	100	99.71	100	99.44	100	99.72	94.13	59.83	94.13	73.16
	Web_Based	99.21	99.60	99.21	99.40	99.40	99.60	99.40	99.50	99.41	100	99.41	99.70	12.30	100	12.30	21.91
	Brute Force	99.02	98.44	99.02	98.73	99.00	99.00	99.00	99.41	97.45	92.55	97.45	94.94	63.40	20.99	63.40	31.54
	Benign	98.48	99.23	98.48	98.85	99.02	99.80	99.02	99.41	96.19	97.87	96.19	97.06	38.16	57.35	38.16	45.83
	DDoS	100	100	100	100	100	100	100	100	100	99.58	100	99.79	94.44	61.82	94.44	74.73
	DoS	100	100	100	100	100	100	100	100	99.64	100	96.64	99.82	33.33	92.27	48.97	99.50
	Mirai	100	100	100	100	100	100	100	100	100	100	100	100	67.93	100	67.93	80.90
	Macro_avg	99.56	99.56	99.56	99.56	99.65	99.65	99.65	99.65	98.60	98.65	98.60	98.61	50.46	61.53	50.46	47.13
FGSM	Recon	100	100	100	100	99.79	99.79	99.79	99.79	96.12	100	96.12	98.02	0	0	0	0
	Spoofing	100	99.60	100	99.80	100	99.80	100	99.90	100	99.80	100	99.90	95.62	46.90	95.62	62.93
	Web_Based	99.79	99.79	99.79	99.79	99.80	99.80	99.80	99.80	98.97	99.18	98.97	99.07	4.71	100	4.71	9
	Brute Force	99.07	97.79	99.07	98.42	99.60	99.60	99.60	99.60	98.32	91.16	98.32	94.60	61.40	22.19	61.40	32.60
	Benign	97.85	99.60	97.85	98.72	99.61	99.80	99.61	99.71	93.55	97.76	93.55	95.61	41.43	47.63	41.43	44.31
	DDoS	100	100	100	100	100	100	100	100	100	99.82	100	99.91	73.63	65.02	73.63	69.05
	DoS	100	100	100	100	100	100	100	100	99.80	100	99.80	99.90	44.39	89.66	44.39	59.38
	Mirai	100	100	100	100	100	100	100	100	100	100	100	100	68.35	100	68.35	81.20
	Macro_avg	99.59	99.60	99.59	99.59	99.85	99.85	99.85	99.85	98.35	98.46	98.35	98.38	48.69	58.92	48.69	44.81
PGD	Recon	99.80	100	99.80	99.90	99.61	99.61	99.61	99.61	99.40	100	99.40	99.70	99.02	100	99.02	99.51
	Spoofing	100	99.80	100	99.90	100	99.80	100	99.90	100	99.61	100	99.80	100	100	100	100
	Web_Based	100	99.59	100	99.79	99.60	99.40	99.60	99.50	100	99.38	100	99.69	99.40	99.60	99.40	99.50
	Brute Force	98.50	98.32	98.50	98.41	99.42	99.22	99.42	99.32	98.50	97.77	98.50	98.14	99.22	97.51	99.22	98.36
	Benign	98.29	98.85	98.29	98.57	99.26	99.81	99.26	99.54	97.91	99.04	97.91	98.47	98.89	99.44	98.89	99.17
	DDoS	100	100	100	100	100	100	100	100	100	100	100	100	99.80	100	99.80	99.90
	DoS	100	100	100	100	100	100	100	100	100	100	100	100	100	99.80	100	99.90
	Mirai	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Macro_avg	99.57	99.57	99.57	99.57	99.74	99.73	99.74	99.73	99.48	99.48	99.48	99.48	99.54	99.54	99.54	99.54

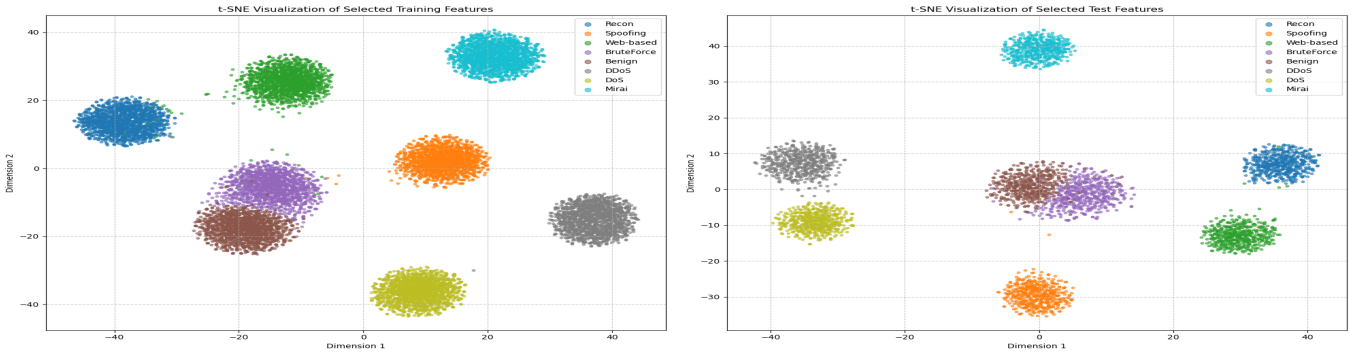
undefended system (Figure 4) exhibit significant entanglement of sample distributions based on the selected features.

For strong perturbations ($\varepsilon = 0.3$), the adversarial distortion becomes highly destructive, severely degrading the visual structure of the images. After applying the CAED defense, accuracy improves to 50.46% for FGM, 48.49% for FGSM, and 99.54% for PGD. Although this recovery demonstrates that the autoencoder can partially restore severely corrupted images, especially for PGD, it remains insufficient for highly damaged inputs. The magnitude of the perturbation at $\varepsilon = 0.3$ overwhelms the reconstruction capacity of the autoencoder, meaning that CAED alone cannot fully counteract such large-scale distortion.

*4) **Performance of the Merged Dual Defense:*** The merged dual defense strategy, combining the Convolutional Autoencoder Denoising and adversarial training, enhances the robustness of the IDS across all tested perturbation levels by leveraging complementary strengths to mitigate adversarial perturbations.

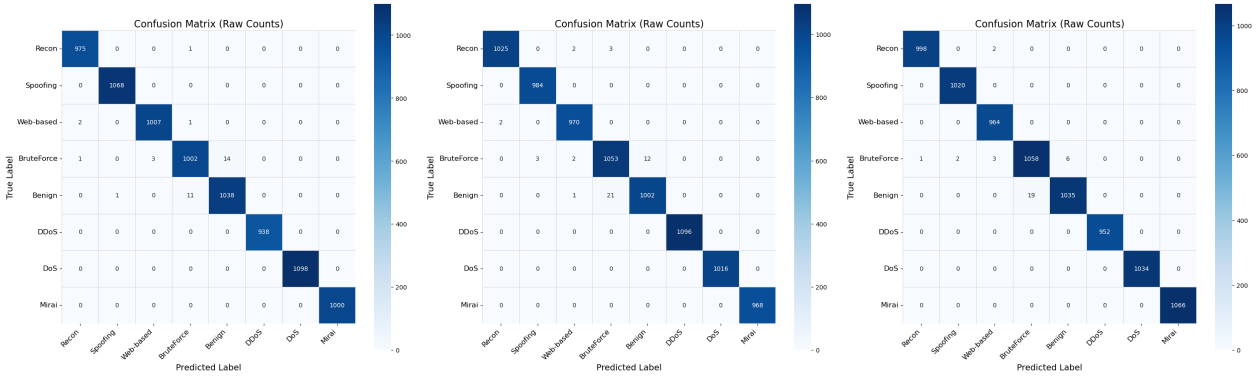
TABLE VII: Class-wise performance of the IDS after applying the dual defense strategy against FGM, FGSM, and PGD attacks for different ε values.

Attack	Class	Acc%	Pre%	Rec%	F1S%	Acc %	Pre%	Rec%	F1S%	Acc%	Pre%	Rec%	F1S%
		$\varepsilon = 0.01$				$\varepsilon = 0.1$				$\varepsilon = 0.3$			
FGM	Recon	99.60	100	99.60	99.80	99.90	99.69	99.90	99.80	98.46	99.07	98.46	98.76
	Spoofing	100	100	100	100	100	99.91	100	99.95	99.32	97.13	99.32	98.21
	Web_Based	100	100	100	100	99.70	99.70	99.70	99.70	98.31	96.59	98.31	97.44
	Brute Force	99.70	99.51	99.70	99.61	98.24	98.72	98.24	98.48	86.30	89.80	86.30	88.02
	Benign	99.53	99.81	99.53	99.67	98.86	98.67	98.86	98.76	90.61	90.25	90.61	90.43
	DDoS	100	100	100	100	100	100	100	100	99.54	99.54	99.54	99.54
	DoS	100	100	100	100	100	100	100	100	99.50	99.50	99.50	99.50
	Mirai	100	100	100	100	100	100	100	100	100	100	100	100
	Macro_avg	99.85	99.85	99.85	99.85	99.59	99.59	99.59	99.59	96.50	96.48	96.50	96.54
FGSM	Recon	99.80	100	99.80	99.90	99.51	99.81	99.51	99.66	99.63	99.63	99.63	99.63
	Spoofing	100	99.80	100	99.90	100	99.70	100	99.85	100	99.64	100	99.82
	Web_Based	100	99.20	100	99.60	99.70	99.49	99.70	99.64	99.64	99.52	99.64	99.58
	Brute Force	98.25	97.49	98.25	97.87	98.24	97.77	98.24	98.09	98.75	98.50	98.75	98.62
	Benign	97.60	99.06	97.60	98.33	98.86	98.82	98.86	98.33	98.50	99.24	98.50	98.87
	DDoS	100	100	100	100	100	100	100	100	100	100	100	100
	DoS	100	100	100	100	100	100	100	100	100	100	100	100
	Mirai	100	100	100	100	100	100	100	100	100	100	100	100
	Macro_avg	99.46	99.44	99.46	99.45	99.45	99.45	99.45	99.45	99.56	99.57	99.56	99.56
PGD	Recon	99.71	99.80	99.71	99.75	99.80	99.90	99.80	99.85	99.80	99.51	99.80	99.66
	Spoofing	100	99.80	100	99.90	100	99.80	100	99.90	100	99.70	100	99.85
	Web_Based	99.70	99.30	99.70	99.50	100	99.48	100	99.74	99.30	99.60	99.30	99.45
	Brute Force	98.35	98.92	98.35	98.63	98.88	98.24	98.88	98.56	99.12	98.45	99.12	98.79
	Benign	99.08	98.99	99.08	99.03	98.20	98.42	98.20	98.81	98.71	99.63	98.71	99.17
	DDoS	100	100	100	100	100	100	100	100	100	100	100	100
	DoS	100	100	100	100	100	100	100	100	100	100	100	100
	Mirai	100	100	100	100	100	100	100	100	100	100	100	100
	Macro_avg	99.60	99.60	99.60	99.60	99.61	99.61	99.61	99.61	99.62	99.61	99.62	99.61



(a) Train data (b) Test data

Fig. 5: t-SNE visualization of sample distributions based on selected features after applying the CAED mechanism on adversarial data.



(a) FGM (b) FGSM (c) PGD

Fig. 6: Confusion Matrix.

Table VII and Figure 6 illustrate the performance of the merged dual defense strategy. In particular, the combined defense does not degrade baseline performance. Against FGM, FGSM, and PGD attacks, it achieves consistently high performance across all evaluation metrics, representing a significant improvement over the undefended baseline and a slight advantage over using adversarial training or CAED alone. For $\varepsilon = 0.3$, it still maintains high accuracy, proving its resilience even under intense adversarial noise. This performance corresponds to a substantial reduction in ASR, reflecting the system's ability to restore adversarial examples to a classifiable state.

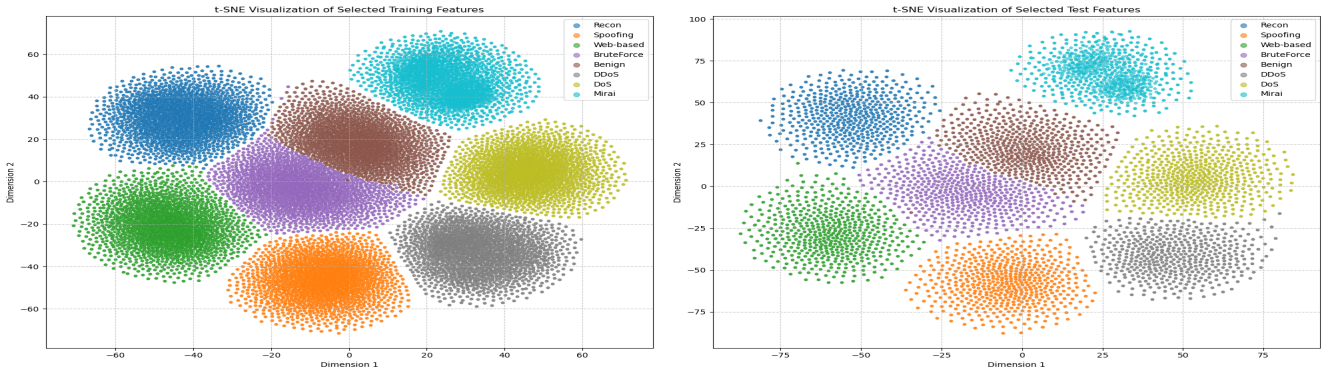
On a class-wise level, DDoS, DoS, and Mirai maintain perfect scores, while the Brute Force and Benign classes also show marked improvement, surpassing the results achieved with individual defenses. These gains are further supported by the t-SNE visualizations: Figure 7(a) displays compact, well-separated clusters for training data after training on a combined dataset of clean and adversarial images. In contrast, Figure 7(b) shows similarly tight clusters for test data with minimal to no overlap. Compared to the plots of the undefended system, the dual defense exhibits a synergistic effect, mapping both clean and perturbed inputs to a highly discriminative feature space. This reduction in feature confusion correlates with higher classification accuracy and improved robustness.

*5) **Summary**:* This section provides an overall synthesis of the performance of the VGG16-BGWO-RF IDS under different adversarial scenarios and defense strategies.

Table VIII presents a consolidated view of the baseline performance, the impact of adversarial attacks, and the effectiveness of the proposed defenses for increasing perturbation strengths ($\varepsilon = 0.01, 0.1, 0.3$). While all attacks progressively degrade the IDS performance as ε increases, the severity of this degradation varies across attack types.

Adversarial training (Defense1) consistently restores robustness under moderate and strong perturbations, whereas the convolutional autoencoder denoising (Defense2) is particularly effective for weak and moderate perturbations but shows limited recovery under severe distortion. In contrast, the merged dual defense achieves the most stable and reliable performance across all attacks and perturbation levels, maintaining near-baseline accuracy even at $\varepsilon = 0.3$.

Figure 8 visually reinforces these results by presenting a line plot that compares accuracy across different attack scenarios and defense mechanisms for varying perturbation values. The trend clearly shows the sharp decline in performance due to adversarial perturbations and the subsequent recovery achieved through each defense, with the merged defense consistently achieving the highest robustness.



(a) Train data (b) Test data

Fig. 7: t-SNE visualization of sample distributions based on selected features after applying the dual defense strategy on combined data.

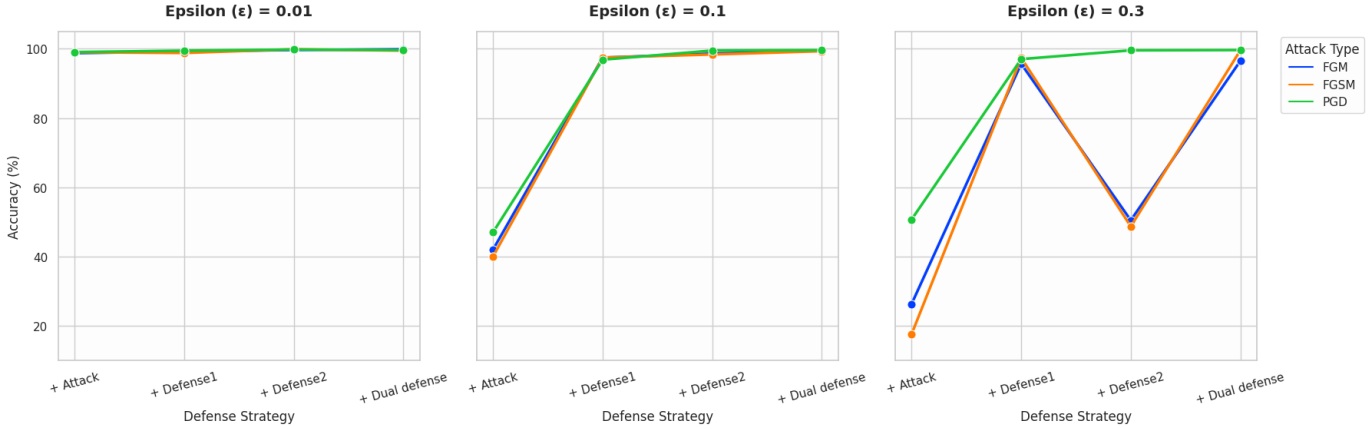


Fig. 8: Accuracy comparison across Attacks and Defenses.

TABLE VIII: Performance metrics per case for different ε values.

Attack	Case	Acc	Pre	Rec	F1S	Acc	Pre	Rec	F1S	Acc	Pre	Rec	F1S
		$\varepsilon = 0.01$				$\varepsilon = 0.1$				$\varepsilon = 0.3$			
	Baseline IDS	99.41	99.41	99.41	99.41	99.41	99.41	99.41	99.41	99.41	99.41	99.41	99.41
FGM	+ Attack	98.73	98.75	98.73	98.73	42.00	59.00	42.00	35.00	25.28	39.38	26.28	21.03
	+ Defense1	99.20	29.20	99.20	99.20	97.43	97.43	97.43	97.42	95.64	95.59	95.64	95.61
	+ Defense2	99.65	99.65	99.65	99.65	98.60	98.65	98.60	98.61	50.46	61.53	50.46	47.13
	+ Dual Defense	99.85	99.85	99.85	99.85	99.59	99.59	99.59	99.59	96.50	96.48	96.50	96.54
FGSM	+ Attack	98.99	99.02	98.99	99.00	40.00	59.00	40.00	32.00	17.55	34.17	17.55	10.91
	+ Defense1	98.78	98.78	98.78	98.78	97.46	97.42	97.46	97.44	97.38	97.38	97.38	97.37
	+ Defense2	99.85	99.85	99.85	99.85	98.35	98.46	98.35	98.38	48.69	58.92	48.69	44.81
	+ Dual Defense	99.46	99.44	99.46	99.45	99.45	99.45	99.45	99.45	99.56	99.57	99.56	99.56
PGD	+ Attack	99.04	99.08	99.04	99.04	47.00	66.00	47.00	41.00	50.69	66.97	50.69	43.93
	+ Defense1	99.45	99.44	99.45	99.44	96.84	96.82	96.84	96.83	97.01	96.98	97.01	96.99
	+ Defense2	99.78	99.73	99.74	99.73	99.48	99.48	99.48	99.48	99.54	99.54	99.54	99.54
	+ Dual Defense	99.60	99.60	99.60	99.60	99.61	99.61	99.61	99.61	99.62	99.61	99.62	99.61

TABLE IX: Class-wise performance results of the multimodal IDS.

Class	Acc%	Pre%	Rec%	F1S%
Recon	100	99.79	100	99.89
Spoofing	100	99.62	100	99.81
Web_Based	99.79	99.79	99.79	99.79
Brute Force	99.61	98.85	99.61	99.23
Benign	98.52	100	98.52	99.25
DDoS	100	100	100	100
DoS	100	100	100	100
Mirai	100	100	100	100
Normal_Crope	100	100	100	100
Macro_avg	99.77	99.78	99.79	99.77

B. Multimodal IDS Evaluation: Numeric Traffic and Native Images

Table IX presents the class-wise performance results of the proposed multimodal IDS, which integrates both network traffic images and native agricultural images. The added class, Normal_Crop, represents real crop images and reflects the smart agriculture domain. As shown, the system achieves very high performance across all classes, with accuracy, precision, recall, and F1-score.

The macro-averaged metrics indicate that the IDS maintains robust performance when handling both numeric traffic and native agricultural images,

1) Vulnerability Assessment and Adversarial Attack Success Rate: To evaluate the robustness of our multimodal IDS against adversarial attacks, we conducted a baseline vulnerability assessment using FGM, FGSM, and PGD. Table X reports the system's performance on both clean and adversarial images, focusing on samples that were correctly classified by the baseline IDS. Standard metrics are presented alongside the ASR metric for each attack type.

For this evaluation, we selected ε **= 0.3**, as previous tests indicated that this value has the most significant impact on IDS performance. The results show that without defenses, the IDS's performance drops drastically under all attack types,

with ASR accuracy values ranging from 77.44% to 88.51%, highlighting the critical need for adversarial defense mechanisms. This assessment provides a clear baseline for evaluating the effectiveness of our proposed dual-defense strategy.

TABLE X: Multimodel IDS performance before and after adversarial attacks.

ART	Acc%	Pre%	Rec%	F1S%
Before Adversarial Attack				
NONE	99.77	99.78	99.77	99.79
After Adversarial Attack				
FGM	11.70	8.77	11.70	3.37
FGSM	11.47	20.87	11.47	2.97
PGD	22.56	33.40	22.56	17.41
Attack Success Rate				
FGM	88.30	91.23	88.30	96.63
FGSM	88.51	79.13	88.51	97.03
PGD	77.44	66.60	77.44	82.59

2) Performance of the Merged Dual Defense: Table XI presents the class-wise performance of the multimodal IDS after applying the merged dual defense strategy against FGM, FGSM, and PGD attacks. The dual defense significantly restores the system's performance compared to the unprotected scenario, demonstrating the effectiveness of the defense mechanism across all attack types.

The added class Normal_crop, representing native agricultural images, maintains perfect performance across all metrics, showing that the defense doesn't compromise the IDS's ability to handle domain-specific crop images. The macro-averaged metrics indicate that, after applying the dual defense, the system achieves high robustness, with Accuracy, Precision, Recall, and F1-Score above 96% for all attacks.

TABLE XI: Class-wise performance of the multimodel IDS after applying the dual defense strategy against FGM, FGSM, and PGD attacks .

Class	Acc%	Pre%	Rec%	F1S%	Acc%	Pre%	Rec%	F1S%	Acc%	Pre%	Rec%	F1S%
	FGM				FGSM				PGD			
Recon	98.96	99.58	98.96	99.27	99.07	99.07	99.07	99.07	99.90	100	99.90	99.95
Spoofing	98.69	97.78	98.69	99.23	98.50	96.79	98.50	97.64	100	99.91	100	99.95
Web_Based	99.28	98.36	99.38	87.88	98.96	97.45	98.96	98.20	100	99.90	100	99.95
Brute Force	85.77	90.10	85.77	87.88	85.48	90.44	85.48	87.89	99.71	98.48	99.71	99.09
Benign	90.32	86.93	90.32	88.59	89.79	87.67	89.79	88.72	98.32	99.79	98.32	99.05
DDoS	98.30	99.55	98.30	98.92	99.01	99.64	99.01	99.33	100	100	100	100
DoS	99.52	98.18	99.52	98.85	99.61	98.94	99.61	99.28	100	100	100	100
Mirai	100	100	100	100	100	100	100	100	100	100	100	100
Normal_crop	100	100	100	100	100	100	100	100	100	100	100	100
Macro_avg	96.76	96.72	96.76	96.73	96.71	96.67	96.71	96.68	99.77	99.79	99.77	99.78

C. Real-Time Applicability and Application Domain

To evaluate the real-time feasibility of the proposed defense-enhanced IDS, we measured the end-to-end latency for each processing stage, including adversarial denoising, VGG16 feature extraction, feature selection, and classification. As shown in Table XII, the denoising and feature extraction stage dominates latency, averaging 42.226 ms per image for clean samples, 42.306 ms for FGSM, 42.285 ms for PGD, and 22.039 ms for FGM samples. Feature selection and classification are negligible (between 0.071 and 0.079 ms per image).

The overall latency per image ranges from 22.118 to 42.385 ms, enabling a throughput of 24 to 45 images per second, suitable for real-time smart agriculture scenarios. Using processing time as a proxy for computational cost, these results demonstrate low overhead, ensuring both efficient real-time operation and robustness against adversarial attacks.

TABLE XII: Average per-image latency for multimodel IDS across clean and adversarial samples.

Processing Stage	Clean	FGSM	FGM	PGD
Denoising + VGG16 Extraction (ms)	42.226	42.306	**22.039**	42.285
Feature Selection + RF (ms)	0.079	0.071	0.079	0.075
Total Latency per Image (ms)	**42.305**	**42.385**	**22.118**	**42.364**

In practical deployment, the IDS operates as a preprocessing and protection layer before the main surveillance system. Cleans and denoises incoming images, ensuring more precise visual data even in critical situations such as wildfires, and protects the infrastructure from adversarial and network-based attacks. This dual role improves both image reliability and cybersecurity in precision agriculture.

Due to its low latency and efficient processing, the multimodel IDS can respond rapidly in emergency scenarios, enabling real-time decision-making in high-risk agricultural environments. Consistent with the proposed three-tier smart agriculture architecture in [18], the IDS is deployed at the Fog layer, where latency-sensitive processing is required. Hosting the denoising, feature extraction, and classification modules at Fog nodes allows local analysis of both network traffic and agricultural images, thereby reducing communication overhead with the Cloud layer. Although the VGG16-based feature extraction and denoising stages constitute the main computational cost, the measured end-to-end latency remains within real-time constraints. Consequently, the Fog layer is well suited to host the proposed IDS using moderate computational resources, while the Cloud layer is reserved for long-term storage and global analytics.

V. Comparison with Related Work

This section compares the performance of the proposed VGG16-BGGO-RF IDS with related works on various datasets, focusing on resilience against adversarial attacks.

The work presented in [21] reports up to 93.5% accuracy under FGSM in a binary classification context. Khan et al. [22] achieve 98.8% accuracy with the HopSkipJump attack, but report lower precision (85.0%) and F1-Score (87.3%). Benyamin et al. [23] report accuracy between 94.0% and 96.3% using GA-based defenses. Paya et al. [24] show a wide range of accuracy (62.6%–93.04%) in different attacks. Bommana et al. [26] achieve an accuracy of 92% and 95%, although their evaluation does not involve specific adversarial attacks.

Compared to these binary-class setups, our multi-class approach demonstrates stronger resilience and significantly outperforms related works.

VI. Conclusion

This study presents a robust multimodel VGG16-BGGO-RF IDS that integrates native IoT images with network-traffic

TABLE XIII: Comparative Results with Related Work.

Reference	Dataset	Adversarial Attack	ML/DL Model	Classification Type	Acc (%)	Pre%)	Rec (%)	F1S (%)
[21]	Hybrid dataset	FGSM	RF	Binary	93.5	N/A	N/A	N/A
		BIM			93	N/A	N/A	N/A
		C&W			92	N/A	N/A	N/A
[22]	Powe System dataset	HopSkipJump	AE-RF	Binary	98.8	85.0	91.4	87.3
		ZOO			97.6	76.6	85.9	79.4
[23]	NSL-KDD	GA	MLP-LSTM	Binary	94.0	93.3	96.9	N/A
	UNSW-NB15				96.3	98.1	95.6	N/A
[24]	CIC-IDS-2017	HopSkipJump	LR, RF, DT, SVM,NB, MLP, FNN	Binary	75.5	N/A	52.6	56.07
		ZOO			93.04	N/A	87.72	88.35
		W-GAN			62.60	N/A	42.50	45.95
[26]	IoT/IIoT	N/A	RCNN-RBM	Binary	92	89	N/A	91
	T-IoT2022				95	92	N/A	95
Our	CICIoT2023	FGM	CAE-VGG16-RF	Multi-class	99.59	99.59	99.59	99.59
		FGSM			99.45	99.45	99.45	99.45
		PGD			99.61	99.61	99.61	99.61

data transformed into image representations, enabling unified image-based analysis across heterogeneous modalities. By converting packet-level numerical features into visual formats, the proposed IDS leverages convolutional feature extraction and transfer learning to identify complex attack patterns in smart agriculture environments.

The work focuses primarily on the system's vulnerability to adversarial attacks by generating adversarial images and evaluating the IDS's resilience. The system's vulnerability to adversarial manipulation was extensively evaluated using adversarial images generated under three perturbation strengths (ε=0.01,0.1,0.3) with FGM, FGSM, and PGD attacks. Tested on both the CICIoT2023 dataset and the crop image dataset, initial results confirmed that the IDS, although highly accurate on clean data, remained sensitive to adversarial perturbations, particularly at higher epsilon values. To mitigate these vulnerabilities, we introduced a dual defense strategy that combines CAED with Adversarial Training. Experiments show that the merged defense substantially improves robustness across all attack types and perturbation intensities. For weak perturbations (ε=0.01), the system maintains nearly perfect accuracy, indicating strong resistance to low-intensity attacks. For moderate perturbations (ε=0.1), the combined defense restores performance to high levels across FGM, FGSM, and PGD. Even under stronger adversarial distortion (ε=0.3), where attacks significantly damage the visual structure, the dual defense continues to improve classification performance, demonstrating substantial resilience despite severe image degradation. These results outperform both the undefended baseline and individual defense mechanisms, substantially reducing ASR. The CAED module demonstrates strong perturbation removal independent of epsilon values. In parallel, adversarial training enhances feature robustness, enabling the system to maintain high classification performance under adversarial conditions. Our multi-class, image-based approach is highly effective, offering a promising, scalable solution for smart agriculture security.

Despite the strong performance of the proposed IDS, several limitations remain. The current study focuses on a limited set of adversarial attacks (FGM, FGSM, and PGD), and results may not generalize to all possible attack types. Additionally, while the system demonstrates low per image latency, a comprehensive evaluation in large-scale real-time deployment scenarios is still required.

Future work may focus on real-time deployment, defense against a broader range of attack methods, and performance optimization to address remaining minor vulnerabilities.

ACKNOWLEDGMENT

This research work is supported by the PHC-Tassili Grant Number PHC 24MDU109, Code Campus France: 51361UE.

REFERENCES

[1] Kushagra Sharma and Shiv Kumar Shivandu. Integrating artificial intelligence and internet of things (iot) for enhanced crop monitoring and management in precision agriculture. *Sensors International*, page 100292, 2024.

[2] Vijaya Choudhary, Paramita Guha, Giovanni Pau, and Sunita Mishra. An overview of smart agriculture using internet of things (iot) and web services. *Environmental and Sustainability Indicators*, page 100607, 2025.

[3] Guma Ali, Maad M Mijwil, Bosco Apparatus Buruga, Mostafa Abotaleb, and Ioannis Adamopoulos. A survey on artificial intelligence in cybersecurity for smart agriculture: State-of-the-art, cyber threats, artificial intelligence applications, and ethical concerns. *Mesopotamian Journal of Computer Science*, 2024:53–103, 2024.

[4] Amina Khacha, Zibouda Aliouat, Yasmine Harbi, Chirihane Gherbi, Rafika Saadouni, and Saad Harous. Landscape of learning techniques for intrusion detection system in iot: A systematic literature review. *Computers and Electrical Engineering*, 120:109725, 2024.

[5] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[6] Octavio Loyola-Gonzalez. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access*, 7:154096–154113, 2019.

[7] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International conference on machine learning*, pages 2484–2493. PMLR, 2019.

[8] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[9] Yun Xiang, Yongchao Xu, Yingjie Li, Wen Ma, Qi Xuan, and Yi Liu. Side-channel gray-box attack for dnns. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 68(1):501–505, 2020.

[10] Rafika Saadouni, Amina Khacha, Chirihane Gherbi, Yasmine Harbi, Zibouda Aliouat, and Saad Harous. Pso-based feature selection for ddos detection using machine learning in smart agriculture. In *2024 International Conference on Advances in Electrical and Communication Technologies (ICAECOT)*, pages 1–6. IEEE, 2024.

[11] Amina Khacha, Rafika Saadouni, Yasmine Harbi, Chirihane Gherbi, Saad Harous, and Zibouda Aliouat. Robust intrusion detection for iot networks: an integrated cnn-lstm-gru approach. In *2023 International Conference on Networking and Advanced Systems (ICNAS)*, pages 1–6. IEEE, 2023.

[12] Amina Khacha, Rafika Saadouni, Yasmine Harbi, Zibouda Aliouat, Chirihane Gherbi, and Saad Harous. Privacy-preserving in agricultural iot: Intrusion detection using federated learning and cnn. In *2024 International Conference on Advanced Aspects of Software Engineering (ICAASE)*, pages 1–7. IEEE, 2024.

[13] Rafika Saadouni, Amina Khacha, Yasmine Harbi, Chirihane Gherbi, Saad Harous, and Zibouda Aliouat. Secure iiot networks with hybrid cnn-gru model using edge-iiotset. In *2023 15th International Conference on Innovations in Information Technology (IIT)*, pages 150–155. IEEE, 2023.

[14] Fatimah Aloraini, Amir Javed, and Omer Rana. Adversarial attacks on intrusion detection systems in in-vehicle networks of connected and autonomous vehicles. *Sensors*, 24(12):3848, 2024.

[15] Md Ahsan Ayub, William A Johnson, Douglas A Talbert, and Ambareen Siraj. Model evasion attack on intrusion detection systems using adversarial machine learning. In *2020 54th annual conference on information sciences and systems (CISS)*, pages 1–6. IEEE, 2020.

[16] Khushnaseeb Roshan, Aasim Zafar, and Shiekh Burhan Ul Haque. Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system. *Computer Communications*, 218:97–113, 2024.

[17] João Costa, Filipe Apolinário, and Carlos Ribeiro. Argan-ids: Adversarial resistant intrusion detection systems using generative adversarial networks. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, pages 1–10, 2024.

[18] Rafika Saadouni, Chirihane Gherbi, Zibouda Aliouat, Yasmine Harbi, Amina Khacha, and Hakim Mabed. Securing smart agriculture networks using bio-inspired feature selection and transfer learning for effective image-based intrusion detection. *Internet of Things*, 29:101422, 2025.

[19] El-Sayed M El-Kenawy, Nima Khodadadi, Seyedali Mirjalili, Abdelaziz A Abdelhamid, Marwa M Eid, and Abdelhameed Ibrahim. Greylag goose optimization: nature-inspired optimization algorithm. *Expert Systems with Applications*, 238:122147, 2024.

[20] Euclides Carlos Pinto Neto, Sajjad Dadkhah, Raphael Ferreira, Alireza Zohourian, Rongxing Lu, and Ali A Ghorbani. Ciciot2023: A real-time dataset and benchmark for large-scale attacks in iot environment. *Sensors*, 23(13):5941, 2023.

[21] Tarek Gaber, Tarek Ali, Mathew Nicho, and Mohamed Torky. Robust attacks detection model for internet of flying things based on generative adversarial network (gan) and adversarial training. *IEEE Internet of Things Journal*, 2025.

[22] Irfan Khan, Syed Wali, and Yasir Ali Farrukh. Radiant: Reactive autoencoder defense for industrial adversarial network threats. *Computers & Security*, 154:104403, 2025.

[23] Benyamin Tafreshian and Shengzhi Zhang. A defensive framework against adversarial attacks on machine learning-based network intrusion detection systems. In *2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 2436–2441. IEEE, 2024.

[24] Antonio Paya, Sergio Arroni, Vicente García-Díaz, and Alberto Gómez. Apollon: a robust defense system against adversarial machine learning attacks in intrusion detection systems. *Computers & Security*, 136:103546, 2024.

[25] Ying-Dar Lin, Wei-Hsiang Chan, Yuan-Cheng Lai, Chia-Mu Yu, Yu-Sung Wu, and Wei-Bin Lee. Enhancing can security with ml-based ids: Strategies and efficacies against adversarial attacks. *Computers & Security*, page 104322, 2025.

[26] Sesibhushana Rao Bommana, Sreehari Veeramachaneni, Syed Ershad, and MB Srinivas. Addressing adversarial attacks in iot using deep learning ai models. *IEEE Access*, 2025.

[27] Khushnaseeb Roshan and Aasim Zafar. Black-box adversarial transferability: An empirical study in cybersecurity perspective. *Computers & Security*, 141:103853, 2024.

[28] Jielong Yang, Rui Ding, Jianyu Chen, Xionghu Zhong, Huarong Zhao, and Linbo Xie. Black-box attacks on graph neural networks via white-box methods with performance guarantees. *IEEE Internet of Things Journal*, 11(10):18193–18204, 2024.

[29] Mingyi Zhou, Xiang Gao, Jing Wu, Kui Liu, Hailong Sun, and Li Li. Investigating white-box attacks for on-device models. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–12, 2024.

[30] Yujie Liu, Shuai Mao, Xiang Mei, Tao Yang, and Xuran Zhao. Sensitivity of adversarial perturbation in fast gradient sign method. In *2019 IEEE symposium series on computational intelligence (SSCI)*, pages 433–436. IEEE, 2019.

[31] Valmik Bhavar, Prakash Kattire, Sandeep Thakare, RKP Singh, et al. A review on functionally gradient materials (fgms) and their applications. In *IOP conference series: materials science and engineering*, volume 229, page 012021. IOP Publishing, 2017.

[32] Harshit Gupta, Kyong Hwan Jin, Ha Q Nguyen, Michael T McCann, and Michael Unser. Cnn-based projected gradient descent for consistent ct image reconstruction. *IEEE transactions on medical imaging*, 37(6):1440–1453, 2018.

[33] Yifei Zhang. A better autoencoder for image: Convolutional autoencoder. In *ICONIP17-DCEC. Available online: http://users. cecs. anu. edu. au/Tom. Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58. pdf (accessed on 23 March 2017)*, 2018.

[34] Gulshan Kumar. Evaluation metrics for intrusion detection systems-a study. *Evaluation*, 2(11):11–7, 2014.

[35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.