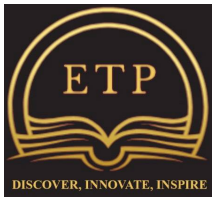
ISSN: 2710-2599



International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https: / / www.svedbergopen.com /

Research Paper **Open Access**

BotVision: Robust BotNet Detection Using Gan-Augmented Transfer Learning on Adversarial Image Data

Mubashir Mohsin[1] and Akinul Islam Jony[2*]

[1]American International University-Bangladesh, Dhaka. E-mail: mubashir.mohsin.42884©gmail.com
[2]Universitat Oberta de Catalunya, Barcelona. E-mail: ajony@uoc.edu

Article Info

Volume 6, Issue 1, May 2026
Received : 18 February 2026
Accepted : 02 May 2026
Published : 25 May 2026
doi: 10.67191/IJDSBDA.6.1.2026.1-16

Abstract

The widespread adoption of Internet of Things (IoT) devices has raised the potential of BotNet attacks, which pose major threats to network security. This study dives into the implementation of deep learning and transfer learning models to detect and categorize such attacks on real-world network traffic. We compare the performance of four popular architectures on image-based representations of the Bot-IoT dataset: VGGNet-16, ResNet50, MobileNetV2, and InceptionV3. To evaluate robustness under adversarial settings, adversarial samples were generated using the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD). Furthermore, Generative Adversarial Networks (GANs) were used to test the models with realistic and deceptive inputs. The findings of the experiments show that although all models were highly accurate under normal circumstances (e.g., 99.89% for ResNet50 and 99.93% for VGGNet-16), they differed greatly in how resilient they were to adverse data. With a robust score of 91.78%, VGGNet-16 outperformed InceptionV3 in terms of adaptation to attacks. ResNet50 and MobileNetV2, on the other hand, demonstrated significant performance decreases when subjected to adversary influence. These results underline the significance of adversarial training and robust model selection for IoT security applications. This research offers useful information for creating more robust deep learning-based Botnet detection systems by highlighting the advantages and disadvantages of various topics. The findings also highlight the gaps of ample research on adaptive defenses that may change to meet new threats in intricate IoT ecosystems.

Keywords: *Adversarial attack, Generative adversarial networks, Transfer learning, Deep learning, Cybersecurity*

1. Introduction

The emergence of IoT has sparked a technological revolution that is currently impacting every aspect of our

* *Corresponding author: Akinul Islam Jony, Universitat Oberta de Catalunya, Barcelona. E-mail: ajony@uoc.edu*

modern civilization. IoT technology has never been easier or more connected, but it also presents a variety of intricate and ubiquitous cybersecurity problems. The diverse range of applications and the complex network of devices that reinforce IoT bring major difficulties to the identification and prevention of malicious activity. This emerging field is often characterized by the mysterious traces of attackers who use vulnerabilities to steal confidential data and tamper with the operations of competitors (Liu *et al.*, 2009; Jony and Arnob, 2024). The variance in processing speeds between edge and core routers creates an inherent vulnerability in the internet infrastructure (Hoque and Bhattacharyya, 2015; Arnob and Jony, 2024). This speed difference, however, only touches the surface of a complex problem that affects the cybersecurity landscape. In this complex digital landscape, the use of malicious behavior takes the shape of Botnet attacks, which are groups of infected devices coordinated by a single harmful entity. Although these malicious entities carry out an array of harmful operations, they are particularly interested in Denial of Service (DoS) and Distributed Denial of Service (DDoS) attacks (Biswas *et al.*, 2024). Both DoS and DDoS attacks rank as the foremost threats within the domain of Botnet attacks (Jony and Hamim, 2024), which are aggravated by the increasing reliance of spammers on using bots to generate spam messages under false identities (Mohsin and Jony, 2024).

The overlap of common features and structural attributes within these attacks complicates their differentiation when they traverse transmission protocols such as TCP, UDP, and HTTP (Biswas *et al.*, 2024). Several works on large-scale datasets, such as the Bot-IoT dataset (Biswas *et al.*, 2024), ISCXIDS2012 (Devi *et al.*, 2023), CIC-IoT2023 dataset (Jony and Arnob, 2024), and numerous more, have been investigated, examined, and evaluated with multiple approaches with the objective to gain insight about the structural attributes of the Botnet attacks. Although deep learning models have demonstrated significant potential in identifying threats in IoT environments (Dong and Sarem, 2022; Wang *et al.*, 2020; Aborujilah and Musa, 2017; Raja Sree and Saira Bhanu, 2017; Asad, 2019; Muraleedharan and Janet, 2021; Alomari *et al.*, 2023), such as Botnet-based DoS and DDoS attacks, they are still susceptible to adversarial perturbations, which are subtle, frequently undetectable changes to input data that can cause models to classify data incorrectly. In real-world IoT systems, where accurate and dependable threat detection is essential, this vulnerability becomes extremely challenging. This study addresses this challenge by transforming network traffic data into visual image representations, enabling classification through advanced image-based deep learning models. Beyond detection, these models are resistant to adversarial image alterations, which increases the reliability and dependability of Botnet attack detection in real-world IoT applications.

This research endeavors to illustrate the intricacies of detecting and classifying Botnet attacks within IoT systems by introducing a novel approach. It harnesses the transformative power of deep learning and transfer learning models to enhance the security of IoT networks. For this experiment, four well-known transfer learning models—VGGNet-16, ResNet50, MobileNetV2, and Inception-v3—have been carefully chosen, each for its distinct qualities in image classification. As cyber threats get more sophisticated, attackers are using deliberately constructed perturbations, or Adversarial Attacks, to take advantage of weaknesses in machine learning models. These attacks can drastically reduce the accuracy of detection. Therefore, our research not only proposes a novel image-based classification approach for detecting Botnet activity but also rigorously tests its resilience against adversarial interference. This research seeks to analyze the adaptability, robustness, and resilience of these models when challenged with adversarial samples generated through potent adversarial techniques like the Fast Gradient Sign Method (FGSM) (Liu *et al.*, 2019), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD) (Arnob and Jony, 2024). The proposed framework focuses on adversarial robustness and detection efficacy promises that it not only accurately detects threats but also retains its integrity in the face of malicious manipulation, improving the overall credibility of IoT cybersecurity systems. Additionally, it explores the use of Generative Adversarial Networks (GANs) not only to simulate adversarial scenarios but also to enhance model resilience. GANs are employed to generate augmented adversarial samples, enabling the models to learn from potential attacks and strengthen their defenses against adversarial manipulation. Using transfer learning models in conjunction with GANs, this study suggests a unique deep learning framework that transforms IoT traffic data into visual representations to both detect Botnet attacks and protect against adversarial tampering

This research focused on the generation and augmentation of image samples, including both original and adversarial ones, to investigate Botnet attack detection and classification. The foundation of this experiment

lies in the Bot-IoT dataset (UNSW Canberra, 2021), a comprehensive collection of real-world IoT network traffic encompassing both benign and malicious activities. To enable the application of image classification models, we transformed the tabular network traffic data into visual representations, converting complex traffic features into vivid and colorful image samples. To ensure a rigorous evaluation, the dataset was systematically partitioned into training, validation, and testing subsets. Deep learning models were trained to identify patterns linked to various traffic classes using the training and validation sets, while the testing set was reserved for assessing the models' generalization capabilities on unseen data, including adversarially modified samples. By simulating actual evasive techniques, these adversarial samples reveal insight into the model's robustness under adversarial conditions. The authenticity and diversity of the Bot-IoT dataset, derived from realistic attack scenarios, lend credibility and real-world relevance to the findings. This structured and reproducible approach thus provides a comprehensive evaluation (Aafaq *et al.*, 2019) of deep learning-based visual classifiers in the context of Botnet attack detection, emphasizing both detection capability and adversarial defense.

The research's broader focus emphasizes how urgently robust and durable defenses are needed to shield IoT systems from the growing threat of botnet attacks as well as adversarial perturbation. This work offers a trustworthy framework for enhancing the security of IoT settings and advances our understanding of their weaknesses by investigating adversarially robust image-based deep learning models. This paper's remaining sections are organized as follows: A comprehensive review of relevant work is given in Section 2, our proposed method is described in Section 3, the experimental results and findings are covered in Section 4, and a critical analysis and discussion of the results are provided in Section 5, which is followed by closing remarks.

2. Literature Review

There are three main topics covered in this part of the review and related works. Initially, it explores the spread of Botnet attacks in the context of IoT, emphasizing the range of attacks and their consequences. In the second section, it delves beyond the use of transfer learning and deep learning models for BotNet attack detection, accentuating distinct model configurations and their implications. Finally, it explores the methods utilized to generate adversarial samples and the use of GANs in cybersecurity, focusing on how they affect model vulnerability. Essential assessment measures are also introduced, including evaluation metrics like accuracy, precision, recall, F1-score, and support.

2.1. Bot-IoT Dataset

The complexity and sheer scale of the IoT landscape render the identification of malicious activities a formidable challenge, often leaving little trace of their origins (Liu *et al.*, 2009). This inherent weakness in networked systems positions them as attractive targets for a spectrum of attacks orchestrated with the intent to illicitly access sensitive data or disrupt the resources of business competitors, leveraging diverse stashes of attack tools (Hoque, 2014). To aid in the investigation of such attacks, the Cyber Range Lab at UNSW Canberra created the Bot-IoT dataset, which is a well-structured and comprehensive resource for training and assessing network security models. The dataset simulates realistic IoT network traffic, incorporating both benign and malicious flows. It includes the four primary types of attacks connected to botnets: Denial of Service (DoS), Distributed Denial of Service (DDoS), information theft, and reconnaissance (Biswas *et al.*, 2024). We explicitly target DoS and DDoS assaults in our work because they are quite common and damaging in actual IoT scenarios (Levy, 2003; Cook *et al.*, 2006).

The dataset includes 26 network flow features that were extracted using the Argus network flow collector; the 14 computed features that Koroniotis *et al*. (2019) later introduced are not included. These characteristics reflect a variety of traffic characteristics, including packet counts, source / destination IPs and ports, and connection duration. Importantly, the attacks in the dataset were created utilizing popular protocols that are frequently used in flooding-based DoS / DDoS scenarios, such as TCP, UDP, and HTTP (Biswas *et al.*, 2024). Because DoS and DDoS attacks share many structural and behavioral characteristics, it is still difficult to differentiate between them using traffic patterns alone. Both attacks rely on flooding targets with too much traffic, which is frequently started by clients controlled by bots and coordinated via agent-handler models. The difficulty of detection is further increased by the growing usage of bots to send automated spam or

malicious traffic while hiding their origin (Kaur *et al.*, 2015). In this study, a filtered subset of the Bot-IoT dataset is utilized, concentrating on DoS and DDoS instances over the protocols. The goal is not only to detect these attacks with a unique image-based classification framework, but also to verify the model's robustness under adversarial perturbations, guaranteeing that detection is reliable even in adversarial environments. The richness and realism of the Bot-IoT dataset make it an ideal benchmark to evaluate model resistance and detection accuracy against contemporary cyberthreats.

2.1.1. Deep Learning and Transfer Learning

Deep learning, a subfield of machine learning, has made immense progress in computer vision, redefining how computers process and interpret graphical data, videos, and images. However, despite its growing vulnerability to changing threats, its full potential is still very little in the domains of cybersecurity and computer networks. Particularly, transfer learning has become an effective technique of identifying and interpreting patterns from network-related data, which presents promising prospects for threat identification. Transforming tabular network traffic data into image representations is a novel visual strategy to detect BotNet assaults using transfer learning models. This visual transformation facilitates the use of complex picture classification algorithms in cybersecurity applications. Among various transfer learning architectures, four renowned models were selected based on their proven effectiveness in image classification:

- VGGNet-16, valued for its deep yet straightforward convolutional architecture,
- ResNet50, which addresses vanishing gradient issues through the use of residual connections,
- MobileNetV2, recognized for its lightweight design and use of inverted residuals with linear bottlenecks, and
- InceptionV3, an optimized iteration of InceptionNet that reduces computational cost without sacrificing performance.Over the years, the cybersecurity landscape has seen the development of various advanced detection frameworks and tools aimed at identifying and mitigating network intrusions. Techniques such as DDAML (Dong and Sarem, 2022), SBS-MLP (Wang *et al.*, 2020), HADEC (Hameed and Ali, 2018), D-FACE (Behal, 2018), MADM (Aborujilah and Musa, 2017), and HADM (Raja Sree and Saira Bhanu, 2017) represent some of the current state-of-the-art mechanisms designed to tackle sophisticated attacks. The commonality among many of these systems lies in their focus on identifying signature patterns, particularly in DoS and DDoS attacks, which continue to pose severe risks to critical infrastructures. Within this context, this study undertakes a rigorous comparative evaluation of several deep learning-based cybersecurity solutions. These include LUCID (Doriguzzi-Corin *et al.*, 2020), RNN-AE (Elsayed, 2020), DeepDetect (Asad, 2019), DNN-SlowDoS (Muraleedharan and Janet, 2021), 7-NN (Dhanya *et al.*, 2023), SySeVR (Li *et al.*, 2022), DNN+LSTM (Alomari *et al.*, 2023), and WAF-LSTM (Devi *et al.*, 2023). Each model is assessed for its capability to detect and classify network attacks, especially those within the DoS and DDoS domain. The comparative analysis provides valuable insights into their performance, strengths, and limitations, ultimately informing the development of more robust and resilient defenses against emerging threats in IoT-enabled networks.

2.2. Generation of Adversarial Samples and Use of GANs

Adversarial attacks in machine learning and deep learning have garnered substantial attention, focusing on their capacity to influence model predictions by introducing undetectable perturbations to input data. Adversarial samples remain a significant research area, especially in the field of computer vision, generative AI, and image processing. Techniques like FGSM, BIM, and PGD efficiently generate adversarial samples, exposing vulnerabilities in deep learning models. FGSM (Liu *et al.*, 2019), as one of the pioneering techniques, efficiently generates adversarial samples by perturbing input data in the direction of the model's gradient. Subsequently, the BIM and the PGD (Arnob, and Jony, 2024) evolved as iterative variants of FGSM, enabling the creation of more robust and steered adversarial samples. These techniques have been instrumental in exposing vulnerabilities in deep learning models, necessitating the development of robust defenses against adversarial attacks.

Incorporating GANs into adversarial sample generation has been a groundbreaking development (Liu *et al.*, 2019). GANs consist of a generator and a discriminator network engaged in a two-player game to create highly realistic synthetic data (Purwono *et al.*, 2925). The integration of GANs in adversarial attack generation broadens the scope of potential threats to deep learning models. GAN-based adversarial samples are more confronting to perceive and have a higher success rate in conning models. The capacity of GANs to generate diverse and realistic adversarial samples introduces a new level of complexity to the arms race between attackers and defenders in the field of cybersecurity. Researchers have utilized GANs to generate synthetic data for training models, bridging the gap in data scarcity for certain tasks.

The development of adversarial sample generation techniques, including FGSM, BIM, and PGD, alongside the integration of GANs, represents a pivotal intersection of deep learning and cybersecurity. While adversarial attacks pose significant challenges, they have also spurred innovations, revealing the adaptability and robustness of modern deep learning models. Future research should prioritize the development of more effective defenses, ensuring the reliability and security of machine learning applications in diverse domains.

2.3. Model Evaluation Techniques

To fully evaluate the prediction power of deep learning models, a strict set of measures must be used. In this area, the most significant assessment measures are Support (S), Recall (R), Accuracy (ACC), Precision (P), and F1-score (F1) (Lisun-Ul-Islam *et al.*, 2023; Arnob and Jony, 2024). As the ratio of accurately anticipated instances to the total number of instances, accuracy is a fundamental statistic. The precision of positive predictions is measured by expressing the ratio of true positive forecasts to all positive predictions. The ratio of true positive predictions to the total number of actual positive cases is called recall, which is often referred to as sensitivity or the true positive rate. It determines how well the model can identify all pertinent events (Jony and Arnob, 2024; Hamim and Jony, 2024). A balanced evaluation is provided by the F1-score, which is a harmonic mean of precision and recall and is especially helpful in the case of unbalanced datasets. Support, which offers information on class distribution, indicates the actual number of instances of each class in the test dataset. These metrics impart a broader overview of the effectiveness of models, reflecting both the accuracy of predictions and the ability to categorize cases across different classes (Hamim and Jony, 2024). A better understanding of the model's performance in tasks like Botnet attack classification is made possible by the analysis of these indicators, which offer helpful insights into the predictions' strengths and limitations.

3. Materials and Methods

3.1. Materials Used

Chief among these materials was the Bot-IoT Dataset (UNSW Canberra, 2021), meticulously curated by UNSW Canberra, comprising an extensive repository of CSV files housing the dataset's multifarious samples. The analysis and computation were primarily conducted using the Python programming language, supported by crucial data preprocessing libraries like Pandas and NumPy. Furthermore, the intricate domain of image processing was navigated with the adept utilization of OpenCV. Deep learning, a cornerstone of our methodology, was realized through the concerted employment of state-of-the-art TensorFlow and Keras frameworks. In this ambitious endeavor, a triumvirate of transfer learning models is harnessed, namely VGGNet-16, ResNet50, MobileNetV2, and InceptionV3. To unveil the vulnerability landscape, diligently adversarial attack techniques are harnessed, notably FGSM, BIM, and PGD. Beyond software, the research journey was bolstered by the robust hardware resources, underpinned by NVIDIA GeForce RTX 3060 Ti GPU, which expedited model training and experimentation. Taken as a whole, these resources served as the framework for extensive research, which revealed the intricate relationships between adversarial samples and GANs in the context of Botnet attack images.

3.2. General Overview of the Model

This research introduces a novel approach that transforms tabular network traffic data from the Bot-IoT dataset into visual image representations. The transformation process involves selecting relevant features (Biswas *et al.*, 2024) and encoding them into distinct color mappings, thereby enabling the use of computer vision techniques for network threat detection. The dataset comprises seven classes, including normal traffic

and BotNet attacks such as DoS and DDoS, each subdivided into three protocol-specific categories (TCP, UDP, and HTTP).

The core objective of this study is to construct a predictive model that not only distinguishes between normal and malicious traffic but also demonstrates robustness against adversarial samples. These adversarial images are crafted using established attack generation techniques, including FGSM, BIM, PGD, and GANs. By incorporating adversarial training and evaluation into the workflow, the proposed framework aims to enhance model resilience, ensuring more reliable and secure detection of BotNet threats in IoT environments. The whole pipeline, including data transformation, adversarial sample generation, and the Deep Neural Network (DNN)-based prediction analysis, is shown in Figure 1 for illustration purposes.

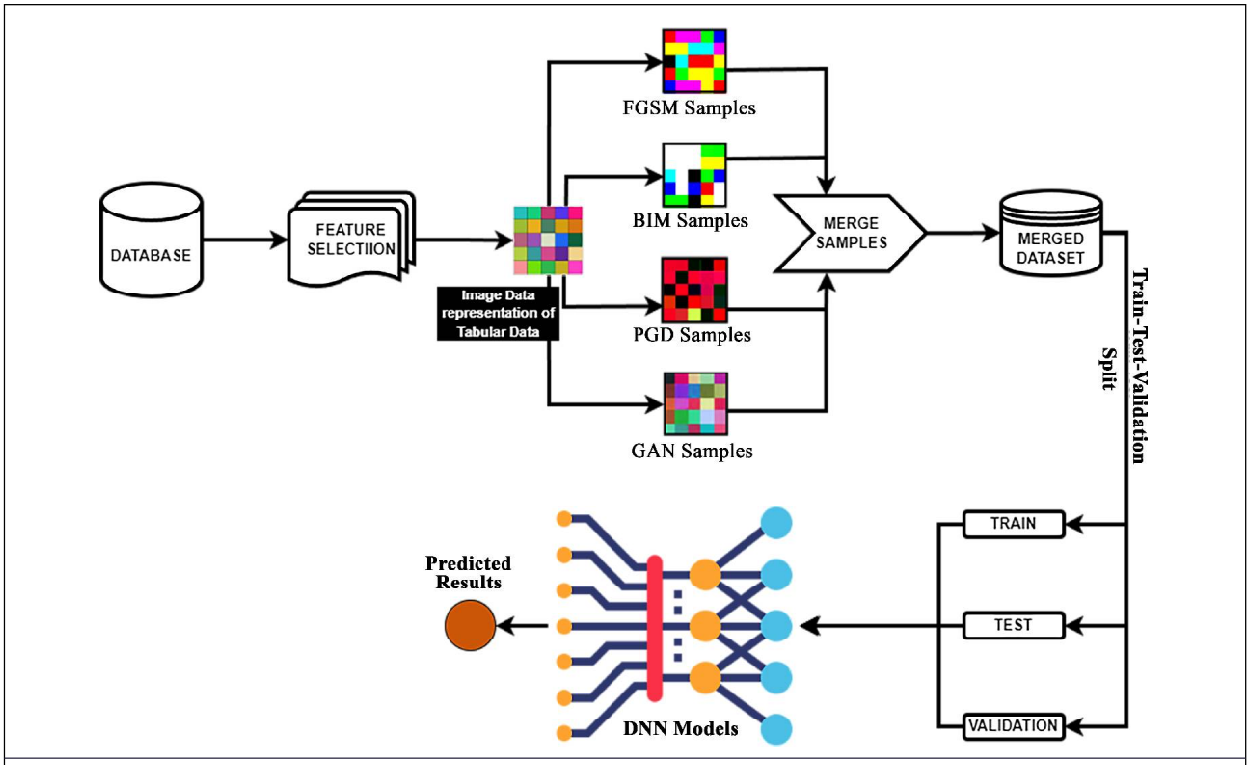


Figure 1: Overview of the Proposed Model Workflow

3.3. Data Preparation and Tabular-to-Image Conversion

This study utilizes the Bot-IoT dataset curated by UNSW Canberra (2021), comprising over 73.9 million records that encapsulate a wide range of network behaviors, including various attack types and normal traffic. Due to the inherent class imbalance in the dataset, we selectively extracted samples belonging to three principal categories: DoS, DDoS, and normal traffic (Ferdous *et al.*, 2024). This balanced subset was further refined into three CSV files, which served as the foundation for the image generation process.

A total of 12,000 images were generated for each of the five types of image categories: original, FGSM, BIM, PGD, and GAN. This resulted in a cumulative dataset of 60,000 image samples (12,000 × 5 = 60,000). These images were partitioned into training, validation, and testing sets to evaluate the model's performance and generalizability. Specifically, 49,750 images were used for training, 2,750 images for validation, and 7,500 images for testing. From the training set, 9,950 images were allocated to each of the five categories (5 × 9,950 = 49,750), ensuring balanced representation across the different adversarial types. Similarly, the validation set contained 550 images per category (5 × 550 = 2,750), and the test set included 1,500 images per category (5 × 1,500 = 7,500). See the Table 1 for the breakdown and distribution of the final dataset.

To enable the use of deep learning models that typically operate on visual data, the tabular network data is transformed into a structured image representation. A total of 25 relevant features were selected based on their predictive value and contribution to the classification task (Koroniotis *et al.*, 2019). These features were then encoded into a visual format using the following approach:

Table 1: Distribution of Image Samples Across Dataset Splits and Adversarial Types

Image Type	Training Samples	Validation Samples	Testing Samples	Total
Original	9,950	550	1,500	12,000
FGSM	9,950	550	1,500	12,000
BIM	9,950	550	1,500	12,000
PGD	9,950	550	1,500	12,000
GAN	9,950	550	1,500	12,000
Total	**49,750**	**2,750**	**7,500**	**60,000**

1. Each feature was assigned a distinct color, mapped based on its normalized value range across the dataset.
2. A fixed grid of 5 × 5 layout was used to represent these 25 features spatially.
3. The final image size was set to 250 × 250 pixels at 300 dpi resolution.
4. Each grid cell within the image corresponds to a single feature and is filled with the color representing that feature's value in a specific sample.

This method results in one RGB image per network record, with consistent spatial positioning of features across all samples. The transformation allows for visual patterns to emerge that are potentially more discriminative when processed by convolutional neural networks. Five image classes were generated to represent both original and adversarial variants of the data: (1) Original Samples, (2) FGSM Samples, (3) BIM Samples, (4) PGD Samples, and (5) GAN Samples.

These classes collectively form the complete dataset used for training and evaluating the deep learning models in this study. Figure 2 illustrates the 25-feature grid layout and the distinct color mapping used in the transformation process, where each feature is represented by a distinct color.

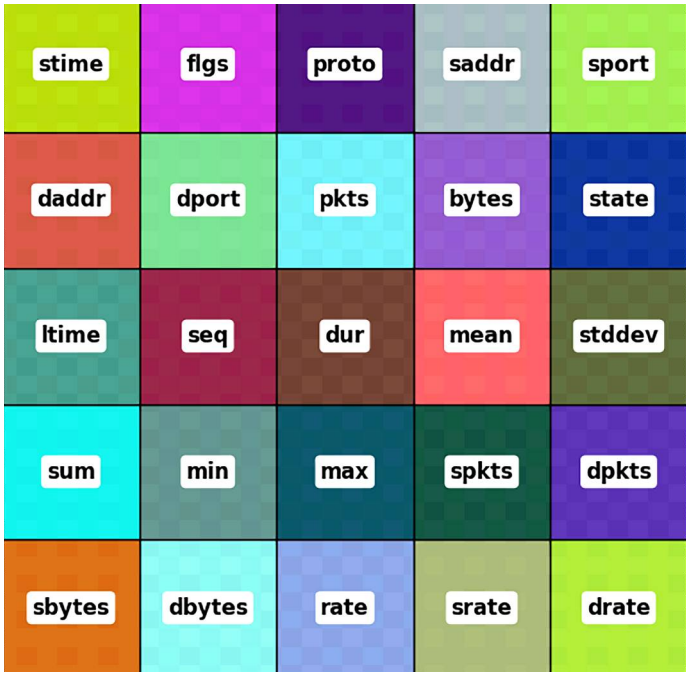


Figure 2: Visual Mapping of 25 Selected Features into a 5 × 5 Grid for Tabular-to-Image Transformation

3.4. Transfer Learning Model Architecture

To construct robust and efficient classifiers for adversarial and original network traffic image data, we employed four prominent transfer learning architectures: VGGNet-16, ResNet50, MobileNetV2, and InceptionV3. These pre-trained models, initially trained on the ImageNet dataset, were fine-tuned to adapt to our specific classification task involving seven distinct classes.

To maintain architectural consistency and ensure fair model comparison, each base model was modified by appending a standardized fully connected (dense) head:

- A dense layer with 512 units and ReLU activation,
- Followed by a dense layer with 128 units and ReLU activation,
- Culminating in a final dense layer with 7 units and Softmax activation to enable multiclass classification.

All models were compiled using the Adam optimizer with the following hyperparameters: a learning rate of 0.0001, $\beta_1 = 0.99$, $\beta_2 = 0.99$, and $\epsilon = 1 \times 10^{-9}$. The loss function employed was Sparse Categorical Crossentropy, which is well-suited for multi-class problems with integer labels. Training was conducted using a batch size of 32, and to prevent overfitting, we implemented early stopping callbacks triggered after ten consecutive epochs without improvement on the validation loss.

These configurations allowed for consistent convergence behavior across all architectures. A detailed overview of the unified architectural framework used for the transfer learning models is illustrated in Figure 3, providing clarity on the shared and model-specific components.

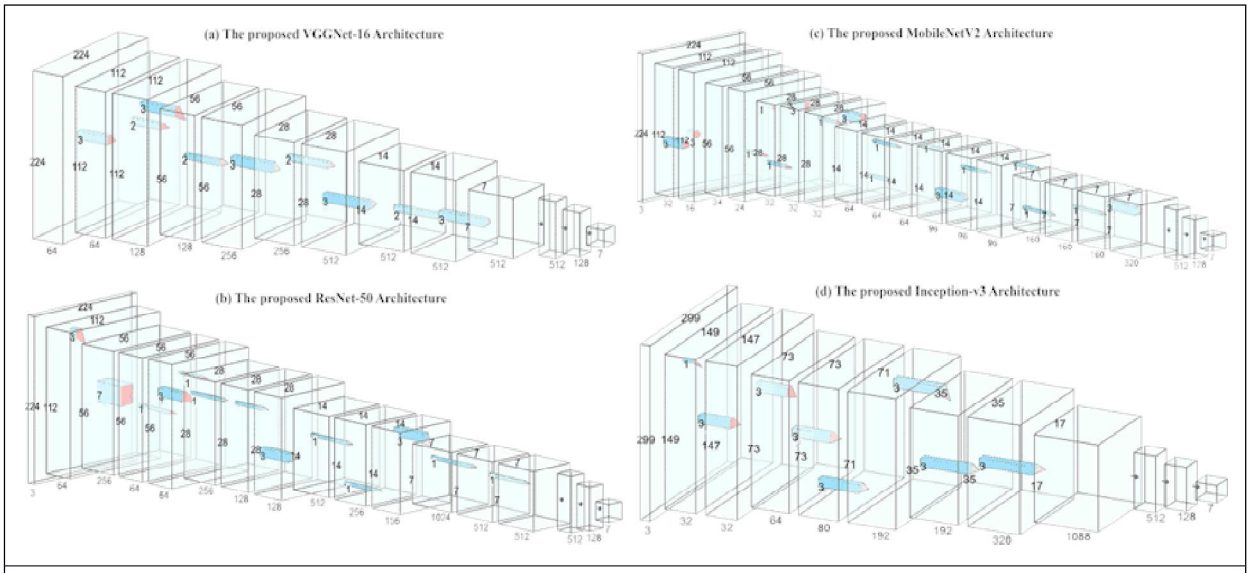


Figure 3: Architectural Diagrams of the Four Transfer Learning Models: (a) VGGNet-16, (b) ResNet50, (c) MobileNetV2, and (d) InceptionV3

3.5. Adversarial Sample Generation

Adversarial attack generation is a pivotal aspect of our research, aiming to evaluate the robustness of image classification models. This section encompasses a diverse set of adversarial techniques, including the FGSM, BIM, PGD, and the application of GANs.

Fast Gradient Sign Method (FGSM): FGSM is a single-step attack that efficiently generates adversarial examples by exploiting the gradients of the loss function with respect to the input image.

Basic Iterative Method (BIM): BIM is an iterative extension of FGSM that generates adversarial examples by repeatedly applying small perturbations in the direction of the gradient. To ensure the validity of image data, the resulting images are clipped to remain within allowable pixel value bounds.

Projected Gradient Descent (PGD): PGD builds upon BIM by introducing a step size parameter, offering more granular control over the perturbation process. Like BIM, PGD performs iterative updates based on the

gradient of the loss. This projection mechanism improves the robustness of generated adversarial examples, making them more difficult for models to defend against.

Generative Adversarial Network (GAN): To complement gradient-based attacks, GAN is employed for generating adversarial samples. During training, these components compete in a minimax game, enabling the generator to learn realistic image distributions. Smoothed labels are incorporated to stabilize the training dynamics and reduce overfitting. By leveraging GANs, this research aims to produce diverse and stealthy adversarial examples that test the model's robustness. Once generated, these samples are systematically organized into labeled directories and partitioned into training, validation, and testing sets to support experimental consistency. Figure 4 illustrates representative images from each of the seven generated sample classes, where each sample is illustrated by 7 images from the 7 classes.

Figure 4: Image representation of the Bot-IoT dataset with original samples along with the BIM, FGSM, PGD, and GAN adversarial samples.

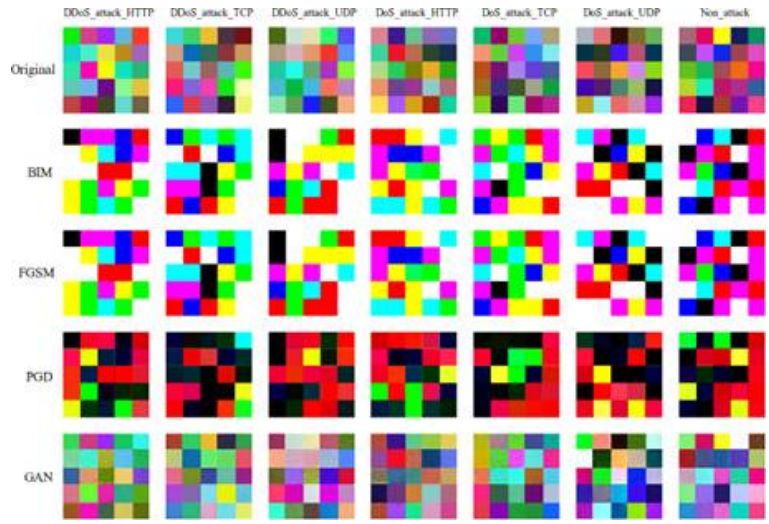


Figure 3: Architectural Diagrams of the Four Transfer Learning Models: (a) VGGNet-16, (b) ResNet50, (c) MobileNetV2, and (d) InceptionV3

3.6. Evaluation Metrics

To assess the predictive performance of the proposed models, a suite of evaluation metrics such as accuracy, precision, recall, F1-Score, and support is employed. These evaluation metrics together offer a complete picture of the model's performance. This performance evaluation is useful for analyzing the strengths and weaknesses of the models' predictions, contributing to a better understanding of their performance in classifying Botnet attacks. A brief description of each of these is given below.

Accuracy (ACC): Accuracy is a fundamental evaluation measure that calculates the percentage of correctly predicted instances in a test set.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad ...(1)$$

Precision (P): Precision is an evaluation measure that calculates the ratio of true positive predictions to the total number of positive predictions.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad ...(2)$$

Recall (R): Recall (also known as sensitivity or true positive rate) is another evaluation measure that calculates the ratio of true positive predictions to the total number of actual positive instances.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad ...(3)$$

F1-Score (F1): The F1-score is an evaluation measure that defines a balance between a harmonic mean of precision and recall. It is particularly a useful measure for dealing with imbalanced datasets.

$$F1-Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad ...(4)$$

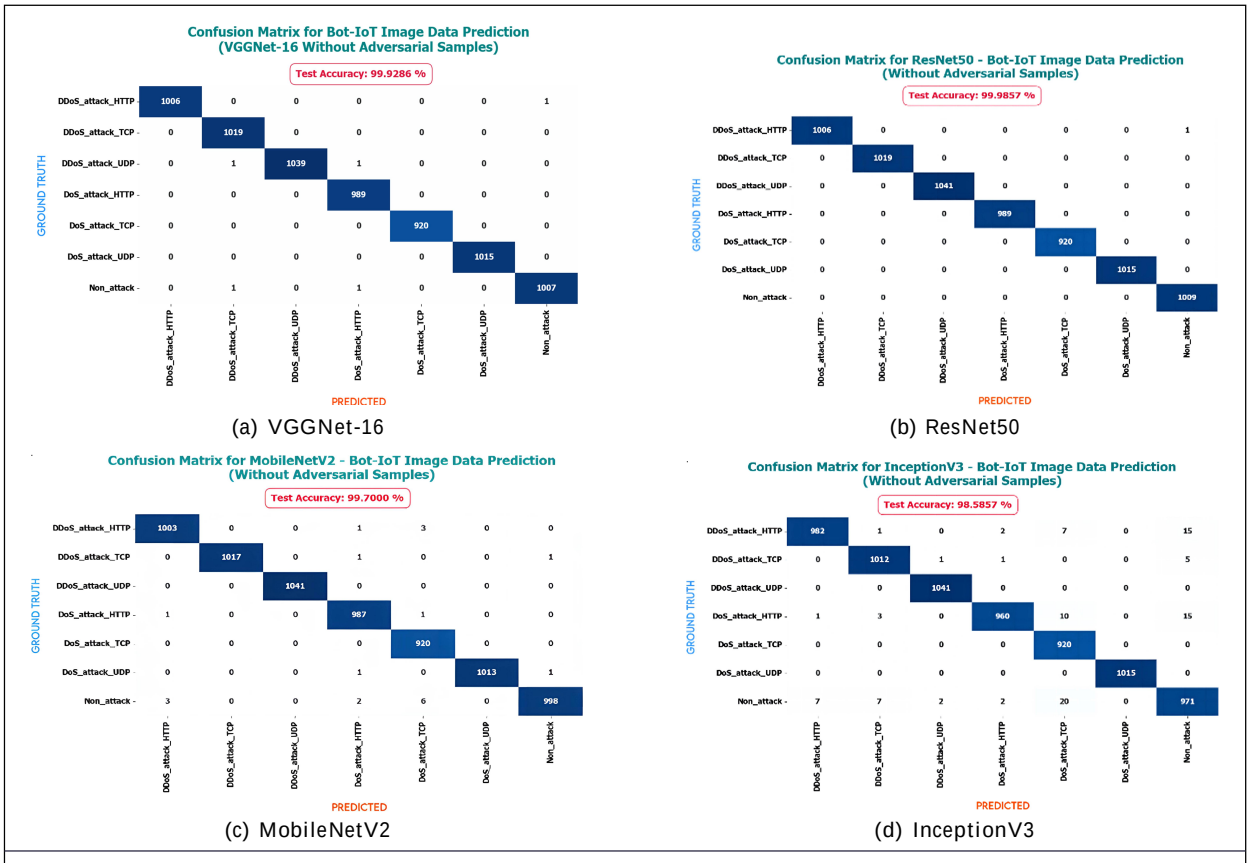
Support (S): Support refers to the number of actual occurrences of each class in the test dataset. It helps contextualize the above metrics by indicating class distribution during evaluation.

4. Results and Findings

The outcomes of our research, including the performance of deep learning models, the impact of adversarial samples, and the effectiveness of GANs in the context of image representations of Botnet attacks, are presented in this section.

4.1. Model Performance without Adversarial Samples

This study initiated by training deep learning models, namely VGGNet-16, ResNet50, MobileNetV2, and InceptionV3, using the image representations of BotNet attacks (DDoS and DoS) and non-attack samples. The models exhibited remarkable accuracy and precision across multiple attack categories. For VGGNet-16 (Figure 5a), the test accuracy reached an impressive 99.93%. ResNet50 (Figure 5b) surpassed this with a test accuracy of 99.99%. InceptionV3 (Figure 5d) demonstrated a slightly lower test accuracy of 98.59%, while MobileNetV2 (Figure 5c) achieved an accuracy of 99.70%. Table 2 indicate the capability of deep learning models to accurately classify BotNet attacks, even in the presence of class imbalance. The confusion matrix of the model's performance without adversarial samples is attached to Figure 5.



(a) VGGNet-16 (b) ResNet50

(c) MobileNetV2 (d) InceptionV3

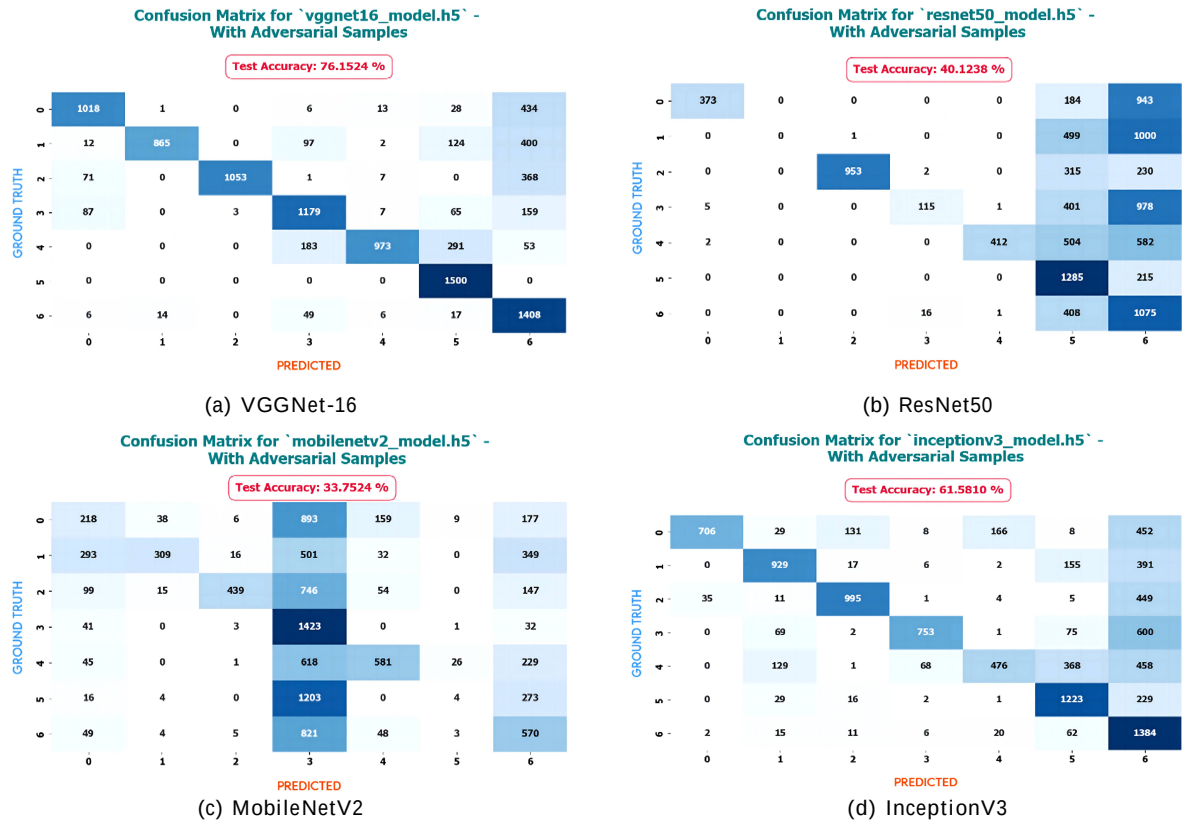
Figure 5: Confusion Matrix without Adversarial Samples

Table 2: Model's Performance without Adversarial Samples

Model Name	Test (Acc) %	Precision	Recall	F1-Score	Support
VGGNet-16	99.93	0.99	0.99	0.99	7500
ResNet50	99.99	1.00	1.00	1.00	7500
MobileNetV2	99.70	0.98	0.98	0.98	7500
InceptionV3	98.59	0.97	0.97	0.97	7500

4.2. Model Performance with Adversarial Samples

In the evaluation of model performance with integrated adversarial samples, we observed varying results across the transfer learning models. VGGNet-16 (Figure 6a) exhibited the highest accuracy, achieving 76.15%, signifying its robustness against adversarial attacks. Notably, it maintained a balanced performance with a precision of 0.83 and a recall of 0.76. InceptionV3 (Figure 6d) also showed substantial performance with an accuracy of 61.58%, supported by a precision of 0.75 and a recall of 0.70. On the other hand, ResNet50 (Figure



(a) VGGNet-16 (b) ResNet50

(c) MobileNetV2 (d) InceptionV3

Figure 6: Confusion Matrix with Adversarial Samples

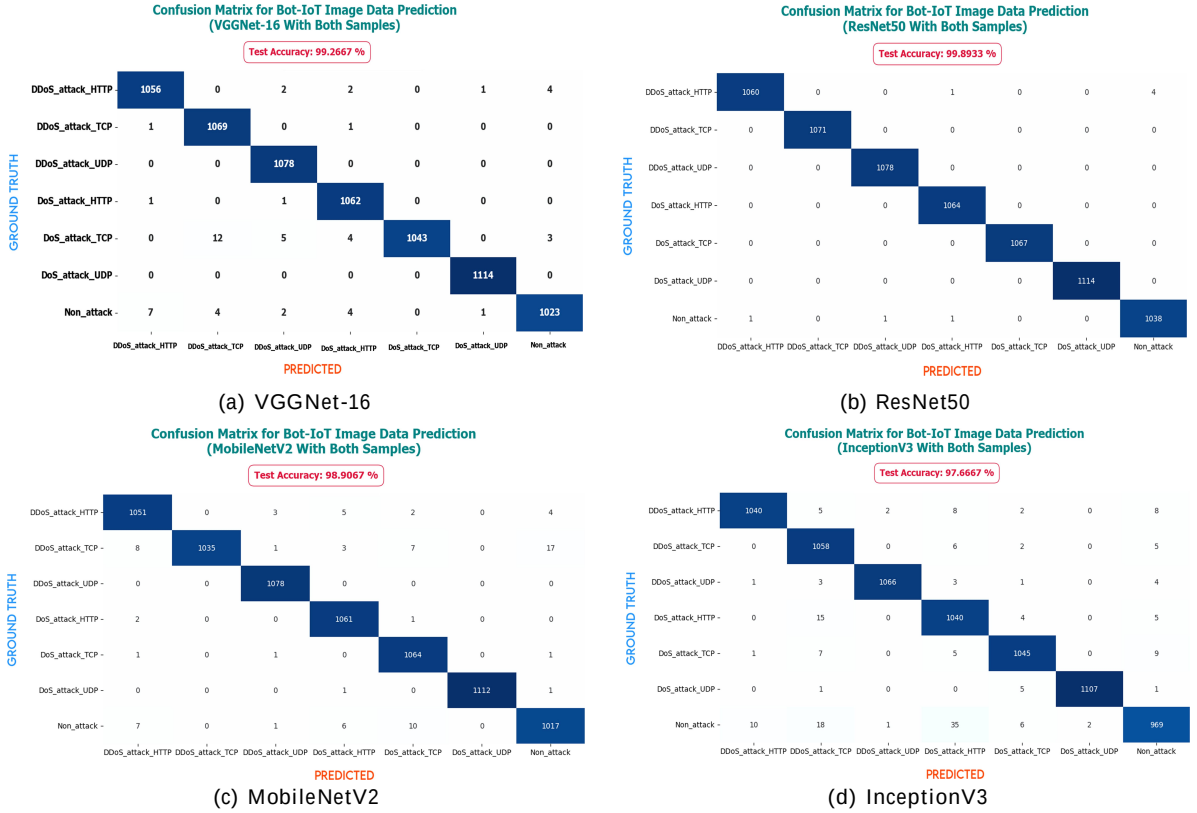
Table 3: Model's Performance with Adversarial Samples

Model Name	Test (Acc) %	Precision	Recall	F1-Score	Support
VGGNet-16	76.15	0.83	0.76	0.76	7500
ResNet50	40.12	0.62	0.39	0.36	7500
MobileNetV2	33.75	0.47	0.38	0.33	7500
InceptionV3	61.58	0.75	0.70	0.67	7500

6b) and MobileNetV2 (Figure 6c) demonstrated lower accuracies of 40.12% and 33.75%, respectively, indicating reduced resilience to adversarial samples. This trend was further evident in their comparatively lower precision and recall values. The support for all models remained constant at 7500 samples. These findings highlight the varying degrees of robustness of the models in Table 3 when evaluated with adversarial inputs. The confusion matrix of the model's performance with adversarial samples is attached to Figure 6.

4.3. Model Performance with both Adversarial and Original Samples

Upon integrating original images with adversarial samples and conducting model retraining, a notable enhancement in model accuracy was observed across all transfer learning models. VGGNet-16 (Figure 7a), ResNet50 (Figure 7b), MobileNetV2 (Figure 7c), and InceptionV3 (Figure 7d) exhibited impressive test accuracies of 99.27%, 99.89%, 98.91%, and 97.67%, respectively, according to Table 4. These results underscore the model's significantly improved resilience to adversarial attacks after training with an integrated dataset. The precision, recall, and F-1 score values were consistently high for all models, emphasizing the robustness achieved through this training approach. This outcome illustrates the effectiveness of integrating adversarial samples into the training process, enabling the models to better handle adversarial inputs in real-world scenarios. The confusion matrix of the model's performance with both original and adversarial samples is attached to Figure 7.



(a) VGGNet-16 (b) ResNet50

(c) MobileNetV2 (d) InceptionV3

Figure 7: Confusion Matrix with both Original and Adversarial Samples

Table 4: Model's Performance with both Original and Adversarial Samples

Model Name	Test (Acc) %	Precision	Recall	F1-Score	Support
VGGNet-16	99.27	1.00	1.00	1.00	7500
ResNet50	99.89	1.00	1.00	1.00	7500
MobileNetV2	98.91	0.99	0.99	0.99	7500
InceptionV3	97.67	0.98	0.98	0.98	7500

4.4. Performance Analysis and Findings

This research uncovers a two-fold revelation regarding the performance of deep learning models when confronted with adversarial samples in the context of Botnet attack detection. The initial evaluation of the models on original samples highlighted their remarkable ability to distinguish between benign and malicious activities, characterized by near-flawless accuracy, precision, recall, and F1-scores. VGGNet-16 demonstrated high resilience and maintained robust performance, underscoring its adaptability in the face of adversarial threats. In contrast, ResNet50 and MobileNetV2 displayed susceptibility to adversarial perturbations, resulting in significant performance degradation. The disparity in model performance highlights a crucial issue about these models' adaptability and resilience in real-world scenarios including adversarial threats.

It is possible to explain this sharp disparity in model performance to the inherent difficulty presented by adversarial data. By introducing minor perturbations that alter the model's decision-making process, adversarial attacks capitalize on the models' vulnerabilities. The resilience of VGGNet-16 and the relatively less accurate performance of InceptionV3 put emphasis on the models' differing levels of robustness, whilst ResNet50 and MobileNetV2 showed an apparent vulnerability to such perturbations. This research highlights how critical it is to strengthen these models against adversarial challenges, demanding a new strategy to guarantee their long-term predicted accuracy and reliability.

The findings presented in this study carry profound implications for the overall success of the project. The vulnerability of deep learning models to adversarial attacks in the realm of Botnet detection highlights the critical necessity for robust defenses. The success of this research relies on its ability to build models that can reliably operate under a variety of conditions, including adversarial ones as well. This effort not only improves the models' defense but also raises the bar for security and reliability in the field of botnet detection by using integrated adversarial samples to train the models and utilizing their adaptability. This is an important factor for assuring the project's feasibility in practical cybersecurity applications. When examining accuracy metrics in the assessment of the outcomes of the models that were employed, Table 5 shows that the models differ in their resilience and robustness.

Table 5: Comparison of Model Accuracies Under Different Scenarios and the Average Model Robustness

Model Names	Without Adv. Samples	With Adv. Samples	With Both	Avg. Robustness
VGGNet-16	99.93	76.15	99.27	91.78
ResNet50	99.99	40.12	99.89	80.00
MobileNetV2	99.70	33.75	98.91	77.45
InceptionV3	98.59	61.58	97.67	85.97

5. Conclusion

The research embarked on a comprehensive journey into the realm of Botnet image representations, meticulously examining the performance and adaptability of deep learning models in this context. This exploration allowed us to uncover intriguing insights into the strengths and vulnerabilities of various models when subjected to a variety of adversarial attacks. The major variances in performance metrics between benign and malicious samples acted as an indicator of the critical need to strengthen the defenses against adversarial threats. The vulnerability of models such as ResNet50 and MobileNetV2 to these attacks emphasized the necessity for enhancing the cybersecurity defenses. This study marks an important milestone in the continuous search for strong and resilient protection against Botnet attacks. It highlights the complex interaction between the intricate nature of adversarial attacks and the strength of deep learning models. Nevertheless, the field of research is broad and complex, and our study has limitations too.

One significant drawback of this study is that it only looks at a limited number of attack methods, a limited number of models, and a dataset that is complete but has certain temporal constraints. This allows for additional research and improvement of this study. The research needs to be expanded to include a wider range of

attacks, more adversarial strategies, and more recent and extensive datasets that correspond with the ever-changing Botnet threat scenario.

Another avenue for future exploration is the practical application of the models in real-world cybersecurity scenarios. This involves assessing these models' performance in real-world settings and their degree of compatibility with current cybersecurity procedures. Furthermore, the dataset used, while extensive, is not exhaustive. A more thorough investigation will benefit from using more recent datasets that contain more complex network traffic information and diverse profiles from numerous IoT devices. These extended datasets can facilitate the enhancement of model generalization and robustness, offering a more accurate reflection of the diversity and intricacy of Botnet attacks.

The future of adversarial defense in BotNet detection necessitates a holistic and forward-looking approach. Upcoming research should emphasize the creation of novel defense mechanisms and training strategies that are not only effective but also explainable and statistically grounded. These approaches must strike a careful balance between enhancing model robustness and preserving high predictive accuracy, leveraging reliable inference techniques and interpretable outputs.

Given the constantly evolving nature of BotNet attacks, models must possess the ability to adapt dynamically, staying ahead of emerging threats. To achieve this, future studies should explore model behavior in real-world cybersecurity environments, exposing them to a diverse range of BotNet image variants. Continuous performance monitoring and strategic refinement will be crucial to developing models that are resilient under varying and unpredictable conditions.

This constant quest for robustness is defined by persistent thinking, adaptability, and critical learning. We hope to make a significant contribution to the development of safe, reliable detection systems by addressing the present limitations of our work and investigating these potential pathways for the future. In the end, we hope that our research will serve as a basis for future developments, assisting in the creation of a cybersecurity environment that is stable in the face of increasingly complex attackers.

Acknowledgment

The authors wish to acknowledge that this research was independently conducted without external funding or support. All materials, including hardware and software, utilized in this study were privately owned and managed by the authors themselves.

References

Aafaq, N., Mian, A., Liu, W., Gilani, S.Z. and Shah, M. (2019). Video Description: A Survey of Methods, Datasets and Evaluation Metrics. *ACM Computing Surveys*, 1-37.

Aborujilah, A. and Musa, S. (2017). Cloud-Based DDoS HTTP Attack Detection Using Covariance Matrix Approach. *Journal of Computer Networks and Communications*.

Alomari, E., Nuiaa, R.R., Alyasseri, Z.A., Mohammed, H.J., Sani, N.S., Esa, M.I. and Musawi, B.A. (2023). Malware Detection Using Deep Learning and Correlation-Based Feature Selection. *Symmetry*, 123.

Arnob, A.K.B. and Jony, A.I. (2024). Comparing Machine Learning Algorithms for Breast Cancer Diagnosis: Wisconsin Diagnostic Dataset Analysis. *International Journal of Data Science and Big Data Analytics*, 4(2), 1-11.

Arnob, A.K.B. and Jony, A.I. (2024). Enhancing IoT Security: A Deep Learning Approach with Feedforward Neural Network for Detecting Cyber Attacks in IoT. *Malaysian Journal of Science and Advanced Technology*, 4(4), 413-420.

Asad, M. (2019). DeepDetect: Detection of Distributed Denial of Service Attacks Using Deep Learning.

Behal, S. (2018). D-FACE: An Anomaly Based Distributed Approach for Early Detection of DDoS Attacks and Flash Events.

Biswas, T., Ferdous, F.S. and Jony, A.I. (2024). Binary and Multi-Class Prediction of DDoS Attack Using Deep Learning Models. *International Journal of Data Science and Big Data Analytics*, 4(2), 49-58.

Cook, D., Hartnett, J., Manderson, K. and Scanlan, J. (2006). Catching Spam Before it Arrives: Domain-Specific Dynamic Blacklists. in *Proceedings of the Australasian Workshops on Grid Computing and EResearch*, 193-202, Hobart, Australia.

Devi, S.K., Albraikan, A.A., Al-Wesabi, F.N., Nour, M.K., Ashour, A. *et al.* (2023). Intelligent Deep Convolutional Neural Network Based Object Detection Model for Visually Challenged People. *Computer Systems Science and Engineering*, 46(3), 3191-3207. https://doi.org/10.32604/csse.2023.036980

Dhanya, K.A., Vajipayajula, S., Srinivasan, K., Tibrewal, A., Kumar, T.S. and Kumar, T.G. (2023). Detection of Network Attacks using Machine Learning and Deep Learning Models. *Procedia Computer Science*, 57-66.

Dong, S. and Sarem, M. (2022). DDoS Attack Detection Method Based on Improved KNN with the Degree of DDoS Attack in Software-Defined Networks. in *IEEE Journals & Magazine,* IEEE Xplore.

Doriguzzi-Corin, R., Miller, S. and Scott-Hayward, S. (2020). Lucid: A Practical, Lightweight Deep Learning Solution for DDoS Attack Detection. *IEEE Journals & Magazine,* IEEE Xplore.

Elsayed, M.S. (2020). DDoSNet: A Deep-Learning Model for Detecting Network Attacks.

Ferdous, F.S., Biswas, T. and Jony, A.I. (2024). Enhancing Cybersecurity: Machine Learning Approaches for Predicting DDoS Attack. *Malaysian Journal of Science and Advanced Technology*, 4(3), 249-255.

Hameed, S. and Ali, U. (2018). HADEC: Hadoop-Based Live DDoS Detection Framework. *EURASIP Journal on Information Security*.

Hamim, S.A. and Jony, A.I. (2024). Enhanced Deep Learning Model Architecture for Plant Disease Detection in Chilli Plants. *Journal of Edge Computing*, 3(2), 136-146. Available from: https://doi.org/10.55056/jec.758

Hamim, S.A. and Jony, A.I. (2024). Enhancing Brain Tumor MRI Segmentation Accuracy and Efficiency with Optimized U-Net Architecture. *Malaysian Journal of Science and Advanced Technology*, 4(3), 197-202.

Hoque, N. (2014). MIFS-ND: A Mutual Information-Based Feature Selection Method. *Expert Systems with Applications*, 41, 6371-6385.

Hoque, N. and Bhattacharyya, D.K. (2015). Botnet in DDoS Attacks: Trends and Challenges. *IEEE Journals & Magazine,* IEEE Xplore.

Jony, A.I. and Arnob, A.K.B. (2024). A Long Short-Term Memory Based Approach for Detecting Cyber Attacks in IoT using CIC-IoT2023 Dataset. *Journal of Edge Computing*, 3(1), 28-42. Available: https://doi.org/10.55056/jec.648

Jony, A.I. and Arnob, A.K.B. (2024). Deep Learning Paradigms for Breast Cancer Diagnosis: A Comparative Study on Wisconsin Diagnostic Dataset. *Malaysian Journal of Science and Advanced Technology*, 4(2), 109-117.

Jony, A.I. and Arnob, A.K.B. (2024). Securing the Internet of Things: Evaluating Machine Learning Algorithms for Detecting IoT Cyberattacks Using CIC-IoT2023 Dataset. *International Journal of Information Technology and Computer Science*, 16(4), 56-65.

Jony, A.I. and Hamim, S.A. (2024). Navigating the Cyber Threat Landscape: A Comprehensive Analysis of Attacks and Security in the Digital Age. *Journal of Information Technology and Cyber Security*, January 10. Available: https://doi.org/10.30996/jitcs.9715

Kaur, H., Behal, S. and Kumar, K. (2015). Characterization and Comparison of Distributed Denial of Service Attack Tools. in *IEEE Conference Publication,* IEEE Xplore.

Koroniotis, N., Moustafa, N., Sitnikova, E. and Turnbull, B. (2019). Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset. *Future Generation Computer Systems*, 779-796.

Levy, E. (2003). The Making of a Spam Zombie Army: Dissecting the Sobig Worms. *IEEE Journals & Magazine*, IEEE Xplore.

Li, Z., Zou, D. and Xu, S. (2022). SySeVR: A Framework for Using Deep Learning to Detect Software Vulnerabilities. *IEEE Journals & Magazine*, IEEE Xplore.

Lisun-Ul-Islam, M., Rahat, M.R.H., Esha, S., Faiyaz, A. and Jony, A.I. (2023). Hourly Air Quality Prediction in Dhaka City Using Time Series Forecasting Techniques: Deep Learning Perspectives. *Tuijin Jishu/Journal of Propulsion Technology*, 44(5), 568-579.

Liu, J., Xiao, Y., Ghaboosi, K., Deng, H. and Zhang, J. (2009). Botnet: Classification, Attacks, Detection, Tracing, and Preventive Measures. *Eurasip Journal on Wireless Communications and Networking*.

Liu, Y., Mao, S., Mei, X., Yang, T. and Zhao, X. (2019). Sensitivity of Adversarial Perturbation in Fast Gradient Sign Method. in *IEEE Conference Publication,* IEEE Xplore.

Mohsin, M. and Jony, A.I. (2024). A Comparative Analysis of Medical IoT Device Attacks Using Machine Learning Models. *Malaysian Journal of Science and Advanced Technology*, 4(4), 429-439.

Muraleedharan, N. and Janet, B. (2021). A Deep Learning Based HTTP Slow DoS Classification Approach Using Flow Data. *ICT Express*, 210-214.

Purwono, A. Wulandari, Ma'arif, A. and Salah, W. (2025). Understanding Generative Adversarial Networks (GANs): A Review. *Control Systems and Optimization Letters*, 3(1), 36-45. doi: 10.59247 / csol.v3i1.170.

Raja Sree, T. and Saira Bhanu, S.M. (2017). Detection of HTTP Flooding Attacks in Cloud Using Dynamic Entropy Method. *Arabian Journal for Science and Engineering*, 6995-7014.

UNSW Canberra. (2021). Bot-IoT Dataset. https: / / research.unsw.edu.au / projects / bot-iot-dataset

Wang, M., Lu, Y. and Qin, J. (2020). A Dynamic MLP-Based DDoS Attack Detection Method Using Feature Selection and Feedback. *Computers & Security*.

Cite this article as: Mubashir Mohsin and Akinul Islam Jony (2026). BotVision: Robust BotNet Detection Using Gan-Augmented Transfer Learning on Adversarial Image Data. *International Journal of Data Science and Big Data Analytics,* 6(1), 1-16. doi: 10.67191 / IJDSBDA.6.1.2026.1-16.