Journal of Information Hiding and Multimedia Signal Processing ©2025 ISSN 2073-4212
Ubiquitous International Volume **16**, Number **1**, March **2025**

Detection of Adversarial Attacks Using Enhanced Density Metrics

Thanh-Tuan Nguyen[1]*, Thuc-Minh Bui[1], Thi-Thom Hoang[1], Thanh-Vinh Nguyen[1], Xuan-Huy Nguyen[1]

[1]Faculty of Electrical and Electronics,
Nha Trang University, Nha Trang 650000, Vietnam
tuannt@ntu.edu.vn, minhbt@ntu.edu.vn, thomht@ntu.edu.vn, vinhnt@ntu.edu.vn, huynx@ntu.edu.vn

Chin-Shiuh Shieh[2], Mong-Fong Horng[2], Thanh-Lam Nguyen[2]

[2]Department of Electronic Engineering,
National Kaohsiung University of Science and Technology, Kaohsiung 807618, Taiwan
csshieh@nkust.edu.tw, mfhorng@nkust.edu.tw, f112152194@nkust.edu.tw

*Corresponding author: Thanh-Tuan Nguyen

Received December 11, 2024, revised January 9, 2025, accepted January 11, 2025.

Abstract. *Adversarial attacks pose significant challenges to intrusion detection systems (IDS) by exploiting vulnerabilities in machine learning models. This study proposes an innovative framework, DDM-CNN, which integrates an enhanced density metric for Out-of-Distribution (OOD) detection with Incremental Learning to address these threats. Adversarial datasets were generated using Conditional Tabular GAN (CTGAN) to evaluate the robustness of the model against challenging attack scenarios. Experimental results on CICIDS2017 and CICDDoS2019 datasets demonstrate the superiority of DDM-CNN, achieving F1 Scores of 0.996 and 0.997 for normal data and 0.9379 and 0.9683 for adversarial data, respectively. The model outperformed baseline approaches, including CNN, RNN, MLP, and AE, in terms of accuracy and resilience. This framework highlights the importance of advanced OOD detection metrics and adaptive learning mechanisms in fortifying IDS against evolving adversarial threats.*
Keywords: Adversarial Attack, Incremental Learning, Enhanced Density Metrics, Intrusion Detection Systems, CTGAN, OOD

1. **Introduction.** The exponential growth of cyber threats has transformed network security into a critical area of research, particularly in the context of machine learning applications. Intrusion detection systems (IDS) powered by machine learning have shown remarkable success in identifying malicious activities across diverse network environments. However, adversarial attacks remain a significant challenge, exploiting the vulnerabilities of machine learning models by crafting malicious inputs to evade detection and compromise system integrity [1]. These attacks not only undermine the reliability of IDS but also highlight the need for more resilient defense mechanisms that can adapt to rapidly evolving threats. Adversarial data, often generated using advanced techniques like Conditional Tabular GANs (CTGAN), is specifically designed to deceive machine learning models by mimicking legitimate patterns while embedding malicious intent [2]. Detecting such data is critical to maintaining the reliability of IDS in real-world scenarios. Out-of-Distribution (OOD) detection has emerged as a promising approach to identify such anomalies by recognizing inputs that deviate from the established data distribution [3].

Despite this potential, traditional OOD detection methods face challenges in scalability and accuracy, particularly when dealing with high-dimensional network data [4].

To address these limitations, this paper introduces a novel defense framework that combines an improved density-based metric for OOD detection with incremental learning. The proposed density metric enhances the ability to distinguish between adversarial and legitimate inputs, while the incremental learning framework allows continuous model updates with newly labeled OOD data. This approach ensures the adaptability of the model without requiring costly retraining on the entire dataset [5]. Furthermore, the integration of CTGAN-generated adversarial samples provides a robust testing ground for evaluating the effectiveness of the proposed framework under realistic attack scenarios. The framework is evaluated on two widely recognized datasets, CICIDS2017 and CICDDoS2019, which encompass a comprehensive range of attack types, including Distributed Denial-of-Service (DDoS), brute force, and web attacks [6]. Comparative analysis with baseline models Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Multilayer Perceptrons (MLP), and Autoencoders demonstrates that the proposed method significantly outperforms these approaches in terms of accuracy, resilience, and the ability to adapt to adversarial inputs.

The rest of this paper is structured as follows: Section 2 provides an overview of related work on adversarial attack defenses and OOD detection. Section 3 details the proposed methodology, including the CTGAN-based adversarial data generation, the improved density metric, and the incremental learning framework. Section 4 presents the experimental setup and results, including comparisons with baseline models. Section 5 gives a deep discussion on experiment ressults. Finally, Section 6 concludes the paper and discusses potential directions for future research.

2. **Related Works.**

2.1. **Advancements in AI for DDoS Attack Detection.**

The rising complexity and frequency of DDoS attacks have prompted significant advancements in artificial intelligence (AI)-based detection mechanisms. Traditional methods often struggle to cope with the dynamic and large-scale nature of these attacks, making AI-powered solutions increasingly essential for maintaining network security.

Machine learning (ML) techniques have been a primary focus in DDoS detection. Supervised learning models, such as decision trees and ensemble methods, have shown effectiveness in analyzing traffic patterns and distinguishing legitimate traffic from attack traffic [7]. Ensemble methods like Gradient Boosting and Random Forest are particularly noted for their ability to handle large, imbalanced datasets commonly found in network traffic logs [8]. Deep learning (DL) approaches further enhance DDoS detection capabilities by modeling complex, non-linear patterns in data. Long Short-Term Memory (LSTM) networks, for instance, excel at capturing temporal dependencies in sequential data, making them suitable for identifying the gradual buildup of DDoS attacks [9]. Additionally, autoencoders, particularly LSTM-based ones, are widely used for anomaly detection by learning normal network behavior and flagging deviations as potential threats [10]

Generative Adversarial Networks (GANs) have emerged as powerful tools in improving DDoS detection. By generating synthetic adversarial samples, GANs enable models to train on diverse attack scenarios, enhancing their robustness and generalization capabilities [11]. This approach allows for better preparation against previously unseen attack strategies, a critical aspect of modern network defense systems.Hybrid models combining ML and DL techniques have also been developed to leverage their respective strengths. For example, models integrating CNN with LSTM networks can extract both spatial and

temporal features, achieving improved accuracy in classifying network traffic [12]. Such hybrid approaches are particularly effective in capturing the multi-faceted nature of DDoS attacks.

AI-based solutions have demonstrated particular success in Software Defined Networking (SDN) environments. By leveraging the centralized control and programmability of SDNs, AI algorithms can rapidly analyze traffic flows and detect abnormal patterns indicative of DDoS attacks [13]. This capability allows for real-time detection and mitigation, making AI an indispensable tool in modern network security frameworks.Despite these advancements, challenges remain. Key issues include the need for large labeled datasets, the susceptibility of AI models to adversarial attacks, and the computational overhead of real-time processing. Addressing these challenges requires ongoing research into unsupervised and semi-supervised learning techniques, the development of robust models resistant to adversarial perturbations, and the optimization of computational efficiency. AI has profoundly transformed the field of DDoS detection, offering innovative solutions that surpass traditional methods. The integration of ML, DL, and GAN-based techniques has significantly enhanced detection accuracy and adaptability, paving the way for more resilient and scalable network defenses. Future advancements in AI research hold the promise of further fortifying defenses against the evolving landscape of cyber threats.

2.2. **Challenges of Adversarial DDoS Attacks.** Adversarial DdoS attacks represent a significant evolution in the tactics used by attackers to compromise network systems. These attacks leverage adversarial machine learning techniques to craft malicious inputs that are specifically designed to evade detection by traditional and machine learning-based IDS. By introducing imperceptible perturbations to malicious traffic, adversarial DDoS attacks make it difficult for models to distinguish between legitimate and harmful traffic, thereby exploiting vulnerabilities in detection mechanisms [14].

GAN have emerged as a prominent tool in the synthesis of adversarial DDoS attack data. These networks are capable of generating high-quality synthetic traffic that mimics the statistical properties of legitimate network flows while embedding malicious intent. Recent work has demonstrated the use of CTGAN and Wasserstein GANs (WGANs) with Gradient Penalty to create adversarial traffic that bypasses conventional detection methods [15]. These advanced models have shown that even state-of-the-art intrusion detection systems can be effectively deceived when exposed to such adversarial data [16]. The adaptive nature of adversarial DDoS attacks presents a dynamic challenge for network defense. Standard IDS often rely on static rule-based systems or pre-trained models, which lack the ability to adapt to new and evolving attack patterns. This limitation is particularly concerning as adversarial techniques continue to become more sophisticated, leveraging advancements in AI to stay ahead of defensive measures [17].

To counter these threats, researchers have proposed various approaches, including adversarial training and the use of hybrid detection frameworks. Adversarial training involves exposing detection models to adversarially crafted samples during the training process, thereby improving their resilience to such inputs. However, while adversarial training can enhance robustness, it often requires extensive computational resources and large labeled datasets, which can limit its scalability in real-world applications [18]. Hybrid detection frameworks, combining traditional machine learning algorithms with deep learning models, have also shown promise. For instance, integrating LSTM networks with CNN enables systems to capture both temporal and spatial features in network traffic, making them more effective against adversarial DDoS attacks [19]. Additionally, the use of dual-discriminator GANs (GANDD) has been explored to simultaneously generate and

detect adversarial traffic, creating a more dynamic and comprehensive defense mechanism [20].

Despite these advancements, there remain critical challenges in detecting and mitigating adversarial DDoS attacks. The ability of these attacks to mimic legitimate traffic patterns with high fidelity complicates the task of distinguishing between normal and malicious flows. Furthermore, as adversarial techniques evolve, the risk of generalization errors in detection systems increases, potentially leading to higher rates of false positives and false negatives [21]. These challenges highlight the need for continued innovation in adversarial defense strategies, with a focus on developing more adaptive and scalable solutions capable of addressing the rapidly changing threat landscape.

2.3. **Open-Set Recognition in Machine Learning.** Open-Set Recognition (OSR) addresses the challenge where models must accurately classify known classes while effectively identifying and managing instances from unknown classes not encountered during training. This paradigm reflects real-world scenarios where the assumption that all possible classes are known a priori is impractical. Traditional closed-set classification systems operate under the premise that all test instances belong to predefined categories, leading to potential misclassification when novel classes are introduced. OSR mitigates this issue by enabling models to detect and appropriately handle previously unseen classes, thereby enhancing robustness and reliability in dynamic environments. Recent advancements in OSR have focused on integrating probabilistic generative models to improve the detection of unknown classes. For instance, Conditional Probabilistic Generative Models (CPGM) have been proposed to incorporate discriminative information, allowing for the effective identification of both known and unknown classes. These models force latent features to approximate conditional Gaussian distributions, facilitating more accurate recognition outcomes [22].

Another significant development is the application of meta-learning techniques to OSR. The PEELER algorithm exemplifies this approach by employing random selection of novel classes per episode and maximizing posterior entropy for those classes. This method enhances the model's ability to generalize from limited data, thereby improving its capacity to recognize and appropriately respond to unknown classes [23]. Hybrid models have also been explored to address the complexities of OSR. By combining discriminative classifiers with generative models, these hybrid systems aim to jointly learn representations that are effective for both classifying known categories and detecting unknown instances. This dual capability is crucial for applications requiring high reliability in the presence of unforeseen data [24]. Despite these advancements, OSR remains a challenging field due to the inherent unpredictability of unknown classes and the need for models to balance sensitivity and specificity. Ongoing research continues to explore innovative methodologies to enhance the efficacy of open-set recognition systems.

2.4. **Density Metric for Out-of-Distribution Detection.** The development of metrics to assess OOD data has evolved significantly, driven by advancements in machine learning and the increasing need for robust detection methods. One of the earliest approaches focused on evaluating generative models through single-score metrics like the Fréchet Inception Distance (FID), which became widely used for comparing the distributions of real and generated data. FID offered insights into the quality of data synthesis but had limitations in addressing fidelity and diversity trade-offs [25]. Recognizing these shortcomings, subsequent research introduced dual metrics, notably precision and recall, to disentangle fidelity (the quality of generated samples) from diversity (the extent to which the generative model covers the variability of real data). However, these metrics

faced practical limitations, including sensitivity to outliers and dependency on hyperparameter settings. To mitigate these issues, improved methods like k-nearest neighbors (k-NN) based evaluation were proposed, enhancing robustness against noise and better capturing the nuances of data distributions [26]. Building on these advancements, density and coverage metrics emerged as more reliable alternatives. These metrics addressed critical flaws in earlier methods by redefining neighborhood-based estimations. Density measures the concentration of generated samples within regions densely populated by real data, offering insights into fidelity, while coverage quantifies the proportion of real data distribution covered by the generated samples, emphasizing diversity. This dual approach provided a more comprehensive evaluation framework for generative models and their OOD handling capabilities [27].

In the context of adversarial and open-set recognition tasks, density-based methods have been pivotal. Recent studies have utilized density estimations to enhance the detection of OOD instances by measuring the likelihood of data points against known distributions. By leveraging these metrics, researchers have been able to distinguish in-distribution and OOD samples with greater accuracy, showcasing the utility of density-driven frameworks in security-sensitive applications. Building upon this foundation, our study proposes an upgraded density metric specifically tailored for detecting unknown data in DDoS attack scenarios. This metric enhances the precision of OOD identification by incorporating novel generative techniques and optimizing neighborhood estimations, thereby addressing gaps in existing detection frameworks.

3. **Methodology.**

In this section, we outline the framework developed to address the challenges of adversarial data and OOD detection in network defense systems. The proposed methodology is designed to systematically construct adversarial datasets and enhance classification models through innovative techniques.

First, a robust adversarial dataset is generated using CTGAN based on two comprehensive and diverse original datasets: CICIDS2017 and CICDDoS2019. These datasets serve as a foundation for simulating realistic attack scenarios. The adversarial dataset is then used to evaluate the defensive capabilities of prominent deep learning architectures, ensuring the assessment of model robustness under adversarial conditions.

Second, the study introduces a novel classification framework that incorporates an upgraded version of the Density metric. This advanced metric is integrated with incremental learning techniques, allowing the model to adapt dynamically to newly identified OOD data. This combination enables the classifier to not only detect adversarial data effectively but also improve its accuracy over time without requiring extensive retraining.

The subsequent sections detail the implementation of these steps, including the adversarial data generation process, the integration of the enhanced Density metric, and the incremental learning mechanism. These components collectively contribute to a robust and adaptive defense model designed to tackle evolving adversarial threats in network environments.

3.1. **CTGAN.**

CTGAN is a specialized generative model designed to address the challenges of synthesizing realistic tabular data, particularly when dealing with highly imbalanced categorical distributions and non-Gaussian continuous features. The workflow of CTGAN, illustrated in the provided diagram, integrates conditional data generation and training strategies to ensure high fidelity and diversity in generated data. At its core, CTGAN utilizes mode-specific normalization, a conditional generator, and the Wasserstein GAN with

Gradient Penalty (WGAN-GP) framework. To tackle the challenges of modeling continuous features, CTGAN employs mode-specific normalization, which leverages a Variational Gaussian Mixture Model (VGM). This approach segments each continuous column into multiple modes, each represented by a Gaussian distribution. For a continuous variable C_i with observed values $c_{i,j}$, the probability density function is modeled as:

$$P_{C_i}(c_{i,j}) = \sum_{k=1}^{m} \pi_k \mathcal{N}(c_{i,j}; \mu_k, \sigma_k^2), \tag{1}$$

where π_k represents the weight of the k-th mode, and $\mathcal{N}(c; \mu_k, \sigma_k^2)$ is the Gaussian distribution with mean μ_k and variance σ_k^2. This normalization ensures that both global distribution characteristics and local variations are preserved, enabling the generator to produce continuous features that align with the original data distribution.

The core of CTGAN is its conditional generator, which allows data generation conditioned on specific categories of a discrete column. Given a dataset $\mathbf{X}$, let D_i be a categorical column with categories $d_{i,1}, d_{i,2}, \ldots, d_{i,k}$. During training, a category $d_{i,k}$ is randomly selected, and rows where $D_i = d_{i,k}$ are used to construct a conditional vector $\mathbf{c}$. For the generator G, this condition vector $\mathbf{c}$ is concatenated with a noise vector $\mathbf{z} \sim \mathcal{N}(0, I)$ to form the input:

$$\mathbf{x}_{\text{fake}} = G([\mathbf{z}, \mathbf{c}]), \tag{2}$$

where G is implemented as a series of fully connected layers interleaved with batch normalization and activation functions.

To address the issue of class imbalance in categorical data, CTGAN introduces a training-by-sampling strategy. A categorical column D_i is selected, and a category $d_{i,k}$ is sampled according to a probability mass function $P(D_i = d_{i,k})$, which is calculated as:

$$P(D_i = d_{i,k}) = \frac{\log(\text{freq}(d_{i,k}) + 1)}{\sum_{j=1}^{K} \log(\text{freq}(d_{i,j}) + 1)}, \tag{3}$$

where $\text{freq}(d_{i,k})$ represents the frequency of category $d_{i,k}$ in the dataset. This ensures balanced training by prioritizing underrepresented categories, preventing the generator from collapsing to dominant classes.

$$L_C = \mathbb{E}_{\mathbf{x}_{\text{real}} \sim P_{\text{data}}} \left[C(\mathbf{x}_{\text{real}}, \mathbf{c}) \right] - \mathbb{E}_{\mathbf{x}_{\text{fake}} \sim P_G} \left[C(\mathbf{x}_{\text{fake}}, \mathbf{c}) \right], \tag{4}$$

with a gradient penalty term:

$$L_{GP} = \lambda \mathbb{E}_{\hat{\mathbf{x}} \sim P_{\hat{\mathbf{x}}}} \left[\left(\| \nabla_{\hat{\mathbf{x}}} C(\hat{\mathbf{x}}) \|_2 - 1 \right)^2 \right], \tag{5}$$

where $\hat{\mathbf{x}}$ is sampled uniformly along straight lines between real and generated samples. The total loss for the critic enforces the Lipschitz constraint and stabilizes training:

$$L_{\text{total}} = L_C + L_{GP}. \tag{6}$$

Figure 1 depicts the overall workflow of CTGAN. A categorical column D is selected, and a specific category is sampled. The corresponding rows from the dataset are used to construct the conditional vector $\mathbf{c}$, which is then concatenated with noise $\mathbf{z}$ to generate synthetic data via the generator G. The critic C evaluates the synthetic samples against real data and provides feedback to refine G.

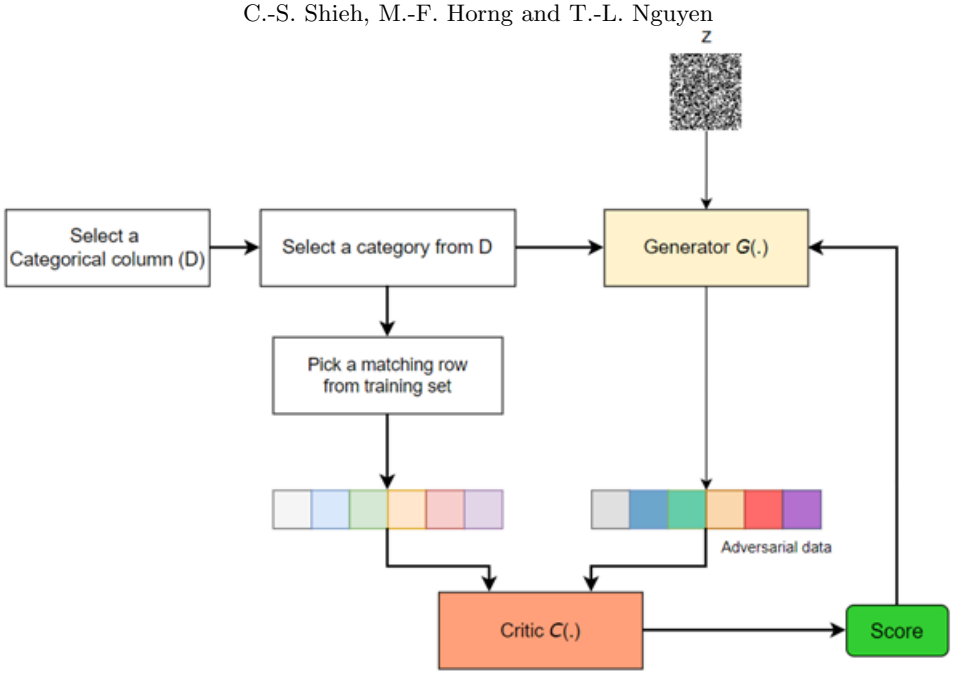

FIGURE 1. The architecture and workflow of CTGAN.

In this study, CTGAN is employed to generate adversarial datasets based on CICIDS2017 and CICDDoS2019, enabling the evaluation of an upgraded Density metric. This enhanced metric is specifically designed to improve the detection of out-of-distribution data in network defense scenarios, demonstrating the practical applicability of CTGAN-generated data in real-world security challenges.

3.2. **Enhanced Density Metric.** Building on the advancements of density-based metrics for OOD detection, we introduce the Distance Density Metric (DDM) that refines traditional approaches by introducing a distance-weighted contribution mechanism. Unlike prior methods that treat all neighbors equally, DDM prioritizes neighbors closer to the evaluation point, ensuring that high-density regions are emphasized while mitigating the influence of sparsely distributed points and outliers. This enhancement allows for more precise OOD identification, particularly in high-dimensional and noisy datasets. The DDM is mathematically expressed as:

$$\text{DDM} = \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{1_{Y_j \in B(X_i, \text{NN}_D^k(X_i))}}{1 + \text{dist}(Y_j, X_i)} \tag{7}$$

where M is the number of known (in-distribution) samples $\{Y_j\}$, N represents the number of unknown samples X_i, and $\text{dist}(Y_j, X_i)$ is the Euclidean distance between a known sample Y_j and an unknown sample X_i. The indicator function $1_{Y_j \in B(X_i, \text{NN}_D^k(X_i))}$ evaluates to 1 if Y_j belongs to the k-nearest neighbors of X_i within the known data distribution and 0 otherwise. The set $B(X_i, NN_b^k(X_i))$ denotes the k-nearest neighbors of X_i in the known data Y_j, determined by a chosen distance metric.

This metric enhances OOD detection by weighting each neighbor's contribution inversely to its distance from the evaluation point. The denominator $1+\text{dist}(Y_j, X_i)$ ensures closer neighbors exert greater influence, while distant ones contribute less, improving sensitivity to local data variations and robustness against noise. Applying DDM involves finding the k-nearest neighbors of each unknown sample X_i from the known samples Y_j,

computing distances, and weighting contributions. The aggregated result captures local density structures, making it effective for OOD detection. By incorporating distance weighting, DDM reduces outlier impact and better represents local densities, addressing challenges in adversarial and high-dimensional datasets where traditional metrics often fail. Integral to the proposed framework, DDM enables precise OOD identification in CTGAN-generated adversarial datasets and, combined with incremental learning, enhances model robustness and adaptability for advanced network defense.

3.3. **Proposed Model.** The proposed model, referred to as *DDM-CNN* and illustrated in Figure 2, combines the DDM with a CNN-based classifier and incremental learning to enable robust OOD detection and continuous adaptability. The workflow is composed of multiple components, including data preprocessing, anomaly detection via DDM, classification through a CNN model, and iterative updating using expert feedback.

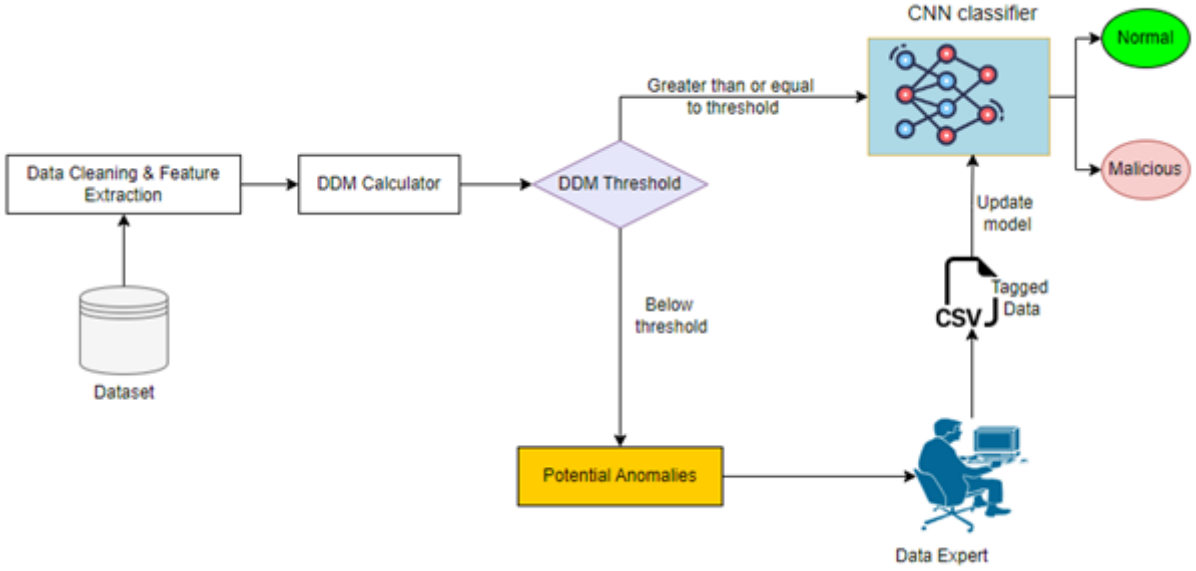

FIGURE 2. Workflow of the proposed model.

The process begins with data preprocessing, where the raw dataset $\mathbf{X}$ undergoes cleaning, normalization, and feature extraction to produce structured data suitable for analysis. The preprocessed samples are then evaluated using DDM, defined in Formula (1). DDM evaluates each sample's density relative to the known data, assigning higher scores to samples closer to dense regions of the distribution. Samples with $\mathrm{DDM}(X_i) \geq \tau$ (where τ is a predefined threshold) are passed to the CNN classifier, while those below the threshold are flagged as potential anomalies.

The CNN classifier receives samples passing the DDM threshold and categorizes them as either normal or malicious. The classifier is trained using cross-entropy loss:

$$\mathcal{L}_{\mathrm{CNN}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log \hat{y}_{i,c}, \tag{8}$$

where $y_{i,c}$ and $\hat{y}_{i,c}$ denote the true and predicted labels for class c, respectively. The CNN's predictions are then validated or refined through a feedback loop involving expert tagging.

Incremental learning is applied to integrate new knowledge without retraining the entire model from scratch. Samples identified as anomalies are reviewed and labeled by a data

expert. These newly labeled samples $\mathbf{X}_{\text{new}}$ are added to the dataset, and the CNN is updated incrementally. The optimization during incremental learning is defined as:

$$\mathcal{L}_{\text{incremental}} = \mathcal{L}_{\text{CNN}} + \lambda \left\| \mathbf{W}_{\text{new}} - \mathbf{W}_{\text{old}} \right\|^2, \tag{9}$$

where $\mathbf{W}_{\text{new}}$ and $\mathbf{W}_{\text{old}}$ represent the updated and previous model weights, and λ controls the trade-off between retaining old knowledge and learning new information.

The iterative feedback loop ensures the continuous improvement of the model by incorporating expert-labeled anomalies, enabling adaptability to dynamic changes in the data. This integration of DDM with incremental learning allows for precise OOD detection and efficient model updates, providing a robust solution for evolving network defense challenges.

4. Experiment and Results.

4.1. Dataset.

In this study, the evaluation of our IDS was conducted using two datasets: CICIDS2017 and CICDDoS2019. Both datasets, developed by the Canadian Institute for Cyber Security (CIC), are widely recognized in the cybersecurity field. Their purpose is to replicate real-world network activities and various cyberattack scenarios in a controlled environment. These datasets include genuine traffic data combined with network configurations, providing a comprehensive setup for examining IDS effectiveness, refining algorithms, and extracting key features.

The CICIDS2017 dataset contains traffic data recorded on Wednesday and Friday, offering a blend of benign and malicious network activities. On Wednesday, it includes 319,186 benign packets (64.26%) and 159,049 packets from DoS Hulk attacks (32.021%), along with smaller volumes of attacks such as DoS GoldenEye (7,647 packets), DoS Slowloris (5,071 packets), and DoS Slowhttptest (5,109 packets). Additionally, 11 packets are related to the HeartBleed vulnerability. On Friday, the dataset logs 128,027 benign packets (56.713%) and 97,718 packets corresponding to DDoS attacks (43.287%).

The CICDDoS2019 dataset, on the other hand, focuses on DDoS attacks and other network threats. A notable example is the LDAP attack, which includes 2,179,928 packets (99.927%) compared to only 1,602 benign packets (0.073%). Other prominent attacks recorded in the dataset are MSSQL (5,071,002 packets), NetBIOS (4,093,273 packets), and UDP (3,134,643 packets), with additional records for NTP and SYN attacks.

These datasets form a crucial basis for training, validating, and benchmarking IDS models. They also enable researchers to compare the performance of various detection algorithms, refine feature selection methods, and test the robustness of IDS models under different conditions.

4.2. Evaluation Metrics.

In machine learning, evaluation metrics are critical tools for determining the effectiveness of a model and assessing its efficiency. They provide a structured way to analyze and compare model performance under specific challenges and tasks. In this research, several key metrics were utilized to evaluate the proposed model:

Accuracy (Acc): This metric calculates the proportion of correct predictions made by the model out of the total number of predictions. It gives an overall sense of the model's performance in distinguishing between classes.

Precision (Prec): Precision measures the ratio of true positive predictions to the total number of positive predictions made by the model, indicating how many of the predicted positives were actually correct.

Recall: Also known as sensitivity, recall measures the proportion of actual positive instances correctly identified by the model, reflecting its ability to capture relevant positive cases.

F1 Score (F1): The F1 score is the harmonic mean of precision and recall, providing a balanced evaluation by considering both false positives and false negatives. This metric is particularly useful when the dataset has imbalanced class distributions.

The mathematical formulations for these metrics are presented below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{10}$$

$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

$$Recall = \frac{TP}{TP + FN}, \tag{12}$$

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{13}$$

where: TP represents the number of true positive predictions, TN represents the number of true negative predictions, FP represents the number of false positive predictions, FN represents the number of false negative predictions.

4.3. **Experiment Environments.** For this research, the experiments were conducted on a Windows 11 system equipped with a high-performance configuration to facilitate efficient model training and evaluation. The system is powered by an Intel Core i7-14700F processor with 20 cores and a turbo boost of up to 5.4 GHz, paired with 64 GB of DDR5 G.Skill Trident Z5 RAM (5600 MHz). Graphics processing was handled by an NVIDIA GTX 4070, ensuring smooth execution of computationally intensive tasks.

The framework employed for implementation and evaluation included PyTorch 2.0.1 + cu118 for model construction and Sklearn 1.3.0 for performance metric computation. Python 3.9 was utilized for programming and testing. To ensure robust results, ten training iterations were conducted, each with a different random seed to introduce variability in the training process. The Adam optimizer was chosen to improve the learning efficiency of the model. Detailed configurations and parameter settings are outlined in Table 1 and Table 2.

Table 1. CNN Classifier configuration

Layer	**Configuration**
Input	(None, 1, 9, 9)
Conv2D	(None, 120, 9, 9)
Dropout	(None, 120, 9, 9)
Conv2D	(None, 60, 9, 9)
Dropout	(None, 60, 9, 9)
Conv2D	(None, 30, 9, 9)
Dropout	(None, 30, 9, 9)
Flatten	(None, 2430)
Dense	(None, 1)
Dropout	(None, 1)
Dense	(None, 1)
Sigmoid	(None, 1)

TABLE 2. Training parameter settings system.

Parameters	Value
Learning Rate	0.0001
Weight Decay	0.0003
Optimizer	Adam
Batch Size	1024
Training Split Ratio	0.7 training, 0.3 testing
DDM Threshold	0.9
Random Seeds	2, 16, 32, 104, 587, 1023, 1502, 2041, 3412, 4105, 5213, 6512, 7103, 8245, 9013, 99532

4.4. **Evaluation Results on Coventional Attack.** Building on the evaluation framework, the proposed DDM-CNN model was compared against four widely recognized baseline models CNN, RNN, MLP, and AE on the CICIDS2017 and CICDDoS2019 datasets. Each baseline was chosen for its established relevance to network intrusion detection. CNN, implemented by Halbouni et al. [28] in 2022, utilize spatial feature extraction to detect malicious traffic effectively. Recurrent Neural Networks (RNN), employed by Ibrahim and Elhafiz [29] in 2023, excel at modeling sequential patterns, making them suitable for detecting time-dependent attacks. Multilayer Perceptrons (MLP), optimized by Ali et al. [30] in 2024 offer robust classification with improved computational efficiency. Autoencoders (AE), as applied by Singh and Jang-Jaccard [31] in 2022, leverage reconstruction-based anomaly detection to identify deviations in network behavior. The evaluation results are summarized in Table 3 and Table 4, which present the performance metrics for CICIDS2017 and CICDDoS2019, respectively.

TABLE 3. Performance Metrics on CICIDS2017

Model	Accuracy	Precision	Recall	F1 Score
CNN [28]	0.944	0.945	0.941	0.943
RNN [29]	0.933	0.935	0.932	0.933
MLP [30]	0.912	0.895	0.921	0.893
AE [31]	0.880	0.885	0.881	0.883
DDM-CNN	**0.9971**	**0.9927**	**0.9994**	**0.996**

TABLE 4. Performance Metrics on CICDDoS2019

Model	Accuracy	Precision	Recall	F1 Score
CNN [28]	0.954	0.955	0.951	0.953
RNN [29]	0.943	0.945	0.942	0.944
MLP [30]	0.922	0.905	0.931	0.913
AE [31]	0.890	0.895	0.891	0.893
DDM-CNN	**0.9994**	**0.9999**	**0.9995**	**0.997**

These results clearly illustrate the superior performance of the proposed DDM-CNN model compared to the baseline models. For CICIDS2017, DDM-CNN achieved the highest F1 Score of 0.996, significantly outperforming CNN (0.943), RNN (0.933), MLP (0.893), and AE (0.883). Similarly, on CICDDoS2019, DDM-CNN recorded an F1 Score of 0.997, demonstrating its capability to generalize effectively across datasets while maintaining robust performance.

4.5. **Evaluation Results on Adversarial Attack.** To further evaluate the robustness of the proposed DDM-CNN model, experiments were conducted using adversarial datasets generated by CTGAN. This technique generates synthetic adversarial samples by mimicking the underlying data distribution while introducing perturbations that create challenging scenarios for intrusion detection. The adversarial datasets were derived from CICIDS2017 and CICDDoS2019, representing diverse and complex network threats. The evaluation results, summarized in Table 5 and Table 6, demonstrate the effectiveness of the proposed model in handling these adversarial challenges, significantly outperforming baseline models.

Table 5. Performance Metrics on Adversarial Data from CICIDS2017

Model	Accuracy	Precision	Recall	F1 Score
CNN [28]	0.1823	0.1917	0.1774	0.1844
RNN [29]	0.1715	0.1832	0.1598	0.1709
MLP [30]	0.1532	0.1628	0.1375	0.1489
AE [31]	0.1029	0.1125	0.0963	0.1038
DDM-CNN	**0.9123**	**0.9095**	**0.9681**	**0.9379**

Table 6. Performance Metrics on Adversarial Data from CICDDoS2019

Model	Accuracy	Precision	Recall	F1 Score
CNN [28]	0.1867	0.1983	0.1851	0.1916
RNN [29]	0.1764	0.1876	0.1614	0.1728
MLP [30]	0.1624	0.1736	0.1428	0.1556
AE [31]	0.1417	0.1513	0.1389	0.1449
DDM-CNN	**0.9694**	**0.9978**	**0.9405**	**0.9683**

The results in Table 5 and Table 6 clearly demonstrate the superior performance of the proposed DDM-CNN model when tested against adversarial data. Notably, the DDM-CNN achieved an F1 Score of 0.9379 on CICIDS2017 adversarial data, significantly surpassing the baseline models such as CNN (0.1844), RNN (0.1709), MLP (0.1489), and AE (0.1038). Similarly, on CICDDoS2019 adversarial data, DDM-CNN recorded an F1 Score of 0.9683, which is far above the baseline models. These finding emphasize two key points: First, CTGAN-generated adversarial datasets create realistic and challenging scenarios, highlighting the need for robust detection. Second, the integration of the DDM-based OOD detection metric with Incremental Learning enhances adaptability and ensures superior performance, establishing DDM-CNN as a strong framework for adversarial intrusion detection.

5. **Discussion.** The evaluation results across the four datasets (CICIDS2017 and CICDDoS2019 with both normal and adversarial data) provide a comprehensive insight into the performance of the proposed DDM-CNN model. The results, summarized in Tables 3 and 4 for normal data, demonstrate that DDM-CNN consistently outperforms all baseline models. Notably, the F1 Scores of 0.996 and 0.997 for CICIDS2017 and CICDDoS2019, respectively, underline the model's robustness in distinguishing legitimate traffic from network intrusions. Tables 5 and 6 extend the analysis to adversarial datasets generated using CTGAN. These adversarial datasets were intentionally crafted to mimic realistic attack patterns, creating significantly more challenging detection scenarios. Despite this, DDM-CNN continues to achieve remarkably high performance, with F1 Scores of 0.9379

for CICIDS2017 and 0.9683 for CICDDoS2019. In contrast, the baseline models: CNN, RNN, MLP, and AE struggle significantly under adversarial conditions, highlighting their vulnerability. These results emphasize the robustness of DDM-CNN, particularly in handling adversarial attacks that exploit traditional classifiers' weaknesses.

The use of CTGAN to generate adversarial data highlights the strength of such attacks in challenging the reliability of intrusion detection systems. Adversarial samples closely mimic the original data distribution while introducing subtle perturbations, making detection significantly harder. The drop in performance for baseline models across adversarial datasets further underscores this point, demonstrating the ability of adversarial attacks to exploit vulnerabilities in conventional network defenses.

The superior performance of DDM-CNN is further amplified when coupled with the enhanced Density-based OOD detection mechanism. With a threshold set at 0.9, the model effectively filters out potential anomalies before classification. This threshold selection allows DDM-CNN to maintain a high true positive rate while minimizing false positives. The incremental learning component further ensures adaptability, enabling the model to refine itself continuously as new adversarial patterns emerge. This dynamic integration of DDM and incremental learning solidifies DDM-CNN's position as a robust framework for intrusion detection.

Figures 3 and 4 present the Receiver Operating Characteristic (ROC) curves for adversarial datasets derived from CICIDS2017 and CICDDoS2019, respectively. The Area Under the Curve (AUC) scores offer a detailed perspective on the detection capabilities of DDM-CNN compared to baseline models.

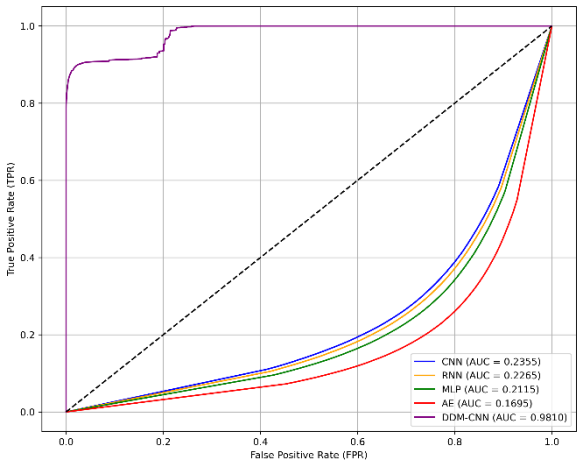

FIGURE 3. ROC curve for adversarial data on CICIDS2017.

For both datasets, DDM-CNN achieves AUC scores of 0.9810 (Figure 3) and 0.9992 (Figure 4), demonstrating its near-perfect ability to distinguish between normal and adversarial traffic. In contrast, baseline models such as CNN, RNN, MLP, and AE exhibit significantly lower AUC scores, reflecting their susceptibility to adversarial perturbations. The steep initial rise in the ROC curves of DDM-CNN indicates a high true positive rate at low false positive rates, affirming its reliability in real-world applications. The comparison across Figures 3 and 4 also highlights a consistent pattern: while DDM-CNN maintains exceptional performance across both datasets, baseline models exhibit varying degrees of degradation under adversarial conditions. This disparity underscores the effectiveness of the proposed model in mitigating adversarial threats, a critical requirement for modern intrusion detection systems.

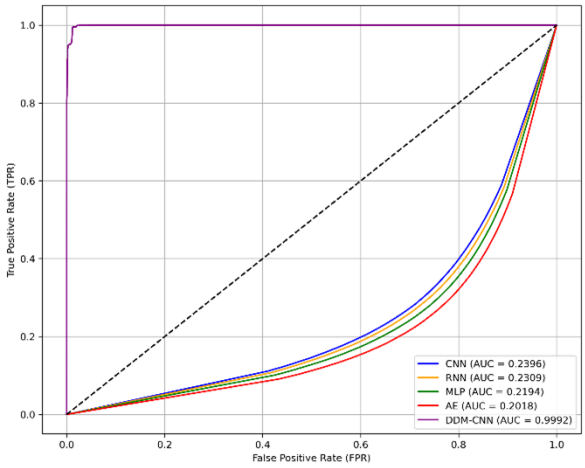

FIGURE 4. ROC curve for adversarial data on CICDDoS2019.

6. **Conclusion.** This study presented a robust intrusion detection framework, DDM-CNN, which integrates an enhanced Density-based OOD detection metric with Incremental Learning. The model demonstrated exceptional performance on both normal and adversarial datasets derived from CICIDS2017 and CICDDoS2019. By leveraging adversarial data generated using CTGAN, the evaluation highlighted the vulnerabilities of traditional deep learning models such as CNN, RNN, MLP, and AE, and showcased the superior adaptability and accuracy of the proposed method. The experimental results underscored the model's ability to maintain high F1 Scores and AUC values across challenging scenarios, confirming its robustness against adversarial attacks. The incorporation of DDM with a threshold-based anomaly detection mechanism effectively minimized false positives, while Incremental Learning ensured continuous adaptability to evolving threats. These characteristics make DDM-CNN a reliable and scalable solution for modern intrusion detection systems.

Acknowledgment. This research is funded by the Nha Trang University (NTU), Vietnam under grant number "TR2023 - 13 - 29".

REFERENCES

[1] N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," Feb. 2018, Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ACCESS.2018.2807385.

[2] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA: Curran Associates Inc., 2019.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," CoRR, vol. abs/1412.6572, 2014, [Online]. Available: https://api.semanticscholar.org/CorpusID:6706414.

[4] P. Cui and J. Wang, "Out-of-Distribution (OOD) Detection Based on Deep Learning: A Review," Electronics (Basel), vol. 11, no. 21, p. 3500, Oct. 2022, doi: 10.3390/electronics11213500.

[5] S. Thrun, "Lifelong learning algorithms," in Learning to Learn, USA: Kluwer Academic Publishers, 1998, pp. 181–209.

[6] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in International Conference on Information Systems Security and Privacy, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4707749.

[7] N. Ahuja, G. Singal, and D. Mukhopadhyay, "DLSDN: Deep Learning for DDOS attack detection in Software Defined Networking," Dec. 2021. doi: 10.1109/Confluence51648.2021.9376879.

[8] T. J. Lucas et al., "A Comprehensive Survey on Ensemble Learning-Based Intrusion Detection Approaches in Computer Networks," IEEE Access, vol. 11, pp. 122638–122676, 2023, doi: 10.1109/ACCESS.2023.3328535.

[9] H. Aydın, Z. Orman, and M. A. Aydın, "A long short-term memory (LSTM)-based distributed denial of service (DDoS) detection and defense system design in public cloud network environment," Comput Secur, vol. 118, p. 102725, 2022, doi: https://doi.org/10.1016/j.cose.2022.102725.

[10] K. Yang, J. Zhang, Y. Xu, and J. Chao, "DDoS Attacks Detection with AutoEncoder," Dec. 2020, pp. 1–9. doi: 10.1109/NOMS47738.2020.9110372.

[11] V. Kumar and D. Sinha, "Synthetic attack data generation model applying generative adversarial network for intrusion detection," Comput Secur, vol. 125, p. 103054, 2023, doi: https://doi.org/10.1016/j.cose.2022.103054.

[12] B. Habib and F. Khursheed, "Time-based DDoS attack detection through hybrid LSTM-CNN model architectures: An investigation of many-to-one and many-to-many approaches," Concurr Comput, vol. 36, no. 9, p. e7996, 2024, doi: https://doi.org/10.1002/cpe.7996.

[13] K.-M. Ko, J.-M. Baek, B.-S. Seo, and W.-B. Lee, "Comparative Study of AI-Enabled DDoS Detection Technologies in SDN," Applied Sciences, vol. 13, no. 17, 2023, doi: 10.3390/app13179488.

[14] Y. Wei, J. Jang-Jaccard, F. Sabrina, W. Xu, S. Camtepe, and A. Dunmore, Reconstruction-based LSTM-Autoencoder for Anomaly-based DDoS Attack Detection over Multivariate Time-Series Data. 2023. doi: 10.48550/arXiv.2305.09475.

[15] Md. S. Rahman, S. Pal, S. Mittal, T. Chawla, and C. Karmakar, "SYN-GAN: A robust intrusion detection system using GAN-based synthetic data for IoT security," Internet of Things, vol. 26, p. 101212, Dec. 2024, doi: 10.1016/j.iot.2024.101212.

[16] M. Abdelaty, S. Scott-Hayward, R. Doriguzzi Corin, and D. Siracusa, "GADoT: GAN-based Adversarial Training for Robust DDoS Attack Detection," Dec. 2022. doi: 10.48550/arXiv.2201.13102.

[17] A. Mustapha et al., "Detecting DDoS attacks using adversarial neural network," Comput Secur, vol. 127, p. 103117, 2023, doi: https://doi.org/10.1016/j.cose.2023.103117.

[18] A. Nazir et al., "A deep learning-based novel hybrid CNN-LSTM architecture for efficient detection of threats in the IoT ecosystem," Ain Shams Engineering Journal, vol. 15, no. 7, p. 102777, 2024, doi: https://doi.org/10.1016/j.asej.2024.102777.

[19] M. S. Raza, M. N. A. Sheikh, I.-S. Hwang, and M. S. Ab-Rahman, "Feature-Selection-Based DDoS Attack Detection Using AI Algorithms," Telecom, vol. 5, no. 2, pp. 333–346, 2024, doi: 10.3390/telecom5020017.

[20] C.-S. Shieh et al., "Detection of Adversarial DDoS Attacks Using Generative Adversarial Networks with Dual Discriminators," Symmetry (Basel), vol. 14, no. 1, 2022, doi: 10.3390/sym14010066.

[21] M. Novaes, L. Carvalho, J. Lloret, and M. Proença, "Adversarial Deep Learning approach detection and defense against DDoS attacks in SDN environments," Future Generation Computer Systems, vol. 125, Jun. 2021, doi: 10.1016/j.future.2021.06.047.

[22] X. Sun, C. Zhang, G. Lin, and K.-V. Ling, Open Set Recognition with Conditional Probabilistic Generative Models. 2020. doi: 10.48550/arXiv.2008.05129.

[23] B. Liu, H. Kang, H. Li, G. Hua, and N. Vasconcelos, "Few-Shot Open-Set Recognition Using Meta-Learning," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8795–8804. doi: 10.1109/CVPR42600.2020.00882.

[24] H. Zhang, A. Li, J. Guo, and Y. Guo, "Hybrid Models for Open Set Recognition," in Computer Vision – ECCV 2020, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 102–117.

[25] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," in Proceedings of the 32nd International Conference on Neural Information Processing Systems, in NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 5234–5243.

[26] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, "Reliable fidelity and diversity metrics for generative models," in Proceedings of the 37th International Conference on Machine Learning, in ICML'20. JMLR.org, 2020.

[27] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6629–6640.

[28] A. Ha, T. Gunawan, M. Halbouni, F. Assaig, M. Effendi, and N. Ismail, CNN-IDS: Convolutional Neural Network for Network Intrusion Detection System. 2022. doi: 10.1109/ICWT55831.2022.9935478.

[29] M. Ibrahim and R. Elhafiz, "Modeling an intrusion detection using recurrent neural networks," Journal of Engineering Research, vol. 11, no. 1, p. 100013, 2023, doi: https://doi.org/10.1016/j.jer.2023.100013.

[30] A. Ali et al., "An optimized multilayer perceptron-based network intrusion detection using Gray Wolf Optimization," Computers and Electrical Engineering, vol. 120, p. 109838, 2024, doi: https://doi.org/10.1016/j.compeleceng.2024.109838.

[31] A. Singh and J. Jang-Jaccard, Autoencoder-based Unsupervised Intrusion Detection using Multi-Scale Convolutional Recurrent Networks. 2022. doi: 10.48550/arXiv.2204.03779.