Addressing Data Security in IoT: Minimum Sample Size and Denoising Diffusion Models for Improved Malware Detection

Chiara Camerota*† Lorenzo Pappone† Tommaso Pecorella* Flavio Esposito†
* Department of Information Engineering (DINFO), University of Florence, Italy
†Department of Computer Science, Saint Louis University, USA

***Abstract*—Machine learning (ML) has emerged as a compelling approach to identify attacks in network traffic security. Existing malware detection strategies often concentrate on specific facets, such as efficient data collection, particular types of malware, or handling data scarcity. While valid, these strategies typically overlook the potential for minimizing sample size, focusing instead on data augmentation. This work introduces a novel method to determine the minimum sample size necessary to achieve a specified accuracy level, measured by the F1 score derived from the confusion matrix. We focus on TCP header traffic data transformed into images through flow-splitting techniques for multi-class traffic classification. In addition, we introduce a diffusion model to generate new synthetic traffic images and show that our method outperforms existing techniques in terms of stability and predictability. This study also compares the effectiveness of synthetic image augmentation using Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPM) in improving image recognition and classification accuracy.**

***Index Terms*—malware detection, traffic classification, deep learning**

I. Introduction

Securing Internet of Things (IoT) applications is important but challenging for several reasons. One of them is the inherent limitations of low-power devices, which cannot adequately address robust protection strategies. Furthermore, consider the scenario in which an attack persists due to a lack of immediate identification: in such instances, the system may be severely compromised, potentially leading to significant financial losses for the network owner, as highlighted by [1]. This is particularly problematic because it compromises sensitive data and threatens physical and network security.

The application of machine learning (ML) has become increasingly prevalent in the detection of malware and other cyber threats [2], given its ability to process large data sets and decipher complex relationships between system variables [3]. Several ML approaches have been proposed to effectively detect previously seen and unseen malware, securing (IoT) networks [4].

While these approaches have merit, an accurate model often requires the availability of extensive and heterogeneous datasets, that are not always available or accessible. In addition, factors such as data scarcity and quality of training samples can affect the performance of classification models. As highlighted in [5], [6], the spurious correlation due to a lack of data and methodological rigor may lead to erroneous conclusions. Even when the ML model is rigorous, having a correct training dataset is the key.

To address a robust model and limit spurious correlation, the literature investigates methods to determine the minimum sample size [5], [6]. In addition, in network security literature, a common solution in case of data scarcity is to implement a data augmentation strategy that enhances the learning model's robustness. Data augmentation involves artificially expanding a training dataset by generating modified versions of the original data without requiring new data collection [7]. Introducing noise and data variability, these techniques enhance the model's generalization ability and reduce overfitting [8]. A strategy to apply this technique is to feed the model with traffic transformed into images. In [9], for example, feature extraction from traffic-based-images is shown to be not only more efficient in terms of accuracy compared to the binary representation of the same packet but also allows the method scalability. As demonstrated e.g., in [10], a low-dimensional representation of the image simplifies its categorization.

Our Contribution. Motivated by these results, we show how new image generation techniques, such as diffusion model-based [11], combined with dataset management, can achieve higher accuracy and lower false positive rates in malware classification, compared to existing approaches. We propose a new approach to find a minimum sample size based only on the confusion matrix to address the lack of sufficient training. Moreover, we adopt Denoising Diffusion Probabilistic Models (DDPM) for the data augmentation and dissect its link with the false positive rate. *We found that the DDPM-generated data have a 7% higher F1 score and less variance (5%) than the Generative Adversarial Networks (GAN) generated data* and a higher average AUROC index along the classes. Also, through explainable AI techniques, we show how DDPM images are more similar to real with respect to those generated with GAN, hence validating our approach.

The rest of the paper is organized as follows. Section II outlines the state of the art, while section III highlights the problem definition and our contributions. In Section IV, we describe the experimental setup, detailing our methodology. Section V discusses the experimental results, providing insightful comments and interpretations. Finally, the paper concludes with a summary of the key findings and their implications.

II. Related Work

Malware attacks pose an increasing threat to both IoT and traditional systems [12]. To mitigate IoT device damage, [13] aim to detect existing and new IoT malware by converting

978-3-903176-66-9 ©2024 IFIP

traffic data into RGB images. However, their feature preprocessing combines behavioral data with malicious activity frequency, which requires extensive data collection and may limit applicability in other environments. Similarly, in [14], Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs) based method is exploited to deal with malware traffic classification in high-variance daily energy consumption scenarios. Moreover, these work highlight the correlation between good prediction performance and the degree of diversity with datasets, in specific scenarios where the data are often unavailable. To the best of our knowledge, a method to achieve high accuracy with reduced data collection for IoT malware detection problems is still missing.

Data scarcity remains a significant challenge in many research domains, prompting the development of various mitigation techniques. Among these, data augmentation has emerged as a particularly effective and rapidly evolving method. Recent literature highlights diverse approaches to this problem, especially in specialized fields such as IoT security. [15] proposes a novel CNN architecture with dilated convolutions, channel squeezing, and boosting to identify IoT-specific malware patterns. While the study includes basic data augmentation methods like image rotation, these prove less effective due to the lack of natural directional orientation in the images. In contrast, [16] demonstrates a more sophisticated approach by integrating Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) to improve malware traffic classification after generating synthetic malware samples for training. While their findings indicate that GANs could improve detection performance, GAN-generated traffic images still produce a high false negative rate. In fact, the synthetic images closely resemble the distribution of the original samples, and fail to introduce sufficient noise into the training set to achieve a better generalization.

To create realistic traffic-based images and identify traffic profiles, the author of [17] introduce a stable diffusion model for data augmentation. They demonstrated that profile detection improves by switching to a text-to-image generative model, requiring two datasets: one for RGB images and one for stable diffusion prompts. While this reveals the potential of denoising models in image generation, the process can be costly. This inspired the use of the DDPM as an image-to-image generative method, which offers high-fidelity images and more precise results in terms of Fréchet Inception Distance (FID) compared to GANs [18], [19].

III. Problem Definition: Addressing Data Scarcity in IoT Malware Detection

In traditional approaches (for IoT attack classification with or without complete or sufficient data), the accuracy of a model is directly linked to the amount of data it is exposed to. Researchers have shown that more data points allow the model to recognize a wider range of patterns and establish a more resilient understanding of the underlying relationships [20]. However, in the context of IoT malware detection, data are frequently lacking and costly to obtain, making it difficult to achieve accurate classification. Our aim is to address this challenge by collecting data over a short period, determined by the identified minimum sample size, and then classifying it over a longer timeframe. The objective is to categorize different types of malware when data collection is limited or expensive. To identify the minimum amount of data to collect, we propose a new method based on the confusion matrix, without distribution assumption, employed to obtain accurate results in terms of model accuracy and false/true positive rates. Data augmentation is also performed using the diffusion probabilistic model to create more stable and high-fidelity images.

IV. Methodology and Solution Design

Data scarcity is a common issue in malware classification. In this work, we introduce a new method based on the confusion matrix and F-score to determine the minimum sample size for malware detection datasets. Our process does not need an assumption on the data distribution and identifies the minimum number of experiments for each class thanks to the McNemar-Bowker statistics test [21]. Furthermore, according to the data scarcity hypothesis, we explore the Denoising Diffusion Probabilistic Model to increase the data variability and make the classification model robust and precise [22].

A. Minimum Sample Size Definition

The literature explores the relationship between the number of examples considered (sample size) and model performance in ML. Studies such as [23] highlight that models with fewer parameters often perform well on smaller datasets. For instance, [24] demonstrates that simpler models can excel with limited data. Additionally, [25] investigated how the training set size impacts accuracy assessment, using a normal distribution assumption to define confidence intervals and calculate the sample size. However, this assumes linearity, which is not always empirically supported; in practice, the index curve often follows a more flexible distribution, like Beta, rather than a linear trend [26].

To address this issue, our method avoids assuming a specific distribution for the performance index. Instead, we use the confusion matrix (CF), from which performance metrics like the F1-score are derived. The CF is treated as the empirical joint probability distribution of the model's performance, and we perform statistical tests based on this matrix to quantify accuracy. Each CF element, excluding the diagonal, indicates the probability of misclassification—for example, a Distributed Denial of Service (DDoS) attack misclassified as a port scanning attack. The diagonal elements reflect the model's correct classifications. We apply the McNemar-Bowker test [21], a nonparametric test suited for comparing two related groups on a dichotomous variable. This test evaluates marginal homogeneity under the null hypothesis and follows a chi-square distribution with one degree of freedom.

Given the imbalanced nature of the malware dataset, the F1 score is an appropriate choice of performance index. We

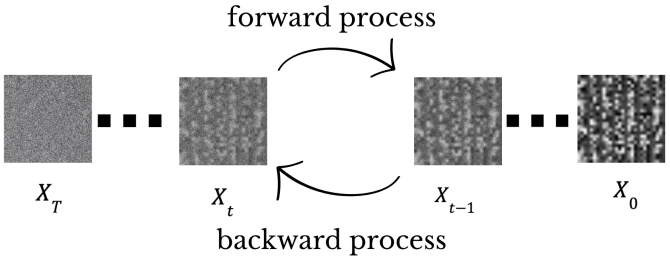

Fig. 1: The Denoising Diffusion Probabilistic Models' mechanism involves adding noise in the forward phase and learning noise removal techniques in the backward phase, without assuming data distribution

can define the F1 score in terms of true positive (TP), false negative (FN) and false positive (FP) values as follows:

$$F_1 \geq \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + 1 \cdot \mathrm{FN} + \mathrm{FP}} \,. \tag{1}$$

With this approach, we can consider the F1 score for each class and determine their sample size by applying the above definition to the test. Since a type I error (α) is a false positive conclusion in a statistic test and a Type II (β) error is a false negative conclusion, these terms are often used interchangeably with the general notion of false positives and false negatives mentioned in the formula.

B. Flow Image Generation

This subsection provides a concise overview of the generative models employed. In particular, the Deep Convolutional Generative Adversarial Network (DcGAN) and the denoising diffusion probabilistic model are considered.

A GAN architecture consists of a generator model that creates fake new samples as input to a discriminator model that determines if the input samples are fake or real. The generator learns to create traffic images as close as possible to real ones, while the discriminator checks their authenticity. The network aims to maximize the discriminator's loss (ensuring similarity between real and fake images) and minimize the generator's loss (improving its ability to mimic traffic behavior).

In our analysis, both losses are defined using Binary Cross Entropy. We apply the DcGAN model for its computational efficiency and ability to capture semantic features [27]. Here, the generator uses a one-dimensional convolutional network, reflecting the simplicity of the black-and-white data and minimizing computational cost.

The DDPM network takes two inputs, X_t (the final traffic-based image) and t (the steps that we want), and outputs a vector $\mu_\theta(X_t, t)$ and a fixed matrix $\Sigma_\theta(X_t, t)$, respectively the mean of the pixel and their variance matrix. This allows each step in the forward diffusion process to be approximately reversed by $X_{t-1} \sim N(\mu_\theta(X_t, t), \Sigma_\theta(X_t, t))$. The process is simplified and illustrated in Fig. 1. This method offers several advantages and addresses all the DcGAN issues listed above, including the ability to easily tune the model and generate more stable and performing outcomes in terms of image fidelity. In addition, the learning model to implement this kind of model can be chosen appropriately for the task.

In the literature, several neural networks are available, but for our purposes, lightweight models that capture key aspects of traffic images are most suitable. Therefore, we use a simple U-Net [28] with a 1-dimensional convolution layer. U-Net's ability to retain spatial information aids in recovering fine details and precise object localization. Its U-shaped architecture effectively combines local and global information, enhancing semantic accuracy compared to Convolutional Networks and common models like YOLO, which are slower and more resource-intensive. Additionally, U-Net reduces parameters, allowing faster training and lower computational demands. Its ability to learn from limited data and handle varying image sizes without preprocessing further adds to its practicality [28].

V. Evaluation

The following sections explain how the application of a new method to define the sample size combined with data augmentation through the adoption of DDPM can achieve high results in terms of F1 score, taking into account the False positive rate. Also, a deep study of synthetic images using explainable AI techniques demonstrates how DDPM helps the classification model.

A. Experimental Setting

Figure 2 shows a summary of the experiment workflow. The packet capture (PCAP) files collected in *EDGE-IIOTSET* [29] and *Malicious Network Traffic PCAPs and binary visualization images Dataset* (MNT) [14] are considered and analyzed as images for this work. Specifically, the files are initially divided into unidirectional flows based on the same 5-tuple: source IP, source port, destination IP, destination port, and transport-level protocol. Subsequently, the TCP headers of each flow are concatenated. Additionally, they are truncated if they exceed 743 bytes; otherwise, zero padding is added, based on [30], where it is shown that the most pertinent information is concentrated in these initial bytes. The resulting files are converted to hexadecimal and then into 28x28 grayscale images. This choice is coherent with the data parsimony and offers a quick transformation that only needs the header information. The generative models, DcGAN and DDPM, are specific for each class considered and are trained on the sub-dataset to achieve superior prediction accuracy. Also, the final datasets include different proportions of synthetic data to explore the impact of synthetic data on resilience to data scarcity across methods. Considering our study and utilizing the formula 1, we can calculate the β value for the sample size. Let's suppose $F_1 = 0.8$ and $\alpha = 0.05$, we can calculate a β value of 0.128 ($\beta = 0.128$). Performing the McNemar-Bowker test, the sample size outcome is 735 examples for each class for EDGE-IIOTSET and 580 for the MNT; these differences are due to the different number of classes considered. After, a CNN with fixed parameters and hyperparameters was employed to investigate the influence of sample size and synthetic data on model performance. Utilizing the identical CNN, the baseline model was trained exclusively on the original data with a fixed sample size. All models were trained for 200 epochs with a

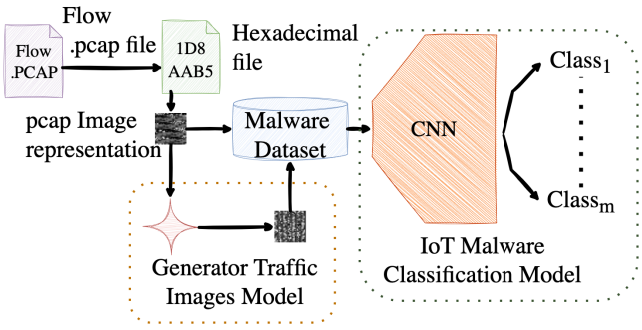

Fig. 2: The data flow starts with creating hexadecimal-based images for both datasets using the unidirectional flows header. After training the Generative model image-to-image for each class, the dataset is input to the CNN classification model to collect classification outcomes for evaluation.

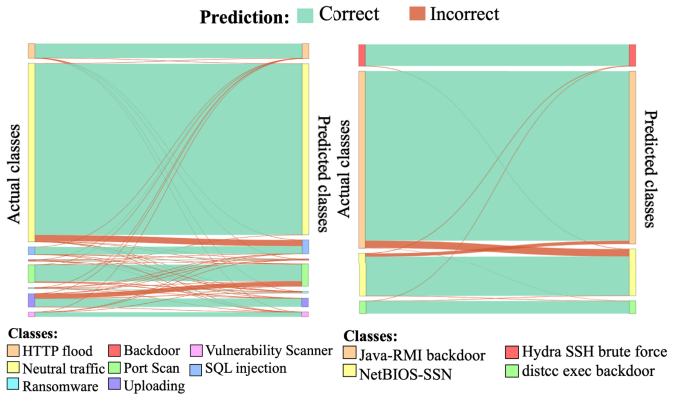

Fig. 3: The Sankey diagrams visualize the confusion matrix for EDGE-IIOTSET (right) and MNT set (left). The critical class to classify for the EDGE-IIOTSET is upload attacks, while for MNT, it's Java RMI backdoor attacks. These diagrams establish the baseline for classification and serve as a reference point for improvement using artificial images.

TABLE I: F1 scores of the test set with a training model fed with the real dataset and the balanced dataset with a sample size less than the thresholds. The performance in other cases is worse than our method.

num. ex. per classes	EDGE-IIOTSET	MNT
$\leq$ threshold	0.6	0.7
real unbalanced train set	0.73	0.58
our train set	0.93	0.97

0.01 learning rate, employing stochastic gradient descent. Performance was evaluated with the F1 score on real, unbalanced test sets to evaluate the impact of dataset configuration on results.

B. Evaluation of Classification Performance

Figure 3 displays the confusion matrix for both baseline models using a Sankey diagram [31]. This diagram visualizes each class with boxes on either side, with green arrows indicating correct classifications and red arrows indicating misclassifications. The graph highlights dataset imbalances and critical classes. In the EDGE-IIOTSET, *uploading attacks* is the least accurately classified, while the *Java-RMI backdoor* is most frequently misclassified in the MNT set. Also, Table I shows that F1 scores degrade compared to our baseline when the number of examples per class falls below the threshold or when using the real unbalanced training set.

EDGE-IIOTSET. Focusing on the EDGE-IIOTSET, there are clear differences between the results obtained using GAN-generated images and DDPM-generated images, particularly regarding the classification of neutral traffic. It should be noted that DDPM consistently accurately tracks neutral traffic, regardless of the volume of synthetic data, thereby reducing misclassification variability. Conversely, models trained with GAN synthetic data exhibit increased misclassification, as shown in Figure 4. The plots highlight the challenge of accurately classifying minority classes, as shown by the curves. For instance, with 30%, 50%, and 60% synthetic data, the DDPM models struggle to classify uploading attacks correctly. However, in other cases, the models perform better across all critical classes. This may be due to increased noise from synthetic data and the limited number of training epochs set based on the baseline experiment. Moreover, Figure 5 displays the F1 score for each model along with the relative standard deviation. The baseline model, trained without synthetic data, exhibits high F1 scores and lower standard deviation values. Focusing on the other models, *we find that the DDPM models generally achieve better F1 scores but tend to have higher standard deviations,* except in the case where 30% synthetic data are used. This suggests that while DDPM models can enhance performance, they tend to exhibit greater variability, as said before. Another interesting aspect of these results is the DDPM-based model's capacity to classify better than the GAN-based one's class in the case of 10% synthetic data. However, the overall performance in terms of the F1 score suggests the opposite.

Malicious Network Traffic PCAPs Dataset. We further analyze the Malicious Network Traffic PCAPs and binary visualization images dataset. The main differences between this dataset and the previous one are the lack of neutral traffic and the smaller number of classes considered. However, the ROC curves shown in Figure 6 combined with the results in Figure 5 confirm the fact that DDPM-based models perform better than the GAN-based models, likewise for the critical class, both in terms of average F1-scores and their variance. The performance of both models declines significantly when the synthetic data percentage reaches 70%, suggesting that the models struggle with noise introduced by artificial datasets. This is particularly interesting given the previous ROC curve, where the same model struggled to discriminate classes with 30% synthetic data. Additionally, the GAN-based model trained with 90% artificial data failed. We extended the training by 200 epochs to validate these findings but observed no significant improvements, indicating that a more complex model may be necessary for this task. In conclusion, the DDPM-based synthetic images yield a more stable and predictable dataset compared to those based on GAN images. Additionally, the classification performance on the imbalanced test set is superior, as evidenced by reduced misclassification confusion, as shown in Figure 5.

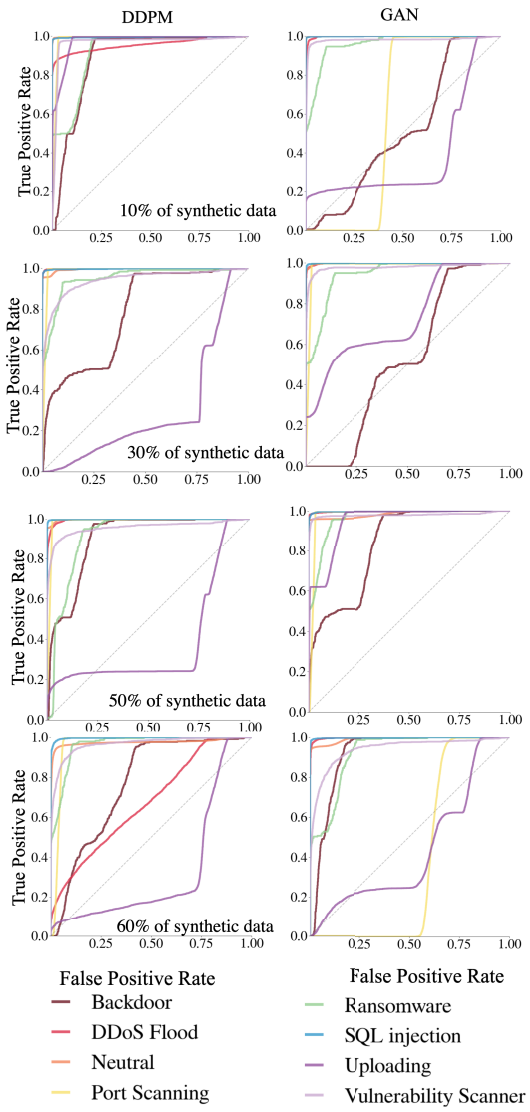

Fig. 4: ROC curve for EDGE-IIOTSET. These graphs plot the true positive rate against the false positive rate at each threshold setting. DDPM performs better than GAN, as the curve below the bisector line indicates. An increase in misclassification highlights the influence of synthetic data percentage on performance degradation.

C. Pixels vs. Packets: Analysis of Synthetic Traffic via XAI

Integrating synthetic traffic data into malware classification models introduces complexities that require a deeper understanding of model behavior. While performance metrics offer a wide measure of accuracy, they do not elucidate how synthetic data affects the classifier's decision-making process. In this context, explainable AI (XAI) is essential for validating the fidelity of synthetic traffic by comparing its impact on model decisions with that of real traffic data. This method helps evaluate synthetic data quality and reveals potential biases or artifacts introduced by various generation techniques. We use Gradient-weighted Class Activation Mapping (Grad-CAM) [32], a leading XAI technique, to analyze how synthetic traffic generation influences classifier decisions. This approach

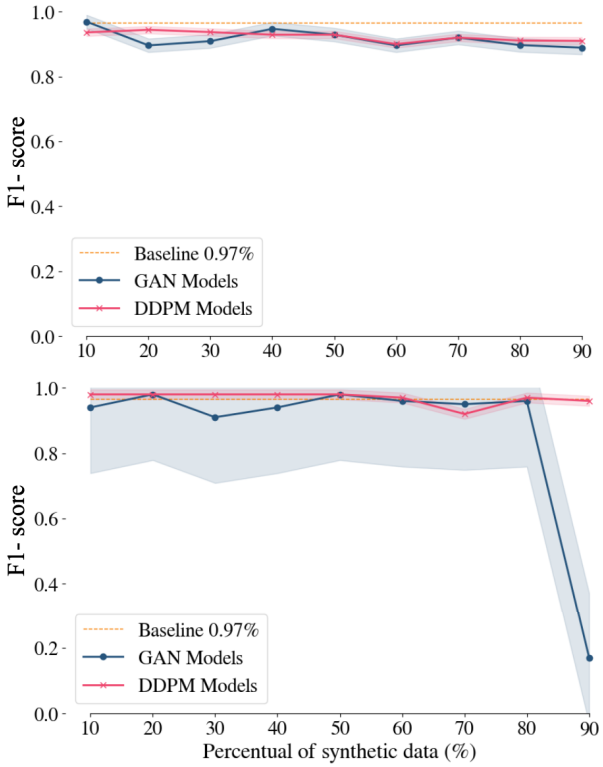

Fig. 5: Trend of class average F1 score for the test set with a *95% confidence interval*. The EDGE-IIOTSET plot shows that the DDPM-generated data have a 7% higher average F1 score and less variability than the GAN-generated data. The MNT F1 score trend curve highlights that DDPM-based models are more stable with minimal performance variance with more synthetic data. The performance of GAN deteriorates at 90% synthetic data.

enables us to visualize and quantify the key features driving the CNN's predictions for both real and synthetic traffic flows.

Grad-CAM Heatmaps. We analyze the generative capabilities of both DDPM and GAN models by exploiting the Grad-CAM heatmaps. Figure 7 shows an example of the heatmaps generated by the Grad-CAM algorithm applied to a correctly classified flow sample of the SQL injection attack. To compare the generative capabilities of DDPM and GAN, we evaluated three CNN-based classifiers trained on different formulations of the EDGE-IIOT dataset: (i) real data only, (ii) DDPM-based data combined with real data, (iii) GAN-based data combined with real data. The heatmaps in Figures 7b, 7c, and 7d reveal that the CNN model considers various parts of the traffic flow header with differing levels of importance depending on the input dataset, suggesting the need for a deeper evaluation.

Fidelity Analysis. In this analysis, we aim to deepen the behavior of the classifier by observing how the synthetic traffic data impacts the classification process. In Figure 8, we observe notable differences between DDPM and DcGAN-generated synthetic traffic. Specifically, the classifier shows dependencies to different header fields for each traffic generation method. For DDPM-generated data, the TCP Acknowledge Number is crucial for most attacks, while no clear pattern emerges with DcGAN-generated data. Additionally, the classifier uses

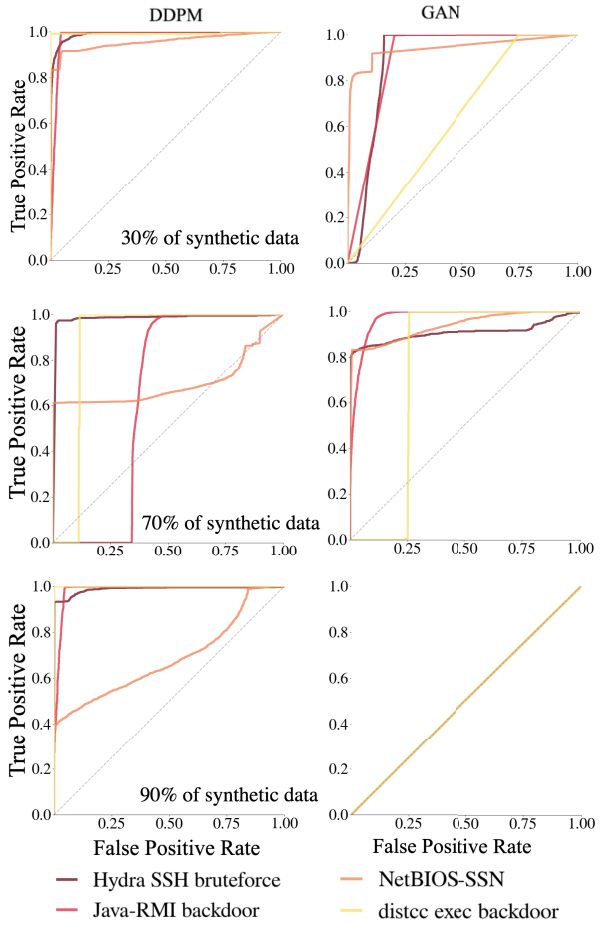

Fig. 6: ROC curve for the MNT Dataset. The DDPM's performances are better than the GAN's, and their curves appear below the bisector line. The DDPM can more effectively identify and classify different classes. When considering 90% of artificial data, the GAN-based model cannot achieve good classification, while the DDPM model can.

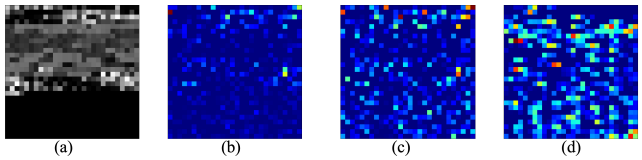

Fig. 7: Grad-CAM heatmaps of the SQL Injection attack. Red elements indicate a strong influence on the prediction, whereas blue elements have little effect on the prediction. (a) Original image of a SQL injection flow. (b) Resulting Grad-CAM of a CNN model trained only using real data (i.e., no synthetic traffic). (c) Resulting Grad-CAM of a CNN model trained with 50% traffic generated by DDPM. (d) Resulting Grad-CAM of a CNN model trained with 50% traffic generated by GAN.

distinct fields for each model: DDPM-generated traffic relies on a combination of TCP flags, TCP window size, IP header length, IP destination, and TCP sequence number to detect port scanning attacks, whereas GAN-generated traffic depends predominantly on TCP flags.

Packet Type Distribution. While header field analysis focuses on specific packet header values, we provide a higher-level view by examining the packet types associated with the most

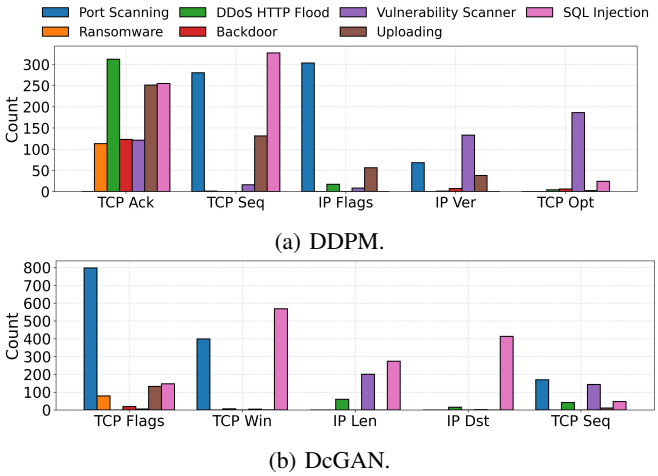

Fig. 8: (a) Distribution of top-5 header fields across attacks over DDPM-generated synthetic malware traffic. (b) Distribution of top-5 header fields across attacks over DcGAN-generated malware traffic.

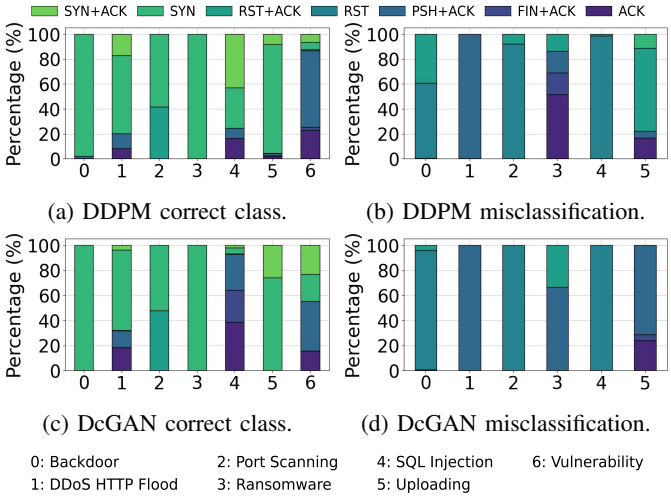

Fig. 9: Distribution of TCP flags for packets determined as most influential in the classifier's predictions by the Grad-CAM algorithm. (a) TCP flag distribution for correctly classified DDPM-generated traffic. (b) TCP flag distribution for misclassified DDPM-generated traffic. (c) TCP flag distribution for correctly classified DcGAN-generated traffic. (d) TCP flag distribution for misclassified DcGAN-generated traffic.

influential bytes using XAI techniques. We analyze TCP flags for each attack to assess classifier behavior with synthetic data from different generative models. Figures 9a, 9c show the distribution of TCP flags for correct classifications (Figures 9b, 9d) and misclassifications (Figures 9b, 9d). SYN packets are key for identifying attacks like backdoor and DDoS, appearing nearly 100% of the time in correct predictions. Exceptions include scanning attacks, where FIN+ACK and ACK packets are more relevant, and port scanning attacks, which feature high percentages of RST+ACK flags. SQL injection attacks show varied packets depending on the generative model. This analysis highlights that both generative models effectively capture attack behaviors, with SYN packets often linked to connection initiation attacks and other flags used for different

types of attacks. Interestingly, Figures 9b and 9d reveal that the classifier mistakenly emphasizes RST packets for backdoor, port scanning, and SQL injection attacks. This suggests the classifier might be overfitting to synthetic data patterns that do not align with real-world attack behaviors. RST packets, often linked to rare abrupt terminations, may be overrepresented in synthetic data. To improve performance, future work should incorporate domain-specific knowledge of network protocols and attack patterns into both data generation and classifier training.

VI. Conclusion

In the IoT domain, data scarcity remains a significant challenge, often addressed by generating artificial data using DL-based generative techniques like GANs and their variants. This study investigates the efficacy of Denoising Diffusion Probabilistic Models (DDPM) and Generative Adversarial Networks (GAN) in augmenting limited training data for large-scale malware traffic image classification. Our findings indicate that DDPM consistently outperforms GAN as the proportion of synthetic data increases. We propose a novel method to determine the minimum sample size needed to achieve the desired F1-score accuracy. Future research directions include validating our methodology across diverse data sources and IoT contexts, as well as refining DDPM to enhance its specificity for IoT-related data generation.

Acknowledgement

This work has been partially supported by the National Science Foundation (NSF) awards # 2133407 and # 2201536. Also, it was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, a partnership on "Telecommunications of the Future" (PE0000001 - program "RESTART").

References

[1] V. Hassija, V. Chamola, V. Saxena, D. Jain, P. Goyal, and B. Sikdar, "A survey on iot security: application areas, security threats, and solution architectures," *IEEe Access*, vol. 7, pp. 82 721–82 743, 2019.

[2] A. Souri and R. Hosseini, "A state-of-the-art survey of malware detection approaches using data mining techniques," *Human-centric Computing and Information Sciences*, vol. 8, no. 1, pp. 1–22, 2018.

[3] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, "Compute trends across three eras of machine learning," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul. 2022.

[4] K. Shaukat, S. Luo, and V. Varadharajan, "A novel deep learning-based approach for malware detection," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106030, 2023.

[5] W. Ye, G. Zheng, X. Cao, Y. Ma, X. Hu, and A. Zhang, "Spurious correlations in machine learning: A survey," *arXiv preprint arXiv:2402.12715*, 2024.

[6] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," *Ieee Access*, vol. 5, pp. 7776–7797, 2017.

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[8] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.

[9] L. Nataraj, V. Yegneswaran, P. Porras, and J. Zhang, "A comparative assessment of malware classification using binary texture analysis and dynamic analysis," in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, ser. AISec '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 21–30.

[10] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, pp. 145–175, 2001.

[11] H. Chen and M. A. Babar, "Security for machine learning-based software systems: A survey of threats, practices, and challenges," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1–38, 2024.

[12] Ö. A. Aslan and R. Samet, "A comprehensive review on malware detection approaches," *IEEE access*, vol. 8, pp. 6249–6271, 2020.

[13] J. Jeon, J. H. Park, and Y.-S. Jeong, "Dynamic analysis for iot malware detection with convolution neural network model," *IEEE Access*, vol. 8, pp. 96 899–96 911, 2020.

[14] R. Chaganti, V. Ravi, and T. D. Pham, "A multi-view feature fusion approach for effective malware classification using deep learning," *Journal of information security and applications*, vol. 72, p. 103402, 2023.

[15] M. Asam, S. H. Khan, A. Akbar, S. Bibi, T. Jamal, A. Khan, U. Ghafoor, and M. R. Bhutta, "Iot malware detection architecture using a novel channel boosted and squeezed cnn," *Scientific Reports*, vol. 12, no. 1, p. 15498, 2022.

[16] M. Abdelaty, S. Scott-Hayward, R. Doriguzzi-Corin, and D. Siracusa, "Gadot: Gan-based adversarial training for robust ddos attack detection," in *2021 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2021, pp. 119–127.

[17] X. Jiang, S. Liu, A. Gember-Jacobson, A. N. Bhagoji, P. Schmitt, F. Bronzino, and N. Feamster, "Netdiffusion: Network data augmentation through protocol-constrained traffic generation," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 8, no. 1, feb 2024.

[18] X. Su, J. Song, C. Meng, and S. Ermon, "Dual diffusion implicit bridges for image-to-image translation," *arXiv preprint arXiv:2203.08382*, 2022.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[20] X. Ying, "An overview of overfitting and its solutions," in *Journal of physics: Conference series*, vol. 1168. IOP Publishing, 2019, p. 022022.

[21] M. Eliasziw and A. Donner, "Application of the mcnemar test to non-independent matched pair data," *Statistics in medicine*, vol. 10, no. 12, pp. 1981–1991, 1991.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[23] C. Andrade, "Sample size and its importance in research," *Indian journal of psychological medicine*, vol. 42, no. 1, pp. 102–103, 2020.

[24] A. Gosiewska, A. Kozak, and P. Biecek, "Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering," *Decision Support Systems*, vol. 150, p. 113556, 2021.

[25] G. M. Foody, "Sample size determination for image classification accuracy assessment and comparison," *International Journal of Remote Sensing*, vol. 30, no. 20, pp. 5273–5291, 2009.

[26] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 3121–3124.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[28] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE access*, vol. 9, pp. 82 031–82 057, 2021.

[29] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-iiotset: A new comprehensive realistic cyber security dataset of iot and iiot applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40 281–40 306, 2022.

[30] *USTC-TK2016 Repository*. [Online]. Available: https://github.com/yungshenglu/USTC-TK2016/tree/ubuntu

[31] P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive sankey diagrams," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE, 2005, pp. 233–240.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.