

KiNETGAN: Enabling Distributed Network Intrusion Detection through Knowledge-Infused Synthetic Data Generation

Anantaa Kotal
C.S.E.E.
University of Maryland,
Baltimore County
Baltimore, USA
anantak1@umbc.edu

Brandon Luton
C.S.E.E.
University of Maryland,
Baltimore County
Baltimore, USA
cl39497@umbc.edu

Anupam Joshi
C.S.E.E.
University of Maryland,
Baltimore County
Baltimore, USA
joshi@umbc.edu

Abstract—In the realm of IoT/CPS systems connected over mobile networks, traditional intrusion detection methods analyze network traffic across multiple devices using anomaly detection techniques to flag potential security threats. However, these methods face significant privacy challenges, particularly with deep packet inspection and network communication analysis. This type of monitoring is highly intrusive, as it involves examining the content of data packets, which can include personal and sensitive information. Such data scrutiny is often governed by stringent laws and regulations, especially in environments like smart homes where data privacy is paramount. Synthetic data offers a promising solution by mimicking real network behavior without revealing sensitive details. Generative models such as Generative Adversarial Networks (GANs) can produce synthetic data, but they often struggle to generate realistic data in specialized domains like network activity. This limitation stems from insufficient training data, which impedes the model’s ability to grasp the domain’s rules and constraints adequately. Moreover, the scarcity of training data exacerbates the problem of class imbalance in intrusion detection methods. To address these challenges, we propose a Privacy-Driven framework that utilizes a knowledge-infused Generative Adversarial Network for generating synthetic network activity data (KiNETGAN). This approach enhances the resilience of distributed intrusion detection while addressing privacy concerns. Our Knowledge Guided GAN produces realistic representations of network activity, validated through rigorous experimentation. We demonstrate that KiNETGAN maintains minimal accuracy loss in downstream tasks, effectively balancing data privacy and utility.

Index Terms—Synthetic data, Mobile and IoT system, Knowledge Guided Learning, GAN

I. INTRODUCTION

Network intrusion detection systems (NIDS) are critical for protecting modern enterprise systems, particularly in IoT-based and mobile environments where attacks can lead to both information loss and physical damage. Distributed NIDS enable real-time monitoring across multiple devices and segments, promptly detecting anomalies and potential threats to safeguard sensitive data. Integrating Machine Learning (ML) models in NIDS enhances their effectiveness in preventing cyberattacks [1]–[7].

However, sharing data across distributed systems raises privacy concerns, especially with intrusive detection methods like deep packet inspection. Federated learning offers a solution by allowing collaborative training without sharing raw data, but it’s challenging to implement across diverse devices with varying architectures and learning infrastructure.

Deep learning-based synthetic data generation emerges as a promising solution. By creating data that mirrors authentic network behavior while protecting sensitive details, it enables secure data sharing and analysis. Among these methods, generative adversarial networks (GANs) stand out for their ability to capture and replicate the underlying distribution of a training dataset.

However, standard GANs face limitations in generating realistic system data, particularly for IoT and mobile networks, due to a lack of explicit domain knowledge. For instance, they may misconfigure attributes like port numbers associated with specific attacks, leading to misleading synthetic data. Moreover, class imbalance in training data can bias models towards prevalent classes, hindering accurate intrusion detection.

To address these challenges, leveraging domain knowledge to guide generative models is crucial. By incorporating specific characteristics of the data into the training process, such as rules governing network traffic, generative models can produce more accurate synthetic data. This approach enhances the efficacy of generative models in specialized domains like network security by ensuring that synthetic data closely resembles real-world scenarios.

This paper introduces a novel Privacy-Driven knowledge-infused Generative Adversarial Network (KiNETGAN) model designed to tackle the obstacles related to synthetic data generation for privacy preservation in distributed network intrusion detection systems. Our innovative approach leverages domain knowledge and employs enhanced GAN training to create realistic representations of network activities on individual devices. The KiNETGAN model addresses the limitations of standard generative models, ensuring a comprehensive understanding of the domain’s rules and restrictions. We

demonstrate the effectiveness of our approach by synthesizing network activity data, validating the synthetic dataset against network-specific constraints, and confirming its suitability through likelihood fitness and high efficacy in downstream intrusion detection tasks.

II. BACKGROUND

Generative Adversarial Networks (GANs) are powerful models widely used for generating synthetic data that closely resembles real data [8]–[11]. They consist of a generative model (G) and a discriminative model (D) trained together in a min-max game framework. GANs have demonstrated high accuracy in generating synthetic data, particularly for images and text. However, teaching GANs to learn from Network Activity Data presents challenges due to its tabular nature, combining discrete and continuous values, and exhibiting sparsity and imbalanced distributions.

Xu et al. proposed a GAN model addressing challenges with tabular data by introducing mode-specific normalization, a conditional generator, and training by sampling [12]. Kotal et al. extended this model for privacy-preserving data generation, enforcing t-closeness in the synthetic data distribution to preserve privacy [13].

Differential Privacy (DP) has been applied to GANs to enhance privacy [14]. Various models combine DP with GANs to generate differentially private synthetic data, introducing noise into the discriminator during training to ensure privacy [15], [16].

However, Network traffic data poses challenges for GAN training due to sparsity and limited size. GANs trained solely on observed network activity lack understanding of network attributes and struggle to adhere to strict rules without explicit constraints. Knowledge guidance becomes essential to convey constraints and enhance the data generation process.

Knowledge Graphs (KGs) offer a versatile data model for knowledge representation and reasoning [17]. They store contextual information crucial for learning in distributed systems and can impose constraints on entities. Integrating KGs enriches the data generation process, improving contextual awareness in ML systems. Its integration has been demonstrated to significantly improve contextual awareness in machine learning (ML) systems [4], [18]–[20]. In a related context, Hui et al. introduced a knowledge-enhanced GAN for generating IoT traffic data [17]. Specifically, for privacy preservation, there is evidence that Knowledge Infusion can help generative models gain the added context needed for zero-shot learning and learning with limited parameters [21]–[24].

In this paper, knowledge about network traffic is injected into GAN training by adding the Knowledge base as an independent discriminator. This approach enhances the GAN’s understanding of network attributes and adherence to protocol rules, improving the accuracy of synthetic data generation in the observed system.

III. PROPOSED FRAMEWORK

Network traffic data presents challenges for synthetic data generation due to sparsity, class imbalance, and strict domain

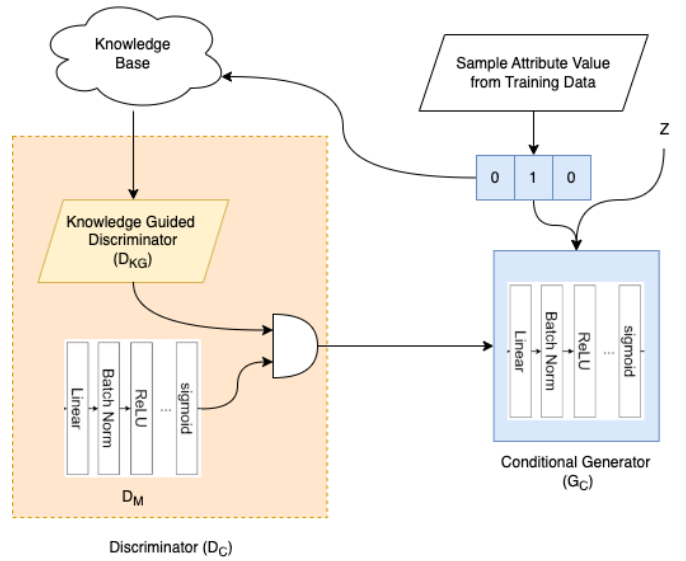


Fig. 1: The KiNETGAN model of Knowledge Infused Synthetic Data Generation

rules. These characteristics hinder accurate model construction solely based on observed data, leading to unrealistic synthetic data. Domain knowledge integration into generative models reduces the need for relearning rules, addressing class and domain restrictions. We propose a knowledge-guided synthetic data generation method, KiNETGAN, leveraging a domain Knowledge Graph (KG) to train Generative Adversarial Networks (GANs) and conditional GANs. This approach tackles class imbalance and attribute cross-correlation issues. The framework’s architecture is detailed in Figure 1.

A. Conditional Generator with Data Balancing

In GANs, training on randomly selected data can lead to underrepresentation of minority categories, affecting generator accuracy. Conditional generation is vital for Mobile Network Activity data, enabling the GAN to learn attribute relationships and address data imbalance. Efficient data resampling ensures balanced category representation during training, maintaining original data distribution fidelity. We propose Conditional GANs and sampling-based training to achieve these goals, preserving model fidelity and accurately reflecting the original data distribution during testing.

1) *Conditional vector*: To introduce the Conditional Attributes as the condition to the Conditional generator, we introduce the condition vector, \mathbf{C} . It is a one hot vector representation of the set of discrete Attributes. Let (c_1, c_2, \dots, c_n) be the list of conditional attributes and is the output from G_C that is the condition for our current generation.

Let us consider the attribute c_1 . Let the range for c_1 be $\{c_{1,1}, c_{1,2}, \dots, c_{1,n}\}$. The chosen value for c_1 is \hat{c}_1 . The one hot vector representation (C_1) of c_1 is defined as follows:

$$C_{1,i} = \begin{cases} 1, & \text{if } c_{1,i} == \hat{c}_1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The conditional vector, C is a concatenation of all C_i 's:

$$C = C_1 \oplus C_2 \oplus \dots \oplus C_n \quad (2)$$

2) **Conditional Generator:** The input to the Generator is a random noise (Z) and the conditional vector C . The objective of the Conditional Generator is to generate realistic synthetic data while adhering to the attribute values specified in C . To ensure that this constrained is met, in addition to the discriminator score, we need to penalise the generator for disregarding the attribute values specified in C . This is done by adding a cross entropy with the condition vector, C to the loss function. Let the generator output for the conditional attribute set be $\hat{c} = (\hat{c}_1, \hat{c}_2, \dots, \hat{c}_n)$, the one hot vector representation of which is \hat{C} . Then we add the following term to the loss function of Conditional Generator: $BCE(C, \hat{C})$ averaged over all the instances of the batch. As the training advances, the generator learns to make an exact copy of C into \hat{C} .

3) **Conditioning on Imbalanced Values:** As discussed, mobile network activity data is often sparse, and attribute values are heavily imbalanced. To ensure that minority attribute values are sufficiently represented during training, we need to compel the generator to consider these minority values. To achieve this, we randomly sample an attribute value from a uniform distribution within the range of the attribute and add it to our condition vector, C . The generator is thus constrained to produce synthetic data where the minority attribute is present. This ensures that sparse attribute values are adequately represented during the training of the generator, enabling it to produce data points within these values.

B. GAN Training with Knowledge-Guided Discriminator

The Knowledge-Guided Discriminator aims to identify external rules restricting attribute values using a Knowledge Graph. It focuses on attributes with relative value restrictions to learn valid and invalid attribute combinations. For example, in the CVE-1999-0003 attack, a valid port address is between 32771 and 34000. Domain knowledge helps rule out invalid combinations, such as ports outside this range. The key difference is that some generator outputs may not just be fake but invalid. Penalizing these instances ensures the generator produces realistic and valid instances. This objective is met by dividing the discriminator into two parts.

1) **Knowledge-Guided Discriminator (D_{KG}):** The objective for the Knowledge-Guided Discriminator is to discriminate between correct and incorrect instances of data according to domain rules. In the case of network activity data, this includes examples of invalid combinations of IP addresses and port numbers for an event. A domain-specific Knowledge Graph (KG) can help rule out invalid combinations of attributes from explicit knowledge. The KG is queried with the output $(\hat{c}_1, \hat{c}_2, \dots, \hat{c}_n)$ of G_C to determine whether the given set of values is valid. Let us represent the KG query as Q . The discriminator's input consists of all valid sets of

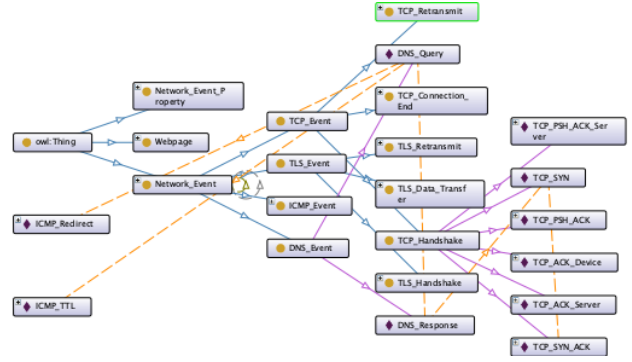


Fig. 2: Ontology for Network Activity Capture

attributes for the conditional vector C queried from the knowledge graph and the output of the generator, G_C .

2) **Regular Discriminator:** The objective here is that of a regular discriminator in a GAN model. It is a standard discriminator tasked with distinguishing real data points from those generated by the generator, G_C . This framework follows the design of a standard GAN. The input to the discriminator D_M includes the output of G_C and real data points from the training set.

The final output of the discriminator (D_C) combines the output of both D_{KG} and D_M :

$$D_C = D_{KG} + D_M \quad (3)$$

Loss function: The generator loss is updated on the output of both D_{KG} and D_M . Thus following from equation 3, the loss for G_C is defined as:

$$\mathcal{L}_{G_C} = E_{z \sim p_z(z)} [\log(1 - D_C(G_C(z)))] \quad (4)$$

IV. EXPERIMENTAL FRAMEWORK

A. Knowledge Graph Creation

The Unified Cybersecurity Ontology (UCO) is a comprehensive framework designed to enhance cyber situational awareness by integrating diverse data and knowledge schemas from various cybersecurity systems and standards. UCO encompasses entities, events, activities, and relationships crucial for cybersecurity analysis. Integrating UCO into machine learning models improves contextual understanding, enhancing their effectiveness in cybersecurity scenarios. This study extends UCO to define concepts in network activity data, introducing entities like "networkEvent" and "domainURL". Each network event is defined by properties such as protocol, source/destination IP addresses, and port numbers. Leveraging this ontology, a Network Traffic Knowledge Graph (NetworkKG) is constructed, guiding data generation. A Knowledge Graph reasoner facilitates queries for valid IP, port, and protocol combinations, assisting the Generative Adversarial Network (GAN) synthesis process. This approach ensures that synthetic data generated aligns with real-world network event attributes, enhancing the quality and relevance of generated data for cybersecurity applications. Figure 2 demonstrates the entities in this ontology.

B. Data Collection

To test our method, we require datasets of real network activity from a system of devices. For this purpose, we utilize two datasets: a network activity dataset collected from a system of IoT devices connected in our lab and the UNSW-NB15 Dataset.

1) *Lab Collected Data*:: In our network setup, we’ve integrated mobile and IoT devices such as a Blink camera, a smart plug, and a motion sensor. Analyzing their communication patterns with Wireshark, we focus on events like motion detection, lamp activation, and tag manager interactions. Collected data includes Source/Destination IP, Ports, and Protocols. Filtering by device IPs, we study their typical communications and simulate attacks like Traffic flooding. The dataset, comprising 14,520 records, is crucial for NIDS training.

2) *UNSW-NB15 Data*:: The UNSW-NB15 dataset consists of 2,540,044 network connection records. This comprehensive dataset includes a wide range of network traffic data, featuring 49 attributes that encompass flow features, basic features, content features, time features, and additional generated features. This size and diversity make it well-suited for training and evaluating machine learning algorithms for intrusion detection systems.

V. EXPERIMENTAL RESULTS

In this section, we will begin by describing and detailing the various measures and tests we used to validate our model. The main goal of network activity datasets is to support Network Intrusion Detection (NIDS) efforts. Machine learning-based NIDS classifiers require high fidelity data for training. As previously mentioned, the bottleneck for security efforts lies in the challenge of obtaining dependable training data. To fulfill its purpose, synthetic data must serve as a viable substitute for the original data in downstream tasks.

To demonstrate that the KiNETGAN framework fulfills these objectives, we present the outcomes of three types of tests:

- **Fidelity Results** to show that the synthetic data is statistically close or similar to the original data.
- **Utility Results** to show that the synthetic data is useful in training downstream ML-based NIDS models.
- **Privacy Results** to show that the synthetic data and KiNETGAN model are resilient against privacy attacks.

We demonstrate that the KiNETGAN framework fulfills these objectives by being distributionally similar to the original dataset and having comparable accuracy in downstream tasks. To validate our model, we compare it with synthetic data generated using other Generative Deep Learning models for tabular data generation: CTGAN [12], OCTGAN [25], PATEGAN [26], TABLEGAN [27], and TVAE [12]. We provide experimental results for synthetic data generated from training on our dataset of lab-collected network traffic data and the UNSW-NB15 data.

	Lab Data		UNSW-NB15	
	EMD	Distance	EMD	Distance
CTGAN	0.06	0.09	0.07	0.2
OCTGAN	1.61	0.95	1.32	1.61
PATEGAN	1.07	0.09	0.53	0.24
TABLEGAN	1.02	0.19	1.21	0.53
TVAE	0.06	0.04	0.13	0.23
KiNETGAN	0.06	0.03	0.07	0.03

TABLE I: Comparison of Distance between Synthetic and Original Data

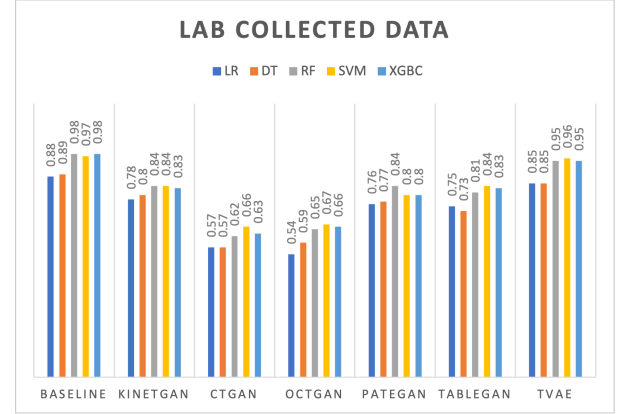


Fig. 3: Comparison of NIDS accuracy for Lab Collected Data

A. Statistical Distance

When assessing the quality of generated data, statistical distance measures help quantify the dissimilarity between the generated data distribution and the original (real) data distribution. We use two distance metrics for our comparison.

- The Earth Mover’s Distance (EMD) or Wasserstein Distance which measures the minimum cost of turning one distribution into another, where the cost is interpreted as the amount of ”mass” that must be moved.
- We use a combination of L_1 norm or Manhattan distance to calculate the distance for categorical variables and the L_2 norm or Euclidean distance to calculate the distance for continuous variables. Since our data is tabular i.e. a mix of categorical and continuous variables, this pragmatic approach is ideal to handle mixed-type data.

The results from our comparison are given in Table I. For the Lab collected data, the KiNETGAN had the lowest EMD distance at 0.06, similar to TVAE and CTGAN. KiNETGAN had the lowest combined distance at 0.032. For UNSW-NB15 KiNETGAN and CTGAN had the lowest EMD distance at 0.007. and the KiNETGAN model has a combined distance of 0.16.

B. Utility Results

Machine Learning (ML) is essential in network intrusion detection, identifying anomalies indicating security threats. Network activity datasets aid ML models by providing extensive training data. We evaluate synthetic data’s efficacy in

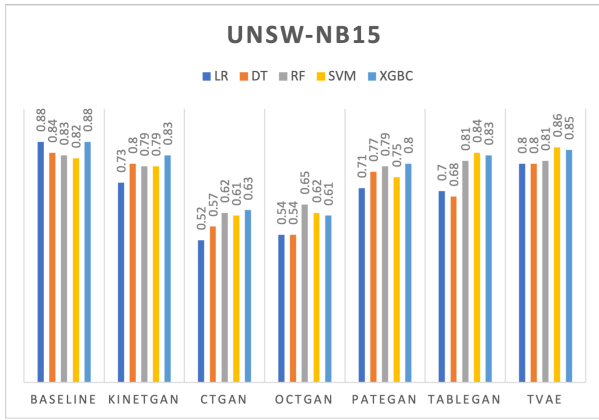


Fig. 4: Comparison of NIDS accuracy for UNSW-NB15

replacing original data by training ML classifiers on both. KiNETGAN, our proposed synthetic data generation model, demonstrates competitive accuracy. Figure 3 demonstrates the accuracy for the baseline classifier with the classifiers trained on synthetic data from generated models including the KiNETGAN model on lab collected data. On lab collected data, KiNETGAN achieves an average accuracy of 0.81, surpassing other models like CTGAN and TableGAN. Figure 4 demonstrates the accuracy for the baseline classifier with the classifiers trained on synthetic data from generated models including the KiNETGAN model on UNSW-NB15 data. For the UNSW-NB15 dataset, KiNETGAN achieves an average accuracy of 0.78, outperforming competing models. These results confirm KiNETGAN’s potential in generating synthetic data for NIDS applications. Notably, KiNETGAN’s accuracy surpasses other tabular data models like CTGAN and OCTGAN. This suggests that KiNETGAN can effectively replace original data in downstream tasks, enhancing the robustness of NIDS classifiers.

C. Privacy Results

Re-identification, Attribute Inference, and Membership Inference attacks are significant privacy threats targeting machine learning models. In our experiments, we observe the effectiveness of KiNETGAN in mitigating these risks.

The Re-identification attack aims to link de-identified data with additional knowledge to reveal sensitive attributes. In Figure 5, we present the results for the accuracy with which the attack model is able to uniquely identify data points assuming it has prior knowledge about 30%, 60% and 90% of the lab collected data. KiNETGAN outperforms other models, achieving an attack accuracy of 0.62 and 0.88 with 60% and 90% overlap with the data respectively.

Attribute Inference attacks deduce sensitive attributes by analyzing seemingly innocuous data points. Figure 6 shows the results for the Attribute Inference attack on synthetic data for the lab collected dataset. KiNETGAN exhibits resilience, with an attack accuracy of 0.3.

Membership Inference attacks determine if a specific data point was part of the model’s training dataset. Figure 7 shows

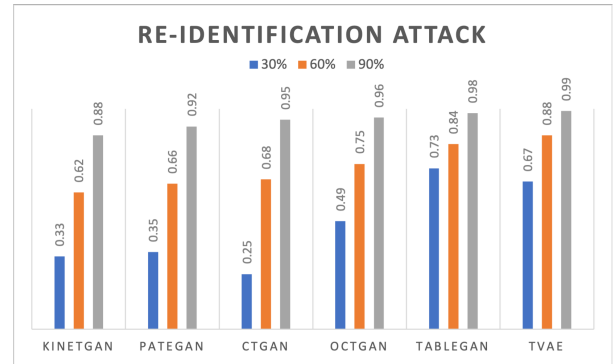


Fig. 5: Comparison of Re-identification Attack with 30%, 60% and 90% overlap on original data

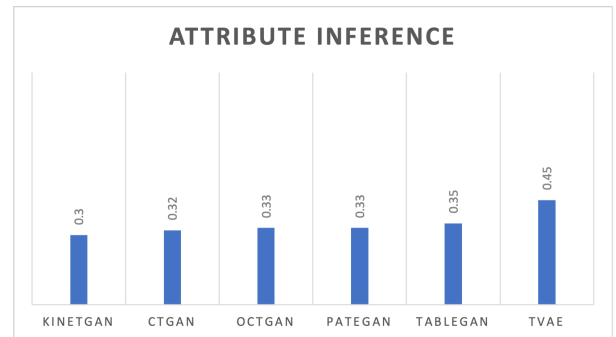


Fig. 6: Comparison of accuracy in Attribute Inference Attack

the results of membership inference attack in WB and FBB setting against lab collected data. KiNETGAN demonstrates resilience in both White-Box (0.54) and Fully Black Box (0.5) settings, outperforming other models like CTGAN and TableGAN. These findings underscore KiNETGAN’s effectiveness in preserving privacy and security in synthetic data generation.

VI. CONCLUSION

To foster collaboration and address privacy concerns, this paper explores a novel knowledge-infused Generative Adversarial Network model for network activity data (KiNETGAN). Leveraging domain knowledge and enhanced GAN training, KiNETGAN overcomes challenges of domain restriction, class imbalance, and privacy preservation, creating realistic representations of network activities. We demonstrate the efficacy of KiNETGAN through synthetic dataset validation and likelihood fitness in our experiments, showing its superiority over other generative models in utility tasks. In future work, we aim to further enhance KiNETGAN’s capabilities and applicability in network intrusion detection. This includes integrating reinforcement learning techniques with KiNETGAN to enable adaptive learning based on real-time threat intelligence, developing algorithms that allow the model to continuously update its parameters in response to new and emerging threats, and conducting extensive field trials to test KiNETGAN’s deployment in various network infrastructures, focusing on scalability, latency, and real-time pro-

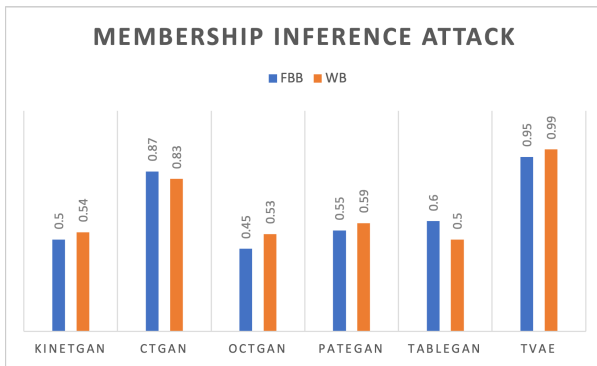


Fig. 7: Comparison of Membership Inference Attack in White Box (WB) and FBB (Fully Black Box) setting

cessing capabilities. This will involve optimizing the model for edge computing environments to ensure low-latency intrusion detection. Additionally, we intend to explore the integration of federated learning with KiNETGAN, enabling collaborative model training across multiple organizations without the need to share raw data, and developing secure aggregation protocols and differential privacy mechanisms to protect individual data contributions. Beyond network intrusion detection, we will adapt KiNETGAN to other critical domains such as healthcare and finance by incorporating domain-specific knowledge into the GAN training process, customizing the model architecture and training algorithms to handle the unique data characteristics and requirements of each domain.

REFERENCES

- [1] K. A. Da Costa, J. P. Papa, C. O. Lisboa, R. Munoz, and V. H. C. de Albuquerque, "Internet of things: A survey on machine learning-based intrusion detection approaches," *Computer Networks*, vol. 151, pp. 147–157, 2019.
- [2] D. Ucci, L. Aniello, and R. Baldoni, "Survey of machine learning techniques for malware analysis," *Computers & Security*, vol. 81, pp. 123–147, 2019.
- [3] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.
- [4] A. Piplai, S. Mittal, A. Joshi, T. Finin, J. Holt, and R. Zak, "Creating cybersecurity knowledge graphs from malware after action reports," *IEEE Access*, vol. 8, pp. 211 691–211 703, 2020.
- [5] A. Piplai, P. Ranade, A. Kotal, S. Mittal, S. N. Narayanan, and A. Joshi, "Using knowledge graphs and reinforcement learning for malware analysis," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 2626–2633.
- [6] S. Dasgupta, A. Piplai, A. Kotal, and A. Joshi, "A comparative study of deep learning based named entity recognition algorithms for cybersecurity," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 2596–2604.
- [7] N. Das, A. Kotal, D. Roseberry, and A. Joshi, "Change management using generative modeling on digital twins," in *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2023, pp. 1–6.
- [8] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.
- [9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

- [10] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.
- [11] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.
- [12] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [13] A. Kotal, A. Piplai, S. S. L. Chukkappalli, and A. Joshi, "Privetab: Secure and privacy-preserving sharing of tabular data," in *Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*, 2022, pp. 35–45.
- [14] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.
- [15] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," *arXiv preprint arXiv:1802.06739*, 2018.
- [16] R. Torkzadehmahani, P. Kairouz, and B. Paten, "Dp-cgan: Differentially private synthetic data and label generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.
- [17] S. Hui, H. Wang, Z. Wang, X. Yang, Z. Liu, D. Jin, and Y. Li, "Knowledge enhanced gan for iot traffic generation," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3336–3346.
- [18] S. N. Narayanan, A. Ganesan, K. Joshi, T. Oates, A. Joshi, and T. Finin, "Early detection of cybersecurity threats using collaborative cognition," in *2018 IEEE 4th international conference on collaboration and internet computing (CIC)*. IEEE, 2018, pp. 354–363.
- [19] L. Elluri, A. Piplai, A. Kotal, A. Joshi, and K. P. Joshi, "A policy-driven approach to secure extraction of covid-19 data from research papers," *Frontiers in big Data*, vol. 4, p. 701966, 2021.
- [20] A. Piplai, A. Kotal, S. Mohseni, M. Gaur, S. Mittal, and A. Joshi, "Knowledge-enhanced neurosymbolic artificial intelligence for cybersecurity and privacy," *IEEE Internet Computing*, vol. 27, no. 5, pp. 43–48, 2023.
- [21] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing, "Rethinking knowledge graph propagation for zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 487–11 496.
- [22] J. Chen, Y. Geng, Z. Chen, I. Horrocks, J. Z. Pan, and H. Chen, "Knowledge-aware zero-shot learning: Survey and perspective," *arXiv preprint arXiv:2103.00070*, 2021.
- [23] A. Kotal, N. Das, A. Joshi *et al.*, "Knowledge infusion in privacy preserving data generation," in *KDD Workshop on Knowledge-infused Learning, 29TH ACM SIGKDD*, 2023.
- [24] A. Kotal, L. Elluri, D. Gupta, V. Mandalapu, and A. Joshi, "Privacy-preserving data sharing in agriculture: Enforcing policy rules for secure and confidential data synthesis," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 5519–5528.
- [25] J. Kim, J. Jeon, J. Lee, J. Hyeong, and N. Park, "Oct-gan: Neural ode-based conditional tabular gans," in *Proceedings of the Web Conference 2021*, 2021, pp. 1506–1515.
- [26] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International conference on learning representations*, 2018.
- [27] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *Proceedings of the VLDB Endowment*, vol. 11, no. 10, pp. 1071–1083, 2018.