Addressing Class Imbalance in Threat Detection Datasets: A Literature Review on Data Augmentation Techniques and Emerging Approaches

Abid Ali Sameer Mohammed
Dept. of Information Technology Muffakham Jah College of Engineering and Technology
(Affiliated to Osmania University)
Hyderabad, India

Zahed Ali Salman Mohammed
Dept. of Computer Science and Engineering
Muffakham Jah College of Engineering and Technology
(Affiliated to Osmania University)
Hyderabad, India

Habib Ayan Aziz Khan
Dept. of Information Technology Muffakham Jah College of Engineering and Technology
(Affiliated to Osmania University)
Hyderabad, India

***Abstract*—Since more and better cyberattacks happen, we need stronger and more reliable intrusion detection systems. Machine learning methods have produced great results so far, but their usefulness is often limited by the fact that there are far more unimportant than harmful incidents in cybersecurity. This literature review examines the challenges posed by imbalanced datasets in AI-driven cybersecurity and explores various approaches to address this issue. This paper reviews the motivation for data- centric threat detection, how to use GANs to make synthetic data and the various approaches for managing unbalanced data. We combine new studies to outline how GAN-based approaches can make datasets more balanced and improve the results of ML- based intrusion detection.**

I. INTRODUCTION

Cybersecurity systems are undergoing a transformation as cyber threats become more advanced, frequent, and adaptive [1], [2]. Traditional rule-based detection methods have become increasingly insufficient in managing these evolving threats, leading to a paradigm shift toward AI- and ML-based solutions [3], [4]. However, the effectiveness of these systems is heavily dependent on the availability of high- quality, diverse, and balanced training data [5].

In real-world cybersecurity datasets, benign traffic tends to dominate while malicious activity is significantly underrepresented, resulting in severe class imbalance [6], [7]. This imbalance causes ML classifiers to become biased toward the majority class, compromising their ability to detect rare and stealthy attacks. Additionally, the confidential nature of cyber incidents often limits access to large, labeled datasets needed for supervised learning [8].

To address these limitations, researchers have increasingly turned to synthetic data generation methods, particularly Generative Adversarial Networks (GANs) and their tabular variants such as CTGAN. These models have demonstrated success in generating realistic, privacy-preserving, and class-balanced datasets that enhance detection performance in imbalanced settings [5], [9]– [10].

II. MOTIVATION: NEED FOR QUALITY DATA IN CYBERSECURITY (AI THREAT DETECTION

The importance of automated and intelligent security systems has been brought to light by the growing complexity of cyberthreats and the dynamic nature of modern networks. Although conventional rule-based methods have been useful, they usually have trouble identifying novel or subtle attacks. As a result, there is a noticeable trend toward threat detection powered by AI. However, the availability of high-quality, balanced, and representative datasets—which remain a serious challenge—is crucial to this progress.

Dhanushkodi and Thejas [1] emphasize this issue in their comprehensive study on AI-enabled threat detection systems. The authors detail how artificial intelligence methods, particularly deep learning and machine learning models, are capable of uncovering complex patterns in network traffic to detect threats. However, the performance of these models deteriorates when trained on real-world datasets characterized by class imbalance, noisy features, and low representation of rare attacks. A notable insight from their work is the identification of data quality as a critical limitation, explicitly stating that "requires high quality data for optimal performance" is a core challenge.

This observation highlights a foundational challenge that motivates further investigation into data augmentation techniques, which seeks to address the data imbalance problem directly.

As a result, Soliman et al. [2] launched the RANK framework, a versatile AI platform that helps detect Advanced Persistent Threats (APTs) by automating the correlation of alerts and extraction of incidents. They group warnings from the network by using threat intelligence and form them into charts, listing the notices by incidents for analysts to look at. Still, according to the authors, these systems are limited in their application because there are not enough well-labeled, diverse data available for training. It further suggests that automation cannot properly work in cybersecurity without resolving the main issue of small datasets to train machines on.

The authors Tellache et.al. [13] also investigate an additional aspect, namely, how flexible AI systems can be. With multi-agent reinforcement learning, their IDS adapts to changes in attacks and the state of the network. They particularly point out that there are classes with smaller samples in the standard CIC-IDS-2017 data. Even though the authors use cost-sensitive learning and weighted loss, they note that the presence of imbalance still makes it difficult to spot underrepresented attacks. It confirms that all adaptive systems must have a stable and proportionate amount of data to operate effectively.

What these studies all agree on is that an unbalanced dataset holds back AI’s use in cybersecurity. The ability of detection systems to perform depends on the relevant training data, regardless of the type of AI used. These studies collectively underscore the relevance of addressing data imbalance in cybersecurity research, as it suggests using generative models (such as GANs and CTGANs) and generating more data artificially to help train models for cybersecurity applications.

III. Synthetic Data Generation Using GANs

Because there is less and less available network attack data compared to legitimate data, IDS systems are losing their effectiveness. Generating synthetic data with GANs is proving to be an effective answer to this problem. In this portion, recent works that support the technical background are examined, particularly CTGANs and their importance for good and realistic generation of tabular data.

A. *Enhancing IDS with GAN-Generated Data*

The Zhao et al. [14] research highlights the use of GANs in network intrusion detection systems (NIDS). To create synthetic network traffic and attack samples, researchers used the CIC-IDS2017 dataset by employing Vanilla GAN, Wasserstein GAN and CTGAN. With the additional data, the classification models showed more significant improvements in precision, recall and F1-score. They showed that comparing artificially generated data and samples from the original data set indicates that using GANs can alleviate scarcity and imbalance of various types of cyber threats. The use of GANs for data resampling in IDS is clearly supported by this evidence. The overall process of synthetic sample generation and feedback using CTGAN for class imbalance mitigation is illustrated in Fig. 1.

Ammara et al. [5] offer a detailed comparison of various generative techniques, which is discussed further in Section III.C in the context of GAN-based augmentation for imbalanced datasets.

B. *Review of Generative Models for Cybersecurity*

The authors Agrawal et al. [9] highlighted a variety of generative models for cybersecurity and concentrated on the strengths and weaknesses of GANs for creating fake attack data. They argued that data realism matters a lot, as it was shown to cause issues like mode collapse, ineffective representation of minority groups and fitting the models to the majority traffic rules. The article covered both traditional and newer GAN architectures, showing how they address the problems seen in older ones. It was found that while GAN-generated data improved deep learning classifiers for IDS tasks, it is vital to ensure that the data and the final classifiers are properly tested before they are used in practice.

C. *Comparative Analysis of Synthetic Data Techniques*

Moreover, in their study, Ammara et al. [5] measured and evaluated synthetic data generation techniques, highlight- ing the benefits of GANs (including CTGAN, CopulaGAN and GANBLR++) over basic approaches (e.g. SMOTE and ADASYN). With data from NSL-KDD and CICIDS2017, the authors observed that CTGAN and CopulaGAN performed better than others, both in terms of the accuracy of their results compared to the real data and their capacity to improve IDS models. The research used mutual information to select features which improved the quality of generated data, proving that utilizing CTGAN is better for cybersecurity when there is major biased data.

D. *CTGAN: Tailored for Tabular Data*

CTGAN is a type of GAN designed by Xu et al. [10] to address specific issues that appear in tabular data, for instance, different types of data, varied distribution patterns and a skewed number of samples for each class. CTGAN relies on a conditioned generator and a training via sampling strategy to address the problem of categorical columns being unevenly distributed. The use of mode-specific normalization in CTGAN results in much better quality and varied data than found in other deep generative models. Testing CTGAN on several datasets, including healthcare and cybersecurity, revealed that it produces much better quality tabular data than other deep learning models and traditional Bayesian methods.

IV. HANDLING IMBALANCED DATASETS IN MACHINE LEARNING

Supervised machine learning often deals with class imbalance, mainly when it comes to intrusion and DDoS detection. Typically, datasets in these areas have a lot of data related to harmless transactions but not much regarding threats, so classifiers handle these poorly. Researchers have examined the

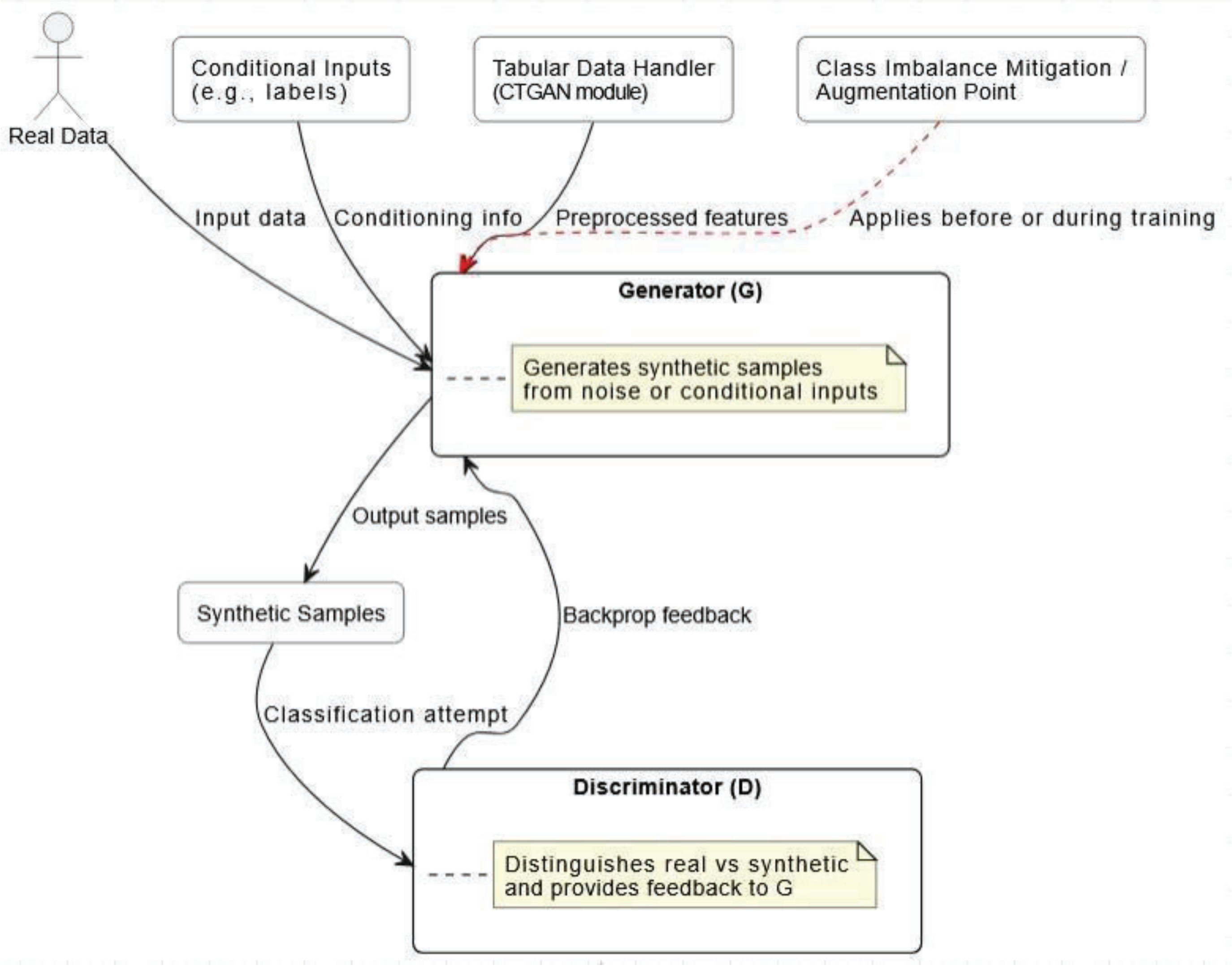
Real Data
Conditional Inputs (e.g., labels)
Tabular Data Handler (CTGAN module)
Class Imbalance Mitigation / Augmentation Point
Input data
Conditioning info
Preprocessed features
Applies before or during training
Generator (G)
Generates synthetic samples from noise or conditional inputs
Output samples
Synthetic Samples
Backprop feedback
Classification attempt
Discriminator (D)
Distinguishes real vs synthetic and provides feedback to G

Fig. 1. GAN-based Synthetic Data Generation Pipeline

drawbacks of an imbalanced learning environment, focusing on unfair treatment of common classes, difficulties in remembering details about uncommon classes and inaccurate measures of learning. This section surveys the relevant literature focused on the origins, consequences and methods of using learning with imbalanced data.

A. Nature and Implications of Imbalanced Data

If the majority class makes up a large part of the data, the resulting model may have problems detecting the minority groups. They He and Garcia [3] state that it occurs because traditional learning algorithms depend on classes being balanced and they fail when the data is skewed. Their work points out that both natural scarcity of some classes and limited sampling affect the way models are sensitive to minority classes. Such methods rely on using precision, recall, F1-score and ROC curves to eliminate the risks of being misled by accuracy.

Guo et al. [7] draw attention to the problems of classic classifiers such as SVMs, decision trees and logistic regression when handling data with uneven class sizes. They consider new techniques in machine learning as well as practical cases like detecting fraud, medical diagnosis and cyber crimes. They offer a single approach to data mining that includes tasks such as preprocessing with SMOTE, classification with ensemble methods and evaluation.

B. Taxonomies and Techniques for Class Imbalance

In [6], Rezvani and Wang divide methods for handling imbalanced data into three types: (1) preprocessing the data, (2) changing the learning algorithm itself and (3) combining methods in both ways. They pointed out that data and the later task should determine which approach performs best, since no technique is the best for every case. They have also applied this taxonomy in regression which is helpful for predicting performance and scoring anomalies in cybersecurity.

Branco et al. [15] expand on the subject by separating imbalance from the general idea of cost-sensitive learning, saying that any good strategy must handle low proportion classes as well as class importance determined by the user. Post-processing techniques are introduced in their taxonomy to modify predictions or thresholds and this allows us to get better recall without re-training. They suggest tailoring objectives to address specific needs, as shown by the target to reduce false negatives in DDoS detection.

C. Comparative Evaluation of Resampling Methods

Çürükogˇ lu [12] presents findings from examining some popular resampling techniques: Random Oversampling (ROS), Random Undersampling (RUS), Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN). Combining SMOTE with Gradient Boosting Classifier results in getting the best F1-scores. Theseoutcomes show that SMOTE is widely used in cybersecurity tasks that rely on creating a representative set of samples from the minority class to support strong learning.

Kotsiantis et al. [11] provide an overview of how different imbalance-handling methods such as cost-sensitive classifiers, bagging/boosting and hybrid techniques, have been used over time. They believe that using ensemble learning and resampling methods together often produces the best results for tuning between learning from all classes and special classes. A side-by-side comparison of traditional resampling methodsand generative approaches for handling class imbalance is provided in Table 1.

D. Benchmarking and Open Challenges

As the area of imbalanced learning develops, new and difficult problems keep appearing. According to Johnson and Khoshgoftaar [4], while deep learning has strong expressive power, it remains sensitive to class imbalance and may benefit from methods like resampling, using special attention or changing the training loss. Issues in cybersecurity are often greater because the attacks and their nature change along with any new developments. This condition in data leads experts to make continual updates to their methods of removing class imbalance, placing more emphasis on approaches that update automatically and those that generate synthetic data.

E. Traditional Approaches to Handling Imbalanced Data

Before advanced functions for generating synthetic data were developed, old methods were used to deal with problems related to imbalanced sets of data. We can organize these approaches into Data-Level Techniques and Algorithm-Level Techniques [3].

1) Data-Level Techniques: At the data level, the imbalance is addressed by changing the proportion of samples in each class [3], [11]. Common ways are:

Oversampling: Part of this is to copy or slightly alter minority class examples in order to give the model more data from the underrepresented group [11]. Random Oversampling copies existing data, whereas SMOTE creates artificial mem- bers of the group by filling the spaces between them [11]. Yet, applying oversampling without care can result in overfit- ting [11].

Undersampling: The number of examples in the majority class decreases by randomly removing samples [11]. This approach can ensure that classes are balanced, yet it might lose helpful examples from the main class group [11]. Among the techniques are Random Undersampling and more sophisticated options that aim to eliminate majority class instances that fit particular rules [11].

2) Algorithm-Level Techniques: The process designed at the algorithm level is adjusted so that it pays more attention to the minority class. Examples include:

When using this approach, each class is given a different penalty for being misclassified, usually with a bigger penalty for misclassifying minority cases [11]. It makes the model work harder to identify examples from the minority class.

Most classifiers provide a probability or score and the label is set as positive or negative based on whether the probability is above or below the default threshold, usually set to 0.5. For classifiers on unbalanced datasets, raising or lowering this threshold may improve the detection of the smaller class [11].

Combining Balanced Random Forest and EasyEnsemble with several classifiers trained on replicated data helps tomake predictions more reliable with imbalanced datasets [11].

Even though these typical measures improved the balance of classes in cybersecurity problems, they often do not work well with advanced and detailed cybersecurity data. Over- sampling may bring about overfitting with little added value, whereas undersampling might result in the loss of potentially informative samples. Tuning algorithm-level techniques can be demanding, as these methods may fail when data is strongly unbalanced [3]. Therefore, the section III examines the use of GANs for generating synthetic data.

V. Summary of Gaps Identified

There are only a limited number of studies combining GANs and techniques for handling inequality in data classes for use in network intrusion detection. Most research works address data augmentation and data imbalance independently, rather than considering their joint use. The literature reveals a lack of integrated approaches that jointly address synthetic data generation and class imbalance in intrusion detection systems using generative models to handle distribution issues in cybersecurity topics.

The different approaches and significant studies across the literature are summarized in Table 2 below. The rows in this table demonstrate how the area of Imbalanced Learning has progressed from observing class imbalance to developing methods for making more data.

It is clear from Table 2 that there are several projects right now involving both new algorithms and data solutions, mainly focused on resolving problems related to uneven data and making top-quality samples. Here, we explore new evaluation methods and trends that demonstrate the improvement and progress of applying data for cybersecurity.

TABLE I
COMPARISON OF TRADITIONAL VS GAN-BASED BALANCING TECHNIQUES

Method	Type	Handles Imbalance?	Pros	Cons
SMOTE	Oversampling	Partial	Simple, fast complexity	May overfit, can't capture
ADASYN	Oversampling	Yes	Focuses on hard examples	Can introduce noise
CGAN	Generative	Yes (with labels)	Class-conditional synthesis	Mode collapse risk
CTGAN	Generative (Tabular)	Yes (indirectly)	Preserves tabular format	Sensitive to tuning
QGAN	Quantum Generative	Theoretical	High complexity capture	Experimental, not scalable

VI. EMERGING TRENDS IN SYNTHETIC DATA FOR CYBERSECURITY

With further advancements in technology and studies, more solutions are appearing to fix important issues in synthetic data such as its accuracy, confidentiality, size and alterability. Here, we explore the most important new trends in synthetic data generation for cybersecurity and see how they relate to data-based threat detection systems.

A. *Digital Twins for Cybersecurity Data Simulation*

Simulation of networks using digital twins is gaining popularity in cybersecurity. They create a reliable, repeatable and customizable setting for generating bots and attack data for many different situations.

Mylrea and Gourisetti [16] show how systems for cyber-physical digital twins can accurately model ICS and carry out test attacks such as command injection and denial-of-service (DoS) for critical infrastructure. These AI simulations provide data that reflects reality and makes it possible for developersto test models without being vulnerable to security threats such as zero days.

Digital twin approaches have also been applied to other areas including enterprise IT, cloud computing and IoT systems. Within these environments, data on normal activities and fake intrusions is collected which helps IDS, anomaly detection and response teams make well-balanced models.

B. *Quantum-Enhanced Data Generation*

Qunatum computing is at an early point, but it is already impacting data generation by means of quantum generative models. Lloyd and Weedbrook [17] developed QGANs that make modeling of high-dimensional probability distributions far more efficient than using conventional networks.

Li et al. [18] conducted a recent study, applying QGANs to cybersecurity data and showing that they can uncover the relationship between important features in network logs. Although quantum generative models are still being developed on small datasets, they might be able to solve the problems of 'mode collapse' and 'sparseness' found in traditional GANs. While working practical hybrids has not gotten far, developers are investing in early systems to shorten training and generation procedures. When we have better hardware, these approaches can produce better and more efficient synthetic datasets used by AI to detect threats.

C. *Privacy-Preserving Synthetic Data (Differential Privacy & Federated Learning*

Since there is tension between wanting to use data and protecting privacy, more attention has been given to methods that protect privacy in data. DP has become widely used since it guarantees that data is not too detailed but still reflects how real data is described in the aggregate.

Jordon et al. [19] introduced PATE-GAN which blends GANs with the Private Aggregation of Teacher Ensembles (PATE) algorithm. This allows us to generate data that protects confidentiality which is very valuable in healthcare and cyber security.

When data is not centralized, data scientists often run GANs on endpoint devices and firewalls to make synthetic information. As a result, privacy is protected and everyone involved can train their models together. Hitaj and et al. [20] discovered that federated GANs are capable of making synthetic intrusions logs for different decentralized enterprises without compromising privacy or effectiveness.

D. *Integration with Reinforcement Learning Environments*

Simulating different cyberattacks using synthetic data is common in training RL agents for cybersecurity. Bedi et al. [21] add synthetic data to cyber ranges to help RL agents train for adaptive security. Thanks to synthetic data, RL agents succeed at dealing with a wide spectrum of cyber threats.

This trend relates to the movement toward using autonomous and active cyber defense strategies that are trained with simulations. In this form of machine learning, synthetic data is used in training and also keeps changing as the learning process goes on.

E. *Automated Evaluation and Verification of Synthetic Data*

Now that synthetic data is crucial in cybersecurity AI, it's necessary to assess these pipelines with stronger techniques. Previously, just distributional similarity and classifier performance were used, but now specialists also look at tasks and their robustness against attacks.

Torkzadehmahani et al. [22] develop a framework that uses statistics (such as Wasserstein distance), privacy measurements and the performance of the data on classification and anomaly detection. As a result, the artificial data resembles real data and at the same time runs smoothly and follows privacy rules.

SDMetrics and CTGAN-Benchmark are being applied for standardized testing by researchers and developers, helping to both create reproducible projects and deploy data without
modification.

TABLE II
COMPARATIVE SYNTHESIS OF KEY STUDIES ON SYNTHETIC DATA FOR CLASS IMBALANCE IN CYBERSECURITY

Study (Authors, Year)	Research Objective	Methodology Used	Key Contribution	Limitations	Open Challenges / Gaps
Zhao et al. (2021)	Improve NIDS performance using GAN-based synthetic data	Vanilla GAN, WGAN, CTGAN on CIC-IDS2017 dataset	Improved precision, recall, and F1 using synthetic attack data	Focused only on CIC-IDS2017; generalization untested	Scalability across diverse settings; real-world validation
Agrawal et al. (2024)	Survey generative models for generating synthetic cyberattack data	Literature review of GAN variants applied in cybersecurity	Identified generative choices and evaluated realism and utility	No benchmarking across datasets	Need standardized evaluation frameworks
Ammara et al. (2024)	Benchmark GANs and conventional models for cybersecurity data synthesis	Compared CTGAN, CopulaGAN, GANBLR++, CastGAN on NSL-KDD, CICIDS-2017	CTGAN and CopulaGAN outperformed others in fidelity and utility	Limited to two datasets	Explore temporal data; integrate with live cyber systems
Çürükoğlu (2019)	Compare resampling methods for class imbalance in classifiers	ROS, RUS, SMOTE, ADASYN applied on classification tasks	SMOTE + GBC achieved highest F1	Excludes generative models	Extend to deep models and cybersecurity datasets
Kotsiantis et al. (2005)	Review imbalance techniques and ensemble methods	Taxonomy of algorithm- and data-level approaches	Summarized hybrid and ensemble strategies	Lacks deep learning and GAN coverage	Integrate modern neural and generative methods
Rezvani & Wang (2023)	Taxonomize imbalance learning for classification and regression	Structured review and comparative analysis	Unified taxonomy including regression, SVM, hybrid methods	Emphasized SVM; lacks GAN/cybersecurity context	Align taxonomy with GAN-based approaches for cyber defense
Branco et al. (2016)	Survey predictive modeling on imbalanced domains	Theoretical and empirical review	Introduced post-processing strategies and new performance metrics	Sparse cybersecurity relevance	Embed methods in real-time threat detection
Guo et al. (2017)	Broad review of imbalance learning across applications	Survey of 500+ studies with technical and applied lens	Broad taxonomy and domain-wise mapping	Synthetic data underexplored	Combine imbalance learning and synthetic data generation
He & Garcia (2009)	Review learning from imbalanced data	Methodological review of sampling, cost-sensitive learning	Established foundational framework for imbalance learning	Predates deep learning and GANs	Recontextualize for modern generative and DL-based models
Johnson & Khoshgoftaar (2019)	Review deep learning techniques for class imbalance	Surveyed 15 DNN studies with imbalanced data	Highlighted limited progress in DL + imbalance	Focus on vision tasks; cybersecurity underrepresented	Benchmark deep models for imbalance in cybersecurity contexts
Xu et al. (2019)	Generate high-fidelity synthetic tabular data with imbalance handling	Conditional Tabular GAN (CTGAN); mode-specific normalization; training-by-sampling	Introduced CTGAN: the first GAN architecture to directly handle imbalanced discrete columns and multimodal continuous features	Focused on general tabular datasets; lacks cybersecurity-specific evaluation	Apply CTGAN to cybersecurity data (e.g., IDS logs); integrate with privacy-preserving methods and real-time systems

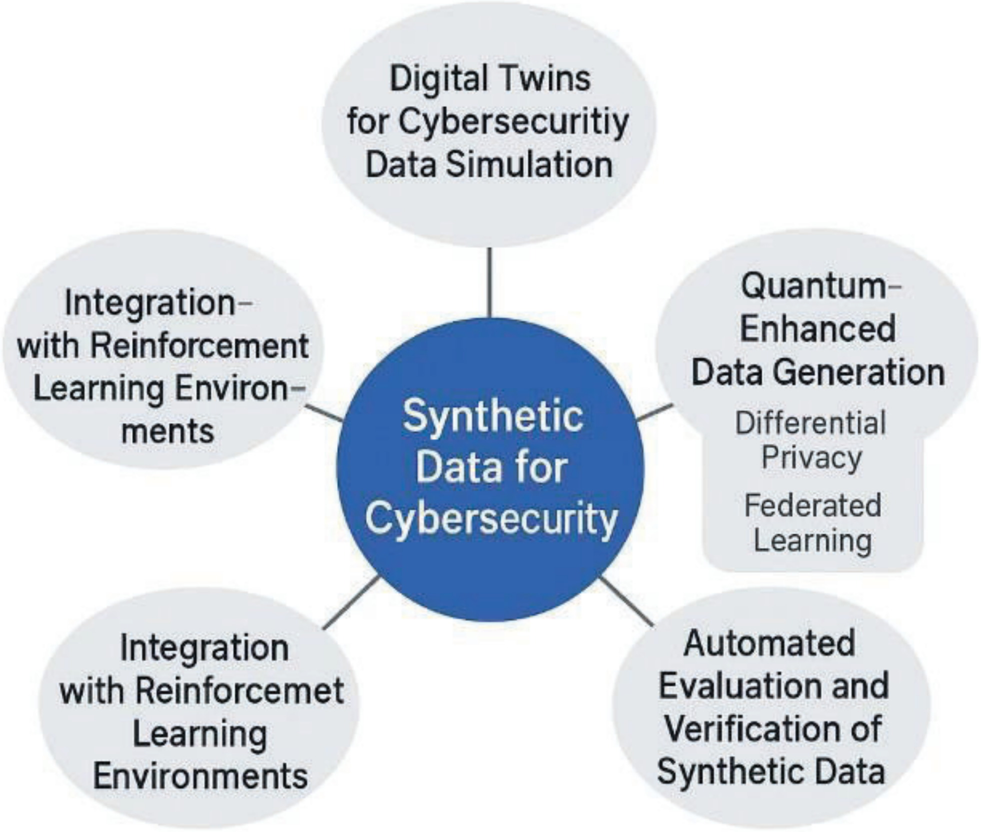
Digital Twins for Cybersecuritiy Data Simulation
Integration- with Reinforcement Learning Environ- ments
Quantum- Enhanced Data Generation
Differential Privacy
Federated Learning
Synthetic Data for Cybersecurity
Integration with Reinforcemet Learning Environments
Automated Evaluation and Verification of Synthetic Data

Fig. 2. Conceptual Map of Emerging Trends

G. Key Takeaways from Emerging Trends

Large-scale data environments can be built using digital twins.

Using quantum-enhanced models makes it possible to represent information more strictly and with increased complexity. Both differential privacy and federated learning help to ensure that data can be useful without becoming insecure.

When used with RL and autonomous systems, synthetic data helps improve cyber security.

Reliable and trustworthy use of synthetic data is possible because of healthy evaluation strategies.

This means that synthetic data is now an indispensable feature of today's intelligently designed and well-protected cybersecurity solutions.

H. Conclusion

Cybersecurity datasets that do not represent all classes continue to reduce the effectiveness of AI in detecting cyber threats. As we have demonstrated, machine learning classifiers that are trained on unevenly distributed data fail to detect unusual but harmful attacks, resulting in a high rate of missed attacks and less useful outcomes.

Several data augmentation techniques have been explored in the research literature to handle this issue. Although traditional approaches like SMOTE and ADASYN have helped, they do not maintain the realism of context in samples. Meanwhile, new advances in generative modeling, especially with GANs and CTGAN, have shown that they can produce more suitable synthetic data from the minority attack classes. They improve the balance within datasets and also help with generalization when included in intrusion detection systems.

In addition, this paper considers novel approaches that move beyond the classic GAN designs. With digital twin simulations, quantum-enhanced data generators, federated synthetic learning and PATE-GAN privacy-preserving frameworks, the stage is being set for better, scalable, secure and aware solutions for data generation. These practices prove that now, the input data's structure and quality matter just as much as the design of the models.

Encouraging facts and improvements, yet many gaps are still evident. The testing of many GAN-based solutions is limited to particular types of attacks or only certain situations. There are also very few investigations that focus on tuning detection models for fast operation or that thoroughly analyze the risk to privacy when artificial data is introduced. To solve these issues, we need cooperation between adversarial learning, privacy engineering and resources focused on threat intelligence.

Overall, generating synthetic data using GANs is a growing and reliable way to handle the issue of class imbalance in datasets for threat detection. Research in the future should concentrate on making data augmentation safe, effective and useful for different domains when integrated into modern cybersecurity solutions.

REFERENCES

[1] D. Kavitha and S. Thejas, "AI enabled threat detection: Leveraging artificial intelligence for advanced security and cyber threat mitigation," IEEE Access, 2024.

[2] H. M. Soliman, D. Sovilj, G. Salmon, M. Rao, and N. Mayya, "RANK: AI-assisted end-to-end architecture for detecting persistent attacks in enterprise networks," IEEE Trans. Dependable Secure Comput., vol. 21, no. 4, pp. 3834–3850, 2023.

[3] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[4] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," J. Big Data, vol. 6, no. 1, pp. 1–54, 2019.

[5] D. A. Ammara, J. Ding, and K. Tutschku, "Synthetic Data Generation in Cybersecurity: A Comparative Analysis," 2024, arXiv:2410.16326. [Online]. Available: https://arxiv.org/abs/2410.16326

[6] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," Appl. Soft Comput., vol. 143, p. 110415, 2023.

[7] Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert Syst. Appl., vol. 73, pp. 220–239, 2017.

[8] V. Shanmugam, R. Razavi-Far, and E. Hallaji, "Addressing class imbalance in intrusion detection: A comprehensive evaluation of machine learning approaches," Electronics, vol. 14, no. 1, p. 69, 2024.

[9] G. Agrawal, A. Kaur, and S. Myneni, "A review of generative models in generating synthetic attack data for cybersecurity," Electronics, vol. 13, no. 2, p. 322, 2024.

[10] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Adv. Neural Inf. Process. Syst., vol. 32, 2019.

[11] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," GESTS Int. Trans. Comput. Sci. Eng., vol. 30, no. 1, pp. 25–36, 2006.

[12] N. Çürükoğlu, "Imbalanced Dataset Problem in Classification Algorithms," in 2019 1st Int. Informatics and Software Eng. Conf. (UBMYK), Nov. 2019, pp. 1–5.

[13] A. Tellache, A. Mokhtari, A. A. Korba, and Y. Ghamri-Doudane, "Multi-agent Reinforcement Learning-based Network Intrusion Detection System," in NOMS 2024—2024 IEEE Netw. Oper. Manag. Symp., May 2024, pp. 1–9.

[14] X. Zhao, K. W. Fok, and V. L. L. Thing, "Enhancing network intrusion detection performance using generative adversarial networks," Comput. Secur., vol. 145, p. 104005, 2024.

[15] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," ACM Comput. Surv., vol. 49, no. 2, pp. 1–50, 2016.

[16] M. Mylrea and S. N. G. Gourisetti, "Cyber-physical digital twins for cybersecurity simulation and training in critical infrastructure," Technologies, vol. 6, no. 4, p. 118, 2018.

[17] S. Lloyd and C. Weedbrook, "Quantum generative adversarial learning," Phys. Rev. Lett., vol. 121, no. 4, p. 040502, 2018.

[18] J. Li, H. Yu, W. Bai, and Y. Tang, "Quantum GAN for synthetic cybersecurity data," Quantum Inf. Process., vol. 20, no. 3, p. 81, 2021.

[19] J. Jordon, J. Yoon, and M. van der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in Proc. Int. Conf. Learn. Represent. (ICLR), 2019.

[20] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the GAN: Information leakage from collaborative deep learning," in Proc. ACM Conf. Comput. Commun. Secur. (CCS), 2017, pp. 603–618.

[21] S. Bedi, S. Shetty, and M. Papa, "Using reinforcement learning and synthetic data to secure autonomous cyber systems," J. Cybersecurity Privacy, vol. 1, no. 2, pp. 370–386, 2020.

[22] R. Torkzadehmahani, P. Kairouz, and B. Paten, "DP-CGAN: Differentially private synthetic data and label generation," 2019, arXiv:1901.02266. [Online]. Available: https://arxiv.org/abs/1901.0226