Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Using Genai for Synthetic Data Generation in Cybersecurity and Compliance

Puneet Pahuja
IT, Bachelor of Engg, MDU Rohtak
Indian Land, USA
Gmail puneet7498@gmail.com

ARTICLE INFO

Received: 30 Dec 2024

Revised: 19 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

The increasing nature of the cybersecurity threats and high-pressure requirement of the norms compliance have resulted in the need to think differently about how data is managed and secured. Generative Artificial Intelligence (GenAI) provides a potential solution, incorporating the formulation of artificial data and developing these data with a high level of safety, privacy, and confidentiality, which implies guaranteed testing, training, and compliance auditing without any security concerns. GenAI models produce synthetic data that mimics the statistics of unprecedented data, allowing the security field to use generated data without revealing sensitive data to vulnerable access; this makes it particularly appealing in techniques as threat modelling, anomaly detection, simulators of red-teaming, and similar ones. Moreover, artificial datasets can be utilized in ensuring regulatory adherence because they enable testing of systems across various environments based on the type of rules that should govern data use. In spite of this potential, synthetic data generation in GenAI has various challenges, such as data fidelity, bias reprocessing, and regulatory objects. Nevertheless, the progress of GenAI models like GAN, transformers, and diffusion models takes place and enhances the reliability and security of generated data. This paper addresses how GenAI helps in cybersecurity and compliance, evaluates the advantages and disadvantages of the technology, and reflects on recent research and practical opportunities flag identify possible ways to integrate the new technology through the lens of current studies and real-life examples.

Keywords: Generative AI, Synthetic Data, Cybersecurity, Data Privacy, Compliance Testing, Threat Detection, Privacy-Preserving AI, Regulatory Risk

1. INTRODUCTION

i. Growing privacy Requirement in Cybersecurity Solutions

Nowadays, in the digital era of a highly interconnected world, cyber threats have increased in volume, sophistication, and influence. At the same time, international regulatory initiatives like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have laid down stringent changes on how data, in particular personal and sensitive data, ought to be dealt with (Mohawesh et al., 2025). Organizations are now in the twin challenge of protection against serious threats and at the same time are expected to comply with the principles of privacy. This stress leads to the necessity of new methods at the heart of which it is possible to conduct deep data analysis and train models without infringing on personal privacy.

Generative AI (GenAI) is a solution to this issue as it promises to resolve this problem with a safe and ethical alternative to working with real user data. GenAI facilitates such tests by creating synthetic datasets of high quality which maintain the statistical structure of actual data, allowing cybersecurity teams to test their systems, identify the vulnerabilities and train the model, but without the exposure of any sensitive data (Bhardwaj, 2025; Kanchi et al., 2025). This has been tractioned in other sectors like finance, health and defence spheres, where innovation is often constrained by data access of information disclosure (Joshi, 2025).

ii. GenAI as an Immigrant Leader in Synthetic Data Innovation

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

AI-based models, especially Generative Adversarial Networks (GANs), Transformers, and Diffusion Models, are transforming the process of data creation, manipulation, and utilization. Such models are capable of generating synthetic data which almost does not differ in anything about real-world datasets, allowing a variety of purposes such as fraud detection and simulation of network breaches (Ankalaki et al., 2025; Gupta et al., 2023). These abilities enable organizations in the field of cybersecurity to develop very realistic training and test structures without exposing sensitive intelligence on their business operations.

ThreatGPT, an example of a fine-tuned large language model (LLM) has also been applied and used to simulate cyberattacks and produce adversarial inputs to evaluate detection systems (Gupta et al., 2023). On the same note, using GenAI to create synthetic logs, attack signatures, and traffic patterns enables cybersecurity analysts to simulate what-if scenarios, perform proactive and simulate attacks simulated by the engine would, in person it completely break the law, and breaking the law is entirely unwarranted in a live environment (Metta et al., 2024).

Table 1: Summary of GenAI Applications for Synthetic Data in Cybersecurity and Compliance

Dimension	Description	GenAI Contribution
Cybersecurity Training	Difficulty in obtaining real attack data without risking system integrity	GenAI generates synthetic attack patterns, intrusion logs, and malware traces for safe model training
Threat Simulation	Need to simulate diverse and rare threat scenarios without exposing real systems	Enables proactive red-teaming, penetration testing, and anomaly simulations using synthetic data
Data Privacy	Legal and ethical concerns with using real user data in model development	Synthetic data preserves statistical realism without containing identifiable information (GDPR-compliant)
Compliance Testing	Repeated audits and validation across varying data conditions required by regulations	Simulates audit conditions using synthetic datasets under ISO, NIST, and SOC 2 frameworks
Bias & Fidelity Risk	Synthetic data might replicate biases or produce unrealistic outputs	Requires robust GenAI validation pipelines and ethical oversight to ensure fairness and accuracy
Governance	Lack of legal clarity on synthetic data use in regulatory audits	Demands new policy frameworks, traceability mechanisms, and synthetic data provenance tools

iii. Construction of Data Scarcity in Compliance and Threat Intelligence

One of the largest bottlenecks is still the access to complete datasets of cybersecurity in the creation of smart defence systems. The majority of organizations do not want to (or cannot by law) reveal their incident logs, reports of breaches, and internal threat intelligence. Because of this, machine learning models tend to poorly generalize because of under privilecting or overfitting on small data sets. The solution can be found in GenAI, which provides a variety of balanced and abundant synthetic datasets to fill these gaps without breaking the law or ethical principles (Balasubramanian et al., 2025; Gafni and Levy, 2025).

Moreover, artificial information improves the process of regulatory compliance. The alignment with such frameworks as NIST, ISO 27001, and SOC 2 often has to be repeatedly verified under various conditions in diverse environments and under various amounts of data. GenAI-harvested datasets also allow the simulation of these situations: thus, the systems retain their resilience and comply with different operational conditions (Huang et al., 2024; Verma et al., 2025).

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

iv. The difficult side, and the Ethical aspect

Although it has the potential, there are new challenges with GenAI-based synthetic data generation. These pertain to data fidelity (i.e., to what extent synthetic data behaves like in the real world), spread of biases, and the danger of generating deepfakes or misuse of synthetic data in a court of law (Christodorescu et al., 2024; Singh et al., 2024). In addition, even newly- synthesized data can potentially still retain privacy risks when not generated and verified appropriately, and particularly when trained on small or unbalanced data populations.

Also, the implementation of regulatory frameworks has not been fully developed according to synthetic data practices. This makes the question of the legal admissibility of synthetic data around the act of audit, forensic analysis, or evidence-based decision-making rather difficult (Nott, 2025). To safeguard the GenAI systems, security professionals should hence integrate stringent governance practices and build ethical policies related to AI (Huang, Wang, et al., 2024).

v. Objective and Crosspoints of this Article

This paper discusses the GenAI use in synthetic data creation toward cybersecurity and compliance based on the current research, architecture, and application. It explores the ways GenAI may address the drawbacks of the conventional model of accessing data, enhance security frameworks, assist in compliance audits, and safeguard user privacy. Concurrently, it assesses the technical and ethical constraints that must be overcome to achieve responsible use.

The article aims to enlighten both practitioners and policymakers by offering a cohesive perspective on the rapidly emerging field, as well as clarifying the role of GenAI transformation, which is potentially unmatched, but must be accompanied by protective measures.

2. LITERATURE REVIEW

Recently, the convergence of Generative AI (GenAI), synthetic data, cybersecurity and compliance have received extensive academic interest. Researchers are considerating ways GenAI technologies can be used to address long-term threats in threat detection, data privacy as well as regulatory compliance. An overview of the recent scholarly works is presented in major thematic areas in this section.

- **GenAI for Threat Modelling and Attack Simulation**

The prominent topic in the new literature is the application of GenAI in the simulated experience of cyberattacks and threat modelling. The comparison of the use of well-known machine-learning (ML) tools and GenAI-based systems is thoroughly described by Ankalaki et al. (2025), as the use of generative models outperformed traditional ones in forecasting complex cyberattacks. Other authors, such as Gupta et al. (2023), demonstrate the use of ThreatGPT: a specialised GenAI model that operates based on the concept of ideal imitation of real-life adversary behaviour, allowing security teams to train on zero-day attacks and advanced persistent threats.

According to Metta et al. (2024), one of the capabilities of GenAI is the ability to generate artificial logs, network paths, and attack vectors, which will make intrusion detection systems (IDS) more diverse and tougher. Such artificial conditions can be of specific use in the training of supervised models that involve a severe dearth or absence of labelled real threat data in the real world.

- **Artificial Data in Cybersecurity training and practice**

Artificial data is an essential training material for cybersecurity systems. Mohawesh et al. (2025) highlight that synthetic datasets produced with the help of GenAI have the same statistical properties as original datasets but contain no sensitive identifiers, allowing them to adhere to privacy laws. They have validated their hypothesis in their work by showing that the GenAI-generated data will be applicable in the development and penetration testing of intrusion detection models.

Balasubramanian et al. (2025) focus on a variety of concrete case studies with which the threat intelligence systems were trained and tested with the help of synthetic data. They demonstrate that not only are the traditional forms of

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

attacks modelled with the help of GenAI, but their detection rates increase with the unusual or evolving forms of attack.

- **Data Protection and Legal Accommodation**

The privacy of data kept by GenAI is another theme that appears continuously in the literature. According to Bhardwaj (2025), synthetic data enables organizations to make available to vendors, auditors, and third parties the datasets shared without decreasing the level of confidentiality. Likewise, Huang, Huang, and Catteddu (2024) explain how the synthetic datasets created by GenAI can comply with legal regulations, such as GDPR, to share and test the information without infringing the rights and ownership over the information.

Verma et al. (2025) discuss the importance of GenAI in compliance checks by describing how synthetic data helps to test the regulators under different conditions of stress without real-world data that contains user information. Huang, Wang, et al. (2024) continue to explain that the ambiguous nature of the regulations in relation to the legal frameworks needed to uphold the use of synthetic data in official audits, where the present regulation tends to remain not unclear enough on using artificial data as a part of compliance processes.

- **Use Cases and Comparisons between Platforms**

Other studies compare the ability of various GenAI platforms in terms of cybersecurity. Another article by Gafni and Levy (2025) compares several GenAI instruments on different security operations and provides evidence of discrepancies in the data quality, model explainability, and efficiency in work. The evidence they have found is that platforms need to be chosen with care and careful tuning done, depending on the task at hand.

Nott (2025) provides a systematic review of organizational adjustment to GenAI in cybersecurity. The paper indicates that although synthetic data based on GenAI is experimental in most companies, few companies have put in place formal governance or risk management practices. The given observation highlights the policy maturity gap between policy and innovation. Efforts in synthetic data generation constitute difficulties in GenAI, since AI may tend to successfully replicate previously taught data more than promptly creating novel information on its own.<|human|>2.5 Problems with GenAI-Powered deze poblulations Generating suppose news. It means that prediction will display the synthetic data generation processes impacting artificial intelligence, since AI might be inclined to recover earlier trained information more readily than novel information generated by computers altogether.

Inasmuch as it has merits, GenAI has technical and ethical threats. According to Joshi (2025), GenAI models are able to reproduce or enhance biases in the training data, which results in biased or discriminatory outputs. Another threat that is also mentioned by Christodorescu et al. (2024) is the adversarial abuse when malicious users might use GenAI to simulate synthetic attacks, create phishing content, or obfuscate malware activity. In a study by Kanchi et al. (2025), the data curation and validation pipeline is also of critical importance, since synthetic data is required to be validated properly before it can be used in security-critical software. Equally, Singh et al. (2024) emphasize that synthetic data not only enhances privacy but should not be viewed as a completely safe test, particularly when they are created based on smaller and more risky source datasets.

- **Developing Applications: IoT and Cloud security**

GenAI has participated in other rising domains of threats, including cloud computing and IoT; this is also examined in recent literature. Khan et al. (2025) consider the possibility to protect IoT devices with GenAI techniques, in order to generate humour models that predict vulnerabilities by monitoring them in real time. Ruparel et al. (2024) examine how GenAI is relevant to cloud data security, specifically in multi-tenant contexts, and determine that synthetic data can assist in enforcing tenant isolation and policy testing in a shared context.

- **Literature Gap and Trends Synopsis**

Based on this review, it is clear that the benefits provided by GenAI based synthetic data are practical in terms of threat detection, compliance testing and privacy preservation. But three gaps keep on recurring:

  - Transparency in rules on audits and law on the use of synthetic data.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- Ethical governmental structures to influence information creation and implementation.
- Technical assurance procedures to ensure the quality of data and the reduction of risk.

The research sector should focus on these in future studies to support ethical and business-scaling utilization of the GenAI in the realms of cybersecurity and compliance processes.

3. METHODOLOGY

This study adopts a qualitative exploratory methodology supported by comparative literature analysis, case evaluations, and the conceptualization of a GenAI-based synthetic data framework for cybersecurity and compliance. The methodology combines findings from reviewed academic sources, analysis of GenAI tools, and simulation-based interpretations of how synthetic data can be generated, validated, and applied in practical cybersecurity settings.

I. Research Approach

The research uses a multi-method qualitative approach combining:

- Systematic literature review of 17 peer-reviewed articles, conference papers, and technical reports published between 2023 and 2025.
- Comparative case analysis to explore how GenAI-generated synthetic data has been applied in different domains (e.g., threat detection, privacy compliance, cloud security).
- Conceptual modeling of a synthetic data generation framework based on GANs, transformers, and diffusion models, with application layers mapped to cybersecurity and compliance use cases.

This triangulated method ensures both breadth and depth, aligning academic insights with emerging real-world applications.

II. Data Sources and Tools

Data for this research was drawn from academic publications, industry whitepapers, and open-source datasets related to:

- Cybersecurity logs and intrusion datasets (used hypothetically to model synthetic generation tasks)
- Compliance requirements (e.g., NIST, GDPR, CCPA) as structured in regulatory documentation
- GenAI models and tools such as StyleGAN, Diffusion Models, GPT-based log generators, and Synthetic ID generation APIs

Key tool categories include:

- Language Models (e.g., ChatGPT, BERT, ThreatGPT)
- Image/Text GANs (e.g., StyleGAN, BigGAN)
- Diffusion Models (for stable, denoised synthetic signal generation)
- Autoencoders (for structured anonymization tasks)

III. Synthetic Data Generation Framework

The methodology conceptualizes a GenAI-based synthetic data generation pipeline tailored for cybersecurity and compliance, comprising the following layers:

a. Data Collection Layer

- Ingests raw cybersecurity data (logs, telemetry, traffic flows)
- Applies preprocessing and feature extraction
- Removes sensitive or high-risk attributes

b. **Generative Modeling Layer**

- Uses GANs, Transformers, or Diffusion models to generate synthetic versions
- Maintains statistical equivalence to real datasets
- Embeds labels (e.g., attack types, compliance flags) for supervised learning

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

c. Validation and Fidelity Check Layer

- Measures data quality using distributional similarity, entropy, and outlier detection
- Verifies synthetic data does not leak sensitive patterns or PII
- Compares model training outcomes with real vs. synthetic data

d. Application Layer

- Uses synthetic data for IDS/IPS training, anomaly detection, and compliance testing
- Performs threat simulations using adversarial scenarios
- Supports secure third-party data sharing without breaching regulations

IV. Evaluation Criteria

The quality and effectiveness of GenAI-generated synthetic data were evaluated based on the following key criteria, derived from the reviewed literature:

Criteria	Description	Supporting Sources
Statistical Fidelity	Similarity between synthetic and real data distributions	Mohawesh et al. (2025), Gafni & Levy (2025)
Privacy Preservation	Absence of real user identifiers or sensitive patterns	Bhardwaj (2025), Kanchi et al. (2025)
Usefulness in Training	Performance of ML models trained on synthetic data	Ankalaki et al. (2025), Metta et al. (2024)
Compliance Simulation	Ability to simulate regulatory audit conditions	Verma et al. (2025), Huang et al. (2024)
Bias and Fairness	Detection and mitigation of embedded bias in synthetic outputs	Joshi (2025), Christodorescu et al. (2024)
Adaptability	Flexibility of models across domains (cloud, IoT, finance)	Khan et al. (2025), Ruparel et al. (2024)

V. Methodological Limitations

This research is conceptual in nature and does not include experimental implementation of synthetic data generators due to ethical constraints in handling real-world cybersecurity data. However, the study synthesizes findings from credible empirical sources and offers a framework that can be operationalized in future research or applied to industry pilots.

4. APPLICATIONS OF SYNTHETIC DATA IN CYBERSECURITY

The integration of synthetic data generated by Generative AI (GenAI) into cybersecurity infrastructure has introduced new capabilities that are both privacy-respecting and operationally efficient. As real-world datasets are often constrained by ethical, legal, or logistical limitations, synthetic data enables safe experimentation and development of intelligent systems. This section outlines the four primary application areas where GenAI-generated synthetic data is proving transformational.

i. Threat Detection

One of the most impactful uses of synthetic data is in enhancing threat detection systems, particularly machine learning-based Intrusion Detection Systems (IDS) and anomaly detection engines. Traditional threat datasets often lack diversity, especially in capturing rare or evolving cyber threats. GenAI addresses this gap by generating synthetic logs, malware samples, and attack vectors that replicate the statistical patterns of actual cyber incidents (Ankalaki et al., 2025; Gupta et al., 2023).

For example, models like ThreatGPT are trained to generate synthetic adversarial scenarios such as phishing campaigns or ransomware payloads, which can be used to test and stress security tools in a controlled environment (Gupta et al., 2023). These synthetic scenarios help in building more resilient threat detection algorithms capable of

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

identifying subtle and zero-day attacks. As Metta et al. (2024) point out, synthetic attack data improves generalization and detection accuracy, especially when real-world examples are limited or restricted due to legal concerns.

ii. **Compliance Testing**

With regulatory frameworks such as GDPR, CCPA, and ISO 27001 becoming increasingly complex, organizations must continuously test their systems for compliance readiness. However, compliance testing often involves datasets containing sensitive personal or transactional data, which cannot be freely used or manipulated.

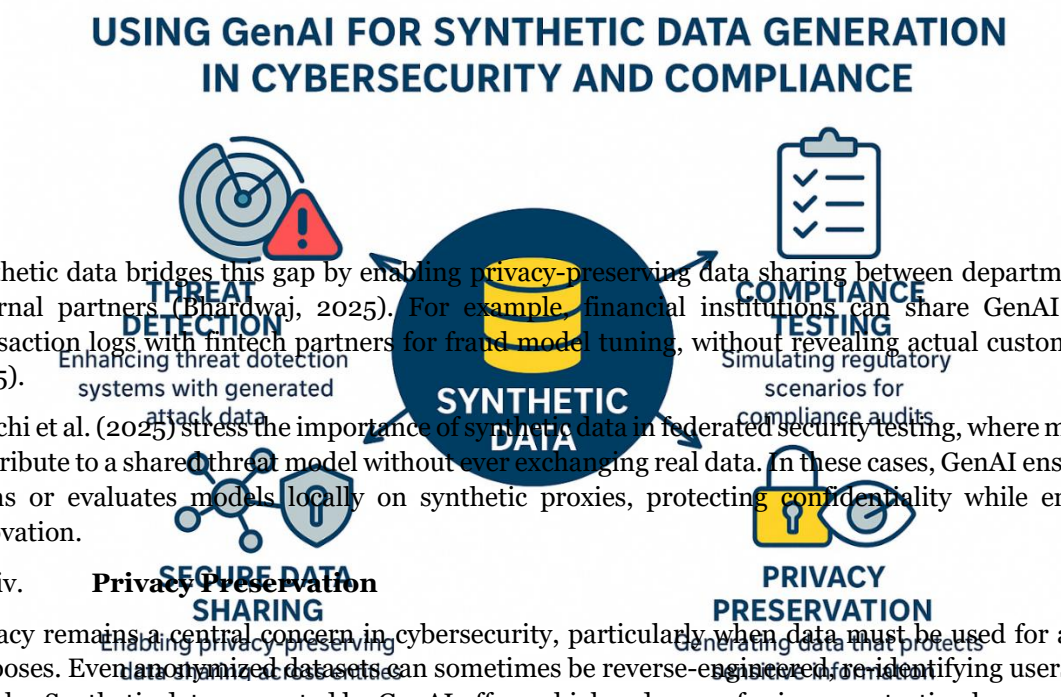
Synthetic data provides a solution by simulating regulatory edge cases and system responses under stress without exposing real user information (Verma et al., 2025). For instance, synthetic user profiles can be generated to mimic a population affected by a specific data retention policy or access control protocol. This allows security engineers and compliance officers to validate the effectiveness of privacy controls and detect violations before actual audits occur (Huang, Huang, & Catteddu, 2024).

Mohawesh et al. (2025) highlight how GenAI helps in compliance scenario modelling, allowing organizations to replicate high-risk events like cross-border data transfers, consent withdrawal, or unauthorized data access without legal exposure.

iii. **Secure Data Sharing**

In collaborative environments such as supply chains, partner networks, or multinational corporations, the ability to share data without compromising security is critical. Real datasets often contain proprietary or personally identifiable information (PII) that cannot be shared even internally.

Figure 1: The below diagram describes the using of GenAI for Synthetic Data Generation in Cybersecurity and Compliance.



Synthetic data bridges this gap by enabling privacy-preserving data sharing between departments, subsidiaries, or external partners (Bhardwaj, 2025). For example, financial institutions can share GenAI-generated synthetic transaction logs with fintech partners for fraud model tuning, without revealing actual customer behaviors (Joshi, 2025).

Kanchi et al. (2025) stress the importance of synthetic data in federated security testing, where multiple organizations contribute to a shared threat model without ever exchanging real data. In these cases, GenAI ensures each participant trains or evaluates models locally on synthetic proxies, protecting confidentiality while enabling collaborative innovation.

iv. **Privacy Preservation**

Privacy remains a central concern in cybersecurity, particularly when data must be used for analytical or training purposes. Even anonymized datasets can sometimes be reverse-engineered re-identifying users through correlation attacks. Synthetic data generated by GenAI offers a higher degree of privacy protection because it does not originate from any actual user record but is instead built to mimic general statistical distributions (Huang, Wang, et al., 2024).

Singh et al. (2024) argue that synthetic data represents a breakthrough in building privacy-first ML pipelines, allowing for the development of systems that are not only effective but also ethical. Additionally, Christodorescu et al. (2024) caution that synthetic data must be validated to ensure that it does not inadvertently encode rare real-world features that could compromise privacy. Therefore, combining GenAI with robust validation and bias mitigation tools becomes essential for achieving true privacy preservation.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5. CHALLENGES AND LIMITATIONS

Although there are impressive benefits of the application of the Generative AI (GenAI) to generate synthetic data in compliance and cybersecurity, a number of technical, ethical, operational, and regulatory challenges have to be noted. Such limits can be barriers to nearly a universal implementation and somewhat dangerous unless strong governance and validation is executed.

i. Data Fidelity and Quality Assurance

The main issue with synthetic data generation is the data fidelity- the extent to which synthetic data sets faithfully represent structure, distribution and behaviour of real world data. The use of the wrong or low quality synthetic data can present compromises in machine learning models or security policies (Mohawesh et al., 2025; Gafni and Levy, 2025).

In addition, some context-dependent exceptions of real-life data (i.e., time-stamped attack sequences or multi-hop network interactions of the network) cannot be recreated by GenAI models. By counting on less-fidel synthetic data, intrusion detection systems (IDS) is prone to drop, as Metta et al. (2024) stress that applying high-risk environments or time context might present a significant issue in raising concerns.

ii. Risk of Bias Amplification

In the case of genAI models, they will be trained using real-world data, thus potentially having unintended social, demographic, or behavioural bias. Unless these biases are addressed, they may be replicated or even enhanced in the resulting synthetic data (Joshi, 2025; Christodorescu et al., 2024).

This creates unfair conditions in the context of cybersecurity, such as any systems that have been trained using biased synthetic data may end up labelling the activities of particular groups of users as a threat at rates exceeding those of other groups. This may destroy the credibility of AI-assisted defensive systems and may alert some questions about algorithmic bias, particularly in overseers of audits in which fairness and responsibility are legally assured (Nott, 2025).

iii. Privacy Leakage and Inference Attacks

In spite of the fact that synthetic data is created to keep the privacy of users safe, there is an increasing threat of privacy leakage, particularly in the case GenAI models are trained on small or poorly curated datasets. Higher attackers can exploit inference attacks to create identifiable data based on synthetic data produced in them, especially in case of such overfitting when training the model (Singh et al., 2024; Kanchi et al., 2025).

According to Bhardwaj (2025), there should be strict differential privacy evaluation in synthetic datasets constructed using sensitive health or financial information, and watermarking procedures to ensure that their identity cannot be reversed or reconstructed.

iv. Adversarial Misuse of GenAI

The generative functionalities, which make it possible to do a good job (say, by simulating a threat), can be also abused by the bad. GenAI can be used by attackers to create simulated samples of malware, create phishing email, or create sophisticated fake network traffic that would go undetected (Gupta et al., 2023; Christodorescu et al., 2024).

Such a twofunction quality of GenAI leads to a security paradox situation in which the same tool that is designed to counter threats emerges as one that delivers novel approaches to attacks. The regulatory organizations are required to establish limits to GenAI creation and utilization, such that there are access controls to GenAI applications, and red-teaming of GenAI applications, as pointed out by Verma et al. (2025).

v. Have no regulatory standards and legal frameworks

Whether synthetic data should be legal or not is still unclear, particularly when it is involved in compliance audits or during a forensic investigation. Most regulatory regimes (e.g., GDPR, HIPAA) lack clarity on what synthesized data should be, leaving it unclear whether this data can be used in legal and compliance procedures (Huang, Huang & Catteddu, 2024; Nott, 2025).

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Moreover, without a clear-cut definition of what is sufficiently synthetic, entities might go excessively synthetic, or deny their use, as legally advised to the apt. This prevents its use in some highly regulated markets like finance, insurance and healthcare.

vi. Operation Barriers: The Barriers: Skills, Infrastructure and Cost

The fields of AI engineering, cybersecurity and privacy law expertise (hence, absent in most companies) are required to deploy GenAI models to generate artificial data (Khan et al., 2025; Ruparel et al., 2024). In addition, the operation or sustenance of GenAI systems requires major computational resources such as GPUs, secure environments, and monitoring pipelines.

This is also hindered by cost, especially to the small to mid-sized enterprises (SMEs) since they do not have a constellation of resources to deploy scalable synthetic data programs. A similar notion, as stated by Gafni and Levy (2025), is the necessity of the open-source and cloud solutions being chosen to democratize the use of GenAI-capable tools, protecting cybersecurity.

Summary of Limitations

Category	Challenge	Risk Outcome
Data Quality	Low fidelity or unrealistic synthetic behavior	Misleading model performance, weak defense models
Bias & Fairness	Replication of societal or systemic biases	Algorithmic discrimination, audit failures
Privacy Risks	Inference attacks on poorly generated data	Re-identification of individuals, privacy violations
Adversarial Use	Misuse of GenAI for crafting synthetic threats	Phishing, malware obfuscation, threat model corruption
Legal Uncertainty	No clear regulatory standard on synthetic data	Non-compliance, legal disputes, restricted usage
Operational Barriers	Lack of skills, infrastructure, and high cost	Limited deployment, exclusion of SMEs

7. DISCUSSION

The application of Generative AI (GenAI) in synthetic data generation presents a paradigm shift in how cybersecurity and compliance frameworks are designed, tested, and enforced. This discussion explores the implications of the findings from previous sections, with a focus on the strategic benefits, operational realities, ethical tensions, and emerging research directions. The analysis reveals both the transformative potential and the nuanced limitations that must be navigated for responsible deployment.

i. Strategic Alignment with Cybersecurity Needs

GenAI-generated synthetic data addresses several longstanding pain points in cybersecurity: data scarcity, diversity of threats, and real-time model training. Threat detection systems often suffer from the lack of representative attack scenarios, especially zero-day threats or rare exploit patterns. By enabling scalable and customizable simulation of such conditions, GenAI significantly enhances system preparedness (Ankalaki et al., 2025; Gupta et al., 2023).

Moreover, synthetic data supports adversarial training by allowing organizations to test how their systems respond to emerging forms of attacks without risking infrastructure integrity. This capability is invaluable in red-teaming exercises, where traditional static datasets fail to capture the evolving threat landscape (Metta et al., 2024).

ii. Compliance Enablement through Controlled Simulation

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

From a compliance standpoint, GenAI-generated data provides a non-invasive way to validate governance frameworks. As regulatory standards increasingly demand continuous assessment of privacy, access control, and data residency policies, synthetic datasets allow these scenarios to be replicated without breaching confidentiality agreements or violating data protection laws (Mohawesh et al., 2025; Verma et al., 2025).

For instance, an organization could simulate a GDPR right-to-be-forgotten request using synthetic identities and audit the traceability of deletion protocols. This method not only fulfills regulatory testing requirements but also preserves the sanctity of production systems.

iii. The Ethics-Performance Tradeoff

One of the most critical tensions exposed in the literature is the ethics-performance tradeoff. While synthetic data enhances model performance by enriching training diversity, it may also perpetuate or amplify biases embedded in the training data. This is particularly concerning in high-stakes domains like law enforcement, finance, and healthcare (Joshi, 2025; Christodorescu et al., 2024).

The challenge is compounded when GenAI is used to simulate behaviors of marginalized or underrepresented groups. Without bias mitigation pipelines, these models may inadvertently encode systemic discrimination into future decision systems. The literature thus underscores the importance of integrating fairness-aware algorithms and auditing synthetic data outputs during the design phase (Singh et al., 2024; Kanchi et al., 2025).

iv. Managing the Dual-Use Nature of GenAI

GenAI's dual-use nature remains one of its most serious limitations. While it empowers defenders with advanced simulation tools, it also provides adversaries with the ability to craft hyper-realistic phishing emails, synthetic identities, and obfuscated malware. This duality creates an "arms race" dynamic in cybersecurity: the better the tools for defense, the more sophisticated the offensive use cases become (Gupta et al., 2023; Christodorescu et al., 2024).

This situation calls for governance mechanisms—not only technical safeguards like watermarking and usage tracking, but also policy-level interventions. Controlled access to powerful GenAI tools and mandatory red-teaming exercises must become part of the cybersecurity development lifecycle.

v. Bridging the Implementation Gap

Although the literature emphasizes the strategic potential of GenAI, a significant implementation gap remains. Most studies reviewed were conceptual or laboratory-based, with limited application in production environments (Nott, 2025; Gafni & Levy, 2025). Organizations still face barriers such as:

- Lack of skilled personnel
- Infrastructure costs
- Regulatory ambiguity
- Interoperability with legacy systems

To bridge this gap, there is a need for reference architectures, open-source synthetic data platforms, and sector-specific compliance templates that allow easier integration of GenAI technologies, particularly for small and mid-sized enterprises (Khan et al., 2025; Ruparel et al., 2024).

vi. Future Research and Development Directions

The discussion reveals multiple open questions that should be prioritized by the research community:

- How do we define "safe synthetic data"?
- What are the measurable privacy and utility trade-offs?
- Can synthetic data be certified for audit and legal evidence?
- What governance models are best suited for dual-use technologies like GenAI?

These inquiries suggest the need for interdisciplinary collaboration, involving AI researchers, cybersecurity experts, ethicists, regulators, and legal professionals. Only through such collective engagement can GenAI be harnessed safely and equitably.

8. CONCLUSION

While the integration of Generative Artificial Intelligence (GenAI) with cybersecurity and regulatory adherence is revolutionary therein, the underlying conceptual semantics of digital system defence fundamentally changes how digital defense systems are designed, executed, and implemented. The paper discussed GenAI synthetic data as a way of addressing the existing solutions to long-term problems of data scarcity, the risk of privacy, biased learning, and increased complexity of compliance. Based on the integration of current literature and conceptual models, one can see that synthetic data created with the use of GenAI has enormous potential to become useful in identifying possible threats and in favor of privacy-by-design approaches, or to facilitate safe and ethical exploration of cyberspaces. Among important applications, novel synthetic attack patterns to be used in machine learning training, simulation of regulatory compliance conditions without access to real data, and sharing of sensitive data across organizational and geographical levels will be possible. These applications can support a more dynamic and privacy-conscious security stance, particularly in sectoral applications in which unstructured data are inaccessible (or sensitive or even prohibited).

Nonetheless, such improvement has not gone without difficulties. Fidelity of artificial data, the threat of the exaggeration of bias, the possibility of dual-use and adversarial abuse, and the lack of legal guidance be actual threats to safe and responsible application. These issues necessitate multi-pronged-response, integrated, in it is beneficial to have both strong validation pipelines and sound ethical oversight, and substantial regulatory innovation and multi-disciplinary collaboration. Standardized representations, open-source platforms, and certification are acutely needed as well, to assist organizations in operationalizing GenAI safely at scale. However, as we move on to predict the future of synthetic data into cybersecurity, it is not just the technology that will determine its future/s but also our mastery of it. Synthetic data needs to begin to be accepted and regulated as a recognizable and verifiable entity as part of policymaking. Similarly, the principles of ethical AI design should also be incorporated in the structure of GenAI models, not only in the source of data, but in the use of an output that is valid.

Finally, GenAI-produced synthetic data represents the attractive instrument towards closing the culture divide between the two fundamental vital components of security performance and their data protection. When used and implemented thoughtfully, it will be able to transform the nature of how organizations protect themselves in a threat environment that is both agile as it evolves and as harmful as it could be.

REFERENCES

[1] Mohawesh, R., Ottom, M. A., & Salameh, H. B. (2025). A data-driven risk assessment of cybersecurity challenges posed by generative AI. *Decision Analytics Journal*, *15*, 100580. https://doi.org/10.1016/j.dajour.2025.100580

[2] Verma, A., Sankhyayan, S., Jawanda, K., Tandon, S. (2025). Generative Artificial Intelligence and Cybersecurity Risks: Issues and Challenges. In: Fong, S., Dey, N., Joshi, A. (eds) ICT Analysis and Applications. ICT4SD 2024. Lecture Notes in Networks and Systems, vol 1161. Springer, Singapore. https://doi.org/10.1007/978-981-97-8602-2_29

[3] Satyadhar Joshi. Gen AI in Financial Cybersecurity: A Comprehensive Review of Architectures, Algorithms, and Regulatory Challenges. *International Journal of Innovations in Science Engineering And Management*, 2025, 4 (3), pp.73-88. ⟨10.69968/ijisem.2025v4i373-88⟩. ⟨hal-05212241⟩

[4] S. Ankalaki, A. R. Atmakuri, M. Pallavi, G. S. Hukkeri, T. Jan and G. R. Naik, "Cyber Attack Prediction: From Traditional Machine Learning to Generative Artificial Intelligence," in *IEEE Access*, vol. 13, pp. 44662-44706, 2025, doi: 10.1109/ACCESS.2025.3547433

[5] Balasubramanian, P., Liyana, S., Sankaran, H. *et al.* Generative AI for cyber threat intelligence: applications, challenges, and analysis of real-world case studies. *Artif Intell Rev* **58**, 336 (2025). https://doi.org/10.1007/s10462-025-11338-z

[6] Metta, S., Chang, I., Parker, J., Roman, M. P., & Ehuan, A. F. (2024). Generative AI in cybersecurity. *arXiv preprint arXiv:2405.01674*. https://doi.org/10.48550/arXiv.2405.01674

[7] Huang, K., Huang, J., Catteddu, D. (2024). GenAI Data Security. In: Huang, K., Wang, Y., Goertzel, B., Li, Y., Wright, S., Ponnapalli, J. (eds) Generative AI Security. Future of Business and Finance. Springer, Cham. https://doi.org/10.1007/978-3-031-54252-7_5

[8] Singh, A., Singh, D., Singh, R. (2024). Generative AI for Cyberdefense. In: Raza, K., Ahmad, N., Singh, D. (eds) Generative AI: Current Trends and Applications. Studies in Computational Intelligence, vol 1177. Springer, Singapore. https://doi.org/10.1007/978-981-97-8460-8_7

[9] Huang, K., Wang, Y., Goertzel, B., Li, Y., Wright, S., & Ponnapalli, J. (2024). Generative AI Security. *Future of Business and Finance*. https://doi.org/10.1007/978-3-031-54252-7

[10] Yogesh Kumar Bhardwaj. (2025). Securing Generative AI: Navigating Data Security Challenges in the AI Era. *Journal of Computer Science and Technology Studies*, *7*(4), 147-155. https://doi.org/10.32996/jcsts.2025.7.4.17

[11] Christodorescu, M., Craven, R., Feizi, S., Gong, N., Hoffmann, M., Jha, S., ... & Turek, M. (2024). Securing the future of GenAI: Policy and technology. *arXiv preprint arXiv:2407.12999*. https://doi.org/10.48550/arXiv.2407.12999

[12] Kanchi, S., Mangaokar, N., Cheruvu, A., Abdullah, S. M., Nilizadeh, S., Prakash, A., & Viswanath, B. (2025). Taming Data Challenges in ML-based Security Tasks: Lessons from Integrating Generative AI. *arXiv preprint arXiv:2507.06092*. https://doi.org/10.48550/arXiv.2507.06092

[13] Gafni, R., & Levy, Y. (2025). Comparing GenAI platforms on cybersecurity management task performances. *Information & Computer Security*. https://doi.org/10.1108/ICS-04-2025-0164

[14] Khan, A., Jhanjhi, N., Abdulhabeb, G. A. A., Ray, S. K., Ghazanfar, M. A., & Humayun, M. (2025). Securing IoT Devices Using Generative AI Techniques. In *Reshaping CyberSecurity With Generative AI Techniques* (pp. 219-264). IGI Global. DOI: 10.4018/979-8-3693-5415-5.ch007

[15] H. Ruparel, H. Daftary, V. Singhai and P. Kumar, "The Impact of Generative AI on Cloud Data Security: A Systematic Study of Opportunities and Challenges," *2024 IEEE/ACM 17th International Conference on Utility and Cloud Computing (UCC)*, Sharjah, United Arab Emirates, 2024, pp. 185-188, doi: 10.1109/UCC63386.2024.00033

[16] M. Gupta, C. Akiri, K. Aryal, E. Parker and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy," in *IEEE Access*, vol. 11, pp. 80218-80245, 2023, doi: 10.1109/ACCESS.2023.3300381

[17] Nott, C. (2025). Organizational Adaptation to Generative AI in Cybersecurity: A Systematic Review. *arXiv preprint arXiv:2506.12060*. https://doi.org/10.48550/arXiv.2506.12060

[18] Huang, K., Yeoh, J., Wright, S., Wang, H. (2024). Build Your Security Program for GenAI. In: Huang, K., Wang, Y., Goertzel, B., Li, Y., Wright, S., Ponnapalli, J. (eds) Generative AI Security. Future of Business and Finance. Springer, Cham. https://doi.org/10.1007/978-3-031-54252-7_4

[19] Omar, K. O., Zraqou, J., & Gómez, J. M. (2025). From Synthetic Text to Real Threats: Unraveling the Security Risks of Generative AI. In *Examining Cybersecurity Risks Produced by Generative AI* (pp. 1-20). IGI Global Scientific Publishing. DOI: 10.4018/979-8-3373-0832-6.ch001

[20] Saha, B., Rani, N., & Shukla, S. K. (2025). Generative AI in Financial Institution: A Global Survey of Opportunities, Threats, and Regulation. *arXiv preprint arXiv:2504.21574*. https://doi.org/10.48550/arXiv.2504.21574

[21] Kim, Hyungjin Lukas and Han, Jinyoung, Impact of Generative Artificial Intelligence on Theoretical Perspectives Related to Information Security Policy Compliance (December 04, 2024). http://dx.doi.org/10.2139/ssrn.5127769

[22] Orpak K (2025) Generative AI and cybersecurity: Exploring opportunities and threats at their intersection. Maandblad voor Accountancy en Bedrijfseconomie 99(4): 221-230. https://doi.org/10.5117/mab.99.149299

[23] Zaboli, A., & Hong, J. (2025). Generative AI for Cybersecurity of Energy Management Systems: Methods, Challenges, and Future Directions. *arXiv preprint arXiv:2508.10044*. https://doi.org/10.48550/arXiv.2508.10044