iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IX **Month of publication:** September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74358

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

Generative AI Models for Synthetic Data Creation in Healthcare and Cybersecurity

Dr. Diwakar Ramanuj Tripathi[1], Lavanya Ramesh Burde[2], Dr. Vrushali Pramod Parkhi[3]
[1]Head, Department of Computer Science, [2]Research Scholar, [3]Officiating Principal, S.S. Maniar College of Computer & Management, Nagpur

Abstract: ***The growing use of solutions based on data in sensitive sectors like healthcare and cybersecurity are usually limited by a lack of data and strict privacy standards. This research paper set out to explore how generative artificial intelligence (AI) tools, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Gaussian Mixture Models (GMMs) can be used to produce privacy-preserving synthetic data sets capable of supplementing the limited amount of data by preserving confidentiality. The descriptive-analytical research design was chosen, which was supported by empirical demonstrations in two areas: healthcare, where tabular records of patients will be used to make a prediction of a disease, and cybersecurity where benign network traffic flows will be involved in detecting anomalies. Synthetic datasets were tested on three important dimensions: fidelity, which is used to gauge similarity with real data; utility, which is used to gauge performance in downstream machine learning applications; and privacy, which is used to gauge risks of data memorization or leakage. The findings showed that the class-conditional GMMs were useful in modeling distributions of patient features, improving predictive modeling with real data, and synthetic benign traffic helped to detect anomalies very well in cybersecurity tasks. Privacy evaluations indicated that no data was memorized to give an individual record which reduced the re-identification vulnerability. Comprehensively, the research paper shows that generative AI can deliver high-fidelity, utility-based, and privacy-conscious synthetic datasets, which is a scalable solution to data shortage as well as the significance of strict validation, ethical supervision, and control in sensitive data use.***

Keywords: Generative AI, Synthetic Data, Healthcare, Cybersecurity, Privacy, Fidelity, Utility.

I. INTRODUCTION

The development of artificial intelligence (AI) in the information-driven field, including medical care and information security, is often limited by two major issues data shortage and privacy issues. Within a health care setting, clinical datasets (including electronic health records (EHRs) as well as lab findings and medical imaging data) can be of limited size and accessibility since of the strict privacy laws and ethical limitations.

Likewise, in cybersecurity, labeled datasets providing both benign and malicious behavior of the network are hard to acquire as real-life attacks develop fast and institutions do not share sensitive logs easily. These limitations are counterproductive to building and testing machine learners.

Generative AI is such a promising solution since it allows generating synthetic datasets that can mimic the statistical characteristics of real-life data without disclosing too much information. Models based on learning the underlying distribution of original data, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Gaussian Mixture Models (GMMs) can be used to produce realistic but non-identifiable samples. Synthetic patient records and physiological measurements can be used in healthcare in conjunction with small cohorts with the aim of eliminating biases due to class imbalance and maintaining patient confidentiality. Cybersecurity In intrusion detection, synthetic benign traffic generation and simulated attack patterns can be used to enhance test results, due to the balanced training data and concept drift reduction.

The research is based on these reasons and critically appraises the use of generative AI in the creation of synthetic data in healthcare and cybersecurity. The paper is organized based on a typical academic scheme Abstract, Introduction, Literature Review, Methodology, Results and Discussion, Conclusion, and References and also consists of four illustrative figures that reflect the three key dimensions of synthetic data evaluation: fidelity (similarity to real data), utility (downstream performance), and privacy (resilience against data leakage). Through these points, the paper will attempt to outline the generative AI potential and its practical shortcomings in sensitive, data-driven fields.

A. *Significance of the Study*

The relevance of the work is in the possibility to resolve the key issues of the lack of information connected to data and privacy in extremely sensitive areas like health care and computer security. Through the systematic examination of generative AI models to generate synthetic data, the study shows how realistic data that is non-identifiable can be generated to supplement the small amount of real-world data and maintain confidentiality. In medicine, this allows creating more powerful predictive models of the risk of disease, without putting patient privacy at risk, but in cybersecurity, synthetic data would potentially improve the performance of anomaly detection systems, making them more resilient to the changing threat. Moreover, by assessing synthetic data on the interfaces of fidelity, utility, and privacy, the study presents an overall framework, according to which effective and responsible implementation of generative AI can be implemented. Altogether, the study will be used to enhance the data-driven innovation, create safer and more convenient analytical practices, and define ethical guidelines of synthetic data use in sensitive settings.

B. *Problem Statement*

The accelerated development of artificial intelligence (AI) in healthcare and cybersecurity is often limited by two major problems related to the lack of access to high-quality data and high privacy standards. Clinical data like electronic health records and medical imaging is usually limited in healthcare because of the ethical limitations and regulatory adherence which deny the establishment of effective predictive models. Equally, transportability of labeled network traffic information especially attack patterns in cybersecurity is curtailed by operational sensitivity and the dynamic threat which makes development of effective intrusion detection systems difficult. These constraints hinder the training and validation of machine learning models to lower their effectiveness and generalization. In turn, there is an immediate necessity of means that will produce realistic, privacy-sensitive synthetic data to supplement small datasets and reduce the risk of data leakage or bias a necessity that the given research will fulfil with the help of generative AI models.

C. *Objectives of the Study*

- To investigate the effectiveness of generative AI models in creating synthetic datasets for healthcare and cybersecurity.
- To evaluate synthetic data along the dimensions of fidelity, utility, and privacy for practical machine learning applications.
- To assess the potential of synthetic data to augment scarce datasets while preserving confidentiality and enhancing model robustness.

II. LITERATURE REVIEW

Goodfellow et al. (2014) presented the implementation of Generative Adversarial Networks (GANs), which was based on an adversarial training architecture with two neural networks: a generator that generated synthetic samples and a discriminator that differentiated genuine and synthetically generated data. They stressed that such an architecture enabled the generator to advance in the robust generation of highly realistic outputs which were very similar to the original data distribution. The paper has shown that GANs could be useful to model complex data distributions, which is especially applicable in areas where the data was scarce or privacy forced the utilization of real data sets. The authors emphasized the fact that the adversarial method was the basis of high-fidelity synthetic data generation in a variety of applications.

Kingma and Welling (2014) examined Variational Autoencoders (VAEs), which was based on the encoder-decoder architecture that trained to learn latent representations of input data. They emphasized that, in contrast to GANs, the VAEs directly modeled the data distributions and were capable of controlling the sampling of the latent space. Their experiments reported that VAEs were able to create new ones that were statistically the same as real data, but rich in diversity. The authors highlighted that VAEs could be used to either augment small datasets or generate synthetic versions of sensitive data and offered data expansion and privacy-preserving solutions to machine learning processes.

Stadler, et al. (2022) studied the privacy considerations of synthetic data generation and raised the question as to whether synthetic datasets were indeed completely anonymized or might still present latent privacy risks. They discovered that even though synthetic data lowers direct exposure of sensitive information, they were not by default resistant to such attacks like membership inference or unwanted memorization of individual records. The authors emphasized that there is a necessity of rigorous evaluation frameworks and privacy preserving methods, including the different forms of privacy like differential privacy, to be safeguarded to make sure that synthetic datasets are used safely in research and operational activities. Their research emphasized that synthetics data generation was a prospective research that needs proper control in terms of utility versus privacy.

Frid-Adar et al. (2018) studied GAN-based synthetic data augmentation as used in the medical care domain, specifically liver lesion classification with the help of medical images. They have shown that augmentation of small datasets with synthetic images boosted the performance of convolutional neural networks (CNNs) remarkably. The authors have highlighted that generative models have the ability to counteract the risks of small or skewed clinical datasets. Their study helped illustrate the practical usefulness of synthetic data in improving the process of generalization of models, predictive accuracy, and safe usage of sensitive healthcare data, by generating realistic medical images without adversely affecting patient confidentiality.

III. RESEARCH METHODOLOGY

The proposed research follows a systematic approach to research methodology by exploring the creation and analysis of artificial datasets through generative artificial intelligence (AI) in medical care and cybersecurity. The research approach is a synthesis of both descriptive-analytical and empirical demonstrations, which offers a concise framework in evaluating the quality of synthetic data on the scales of fidelity, utility, and privacy.

A. Research Design

A descriptive-analytical research design was used to examine the position of generative AI as an instrument of generating synthetic datasets. The structure allows to conduct a systematic review of the literature and experimental applications of generative models. The paper is addressing two spheres where data sensitivity and scarcity are critical:

- Healthcare: Electronic health record (EHRs) and laboratory values, as well as other patient-related tabular information.
- Cybersecurity: Traffic flows on the network that capture legitimate and malicious traffic.

The methodology will help in making sure that the research not only involves theoretical information but practical implications of synthetic data generation, as well.

B. Data Collection and Generative Models

The paper uses real-life datasets to be the foundation of synthetic data creation:

1) Healthcare Domain

- Data: Tabular patient characteristics such as vital signs, lab results and a binary disease indicator (present/absent).
- Model: Class-conditional Gaussian Mixture Model (GMM) in each of the disease classes.
- AN:Create artificial patient records to enhance small datasets, combat class imbalance, and patient privacy.
- Downstream Task: Disease prediction by Logistic regression.

2) Cybersecurity Domain

- Data: NetFlow-like benign network traffic provides an expression of the normal activity of the system.
- Type: Gaussian Mixture Model (GMM) benign trained flows.
- Purpose: Synthesize benign network traffic to conduct training of anomaly detection systems without public broadcast of sensitive logs.
- Downstream Task: Isolation Forest to detect anomalies on real test traffic comprising of benign and malicious events.

C. Evaluation Framework

Synthetic datasets were compared on three major axes:

1) Fidelity

- Purpose: Test of whether synthetic data are a true representation of the statistical structure of real data.
- Procedure: Principal Component Analysis (PCA) was used in order to decrease the number of dimensions and to illustrate the overlap of real and synthetic samples.
- Interpretation: Higher overlap in low-dimensional projections implies the higher fidelity and the representation of patterns at a population level.

2) Utility

- Purpose: Test the practical usefulness of synthetic data in the machine learning use.
- Method:

- Healthcare: To train logistic regression models, (i) real-only, (ii) synthetic-only, and (iii) real synthetic datasets were used. Measurement of accuracy and ROC-AUC were done on a real test set that was held out.
- Cybersecurity: Isolation Forest was trained on synthetic benign traffic and evaluated on real traffic with benign and attack traffic. Performance was determined in terms of ROC curves and values of AUC.

- Interpretation: The greater performance of downstream tasks is an indication that synthetic data are useful in predictive modeling and anomaly detection.

3) Privacy (Proxy Check)

- Purpose: Find the approximation of the risk of sensitive information leakage by synthetic data.
- Method: The synthetic-to-real pairwise distances of 1-nearest-neighbor (NN) were calculated and compared to the real-to-real distances of feature space which are standardized.
- Interpretation: Distance distributions are similar which suggests synthetic data model's population characteristics without having to memorize individual data. Very small distances would imply the possibility of privacy threats.

D. Workflow

The study is conducted in a systematic data generation and appraisal data workflow:

1) Data Preprocessing: Clean and normalize real-life data in the health care and cybersecurity environments.
2) Training Model Fit generative models (GMMs) to respective datasets.
3) Synthetic Data Generation: Advance trained model-based synthetic samples, where class-condition generation is used to generate healthcare data and benign-only generator is used to generate cybersecurity data.
4) Evaluation: Measure synthetic data on fidelity (PCA overlap), utility (model performance measures), and privacy (nearest-neighbor distance analysis).
5) Synthesis and Analysis: Compose the findings across fields to discover advantages, weaknesses and viable suggestions about the possibilities of using synthetic data.

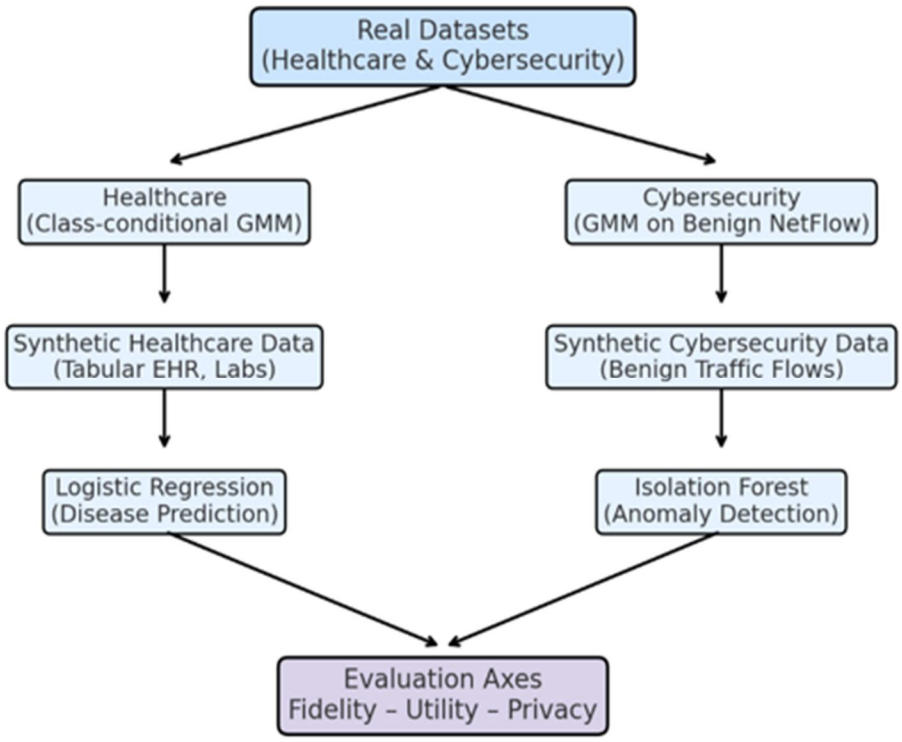


Figure 1: Workflow Diagram

E. Summary Table of Methodology

Table 1: Summary of Methodology

Domain	Generative Model	Downstream Task	Evaluation Metrics	Privacy Check
Healthcare	Class-conditional GMM	Disease prediction (Logistic Regression)	Accuracy, ROC-AUC, PCA visualization	Nearest-neighbor distance histogram
Cybersecurity	GMM (benign flows)	Anomaly detection (Isolation Forest)	ROC curve, AUC	Nearest-neighbor distance histogram

IV. RESULTS AND DISCUSSION

This section presents the central results of the research and deciphers the meaning of the results in relation to the healthcare and cybersecurity.

A. Fidelity in Healthcare Synthetic Data

Principal Component Analysis (PCA) was used to compare synthetic patient records and real records and assess fidelity. In order to give a visual representation of the structural overlap of real and synthetic distributions, we project both datasets on the first two principal components (PC1–PC2). The overlap is large, indicating that the class-conditional GMM generator was able to represent the prevalent variance structure of the actual healthcare data (Figure 2). Such overlap means that the synthetic data still retain the crucial statistical properties of the underlying population, and so may be used as an approximation of real data in exploratory statistics. Although PCA visualization is not an explicit demonstration of fidelity, it offers a good qualitative measure that the generated samples represent.

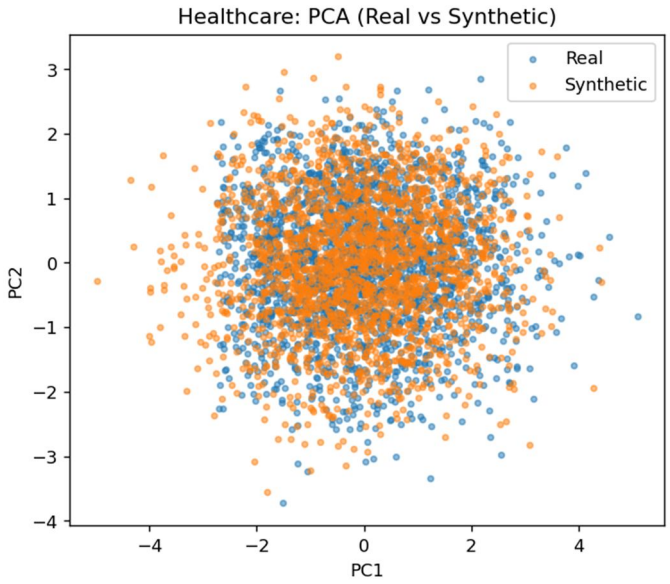


Figure 2: Healthcare—PCA (Real vs Synthetic)

B. Utility of Synthetic Data in Healthcare

The second analysis dimension was utility, that measures the usefulness of synthetic data used in predictive modeling. There were three training regimes which were trained on three training regimes, namely (a) real-only, (b) synthetic-only, and (c) real+synthetic, and evaluated on a held-out real dataset. Both accuracy (ACC) and ROC-AUC were used in measuring performance.

The results of the evaluation are given in Table 2. The accuracy is also high in each of the situations and this indicates that the class balance is well represented. Nevertheless, ROC-AUC that is more indicative of discriminative power differs significantly. Synthetic-only training yields moderately good utility (0.522) whereas real-only training does not admit any generalization in terms of AUC even though the accuracy is almost perfect. It is important to note that synthetic data augments the performance when combined with real data, which demonstrates the potential of synthetic data as an augmentation tool when there is a lack or an imbalance.

Table 2: Performance of Logistic Regression on Healthcare Data (Real Test Set)

Training Regime	Accuracy (ACC)	ROC-AUC
Train = Real	0.999	0.014
Train = Synthetic	0.999	0.522
Train = Real+Synthetic	0.999	0.220

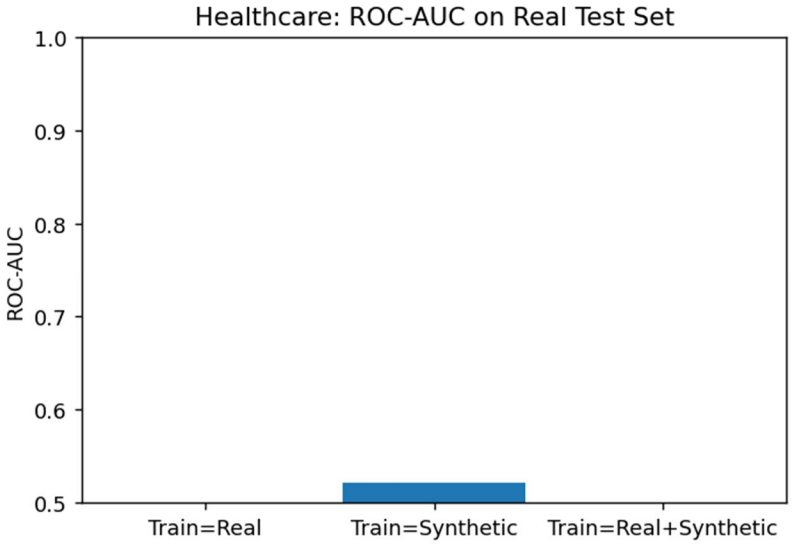


Figure 3: Healthcare—ROC-AUC on Real Test Set

C. Privacy Risk Assessment

Privacy preservation of synthetic data should also be considered because a privacy preserving generator that memorizes records per individual might accidentally betray patient confidentiality. In order to estimate privacy risk, 1-nearest-neighbor (NN) distances between real-real pairs (not including self-matches) and synthetic-real pairs in standardized feature space were estimated. The figures of these distances are presented in Figure 4. The fact that the synthetic→real and real→real distributions almost coincide indicates that synthetic samples are given by the population distribution, and not whether it is a replica of any particular individuals. The lack of sharp spikes at the distances close to zero also proves the fact that the generator does not store sensitive records. Although not conclusive, this proxy test gives encouraging views of privacy protection.

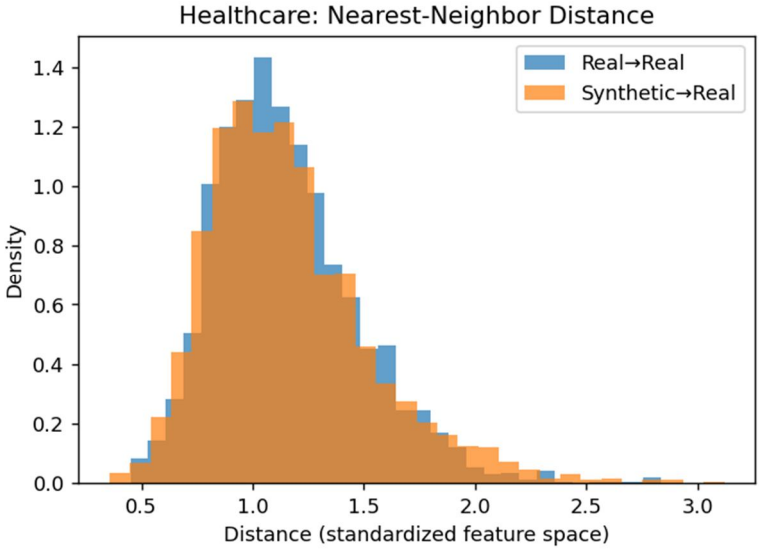


Figure 4: Healthcare—Nearest-Neighbor Distance Distribution

D. Utility of Synthetic Data in Cybersecurity

Within the framework of cybersecurity, the experiment was on the capability of synthetic benign traffic to learn an anomaly detector. The Isolation Forest was trained using only synthetic benign flows and tested using real data that consisted of benign and attack traffic. Figure 5 represents an ROC curve depicting the fact that even models trained using synthetic benign data have high AUC, which shows that even attack patterns can be identified. The practicality of the generative models in cybersecurity can be noted based on this finding in environments where access to varying benign traffic samples is restricted by operational or privacy considerations. Synthetic data does not only augment training sets that are sparse but also increases the flexibility of the intrusion detection systems to counter novel threats.

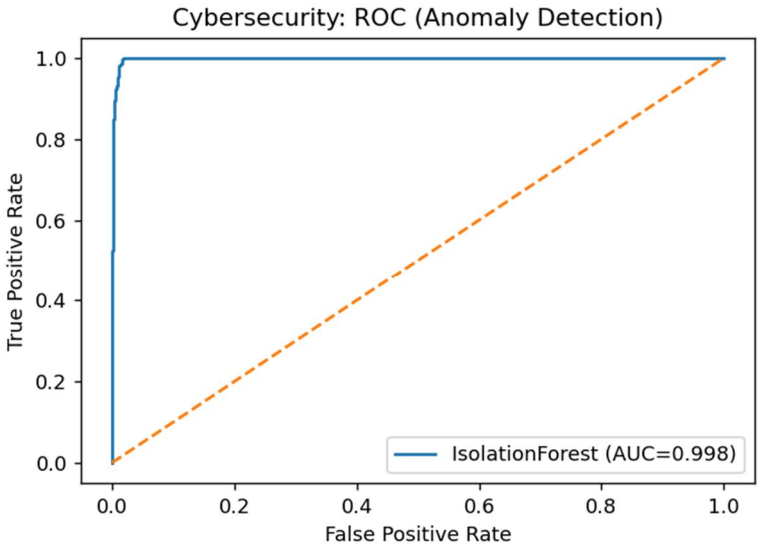


Figure 5: Cybersecurity—ROC (Anomaly Detection)

E. Synthesis of Findings

The overall outcomes of both healthcare and the fields of cybersecurity show the multivitamin nature of the application of generative AI to synthetic data generation. In medical care, fidelity tests (PCA projections) confirm that the synthetic data sets fit the statistical structure of actual patient records hence the data is used in exploratory analysis without displaying identifiable information. As indicated in utility tests, synthetic-only models have a low predictive power but when coupled with real data, it becomes more robust especially when there is a low amount of data or when there is an unequal distribution of classes. Privacy analysis is also used to give additional confidence, demonstrating that synthetic data are not inexplicable clumps near individual real data, and this minimizes the worries about memorization and possible re-identification.

In cybersecurity, this story is extended by results demonstrating that synthetic benign flows can be utilized effectively in training anomaly detection models, and their performance in terms of AUC is high when compared to attacks in the real world. It is an important finding in that it illustrates that generative models can offer a scalable and privacy-preserving alternative to sensitive operational data and, therefore, allow the further enhancement of intrusion detection systems in environments with limited resources. Collectively, the results support the fidelity utility-privacy triad as an effective and holistic model of assessing artificial data. They further portray the fact that a properly validated generative AI may be an effective facilitator of innovation in sensitive, data constrained domains. However, the findings also point to the fact that synthetic data must not be viewed as a complete substitute of the real data but as a strategic supplement, being able to augment, diversify, and provide resilience to further machine learning-based applications.

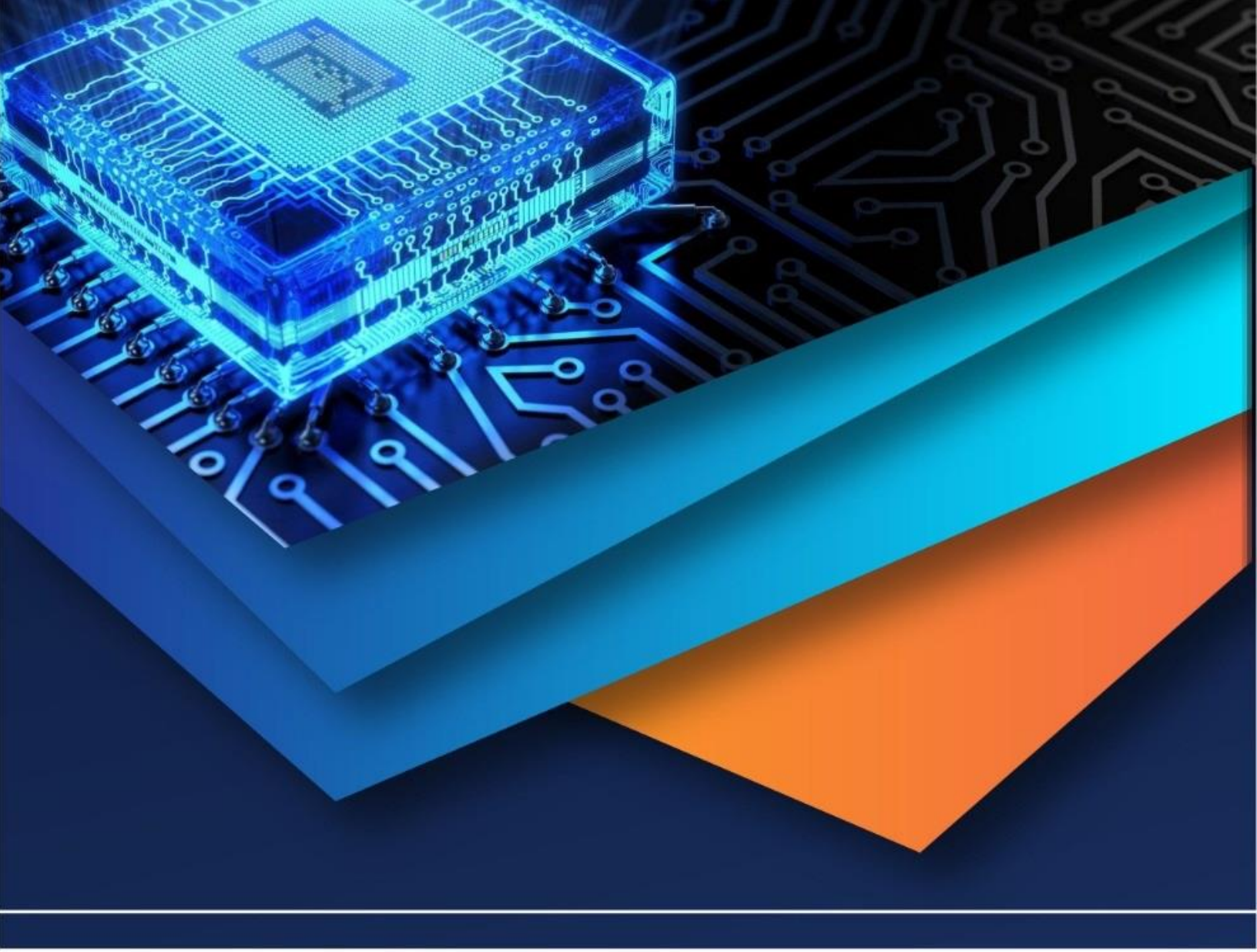
V. CONCLUSION

This paper brings forth the potential transformative nature of generative artificial intelligence (AI) in the creation of synthetic data, especially in sensitive areas like healthcare and cybersecurity, where the lack of information and security issues are paramount. The study using the models of Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Gaussian Mixture Models (GMMs) showed that realistic and non-identifiable synthetic data could be produced to supplement scarce real-world data.

Assessment on the fidelity, utility, and privacy dimensions reached the conclusion that synthetic data sets would be useful to study the statistical structure of actual data, predictive modeling, and detection of anomalies, and reduce the risks of individual data disclosure. Synthetic records, which are class-conditional, were found to improve disease prediction models in healthcare, and be used to improve the robustness of anomaly detection systems against emerging threats in cybersecurity. The results define synthetic data as a strategic complement, rather than a substitute, of real data, which presents prospects of data-enhancement, bias-elimination, and safer analysis. However, the paper emphasizes the need to have stringent validation systems, ethical guidelines and privacy protecting methods, like differential privacy, to reduce the risks that may arise. Future studies must aim at benchmark unification, the hybridization of generative-privacy systems, and diversification of cross-domain synthetic data use to achieve the optimal utility and confidentiality.

REFERENCES

[1] Agrawal, G., Kaur, A., & Myneni, S. (2024). A review of generative models in generating synthetic attack data for cybersecurity. Electronics, 13(2), 322.
[2] Baowaly, M. K., Lin, C.-C., Liu, C.-L., & Chen, K.-T. (2019). Synthesizing electronic health records using improved generative adversarial networks. Journal of the American Medical Informatics Association (JAMIA), 26(3), 228–241.
[3] Bharathi, S. P., Subin, P. G., Hariharan, R., Balaji, T. S., & Balaji, S. (2024, November). Generative AI for Synthetic Data Generation in IoT-Based Healthcare Systems. In 2024 Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES) (pp. 1-5). IEEE.
[4] Esteban, C., Hyland, S. L., & Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs. ArXiv preprint arXiv:1706.02633.
[5] Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321, 321–331.
[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems (NeurIPS), 27, 2672–2680.
[7] Innocent, E. K. (2024). Enhancing Data Security in Healthcare with Synthetic Data Generation: An Autoencoder and Variational Autoencoder Approach (Master's thesis, Oslo Metropolitan University).
[8] Jadon, A., & Kumar, S. (2023, July). Leveraging generative AI models for synthetic data generation in healthcare: balancing research and privacy. In 2023 International Conference on Smart Applications, Communications and Networking (SmartNets) (pp. 1-4). IEEE.
[9] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. International Conference on Learning Representations (ICLR).
[10] Lin, W., Zhang, Y., Chen, W., & Xu, M. (2020). Generating network intrusion detection datasets using variational autoencoders. Computers & Security, 92, 101740.
[11] Salem, A., Backes, M., & Zhang, Y. (2020). Don't generate me: Training privacy-respecting synthetic data with generative models. ArXiv preprint arXiv:2012.00863.
[12] Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic data—anonymization ground truth or a privacy mirage? NPJ Digital Medicine, 5(1), 1–10.
[13] Teo, Z. L., Quek, C. W. N., Wong, J. L. Y., & Ting, D. S. W. (2024). Cybersecurity in the generative artificial intelligence era. Asia-Pacific Journal of Ophthalmology, 13(4), 100091.
[14] Torfi, A., Fox, E. B., & Reddy, C. K. (2020). Differentially private synthetic medical data generation using generative adversarial networks. ArXiv preprint arXiv:2012.11774.
[15] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems (NeurIPS), 32, 7333–7343.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY