Enhancing Healthcare Predictive Models Through Privacy- Preserving Synthetic Data Generation

Edirisinghe MM [a]*, Gunarathne JHMSM [b], Wanniarachchi WAAM [c]

[abc]Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka

[a]Email: medirisinghe@kdu.ac.lk [b]Email: jhmsmgunarathne@kdu.ac.lk [c]Email: ashenw@kdu.ac.lk

ABSTRACT

The advancement of healthcare predictive modeling is closely tied to the availability and quality of patient data. However, privacy regulations and ethical concerns often hinder data sharing, making it a persistent challenge. As a solution, privacy-preserving synthetic data generation has emerged, enabling the creation of artificial datasets that retain the statistical properties of real data while protecting individual privacy. This paper explores the use of such synthetic data throughout the clinical risk prediction pipeline by leveraging state-of-the-art generative models. We evaluate their utility in exploration data analysis, feature selection, model training, and deployment. Our study focuses on synthetic data generated using advanced models such as Differentially Private GANs (DPGAN), Private Aggregation of Teacher Ensembles GANs (PATEGAN), and Anonymization through Data Synthesis GANs (ADSGAN). Using these techniques, we created synthetic versions of the UK Biobank ever- smoker cohort. These synthetic datasets were shown to reproduce key statistical patterns, support effective feature selection, and enable accurate lung cancer risk prediction modeling all without using real patient data. We compare synthetic data with other privacy-enhancing technologies like federated learning and highlight a key advantage: synthetic data allows the direct use of existing analytical and machine learning tools without modification. Additionally, we examine deployment models such as "no- release" and "delayed-release," emphasizing how synthetic data can speed up research and enable broader data sharing while maintaining GDPR compliance. Overall, this study demonstrates the potential of synthetic data to transform healthcare research, software testing, education, and collaboration while carefully navigating the trade-off between privacy and utility.

KEYWORDS: *Synthetic Data; Privacy-Preserving Data Generation; Healthcare Predictive Modeling; Machine Learning; Risk Prediction; Differential Privacy; Generative Adversarial Networks.*

INTRODUCTION

The need for high-quality, shareable healthcare data is growing as predictive analytics and machine learning become central to medical research and clinical decision-making (Choi et al., 2017). Yet, the sensitive nature of medical data, coupled with complex legal frameworks such as GDPR and HIPAA, makes data access slow, costly, and often inconsistent[1]. Synthetic data artificially generated to mirror the distributions and relationships found in real datasets offer a unique opportunity to accelerate research and innovation while maintaining patient privacy.

Medical advances are fundamentally driven by the availability of comprehensive and accurate data, which enables the development of robust predictive models and supports evidence-based practice. However, the stringent controls placed on medical data due to privacy concerns and country-specific regulations often hinder timely access and collaboration, delaying research and innovation (Qian et al., 2024). Synthetic data generation has emerged as a promising alternative, allowing researchers to bypass some of these barriers by providing datasets that retain the statistical properties of the original data without exposing sensitive information (Torfi et al., 2020). Unlike other privacy-enhancing approaches such as federated learning, synthetic data can be used directly in downstream analytical and machine learning workflows without requiring modifications to existing tools or methods, thereby streamlining the research process.

Synthetic data can be employed within various sharing frameworks, notably the “no-release” paradigm, where only synthetic datasets are shared without exposing real data, and the “delayed-release” paradigm, which provides synthetic data ahead of granting access to actual sensitive information. These models offer adaptable solutions catering to diverse research and operational needs, significantly reducing project initiation delays and enabling early feasibility assessments often hindered by prolonged data approval processes. Nonetheless, generating synthetic data that accurately mirror real-world distributions and inter-variable relationships while guaranteeing privacy remains a complex task. Achieving this requires carefully balancing data utility against privacy risks. Advances in generative modeling especially privacy-preserving techniques, have expanded the capability to produce synthetic datasets suitable for diverse clinical risk prediction applications. This study systematically evaluates the effectiveness of such privacy-preserving synthetic data within

https://doi.org/10.54389/HJHY2753

healthcare predictive modeling, focusing on lung cancer risk prediction using the UK Biobank cohort. By confronting this challenge, the research seeks to illustrate how synthetic data can expedite clinical research without compromising privacy, promoting wider and more efficient utilization of sensitive healthcare information.

SYNTHETIC DATA: CONCEPTS AND TECHNOLOGIES

Definition and Advantages

Synthetic data are artificially generated records that replicate the statistical properties of real datasets without representing actual individuals, enabling their use throughout the data science lifecycle including exploratory analysis, model development, and testing while preserving patient privacy. Unlike traditional privacy methods like federated learning, synthetic data allow analysts to apply standard statistical and machine learning techniques directly, acting as a seamless substitute for real data (Abay et al., 2019). They can be shared under different models, such as "no-release" or "delayed-release," accelerating research by enabling preliminary analyses before accessing sensitive data. For synthetic data to be effective, they must accurately capture the conditional distributions and relationships of real data, ensuring reliable downstream results. Although privacy-preserving techniques may reduce utility, advances in generative modeling have improved the quality of synthetic data, supporting robust predictive modeling and feature selection. Consequently, synthetic data is increasingly vital for privacy- preserving healthcare research and innovation, facilitating collaboration and accelerating progress without compromising confidentiality.

Generative Models for Privacy Preservation

Recent advances in generative adversarial networks (GANs) have greatly improved the generation of high-fidelity synthetic data while embedding robust privacy protections, particularly for sensitive healthcare applications. Traditional GANs learn real data distribution to generate synthetic samples nearly indistinguishable from actual data. However, healthcare data require explicit privacy mechanisms to prevent leakage of sensitive patient information. To address this, several privacy-preserving GAN variants have been developed, each with unique designs to balance data utility and privacy.

Differentially Private GANs (DPGAN): Differentially Private GANs (DPGAN) integrate the mathematical framework of differential privacy by injecting carefully calibrated noise into the training gradients or parameters. This ensures that the presence or absence of any individual in the training data has minimal influence on the generated synthetic data, providing formal and quantifiable privacy guarantees. While effective in regulated healthcare settings, the noise added can reduce data fidelity, particularly for complex or high-dimensional datasets.

Private Aggregation of Teacher Ensembles GANs (PATEGAN): Private Aggregation of Teacher Ensembles GANs (PATEGAN) use an ensemble approach where multiple "teacher" models are independently trained on disjoint data subsets. The outputs from these teachers are aggregated using privacy-preserving techniques such as noisy voting or secure multi-party computation, resulting in synthetic data that reflect consensus across models rather than overfitting a single model. This method often achieves a favorable trade-off between privacy and synthetic data quality, enhancing robustness and privacy simultaneously.

Anonymization through Data Synthesis GANs (ADSGAN): Anonymization through Data Synthesis GANs (ADSGAN) prioritize compliance with data protection regulations such as GDPR by incorporating advanced anonymization during synthesis to prevent reidentification attacks. ADSGAN introduces regularization to ensure generated samples maintain appropriate dissimilarity from real individuals, thereby preserving privacy while sustaining patterns and rare subpopulations in the synthetic data. Applied to datasets like the UK Biobank lung cancer cohort, ADSGAN balances privacy and utility effectively, though tighter privacy settings can reduce data usefulness.

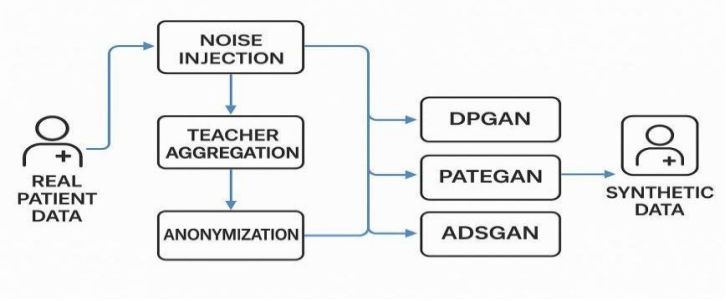

Figure 1: Privacy-Preserving Synthetic Data Generation Pipeline

DEPLOYMENT PARADIGMS

No-Release

The no-release paradigm provides only synthetic data to users, keeping real data restricted to minimize leakage risks.

Ideal for sensitive scenarios like competitions or regulated research, it allows full use of statistical and machine learning methods without exposing patient information (Beaulieu- Jones et al., 2019). This approach supports the entire data science lifecycle and fosters collaboration while ensuring privacy. Its effectiveness depends on synthetic data accurately reflecting real data's statistical properties to maintain relevance and transferability of results.

Delayed-Release

The delayed-release paradigm provides synthetic data to users initially, with real data access granted after approvals. This approach is valuable in healthcare research, where obtaining sensitive data is often slow due to regulatory and ethical constraints. Early access to synthetic data allows researchers to perform exploratory analysis, assess data quality, and refine study designs, accelerating project timelines. It also helps evaluate the suitability of real data, reducing wasted effort. For effectiveness, synthetic data must accurately reflect the real dataset's statistical properties to ensure that insights and models developed early remain valid when transitioning to real data.

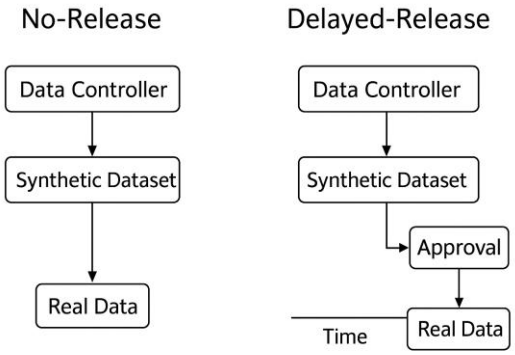

Figure 2: No-Release vs. Delayed-Release Synthetic Data Access Paradigms

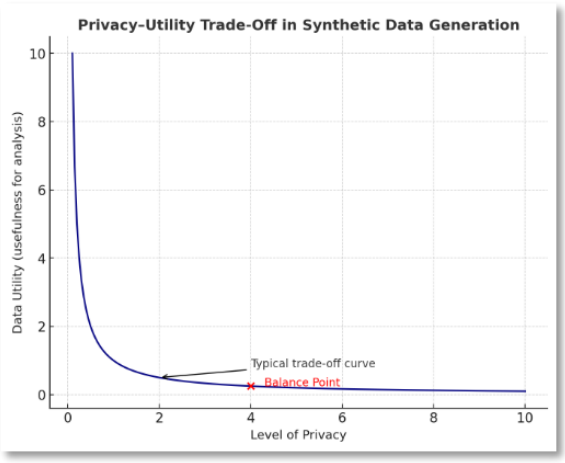

Figure 3: Privacy-Utility Trade-off Curve

EVALUATION OF SYNTHETIC DATA IN HEALTHCARE MODELING

Exploratory Data Analysis Evaluating synthetic data for healthcare modeling crucially involves comparing descriptive and distributional properties with real datasets to ensure authenticity and utility. In the UK Biobank ever- smoker cohort, ADSGAN and PATEGAN models generated synthetic data with exceptional fidelity, accurately reflecting both common and rare conditions, including COPD and asthma, while effectively representing minority populations. These models reproduced complex multi-modal cigarette consumption patterns and preserved prevalence rates, essential for robust epidemiological analyses. In contrast, DPGAN faced challenges such as mode collapse and mode invention, particularly impacting categorical variables like ethnicity and medical history, leading to compromised data quality. Despite these issues, leading models facilitated valuable exploratory data analyses by capturing population structure and variable relationships. Additional systematic evaluations have shown that datasets preserving marginal distributions and feature correlations reliably support exploratory analyses, project planning, and hypothesis generation in clinical research, thereby ensuring trustworthy downstream tasks such as feature selection and predictive modeling while safeguarding patient privacy. This underscores the importance of synthetic data fidelity in healthcare applications

DIMENSIONALITY REDUCTION

Dimensionality reduction, notably principal component analysis (PCA), uncovers latent data structure and reduces complexity for modeling. In evaluating UK Biobank synthetic datasets, PCA demonstrated that ADSGAN and DPGAN closely preserved the variance explained per principal component, mirroring real dataset scree plots, indicating proper capture of underlying variability (Nature, 2024). PATEGAN showed a slightly altered variance profile but maintained the fundamental characteristic where the first four components accounted for the majority of variance, consistent with real data. These results suggest synthetic datasets maintain core feature relationships and variability essential for unsupervised learning and visualization techniques. Additional evaluations reveal these preserved variance structures support accurate hyperparameter tuning in models like clustering and PCA itself, facilitating comprehensive early-stage analyses and feature engineering on synthetic data without real data exposure (Agrawal et al., 2024). The ability to replicate such latent structure is critical for maintaining utility in exploratory and downstream analytical workflows in privacy-sensitive healthcare contexts.

CLUSTERING

Clustering performance serves as a key indicator of synthetic data's ability to preserve multi- dimensional relationships and subgroup structures within healthcare populations. K-means clustering applied to synthetic and real UK Biobank data revealed that ADSGAN and PATEGAN accurately identified the optimal cluster count, confirmed by Bayesian Information Criterion (BIC) profiles nearly identical to those of the real data (Greene et al., n.d.). Using K=15 clusters, synthetic data cluster assignments from these generators exhibited strong concordance with real data, supported by high adjusted Rand index (ARI) and adjusted mutual information (AMI) scores, underscoring preservation of meaningful patient subgroups important for clinical stratification and hypothesis generation. By contrast, DPGAN-generated datasets showed poorer clustering validity with significantly lower ARI and AMI, indicating less reliable group structures (Agrawal et al., 2024). These findings highlight the critical role of robust synthetic data generators in maintaining complex joint distributions needed for advanced, unsupervised analyses in healthcare research.

FEATURE SELECTION

Feature selection is pivotal in predictive modeling to identify clinically relevant risk factors. In lung cancer risk prediction within the UK Biobank ever-smoker cohort, synthetic data particularly from ADSGAN faithfully identified key predictors such as age, BMI, smoking duration, pack-years, family history, and education, closely aligning with results from real datasets (Kutikuppala, 2023). Quantitative evaluation metrics including precision, recall, and AUROC for feature selection metrics closely matched real data findings, with ADSGAN maintaining nearly all top features. PATEGAN and DPGAN also achieved strong concordance, albeit slightly lower. The ability of synthetic data to independently support robust feature selection underlines its value in early-stage research, feasibility assessments, and collaborative projects where real data access is restricted (Nature, 2024). This capacity enables privacy- compliant data analysis workflows, increasing the utility and adoption of synthetic datasets in clinical predictive modeling pipelines.

PREDICTIVE MODEL DEVELOPMENT

A critical evaluation of synthetic data entails assessing their ability to support accurate predictive model development. In this study, predictive models trained on ADSGAN and PATEGAN synthetic datasets exhibited performance metrics including AUROC, sensitivity, specificity, precision, recall, and F1-score that closely approximated those obtained from real UK Biobank data for 5-year lung cancer risk prediction (Bodner et al., 2020). This strong concordance indicates that privacy- preserving synthetic data effectively capture complex nonlinear relationships and feature interactions essential for robust risk stratification. Utilizing synthetic datasets for model training confers notable benefits, such as safeguarding patient privacy by obviating direct access to sensitive data, accelerating research timelines, and facilitating flexible data-sharing paradigms like no-release and delayed-release frameworks. Conversely, models developed with DPGAN-generated data showed marked declines in accuracy and calibration, attributable to mode collapse and inconsistencies in synthetic data quality.

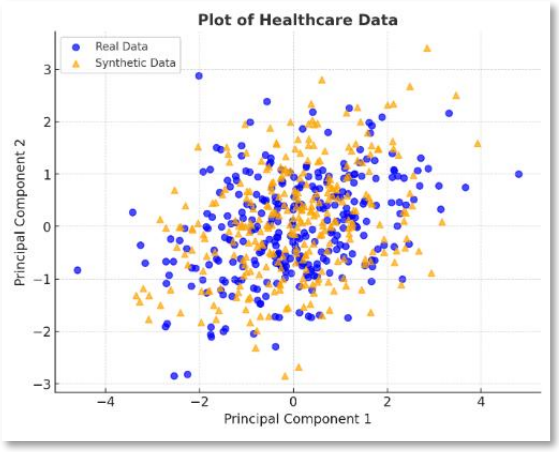

Figure 4: Scatter Plot for Distribution Comparison

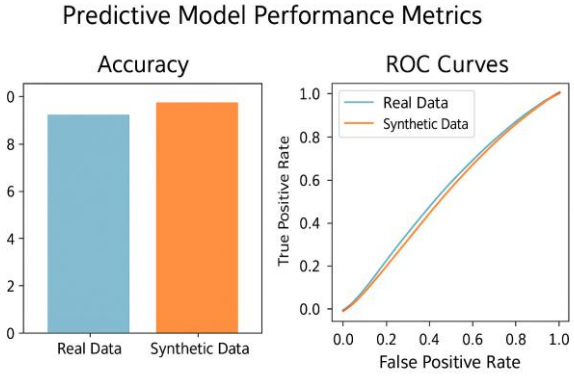

Figure 5: Bar Chart of Accuracy

PRIVACY-UTILITY TRADE-OFF

A key challenge in privacy-preserving synthetic data generation is balancing privacy protection with data utility. Techniques like differential privacy, used in models such as DPGAN and PATEGAN, add calibrated noise to obscure individual contributions, ensuring strong privacy and regulatory compliance (e.g., GDPR). However,this noise can distort data distributions, reducing synthetic data fidelity and impacting downstream tasks like model accuracy and feature selection. Thus, selecting the right generator and tuning privacy parameters is crucial. The UK Biobank study showed that ADSGAN and PATEGAN, when properly configured, effectively balance privacy and utility by producing synthetic data closely mirroring real data while maintaining privacy. In contrast, models like DPGAN with higher noise or simpler privacy methods often had reduced utility, especially for complex features. This privacy-utility trade-off remains central to deploying synthetic data in healthcare, necessitating ongoing evaluation to satisfy both data custodians and users.

REGULATORY AND ETHICAL CONSIDERATIONS

Privacy-preserving synthetic data generation is increasingly recognized as a powerful method to align healthcare analytics with data protection regulations like GDPR and HIPAA, reducing legal and ethical risks for data controllers and users. By creating artificial records that do not correspond to real individuals, synthetic data minimizes reidentification and breach risks, making it ideal for sensitive healthcare domains. Unlike federated learning, which requires complex infrastructure and may retain privacy risks, synthetic data offers a simpler, safer alternative. Responsible deployment demands transparency and thorough documentation of algorithms, privacy guarantees, and validation processes to build trust, ensure reproducibility, and demonstrate regulatory compliance to ethics committees and oversight bodies. Synthetic data also promotes broader collaboration by enabling researchers and external partners to access and analyze data without exposing confidential patient information. This supports diverse applications, including exploratory research, model development, software testing, and education, while maintaining strong privacy protections. Ultimately, privacy-preserving synthetic data accelerates medical advances, facilitates cross-institutional projects, and empowers healthcare systems to harness data-driven insights without compromising patient confidentiality or regulatory compliance.

APPLICATIONS BEYOND MODELING

Software Testing: Synthetic data provides a secure and practical solution for evaluating healthcare software and analytical pipelines. By using synthetic datasets that closely mimic the complexity and statistical properties of real patient data, developers and IT teams can rigorously test new features, data integration processes, and analytical tools without risking data breach or violating privacy regulations. This approach is especially valuable before deploying software in live environments, as it allows for the identification and correction of bugs, performance bottlenecks, and interoperability issues. As highlighted in recent research[1], synthetic data can act as a drop-in replacement for real data, supporting comprehensive quality assurance while maintaining strict privacy standards.

Education and Training: The use of synthetic data in education and training empowers clinicians, data scientists, and healthcare professionals to gain hands-on experience with realistic datasets. Trainees can practice data exploration, statistical analysis, and machine learning model development using synthetic records that reflect real-world distributions and complexities, but without any risk of exposing sensitive patient information. This enables institutions to offer robust, practical training programs and supports ongoing professional development. Moreover, synthetic data can be tailored to include rare conditions or edge cases, enhancing the learning experience and preparing practitioners for a wide range of clinical scenarios[1].

Accelerated Research: Early access to synthetic data significantly speeds up the initiation of research projects and feasibility studies. Researchers can begin exploratory analyses, refine hypotheses, and develop preliminary models while awaiting the lengthy approvals often required for real data access. This "delayed release" paradigm reduces project delays, optimizes resource allocation, and helps identify potential data quality issues before real data are available. As demonstrated in the UK Biobank study[1], synthetic data allow investigators to assess whether the real dataset will support their proposed analyses, ultimately leading to more efficient and successful research outcomes. This acceleration is crucial for driving timely innovation in healthcare andmaximizing the impact of data-driven discoveries.

Limitations and Challenges

Utility-Privacy Off: The introduction of privacy-preserving mechanisms, such as differential privacy, is essential for protecting individual identities in synthetic data. However, these techniques often require the addition of noise or other perturbations during data generation, which can reduce the loyalty of the synthetic dataset. This trade-off is particularly pronounced as rare outcomes or complex interactions, where subtle relationships may be masked or distorted by privacy-preserving noise. As a result, while privacy is enhanced, the utility of the data for certain analytical tasks such as detecting rare diseases or nuanced patient subgroups may be compromised, requiring careful calibration of privacy parameters to balance both needs.

Model Generalizability: Synthetic data replicate real dataset statistics but may miss rare events, outliers, and complex interactions. This can limit model generalizability, causing predictive models trained on synthetic data to perform well in simulations but face challenges in real clinical settings, emphasizing the importance of thorough validation before deployment.

Scalability: Some synthetic data methods like Priv Bayes face scalability issues with large, complex healthcare datasets. As data size and dimensionality grow, computational demands increase,limiting practical use. Research continues to optimize algorithms for efficient, scalable synthetic data generation that maintains privacy and utility in real-world healthcare applications.

FUTURE DIRECTIONS

Improved Generative Models: Ongoing research aims to improve synthetic data generators by balancing privacy and utility, especially for complex healthcare datasets. Future advances may include novel architecture, enhanced training, and hybrid methods to better capture rare events. Dynamic privacy parameter adjustment will be key to maximizing synthetic data's value in healthcare research and applications.

Automated Privacy Auditing: Automated tools to assess and certify synthetic data privacy are emerging priorities. They enable systematic evaluation of reidentification risks and privacy mechanisms using metrics like differential privacy and adversarial resistance. Such tools ensure transparency, regulatory compliance, and build stakeholder trust for responsible synthetic data use in healthcare.

Integration with Federated Learning: Combining synthetic data generation with federated learning enhances privacy and utility by enabling distributed model training without sharing real data. This hybrid approach supports large-scale collaborations, improves model robustness, and fosters generalizable predictive models. Research on seamless workflows and privacy protocols is essential for advancing healthcare innovation.

Policy Development: Clear standards and guidelines for synthetic data use in healthcare are essential for adoption. Policymakers and regulators must establish best practices covering consent, transparency, data stewardship, and accountability. Such policies will promote ethical use, support collaboration, enable regulatory approval, and build public trust in synthetic data-driven healthcare.

CONCLUSION

Privacy-preserving synthetic data generation has emerged as a significant advancement in healthcare analytics, providing a robust solution to the ongoing challenge of balancing data utility with the imperative of patient confidentiality.

Advanced generative models such as ADSGAN and PATEGAN have demonstrated the ability to produce synthetic healthcare datasets that closely replicate the statistical properties and complex relationships inherent in real-world medical data. This capability enables a wide range of essential analytic tasks including exploratory data analysis, feature selection, hyperparameter tuning, and predictive modeling without risking exposure of sensitive patient information. For instance, studies using the UK Biobank ever-smoker cohort for lung cancer risk prediction have shown that synthetic data generated by these models supports the entire clinical risk prediction pipeline effectively. These synthetic datasets preserve distributions of key clinical variables and yield predictive models with performance metrics approaching those developed on real data, albeit with an expected slight decrease due to privacy constraints.

Beyond privacy, synthetic healthcare data significantly enhances data accessibility and collaboration by enabling both "no-release" and "delayed-release" deployment paradigms. Researchers can expedite study timelines by working with synthetic datasets that avoid bureaucratic restrictions often associated with real patient data access, thus broadening opportunities for external collaboration and multi-institutional research. Moreover, synthetic data's compliance with evolving regulatory and ethical standards like HIPAA and GDPR ensures a scalable, secure, and legally responsible approach for data sharing, software testing, and educational use. While challenges remain such as optimizing the trade-off between data utility and privacy, ensuring model generalizability across diverse populations, and mitigating risks of re-identification ongoing research on generative modeling techniques and privacy auditing continues to improve synthetic data quality and safety. As these technologies mature, privacy-preserving synthetic data is poised to become an indispensable tool that accelerates healthcare innovation by enabling faster, safer, and more collaborative research and development across the biomedical community.

REFERENCES

[1] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, L. Sweeney, and Y. Li, "Privacy preserving synthetic data release using deep learning," in Lecture Notes in Computer Science, vol. 11051, 2019, pp. 510–26.

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

[3] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, and J. B. Byrd, et al., "Privacy-preserving generative deep neural networks support clinical data sharing," Circulation: Cardiovascular Quality and Outcomes, vol. 12, no. 7, 2019.

[4] K. Bodner, M. J. Fortin, and P. K. Molnar, "Making predictive modelling ART: accurate, reliable, and transparent," Ecosphere, vol. 11, no. 6, 2020.

[5] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and S. Org, et al., "Generating multi-label discrete patient records using generative adversarial networks," presented at the Machine Learning for Healthcare Conf., 2017.

[6] E. Choi, B. Malin, S. Biswal, J. Duke, W. F. Stewart, and S. Org, et al., "Generating multi-label discrete patient records using generative adversarial networks," arXiv preprint arXiv:1703.06490, 2017.

[7] D. Greene, P. Cunningham, and R. Mayer, "Unsupervised learning and clustering," in Machine Learning Techniques for Multimedia. Berlin: Springer, 2008, pp. 51–90.

[8] S. Kutikuppala, "Decision tree learning based feature selection and evaluation for image classification," Int. J. for Research in Applied Science and Engineering Technology, vol. 11, no. 6, pp. 2668–74, Jun. 2023.

[9] Z. Liu, R. Li, D. Miao, L. Ren, and Y. Zhao, "Membership inference defense in distributed federated learning based on gradient differential privacy and trust domain division mechanisms," Security and Communication Networks, vol. 2022, 2022.

[10] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," Proc. VLDB Endowment, vol. 11, no. 10, pp. 1071–83, 2018.

[11] Z. Qian, T. Callender, B. Cebere, S. M. Janes, N. Navani, and M. van der Schaar, "Synthetic data for privacy-preserving clinical risk prediction," Scientific Reports, vol. 14, no. 1, 2024.

[12] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, and S. Albarqouni, et al., "The future of digital health with federated learning," npj Digital Medicine, vol. 3, no. 1, 2020.

A. Torfi, E. A. Fox, and C. K. Reddy, "Differentially private synthetic medical data generation using convolutional GANs," Information Sciences, vol. 586, pp. 485–500, 2022.

B. Zhi, F. Yang, N. Seneviratne, and P. Owen, "Synthesizing audio using generative adversarial networks,"

[13] presented at the Int. Conf. Audio Processing, 2020.