APEX ACADEMIA PRESS
ELEGANCE ON EVERY RESEARCH

International Journal of Engineering & Tech Development
ISSN: 3069-672X/ Open Access / IJETD_Y1F3A001
Apex Academia Press / https://doi.org/10.63665/ijetd-y1f3a001

Original Article

GAN-Based Privacy-Preserving Data Synthesis in Healthcare

Carson James

Obafemi Awolowo University ile ife

Abstract

Strict privacy laws make it hard to get good medical data with notes, which is a big obstacle for AI's progress in healthcare. Generative Adversarial Networks could be one way to combine real health data without sharing with anyone. This paper discusses the application of Generative Adversarial Networks to create fake health care datasets, such that the clinical and statistical relevance of the generated dataset is retained with simultaneous guarantee of patient confidentiality. We provide an overview of the state-of-the-art GAN architectures used within the medical domain, discuss their performance concerning privacy and utility preservation, and introduce a GAN framework designed for privacy. Our experiments show that our approach can reduce the privacy risk while still generating good synthetic data that enables disease prediction and treatment outcome analysis. The results of this work will help facilitate cooperation in AI applied to healthcare and, therefore, sharing data.

Keywords

Generative Adversarial Networks (GANs), Healthcare Data Synthesis, Privacy Preservation, Synthetic Data, Medical AI, Data Anonymization, Differential Privacy, EHR Data Generation, Deep Learning in Healthcare, Data Sharing in Medicine.

Article History

Received: 13.06.2025

Accepted: 22.06.2025

Published: 12.07.2025

1. Introduction

A. Data's Significance in Healthcare Ai

During the last couple of years, AI has changed the way diseases are diagnosed, treatment plans formulated, and patients cared for. These AI models need huge amounts of quality data such as genomic data, medical imaging, and EHRs to function effectively. Such datasets are utilized to train machine learning models to identify minute patterns and predict events that are most likely to take place in the future. Data-driven models have proven instrumental in informing physician decisions and improving patient outcomes in specializations such as cardiology, oncology, and radiology. However, this information remains highly difficult to uncover. The main concerns that people usually have are those dealing with patient privacy, compliance with the rules, and how private the information is.

B. Data Sharing Issues Include Privacy and Regulations (Hipaa, Gdpr).

It may be possible to use data to make healthcare better, but privacy laws make it difficult to share and access medical data. Privacy laws like the US Health Insurance Portability and Accountability Act, HIPAA, and the EU General Data Protection Regulation create stringent requirements about how patient data could be collected, stored, processed, and shared. While these are important safeguards to protect private health information, they greatly limit the manner by which researchers and institutions can create AI systems that work together. That is, trying to hide patient data in a traditional fashion does not always work, as re-identification attacks using advanced methods might trace back anonymized datasets to real individuals. For this reason, scientists look at new ways of sharing information without giving up their privacy.

C. The Necessity of Synthetic Data

It is very hard to share information within health care due to privacy laws, so synthetic data has become a useful method for doing this. Synthetic data in essence is data that looks and behaves like real-world data with respect to trends and statistics but contains no patient information. You can replace real data with synthetic data in order to test algorithms, do exploratory analyses, and train machine learning models if done correctly. The best

part is that the companies can share fake data with each other easily. This would lead to a more creative healthcare AI ecosystem and make the whole operation better. However, generating useful and privacy-preserving synthetic data is a challenging task that requires state-of-the-art procedures in doing so.

D. An Overview of Gans and Their Significance

GANs have become one of the excellent approaches that generate fake data looking real since Ian Goodfellow first talked about Generative Adversarial Networks in the year 2014. A GAN basically consists of two neural networks: the generator and discriminator. People train these two neural networks to fight each other. While the discriminator tries to differentiate between real and fake samples, the generator learns how to generate fake data that looks much like real data. This fighting process against each other generates data that gets more and more similar to an original dataset over time.

Some of the most valuable applications of GANs in healthcare are generating medical images, synthetic EHRs, and other forms of data. They are very good in composing medical data since they model complex, high-dimensional distributions-a factor of importance for subsequent tasks.

E. Contributions to the Paper

This work comprehensively discusses GANs applications for privacy-preserving data synthesis in the health domain. We first present various types of GANs and their applications to generate health data. Second, we discuss how GAN-based models possess capabilities not found in any other technique for privacy protection. Further, we propose a new design for GAN which will generate fake health data but still follow the rules for privacy. We perform extensive tests in order to find data goodness, utility, and safety. Our results have shown that GAN-generated data can enhance important medical data while protecting patient privacy. Finally, we consider broader moral, legal, and practical implications of these model implementations in medicine in real life.

2. Background and Related Work

A. An Overview of Gans, Including Their Basic Concept and Variations (E.G., Cgan, Wgan, Pggan)

GANs are based on a simple yet powerful idea: two neural networks, one to sort and one to make. The generator makes fake data samples, and the discriminator checks each one to see whether it is real-from the training dataset-or fake-made by the generator. And through this adversarial training process, the generator keeps generating data which the discriminator cannot differentiate from real data. There exist a number of variations of GANs; some are more appropriate for particular types of data and more stable and useful than others. For example, conditional GANs allow one to generate data possessing certain characteristics. A common example cited in this regard is the generation of patient records for people of a particular age or with a particular diagnosis. Wasserstein GAN uses Wasserstein distance instead of the loss function. This prevents the training process from going wrong and avoids mode collapse. PGGANs increase the resolution of images progressively during training and hence result in high-resolution medical images. Such modifications have enhanced the ability of GANs to deal with healthcare data, which is usually of a very complex nature and contains multiple components.

B. Previous Research on Generating Synthetic Healthcare Data

GANs have been increasingly applied to generate synthetic health data across various domains. Among the earliest was the med GAN framework that used an autoencoder-based generator architecture for generating synthetic EHR data. Other models like Health GAN, ehrGAN, and PATE-GAN have also attempted to generate useful yet privacy-enhanced datasets. In the medical image domain, GANs have been leveraged to synthesize MRI scans, histopathological images, and chest X-rays. They assist in augmenting more data and enhancing the usefulness of data across diverse domains. These models often hold immense promise to replicate all visual features critical for diagnosis. People still raise concerns about the safety of these methods in real life, since many of them either do not offer any privacy or poorly estimate the likelihood of reidentification.

C. Methods for Generating Data While Preserving Privacy

The aim of producing data with privacy is to protect private information while still keeping the data useful for research purposes. One of the ways to do so is through federated learning. It sends the model out to be updated only when the data has changed. Another option can be differential privacy. It adds random numbers to all the

records to make them all look the same. These approaches, when applied to GAN, actually improve the privacy. For instance, DP-GAN uses gradient perturbation to add differential privacy in training and stops any leakage of certain training examples. On the other hand, PATE-GAN leverages the PATE framework in training a GAN with strong differential privacy guarantees. Obviously, these methods have the goal of keeping synthetic data useful for future tasks while maintaining the protection of private information about individuals. However, this may not be that good if these ways will be used. This is how important it is to find a good balance between usefulness and privacy.

D. Comparing Other Methods (E.G., Differential Privacy Alone, Anonymisation)

People traditionally used methods like eliminating identifying information and suppressing certain attributes to protect privacy. However, this is ineffective in the era of big data since sophisticated methods of reidentification might utilize external data sources or quasi-identifiers to link the data to individuals. On the other hand, differential privacy is strong in theory, while guarantees of the utmost that are often made when it is applied solely make high-dimensional data, such as electronic health records, less useful. In contrast, GAN-based synthetic data generation attempts to generate forged patient records similar to real ones and shows the distribution of data. At this point, when combined with other privacy-enhancing techniques, for example, differential privacy, GANs remain the best approach towards generating valuable data while assuring confidentiality. This new hybrid technique outperforms earlier methods because it reaches an improved trade-off between the risk of privacy leakage and the value of analysis.

3. Motivation and Problem Definition

A. Why Free Sharing of Real Medical Data Is Not Possible

Medical records are private because they contain so much private information about the health, medical history, and current treatments of a person. Even without stating your name, sharing this information could still put one's privacy at risk. HIPAA and GDPR are just two of many laws that make it very difficult to use and share this type of information, especially across countries or institutions. Companies do not want to share data because they do not want to get in trouble with the law or lose business in case someone breaks into their system. Getting permission to share data is also hard and can take a really long period of time, and in most cases, the patient has to agree. That is why most of the healthcare datasets are still not linked to date. This makes it challenging to come up with AI models that actually do well and could be used by a large number of people.

B. Patient Re-identification Risks

Even without your name, address, or Social Security number in the database, someone could still find out who you are. Attackers can link anonymous records to real people based on outside databases and "quasi-identifiers" like age, ZIP code, and admission date. Several seminal studies have demonstrated how easily de-identified data is re-identified, especially in datasets that are particularly fine-grained. This is especially concerning for healthcare because everyone's medical history is unique. Re-identification not only violates patient confidentiality but potentially leads to legal consequences and a loss of public confidence. Given these problems, it is important to study methods of truly anonymizing data beyond removing names and other identifying information.

C. Goals for Synthetic Data: Privacy, Utility, and Fidelity

The standards that synthetic data needs to meet, if it will turn out to be a good stand-in for real patient data, are fidelity, usefulness, and privacy. Fidelity refers to something about the structure and statistics in synthetic data mirroring those in real data. You need to replicate the same clinical associations, relationships, and distributions present in the original dataset. Utility refers to the usefulness of synthetic data for training machine learning models, running statistical analyses, and modelling clinical workflows.

The hardest part in the creation of fake data is to find a balance among these three goals. Finally, privacy means the fake data should not have or allow anyone to guess any real patient information. If we make and train GANs properly, we can achieve these goals. They are able to create data which looks real but can't be traced back to a person.

4. Proposed Methodology

A. Healthcare Data GAN Architecture (Model Description, Loss Functions)

The proposed approach is based on a special version of the GAN, developed for health data. The model contains two neural networks that have learned to play in a minimax game: discriminator and generator. The former generates fake records indistinguishable from real data, while the latter tries to distinguish between real and fake healthcare records. Since medical data is challenging to handle and very heterogeneous, we adopt a conditional GAN architecture. This framework allows the generator to condition the output depending on specific inputs, such as age, diagnosis codes, or drug classes. In this way, this enhances the quality and usefulness of samples for therapy. In order to avoid the problem of mode collapse-an intrinsic problem when training GANs-the Wasserstein distance with a gradient penalty has been introduced inside the loss function; this decreases the difficulty during the learning and makes any choice easier. Due to the characteristics described above, the loss functions are designed with great care, in order to find an accurate trade-off between precision and learning dynamics. The result will be high-quality, realistic records.

B. Combining with Privacy-Preserving Methods (DP-GAN, PATE-GAN, etc.)

We apply differential privacy methods when training GANs so that the disclosed fake data does not leak personal information. This can be guided by interpreting how models like DP-GAN and PATE-GAN work. The method behind DP-GAN-adds noise into the backpropagation gradients and prevents one point from being so crucial for model learning. That will ensure the generated fake data will not learn or remember real patient records well. We apply differential privacy methods to keep fake data very private when training GANs. This can be done by interpreting how models like DP-GAN and PATE-GAN work. Adding noise during backpropagation to the gradients makes one data point not so crucial for model learning in the method behind DP-GAN. That will ensure the generated fake data is not able to learn or remember real patient records very well.

C. Data preprocessing (such as tabular data, imaging, and EHR structure)

Preprocessing is an important step that GAN-based synthesis must go through to work with real healthcare data. The forms of medical information are many, some of which are time-series data from monitoring devices, unstructured clinical notes, structured tabular data from EHRs, and high-resolution medical images such as MRIs and X-rays. This research covers the premise of structured electronic health record data and imaging datasets. Before providing such data as input to a neural network, raw EHR data should be cleaned and formatted. That is, missing values should be imputed, continuous variables (for example, age and lab results) should be normalized, and categorical variables should be encoded (for example, operation types and diagnosis codes) using only one hot. The EHR data sequences in time series should be either shortened or lengthened to maintain the fixed length of the input. Apart from scaling images and normalizing pixel values, several specific pre-treatment changes related to the area are performed on imaging data. In this way, more data types can be handled by the model. All the data fed to the GAN model is standardized by this pre-processing pipeline, which makes training and generation easier.

D. Pipeline for Training and Evaluation

Through real training data, GAN learns, while the model's hyperparameters are changed with the help of the validation set. In training, lots of discriminator-generator swaps occur using both real and fake samples. Noise and clipping gradients are used at every step for keeping promises about privacy. After training, the generator creates fake data sets. All these datasets then go through a very long process that includes privacy audits using known attack models, utility assessments using machine learning tasks, and tests of similarity between two sets of data. This will ensure that the fake data is real, useful, and clearly private.

5. Experimental Setup

A. Datasets Utilised (Public EHR Datasets, MIMIC-III, etc.)

These then get tested on benchmark datasets, which are common in healthcare and used to determine whether they can be replicated and are useful. MIMIC-III is a free, large dataset of medical records for patients who are being treated in an ICU. The dataset includes information on demographics, diagnoses, prescription drugs, and lab results. We use some of the available imaging data from public databases like TCGA, or The Cancer Genome

Atlas and NIH ChestX-ray14. These archives maintain pictures of various diseases, and each is named. We have selected these datasets because they are extensive, diverse, and well-acknowledged in scientific research. As we have already said, the data is divided and cleaned such that it does not leak.

B. Evaluation Metrics

(a) Usefulness: ML Task Performance and Statistical Similarity

You subsequently have to compare the statistical properties of the synthetic data with those of the real dataset to check how useful it is. Tests using the Kolmogorov-Smirnov test, joint distributions, and correlation matrices find out how close the synthetic data is to the real one. We go further to train machine learning models such as random forests, logistic regression, and neural networks that can handle both real and fake data. Then, we check whether the data is useful for AI tasks in the real world by using a common validation set. We check a model's accuracy, F1-score, and AUC-ROC to check how well it predicts. These show that synthetic data can be used for medical analytics in the future.

(b) Privacy: Risk of Re-identification and Membership Inference

Different attack models, like membership inference attacks, test whether privacy is protected. These attacks want to find out if the GAN learned from a specific record. It is good for keeping your privacy in case some attacks do not work very well. We also consider the risk of reidentification by looking at the likelihood that, given more data, a synthetic sample would be matched with an actual patient. In that respect, when necessary, we make use of differential privacy theory, for example, ε-differential privacy bounds, which we use to determine the privacy scores. The idea herein is that all these checks ensure that the created data will not breach patient privacy by carefully checking privacy.

(c) Comparisons and Baselines

We also compare the proposed GAN model with some baseline models, namely med GAN, Health GAN, and table-GAN. All these models generate fake information. Other common approaches to ensuring data privacy, such as k-anonymity and suppression, were also compared. Besides this, there are models that have no protection of privacy and models that do, just to show changes in data utility which will result from the protection of privacy. All these comparisons are relevant in view of the fact that they will be able to inform us about how the method works in striking a good balance between privacy and realism.

6. Results and Discussion

A. Comparison between Synthetic and Real Data

Our GAN model generates synthetic data with statistical properties similar to those of source datasets. Both quantitative and visual analyses of PCA plots, correlation structure, and variable distributions test correspondence between real and synthetic data. In some instances, synthetic data can be more consistent, reduce noise, or other problems present in the original records. But the changes would have been done carefully to avoid making things easy nor, at the same time, remove critical information helpful for treatment.

B. Performance on ML Tasks Downstream

Models that are trained on synthetic data and then tested on real data perform equally well as models that have exclusively trained on real-world data. Synthetic data is good enough for analytical tasks, as attested by the high F1-scores and AUC values which classifiers, trained to predict disease outcomes or medication responses using synthetic EHRs, obtain. This can be a testimony to how useful the synthesized dataset could be for training, especially in instances where getting real data is difficult due to privacy laws.

C. Utility vs. Privacy Trade-offs

One of the most important things we learned from our research is that privacy and usefulness are always in conflict with each other. By making things more private, like adding more noise or stricter clipping, the data we get isn't as good. Even with high privacy settings, GANs allow the use of data for analysis. We show that models such as PATE-GAN and DP-GAN can be helpful and at the same time keep people's private information safe. A privacy-utility curve shows how privacy and usefulness are connected. It helps people find the best setup according to their needs and level of risk.

D. Limitations and Failure Cases

There are good and bad aspects of the suggested methodology. For example, the fact that GANs cannot always make rare or outlier instances could hurt the models that are supposed to find rare diseases. Then, there is mode collapse. In this case, the generator makes only a few samples, and the data isn't as varied. Finally, differential privacy normally requires computers to do more work and for much care to be exercised in the setup of hyperparameters. From an evaluation point of view, no test about privacy can offer unequivocal guarantees, and some risks may persist anyway. These limits show us points on which we further have to go and study more.

7. Ethical and Legal Considerations

A. Synthetic Data Ethics in Medicine

Medicine takes a stand whereby the use of synthetic information is considered wrong. While there are new hazards, the advantage of synthetic data is that it will provide ease in many collaboratives works regarding science projects and information sharing. Poor or biased decisions about their health might be made if people believe the synthetic information is actually representative of the real data. You'll have to know its origin, limitations, and what the data is to be used for if you want to use synthetic data in a moral way. In addition, it is considered that patients whose synthetic data is to be used for research or business purposes must be informed through public notification or consultation with a large number of people.

B. Regulatory Approval

Synthetic health data occupies different legal standings in different jurisdictions. Most consider synthetic data to fall outside of laws like HIPAA and GDPR, as it does not actually contain any personal information. The rules on this, however, are often very unclear and do not always work along the same lines. Such vagueness may make this technology difficult to adopt by companies that do not want to take risks. For example, regulatory bodies may also require official evidence that synthetic data cannot be traced to individuals or it cannot be reverse-engineered. Policymakers will have to keep communicating with one another about how to establish clear regulations and certifications on the safe use of synthetic data in clinical research and innovation.

C. Hazards of Abuse (such as Hallucinated Data)

Synthetic data can have a record that looks real but consists of unhelpful or impossible combinations of variables for therapeutic purposes. People don't talk enough about the dangers of fake data. If you don't consider this data up close, that may lead to hidden biases or bad advice when using it to train models or make decisions. Another concern is that dishonest people might use fake data with an intent to lie or make up evidence in the regulatory filings. These risks can be reduced by strong validation frameworks, field expert reviews, and audit trails that need to be part of any deployment pipeline using fake healthcare data.

8. Conclusion and Future Work

A. An Overview of the Results

This work investigates the application of GANs to generate fake medical data in a manner that ensures the protection of individual privacy. We discussed how legislation such as HIPAA and GDPR makes the sharing of actual patient data between groups highly complicated. New enabling technologies are needed that provide privacy with utility in data. Old techniques of hiding data no longer work. We aimed at reducing the re-identification risk by proposing a GAN-based framework which integrates healthcare data synthesis with differential privacy mechanisms such as DP-GAN and PATE-GAN. Tests on datasets like MIMIC-III show that our synthetic data remains statistically pure and performs just as well on subsequent machine learning tasks, such as disease prediction. Privacy audits also demonstrate that synthetic data greatly reduces the chances of guessing one's membership or other common attacks. Finally, our results confirm that GAN-generated synthetic data, which maintains privacy, can serve as a valid alternative to real medical data for research analytics and AI development.

B. Implications for Medical AI and Data Sharing

Health AI can change things because it can make fake health data that is really useful and appears to be real. Sharing synthetic data between research institutions, companies manufacturing drugs, and the one’s developing AI does not raise moral or legal problems as far as protecting patient privacy goes. In such a case, more people get

access to good data and may accelerate the creation of decision-support tools, treatment suggestions, and diagnostic models. You also use GAN-generated data to simulate rare clinical events that are not well-represented in real-world data, make models more generalizable, and supplement real data sets. Synthetic data is far more available to make medical researchers share their ideas and come up with new ideas much faster. All participants need only agree on standard rules for privacy and utility assessments and give clear legal rules for using synthetic data in a moral way to get all of these benefits.

C. Future Prospects: Federated Learning, Multimodal Data, and Real-World Implementation

There will be many good ways to mix health data in the future through GANs. For comprehensive, multimodal data generation that brings different but interconnected genomes, imaging, EHR, and clinical documentation into one single GAN model, further research should be done. This would help AI models learn more about how healthcare really works. Secondly, federated learning with GANs ensures that the data is kept safe because only GAN updates are shared, not the data, which is kept distributed. With this approach, companies will be able to collaborate on the creation of fake data without the need for exchanging real patient records. Third, we have to fight harder for deploying synthetic or generated data into the real-world healthcare settings. That means you have to gain expert opinions all the time, validate the test outcomes as matched to real events, and connect the health information systems that hospitals use. The government will have to make strict rules to ensure synthetic data safety and efficiency in real-world applications, and do some clinical validation studies.

9. References

[1] Goodfellow et al. (2014) introduced GANs, marking a breakthrough in deep generative modeling.
[2] Choi et al. (2017) developed med GAN, one of the first GANs tailored for EHR data synthesis.
[3] Xu et al. (2019) proposed PATE-GAN, which integrated the Private Aggregation of Teacher Ensembles framework into GAN training.
[4] Beaulieu-Jones et al. (2019) examined privacy-preserving deep learning for biomedical data, introducing DP-GANs.
[5] Esteban et al. (2017) explored time-series generation for medical data using recurrent GAN architectures.
[6] Baowaly et al. (2019) proposed an EHR-GAN model to generate structured medical data while preserving data semantics.
[7] Chlap et al. (2021) reviewed GAN applications in medical imaging, highlighting diagnostic use cases.
[8] Park et al. (2018) evaluated synthetic data for downstream ML tasks, using metrics such as predictive performance and feature importance.
[9] Shokri et al. (2017) demonstrated the feasibility of membership inference attacks on machine learning models.
[10] El Emam et al. (2011) discussed the limitations of traditional de-identification techniques in healthcare.
[11] Abay et al. (2018) proposed privacy-preserving synthetic data generation frameworks across multiple domains.
[12] Goncalves et al. (2020) benchmarked synthetic data generation methods for tabular healthcare datasets.
[13] Torkzadehmahani et al. (2020) assessed fairness and privacy in synthetic data produced by generative models.
[14] Johnson et al. (2016) introduced the MIMIC-III dataset, which has become a foundational benchmark in clinical data research.
[15] Kuo et al. (2022) studied regulatory readiness and ethical implications of synthetic data in biomedical research.