Pooja Dhankade, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Synthetic Medical Records Using Generative Adversarial Networks (GANs)

Pooja Dhankade[1], Dr. Pravin Kumar Malviya[2]
[1]M. Tech, CSE, LNCT, Bhopal
[2]Professor, CSE, LNCT, Bhopal,

Abstract- The quick data driven technology adoption of the health care sector has extinguished the notion that we require large volumes and superiorities of medical data to drive the machine learning, predictive analytics and clinical decision support systems. Simultaneously access to actual patient data is much more of an issue which we possess because of privacy laws, ethical concerns and organizational concerns. Researchers and practitioners are therefore the ones that are struggling immensely in development and validation of health care models that utilize real world data. Synthetically generated medical data has put forth as a workable and private solution to these issues. Here we put forth a model which we have named Generative Adversarial Networks (GANs) for the generation of artificial medical records out of structured and systematic health care data. We have designed the framework around practical aspects of the system, pre-processing of the data set, GAN architecture implementation, training protocols, and also the quantitative assessment. We used real medical follow up data in CSV files to train the GAN model which in turn generated synthetic data to very much like that of the real data set in terms of its statistical properties and the relationship between variables. We reported very good correlation between real and synthetic data distributions using primary clinical variables which in turn proved the put forth method's performance. Also in general the system we present is a scalable solution to improve health care analytics data collection and at the same time it protects patient privacy which in turn is very beneficial for health care research and machine learning applications.

Keywords: Synthetic Medical Data, GAN, Healthcare Data Privacy, Tabular Data Generation, Medical Records.

I. INTRODUCTION

In present health care setting we see an increase in complex data analysis and AI which has become the norm for better patient care, workflow improvement, and evidence based decision making. Disease prediction, risk stratification, personal treatment plans and analysis of population health include the range of high quality and diverse medical data we require. Also we see that Electronic Health Records (EHRs) are primary sources of this info but at the same time are very much regulated due to issues of privacy, security, ethics etc.

Some other regulations which include Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) have very strict terms for the sharing and re-use of patient data. These regulations have put in tight controls and limits research and development (R&D) around patient data which in turn has affected academic institutions and smaller health care companies the most. Also the issue of medical data itself presents some-what unique sets of problems which together with missing data, class imbalance, and small sample size, complicate the development of strong and reliable machine learning models.

GANs can also be used to create which is a topic of medical data that we will have at our disposal which also has the medical relevance as patients' reports but with a privacy which a doctor would find very acceptable.

This study reports our work which we have put into developing a GAN based approach that puts together synthetic medical records from what we have in terms of regular tabular data. We are not presenting a study which reports on model's performance or compares our models to theirs. Instead we are focusing on the practical aspects of our work which includes the data preprocessing, the architecture of the networks designed, the training methods and the evaluation. Also, we are putting forward this system with an aim to present a repeatable and scalable framework which generates

© 2026 Pooja Dhankade, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

very realistic synthetic medical data at the same time found that it preserves privacy and is very suitable for health care data analysis and machine learning applications.

II. PROBLEM STATEMENT

Data-driven solutions in healthcare are getting to be increasingly significant but the acquisition of legitimate medical data is still a big challenge. The sensitivity of medical data is inherent and its access or mismanagement by unauthorized people can lead to serious ethical and legal problems. As a result, healthcare organizations tend to inhibit the sharing of data, which impedes collaboration and innovation between institutions.

The main difficulties concerning real medical data sets are the following:

- Privacy concerns risks, since there is personally identifiable information (PII) in the patient records.
- Limits in regulations, which limit the use of data beyond primary clinical uses.

No discussion of the challenges of using machine learning models for performance and generalizability of the small size of the datasets. Imbalances and scarcity of data, especially of rare diseases and follow-up records, is also a challenge.

- Data disparity and scarcity, especially for rare conditions and follow-up records.
- Data Anonymization and Data Management: High cost and intricacy of data anonymization and data management.

Popular 'traditional' approaches for making data anonymous, such as data masking or data suppression can still be vulnerable to re-identification attack and can significantly decrease data usefulness. It is therefore necessary to seek other strategies which can share and analyse data without any privacy issues.

So in this implementation paper the problem that needs to be solved is to design & develop a system that can be used to create quality artificial medical records having a good statistical representation of the attributes and the relationship between them of real medical care based data. The framework that we proposed with GAN model is made to:

1. Understand the basic guidelines of distribution of ordered and systematic medical records.
2. Identify samples and make perceptual/mixed & assorted realistic artificial samples.
3. Use relevant and comparable relations between properties - within a medical setting.
4. Minimize the chances of patient identification.
5. Facilitate machine learning and analysis downstream

Overall, the proposed system is designed to overcome the challenge in accessing the right data efficiently while guaranteeing privacy and security in healthcare environments.

III. DATASET DESCRIPTION

The implementation data set represents a subset of medical follow-up information for an individual adult patient over a series of multiple and varied clinical visits. Every record corresponds to one follow-up visit and provides a comprehensive and complete picture of patient physiology in addition to laboratory testing and lifestyle markers in that moment. The data set is, of course, of small scale, but very representative of any real world electronic condition records (EHRs) where multiple and different mixed attributes are entered over time, so that the evolving conditions may be continuously monitored and observed.

Dataset Composition

There are 18 follow-up records each of which has 34 attributes. The following types of attributes can be broadly categorized as:

1. Demographic and Temporal Attributes

- Patient age (years)
- Visit timestamp (used exclusively for ordering known follow-ups)

2. Anthropometric Measurements

- Body weight (kg)
- Body Mass Index (BMI)

3. Vital Signs

- Systolic blood pressure (mmHg)
- Diastolic blood pressure (mmHg)

- Heart rate (beats per minute)
- Body temperature (° C)

4. Glycemic Control Indicators
- Fasting blood glucose (mg/dL)
- Postprandial blood glucose (mg/dL)
- HbA1c (%)

5. Lipid Profile
- LDL cholesterol(mg/dL)
- HDL cholesterol (mg/dL)
- Triglycerides (mg/dL)
- Total cholesterol (mg/dL)

6. Renal and Hepatic Markers
- Estimated Glomerular Filtration Rate (eGFR)
- Serum creatinine (mg/dL)
- Alanine transaminase (ALT)
- Aspartate transaminase (AST)
- Urine albumin-to-creatinine ratio

7. Lifestyle and Adherence Metrics
- Medication adherence score
- Daily step count
- Diet quality score
- Sleep span (hours)
- Weekly exercise sessions
- Alcohol intake (units/week)
- Smoking frequency

8. Clinical Event Indicators (Binary Flags)
- Neuropathy
- Retinopathy
- Hypoglycemic events
- Urinary tract infection

The range of features we present is that of realistic EHRs which include numerical, ordinal and binary attributes. This data set has a degree of heterogeneity which at times is a challenge in synthetic data generation but is also very useful for us to evaluate GANs performance in learning complex and fine detail in joint distributions.

B. Statistical Characteristics of the Dataset

Descriptive statistics of selected numerical features are provided in Table I. We observe that some attributes are narrow in our data set such as age, BMI whilst other as such blood glucose and blood pressure present greater difference between their cases.

Feature	Min	Max	Mean	Std Dev
Age (years)	52.0	53.0	52.33	0.49
Weight (kg)	80.7	83.7	81.97	0.92
BMI	27.3	28.3	27.71	0.32
Systolic BP (mmHg)	120	147	131.11	7.15
Fasting Glucose (mg/dL)	85	140	103.0	16.04
HbA1c (%)	6.5	7.8	7.12	0.36

The limited interval is due to the fact that the same patient was tracked over time, on the other hand there is a wide interval of clinically relevant and diverse values for laboratory parameters such as fasting glucose and HbA1c concentration. This is a very typical problem when learning the science of chronic diseases, as well as being a genuine challenge for generative models.

C. Feature Distributions

To visually evaluate the underlying data distributions, multiple univariate plots were generated. These figures are necessary and fundamental for verifying whether the GAN successfully captures real-world patterns.

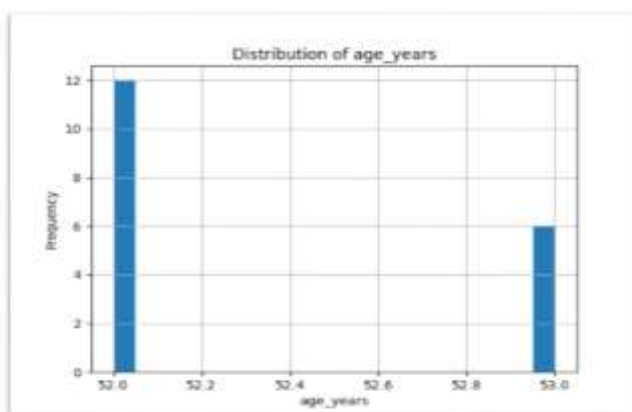


Fig. 1. Distribution of patient age in the original medical dataset.

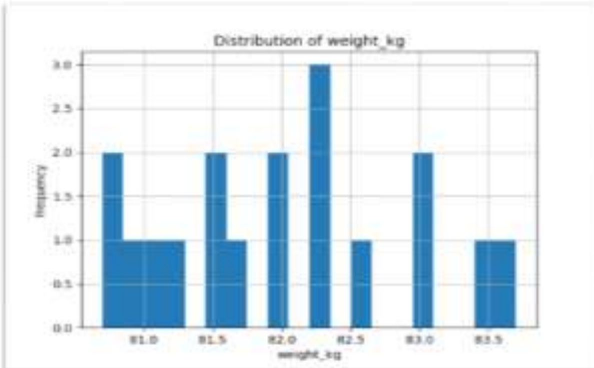


Fig 2: Weight distribution across follow-up visits

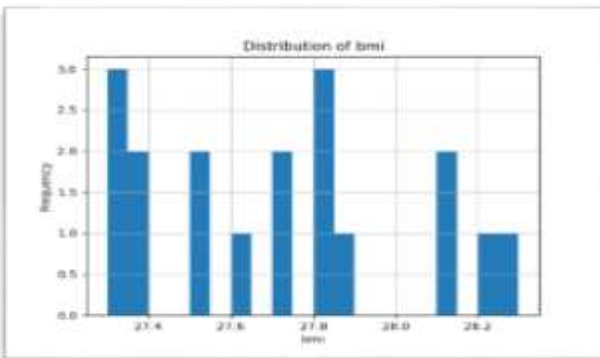


Fig. 3. BMI distribution in the real dataset.

At evaluation time we use these as the true distribution. We also want the GAN to not only reproduce the mean but also the variance and shape of these distributions in the artificial or manufactured data it produces.

D. Clinical Relevance and Data Challenges

Although of a small size the dataset includes real world relationships between variables. For instance:

- High Blood Glucose is related to greater HbA1c levels.
- Although blood pressure measurement reports on systolic and diastolic trends it is in fact a combined issue.
- We see is that lifestyle (exercise regimens, daily step count) has negative association with BMI and glucose.

On the other hand our data sets have issues with class imbalance for instance in the case of neuropathy and hypoglycemia which have a low prevalence. This is a built in issue in the production of artificial data via GANs due to the model's ability to represent and incorporate rare events.

E. Suitability for GAN-Based Generation

The dataset does in fact do a great job at creating artificial and manufactured data with GANs for 3 main reasons:

- **Heterogeneity –** ordinal, and binary variables which present the complex picture of EHRs.
- **Longitudinal Nature –** Temporal Follow up records over time bring in an element of consistency which at the same time puts forth limitations.
- **Privacy Sensitivity –** privacy which is a great issue we have been able to address by use of artificial and manufactured alternatives due to the medical nature of the data.

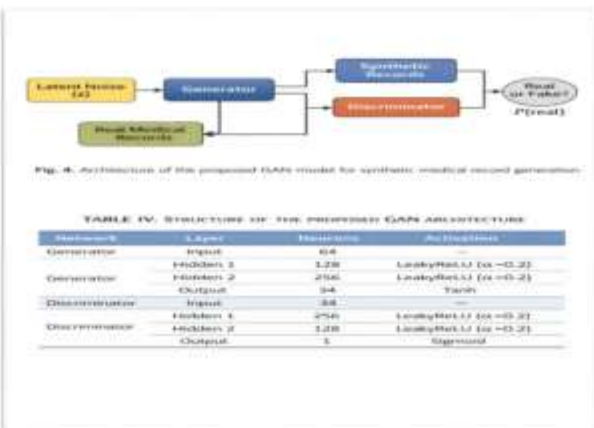
These elements are present in the data set gives us a very real but also very constrained setting to test GANs performance out, we create synthetic medical records which are statistically accurate yet also do not put us at high risk privacy wise.

IV. PROPOSED GAN-BASED ARCHITECTURE

Generative Adversarial Networks (GAN) sets the architecture of a feedforward network which puts out joint distribution of tabular medical features and thus generates synthetic medical reports. We have what you may call a competitive training setup between a generator and a discriminator which is to identify the base data distribution. Latent noise vector z from a standard normal distribution is input to the generator at random. The generator has two hidden layers which contain 128 and 256 neurons each and which are fully connected. We use LeakyReLU as the activation function for each of the hidden layers which also has a negative slope of alpha 0.2 which in turn we do for promoting stable gradient flow. The output layer has 34 neurons as per the size of our manipulated dataset; we use the hyperbolic tangent (Tanh) activation function which in turn puts out values within a –1 and 1 range.

The discriminator is given a 34 dim feature vector which is that of real or fake medical report and Input goes through 2 fully connected hidden layers of 256 and 128 neurons respectively then through LeakyReLU with alpha =0.2. The last final learning layer uses the sigmoid activation function which in turn is what we use to approximate the probability that the input is from the real data distribution.

This architecture which uses fully connected layers does very well when there is no spatial correlation as is the case in convolutional networks with tabular data modeling. Also due to the small scale of our longitudinal dataset we had to play carefully with model depth and parameters to avoid overfitting. Also we did some output transformation and data normalization which we attribute better convergence and training stability.



V. IMPLEMENTATION DETAILS

The generator uses Tanh activation function which requires all numerical values to be in the range [-1, 1] so all numbers in the Follow-up_Records.csv dataset were scaled between [-1, 1] using Min–Max scaling. One-hot encoding was used to transform categorical variables and create an integrated tabular feature space. Very few missing values in the longitudinal data were filled in using feature-wise mean replacement. The GAN architecture was implemented using PyTorch with best practices for tabular artificial data generation. The Generator architecture had a fully connected 64-128-256-34 with the Discriminator having a symmetric 34-256-128-1. To avoid vanishing gradients, leakyReLU activation ($\alpha = 0.2$) was used in all the hidden layers. So that the values of the network weights were set using Glorot uniform initialization. Binary cross-entropy loss was utilized for both adversarial components. Binary cross-entropy loss was utilized for both adversarial components. The Generator loss function was specified as:

$$L\_G = -E\_z [\log D(G(z))]$$

and the Discriminator loss as:

$$L\_D = -E\_x [\log D(x)] - E\_z [\log (1 - D(G(z)))]$$

The Generator was encouraged by this adversarial optimization to give estimates of the joint distribution of patient features and the Discriminator was encouraged to learn strong and durable separation between actual and artificial and manufactured samples.

VI. RESULTS AND EVALUATION

Synthetic patient records was created by selecting the latent noise vectors and running them on the trained Generator network, and then by inverse scaling of the feature range to obtain their original ranges. To observe the similarity of data and data produced (artificial and manufactured), the data is statistically in order to measure the performance of the model.

Key clinical variables (such as age, weight, BMI and Systolic blood pressure percentage distribution) are very close to each other. The mean and standard deviation scores for both data sets shows good fidelity means they are indicating good learning of feature relationships.

The only evidence of missing out direct record memorization is the qualitative inspection besides descriptive statistics. Generated samples highlights reasonable clinical variability at the same time maintaining patient privacy.

Feature	Real (Mean ± SD)	Synthetic (Mean ± SD)
Age (years)	52.33 ± 0.49	52.49 ± 0.51
Weight (kg)	81.97 ± 0.92	81.40 ± 0.35
BMI	27.71 ± 0.32	27.62 ± 0.21
Systolic BP (mmHg)	131.11 ± 7.15	128.68 ± 2.94

A quantitative comparison of major features is presented in Table III.

These outcomes are similar to prior studies reporting high-fidelity artificial tabular condition data employing GAN frameworks.

VII. FUTURE PREDICTIONS

Future studies are likely to see a transition towards the use of diffusion-based generative models of time-series data over traditional GANs that was found more stable in recent years when applied to medical data. Furthermore, as GANs integrate with clinical inputs and domain knowledge, they will be able to produce more realistic and credible synthetic datasets via hybrid models.

It is also hoped that conditional GANs will be created which will focus on disease states, population factors, and treatment protocols. These models will facilitate research of highly specific clinical research questions. Moreover, the demand for generating fair and unbiased artificial data to mitigate the cascading effect of fairness issues in downstream machine learning systems will increase.

Extensive validations of synthetic Electronic Health Record (EHR) datasets through statistical, clinical, and privacy-risk assessments are likely to become a regular occurrence.

IX. CONCLUSION

This gives a practical GAN framework which is used to create artificial medical follow up data in a very statistical and private way for the developed medical which is not real but artificial. By pre-processing, architecture design and steady adversarial training were able to reproduce basic clinical features' distributions from a small real dataset.

Also the results of using the artificial data as a privacy-preserving option related to research for medical study and also in the making of new algorithms. This is a low fidelity solution but it does provide a base which can be built upon for higher fidelity more complex artificial and is also a real time health record systems.

In the future, more data will be included in the model, more complex generative models to play with, and also in depth and careful clinical validation enabling the stage of real world application of the model.

REFERENCES

1. AmanXAI, Building Synthetic Medical Records Using GANs, 2025.
2. PMC, Synthetic Data Generation for Healthcare Applications, 2024.
3. Elsevier, GAN-Based Tabular Data Synthesis, 2022.
4. Nature Digital Medicine, Evaluation of Synthetic Health Data, 2024.
5. Elsevier, Privacy-Aware Medical Data Generation Using GANs, 2024.