Anonymization & Synthetic Data Generation

Pramod K C

22085342

A project proposal submitted for INFO7016 Postgraduate Project A

in partial fulfilment of the requirements for the degree of

Master of Data Science

Supervisor: Rosalind Wang

Co-Supervisor: Dr. Jim Basilakis

School of Computer, Data and Mathematical Sciences

Western Sydney University

August 2024

Table of Contents

Summary

The project's goal is to create a reliable framework for anonymizing medical information and producing high-quality synthetic data using advanced deep learning algorithms. The key goals are to protect patient privacy by anonymizing personally identifiable information (PII) and to generate synthetic data that closely resembles the statistical features and patterns of real medical information. This will be accomplished by utilizing Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and transformer models. The synthetic data created will be realistic and useful for training machine learning models, conducting research, and testing, all while following to rigorous privacy rules.

The project involves the process like collecting the real medical datasets, the development of anonymization and privacy of datasets to save confidentiality. Exploratory Data Analysis and Feature Engineering is implemented to make dataset clean, structured and free from noise. We use deep learning models which are trained and fine-tuned for synthetic data production, followed by thorough evaluation to assure data quality, value, and privacy. The successful completion of this project will equip healthcare organizations with tools for safely using synthetic medical data, thereby improving data-driven research and innovation while protecting patient anonymity. Furthermore, this project addresses the issue of insufficient datasets, allowing for better results.

Introduction

The digital revolution has transformed healthcare along with virtually every other industry. From telemedicine to digital health data, providers now have access to innovative solutions that have the potential to make healthcare more accessible and effective for all. An electronic health record (EHR) is a digital repository of a patient's medical information that documents their entire healthcare journey in real time. Electronic health records (EHRs) have revolutionized healthcare by enabling data-driven decision-making, individualized medication, and sophisticated research. Electronic Health Records (EHRs) improve patient care by giving healthcare providers real-time access to detailed patient information, lowering medical mistakes and increasing diagnosis accuracy. They improve cooperation among providers, increase efficiency by reducing paperwork, and reduce expenses. EHRs also provide useful data analytics for research and public health monitoring, empowering people to participate in their healthcare by providing access to their records and facilitating the use of telemedicine. EHRs have revolution the ancient health care system but there are also many challenges. While encryption and cybersecurity offer sophisticated safeguards to guarantee EHR privacy, digital records always carry a risk of violation. The increase in health data brings with it the risk of patient confidentiality being leaked. EHRs store vital and sensitive patient health information. Regardless of the circumstances, those data must always be protected from strangers. But we also need to study that data to analyze and research the findings, predict diseases that will emerge in the future, and prevent them in the present.

This project aims to create a framework for anonymizing patient health data in order to prevent patient important information from leaking and to maintain confidentiality on EHRs. So, in our first section, we look at how we might secure patient data by masking sensitive data and only sharing useful patient data for research reasons.

For the second part, we will develop synthetic data that is likely to have the same information as real data. The benefit of synthetic data generation is that it has similar information to genuine data, but it is phony data that bears no resemblance to any patients. It is also useful in situations where there are insufficient datasets for research purposes to create models and identify results.

Needs/Problems

- The increasing digitization of medical records raises concerns about the privacy and confidentiality of sensitive patient information.
- Researchers and developers often face challenges in accessing real medical data due to privacy concerns and legal restrictions.
- While anonymization is essential, it often leads to a loss of data utility, rendering the data less useful for analysis and model training.
- Producing synthetic data that closely resembles real-world medical data is difficult.
- Even with synthetic data, there is a risk of inadvertently revealing sensitive information if the data generation process is not carefully controlled.

Objectives

- Create a method for anonymizing genuine medical records by deleting or obscuring personally identifiable information (PII) and sensitive features while preserving the data's usefulness.
- Ensure that the synthetic data generated does not mistakenly expose any real patient information, hence protecting people' privacy and confidentiality.
- Create high-quality synthetic medical data that reflects the statistical features and patterns seen in real medical records. This data should be valuable for training machine learning models, doing research, and performing other analyses.

Procedures

Phase 1: Data Collection and Preprocessing

During this phase, we gather massive health clinical datasets while adhering to legal and ethical requirements. Datasets can be structured, unstructured, image-based, and so on. We then undertake exploratory data analysis, such as viewing dataset features, to identify target variables. We will also undertake feature engineering and data cleaning to ensure that datasets are in the correct format. Clean and preprocess the data to address missing values, normalize characteristics, and remove personally identifiable information. Text data will be tokenized and cleaned to eliminate noise.

Data Used

Our project involve anonymization and generating synthetic data we choose highly practical and diverse healthcare data.Each column provides specific information about the patient, their admission, and the healthcare services provided, making this dataset suitable for various data analysis and modeling tasks in the healthcare domain. Here’s a brief explanation of each column in the dataset -

Name: This column represents the name of the patient associated with the healthcare record.

Age: The age of the patient at the time of admission, expressed in years.

Gender: Indicates the gender of the patient, either ”Male” or ”Female.”

Blood Type: The patient’s blood type, which can be one of the common blood types (e.g., ”A+”, ”O-”, etc.).

Medical Condition: This column specifies the primary medical condition or diagnosis asso- ciated with the patient, such as ”Diabetes,” ”Hypertension,” ”Asthma,” and more.

Date of Admission: The date on which the patient was admitted to the healthcare facility. Doctor: The name of the doctor responsible for the patient’s care during their admission. Hospital: Identifies the healthcare facility or hospital where the patient was admitted.

Insurance Provider: This column indicates the patient’s insurance provider, which can be one of several options, including ”Aetna,” ”Blue Cross,” ”Cigna,” ”UnitedHealthcare,” and ”Medicare.”

Billing Amount: The amount of money billed for the patient’s healthcare services during their admission. This is expressed as a floating-point number.

Room Number: The room number where the patient was accommodated during their admis- sion.

Admission Type: Specifies the type of admission, which can be ”Emergency,” ”Elective,” or ”Urgent,” reflecting the circumstances of the admission.

Discharge Date: The date on which the patient was discharged from the healthcare facility, based on the admission date and a random number of days within a realistic range.

Medication: Identifies a medication prescribed or administered to the patient during their admission. Examples include ”Aspirin,” ”Ibuprofen,” ”Penicillin,” ”Paracetamol,” and ”Lipi- tor.”

Test Results: Describes the results of a medical test conducted during the patient’s admission. Possible values include ”Normal,” ”Abnormal,” or ”Inconclusive,” indicating the outcome of the test.

Phase 2: Anonymization & Privacy

The purpose of the personal identifiable information (PII) de-identification tool is to help scrub sensitive data out of datasets. PII might be used alone or in tandem with other relevant data to identify an individual and can incorporate direct identifiers that can identify a person uniquely, such as passport information, or quasi-identifiers that can be combined to successfully recognize an individual, such as race or date of birth. We develop methods for detecting and removing personally identifiable information (PII) from medical records using natural language processing (NLP) techniques such as named entity recognition.

Data obfuscation is the process of disguising confidential or sensitive data to protect it from unauthorized access. Data obfuscation tactics can include masking, encryption, tokenization, and data reduction. Data obfuscation is commonly used to protect sensitive data such as payment information, customer data, and health records. Data masking, encryption, and tokenization are three common data obfuscation techniques. We can use strategies to conceal sensitive attributes while preserving the data's utility. Techniques such as attribute switching, noise addition, and generalization will be investigated.

Phase 3: Synthetic Data Generation

The primary objective is to build and refine machine learning models that can produce high-quality synthetic medical data. We have chosen top three models that can be implement for synthetic data generation.

We select GANs because of their ability to provide data that is realistically generated through competition between a discriminator and a generator. A generative adversarial network (GAN) is a class of machine learning frameworks and a prominent framework for approaching generative AI. A Generator and a Discriminator, two neural networks that compete in a zero-sum game, are the components of a GAN. The Generator's goal is to produce realistic data, while the Discriminator's goal is to distinguish between real and synthetic data. You can produce synthetic data that bridges the gap between data accessibility and privacy issues in healthcare by carefully adjusting and assessing these models. This synthetic data can then be used for training, testing, and research purposes.

Variational Autoencoders (VAEs) for synthetic data generation is an effective approach due to their ability to model complex data distributions and generate high-quality, realistic data samples. In the fields of artificial intelligence and machine learning, variational autoencoders, or VAEs for short, are a kind of generative model. Finding the underlying structure in the supplied data and using that structure to create new data samples is their main task. In situations when the dataset is complicated and high-dimensional, VAEs are thought to be an extremely effective tool for unsupervised learning tasks. VAEs may produce realistic and high-quality synthetic data for training, testing, and research while maintaining privacy and regulatory compliance by concentrating on the architecture, loss functions, training, assessment, and optimization.

Transformer models are a strong and cutting-edge method for creating synthetic data, especially for unstructured text data such as clinical notes. Transformers' capacity to identify intricate patterns and correlations in text data has transformed natural language processing (NLP) applications. A transformer model is a type of neural network that tracks relationships in sequential data, such as the words in this sentence, to determine context and meaning.Transformer models detect the intricate ways in which even distant data pieces in a series impact and depend on one another through the application of a growing collection of mathematical approaches known as attention or self-attention. In order to create realistic, high-quality medical text data, developing Transformers for synthetic data production requires utilizing cutting-edge NLP approaches. Transformers can create synthetic medical records that are realistic and helpful for research, development, and clinical applications while protecting patient privacy. This is achieved by carefully fine-tuning pre-trained models, optimizing training, and rigorously evaluating the quality of generated data.

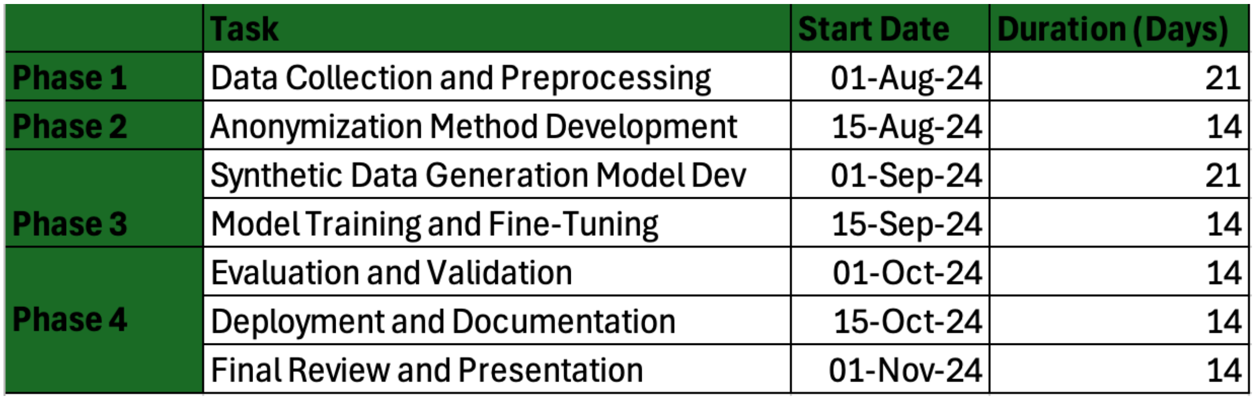
Phase 4: Evalution & Validation

Creating a model is not the only crucial step. Verifying its accuracy and assessing it are equally crucial. We must assess the correctness of the model. We validate the model using testing datasets once it has been trained using train data. Machine learning utility, privacy metrics, and statistical similarity can all be used to evaluate the model for this project. Using measures like mean, variance, and distribution overlap, we compare the statistical distributions of features in synthetic data with those in real data. We are utilizing simulated data, train machine learning models and assess their efficacy using actual test data. To evaluate utility, use measurements like precision and accuracy. We Utilize privacy-preserving measures, such as differential privacy, to measure the possibility of re-identification and make sure that patient privacy is not jeopardized by synthetic data.

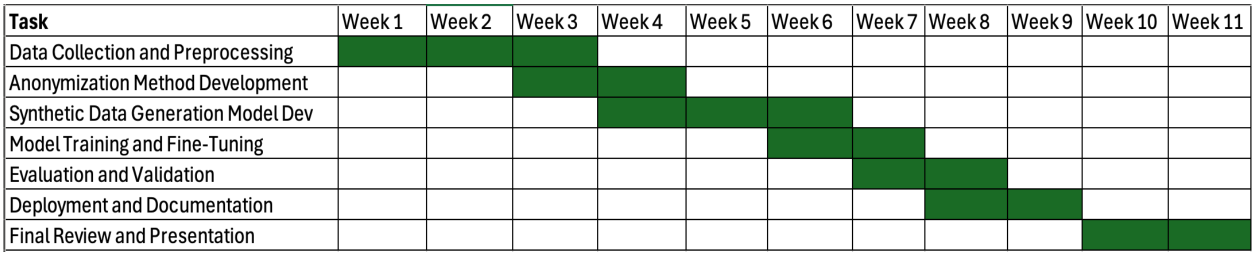
Expected Outcome

Obtained dataset should be strong while maintaining high privacy. While performing different data analysis and Machine learning algorithm, for instance, linear regression,Support vector regression, MLP regression, synthesised data should obtain statistical property closer to the original data. Minimise over fitting and make the data less vulnerable to Membership Inference attack, where an attacker attempts to determine if a specific individual’s data was used to train a model. The susceptibility of different generative models to such attacks is assessed. Maintain a decent level of utility, while retaining high privacy. Synthetic data should achieve high score in discriminator testing. It is a method to evaluate whether the generative models overfit the training data, which could lead to privacy breaches.

Timetable



	Task	Start Date	Duration (Days)
Phase 1	Data Collection and Preprocessing	01-Aug-24	21
Phase 2	Anonymization Method Development	15-Aug-24	14
Phase 3	Synthetic Data Generation Model Dev	01-Sep-24	21
	Model Training and Fine-Tuning	15-Sep-24	14
Phase 4	Evaluation and Validation	01-Oct-24	14
	Deployment and Documentation	15-Oct-24	14
	Final Review and Presentation	01-Nov-24	14



Task	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11
Data Collection and Preprocessing	🟩	🟩	🟩								
Anonymization Method Development			🟩	🟩							
Synthetic Data Generation Model Dev				🟩	🟩	🟩					
Model Training and Fine-Tuning						🟩	🟩				
Evaluation and Validation							🟩	🟩			
Deployment and Documentation								🟩	🟩		
Final Review and Presentation										🟩	🟩

References

ISO. (n.d.). *Electronic health records explained*. [online] Available at: https://www.iso.org/healthcare/electronic-health-records.

Nvidia.com. (2023). *PII Identification and Removal — NVIDIA NeMo Framework User Guide latest documentation*. [online] Available at: https://docs.nvidia.com/nemo-framework/user-guide/latest/datacuration/personalidentifiableinformationidentificationandremoval.html [Accessed 1 Sep. 2024].

SearchSecurity. (n.d.). *What is PII (Personally Identifiable Information)? Definition from SearchSecurity*.[online]Availableat: https://www.techtarget.com/searchsecurity/definition/personally-identifiable-information-PII#:~:text=Personally%20identifiable%20information%20(PII)%20is.

crowdstrike.com. (n.d.). *What is Data Obfuscation? – CrowdStrike*. [online] Available at: https://www.crowdstrike.com/cybersecurity-101/data-obfuscation/.

Wikipedia Contributors (2019). *Generative adversarial network*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Generative_adversarial_network.

Larksuite.com. (2024). Available at: https://www.larksuite.com/en_us/topics/ai-glossary/variational-autoencoders-vaes.

Merritt, R. (2022). *What Is a Transformer Model?* [online] NVIDIA Blog. Available at: https://blogs.nvidia.com/blog/what-is-a-transformer-model/.

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A. and Bennett, K.P. (2019). Assessing privacy and quality of synthetic health data. doi:https://doi.org/10.1145/3359115.3359124.