Reinforcing Synthetic Data of Patients Suffering from Liver Cirrhosis: A Project Proposal

1. Introduction

Liver cirrhosis is a major global health issue. It causes irreversible liver damage and affects millions of people worldwide. In Nigeria, the number of cases is rising due to hepatitis B and C infections, alcohol abuse, and poor healthcare access. The World Health Organization (WHO) estimates that over 18 million Nigerians have hepatitis B, making it one of the highest burdens globally [1]. Early diagnosis and treatment are crucial to improving survival rates. However, research and clinical management of liver cirrhosis are difficult due to a lack of high-quality patient data [2].

One of the biggest challenges in liver cirrhosis research is limited access to reliable medical data. Ethical concerns, privacy laws, and hospital policies restrict access to real patient records. In Nigeria, the problem is worse because many hospitals lack proper digital record-keeping systems. This leads to incomplete or lost patient records, making it hard to build accurate AI models for disease prediction. A study at the University College Hospital in Ibadan found that

poor documentation was a major barrier to liver disease research [3]. These data limitations slow down progress in medical research and AI-driven healthcare solutions.

Synthetic data generation offers a possible solution. It creates artificial patient records that resemble real ones while protecting patient privacy [4]. However, existing synthetic datasets often lack depth and realism, making them less useful for AI models [5].

This research aims to improve synthetic data using advanced AI techniques, including generative adversarial networks (GANs) and reinforcement learning. Better synthetic datasets will help train AI models for liver cirrhosis diagnosis and prognosis, improving healthcare outcomes [6], and ultimately bridging the gap between healthcare and artificial intelligence [7].

2. Problem Statement

Liver cirrhosis is a serious health problem worldwide. It causes permanent liver damage and is responsible for over 1 million deaths every year [8]. In Nigeria, the number of cases is rising due to hepatitis B and C infections, alcohol abuse, and poor healthcare access. The World Health Organization (WHO) reports that Nigeria has one of the highest burdens of hepatitis B, with over 18 million people infected [9]. Many cases of liver cirrhosis are not diagnosed early, leading to complications such as liver failure and liver cancer, both of which have high death rates [10]. Early detection and treatment can improve patient survival, but there are many challenges in studying and managing this disease.

One of the biggest problems is the lack of reliable medical data. Doctors and researchers need high-quality data to understand the disease, build predictive models, and improve treatment plans [11]. However, real-world medical data are hard to access, especially in Nigeria. Hospitals have

strict privacy rules to protect patient records. Ethical concerns also prevent the sharing of sensitive medical information. Many hospitals do not have digital record-keeping systems, making it difficult to store and retrieve patient data [12]. A study at Lagos University Teaching Hospital found that researchers faced strict ethical restrictions when trying to use patient data for an AI-driven hepatitis B detection project. This shows how difficult it is to obtain real patient records for AI research [13].

Without enough data, it is hard to train AI models for disease prediction and diagnosis. AI has the potential to improve liver cirrhosis detection, but these models require large amounts of diverse and high-quality data. Incomplete or low-quality data leads to weak AI models that cannot make accurate predictions. This is a major problem in medical AI research. If AI models are trained on limited or biased data, they may not work well in real-world medical settings. This can lead to incorrect diagnoses and poor treatment recommendations.

Synthetic data provides a possible solution. It creates artificial patient records that resemble real ones while protecting patient privacy [14]. Researchers can use synthetic data to train AI models without violating ethical rules. However, current synthetic datasets have major limitations. Many lack realism and diversity. They do not accurately capture the full range of liver cirrhosis cases, including different patient backgrounds, disease stages, and complications. This limits their usefulness in AI model training. If synthetic data do not match real-world conditions, AI models trained on them will not perform well in hospitals and clinics.

This research will address these issues by improving synthetic data quality. It will use generative adversarial networks (GANs) to create synthetic patient records that closely mimic real ones. It will also apply reinforcement learning to enhance the accuracy and diversity of these datasets.

The goal is to develop synthetic data that can effectively train AI models for liver cirrhosis diagnosis and prognosis. This will help improve healthcare outcomes, especially in Nigeria, where access to real medical data is limited.

3. Aims and Objectives

Aim of the Study

The goal of this study is to improve synthetic patient data for liver cirrhosis research. Liver cirrhosis is a serious health problem, especially in Nigeria, where many cases go undiagnosed due to poor healthcare access and limited medical data. AI can help in diagnosis and treatment, but it needs large, high-quality datasets to work well. Real patient data are difficult to access due to privacy rules, ethical concerns, and poor record-keeping. This project aims to create better synthetic data using AI techniques, so researchers can study liver cirrhosis without needing real patient records.

Research Questions

This study will answer the following questions:

1. **How can AI improve the quality of synthetic liver cirrhosis data?**
2. **Can reinforcement learning make synthetic data more realistic?**
3. **How do AI models trained on synthetic data compare with those trained on real data?**
4. **Can synthetic data help in liver cirrhosis diagnosis and prognosis in Nigeria?**

Principal Problem

The main problem is that real medical data are difficult to access, making AI research on liver cirrhosis challenging. Current synthetic datasets are not realistic enough, as they lack clinical and demographic diversity. If synthetic data do not reflect real-world conditions, AI models trained on them will not be reliable. This study aims to solve this by improving the accuracy and diversity of synthetic patient data using advanced AI techniques.

Objectives

1. **Analyze real-world liver cirrhosis data** – First, the researcher will study existing datasets (where available) to understand key patient characteristics. This will help in designing synthetic data that accurately reflects real cases.
2. **Generate synthetic data using GANs** – Afterwards, generative adversarial networks (GANs) will be trained to create synthetic liver cirrhosis records. These AI models will learn from real data to produce artificial patient profiles.
3. **Improve synthetic data using reinforcement learning** – Reinforcement learning comes in next to refine the synthetic datasets. Thus, the study captures a wide range of patient variations.
4. **Validate the synthetic data** – Going further, the researcher compares the performance of AI models trained on synthetic data with those trained on real data to check if they are equally effective.

5. **Test the use of synthetic data in liver cirrhosis diagnosis and prognosis** – Lastly, the study will assess whether synthetic data can help develop AI models for early detection and treatment planning.

Approach and Methods

To achieve these objectives, the research will use the following methods:

- **Data Analysis** – It will examine available liver cirrhosis datasets to identify important features such as age, gender, disease severity, and risk factors.
- **Machine Learning (GANs)** – It will use GANs to generate synthetic liver cirrhosis patient records that mimic real-world data.
- **Reinforcement Learning** – The research will then refine the synthetic data using reinforcement learning to improve accuracy and diversity.
- **Model Validation** – Finally, it will train AI models using both real and synthetic data, and then compare their performance in predicting liver cirrhosis.

Technologies Used

For this project, Python-based AI tools will be used. These include TensorFlow and PyTorch for building GANs and reinforcement learning models. The study will also use statistical analysis tools like Pandas and NumPy to assess data quality.

Significance of the Study

This study will help researchers and doctors use AI for liver disease diagnosis without needing large amounts of real patient data. If successful, it will provide a way to improve healthcare research in Nigeria and other regions with limited medical records

4. Legal, Social, Ethical, and Professional Considerations

This research involves creating synthetic medical data, which raises legal, ethical, social, and professional concerns. Even though synthetic data do not contain real patient information, they must follow strict guidelines to ensure safety and fairness.

Legal Considerations

Medical data privacy laws, such as Nigeria's National Health Act and global laws like the General Data Protection Regulation (GDPR), require strong data protection. Failure to comply with these regulations can lead to significant legal repercussions, including fines and loss of research credibility [2]. Thus, if the research uses real-world medical data for training its AI models, it must ensure proper anonymization and security. It will also follow all rules to protect patient privacy and avoid legal issues.

Ethical Considerations

One key ethical issue is the risk of synthetic data being inaccurate or biased. If the AI does not generate realistic patient records, it could lead to wrong conclusions in research. This could harm future AI models used for diagnosis. To prevent this, the researcher will carefully validate the synthetic data before using it. This validation ensures that the data accurately reflect real-world

patient populations and clinical scenarios [2]. They will also be open about how the data are created to avoid misreading other researchers.

Social Considerations

This project is meant to help improve healthcare in Nigeria, where liver cirrhosis is a big problem. However, AI models trained on synthetic data must be tested well before use. If the data are not accurate, AI predictions could be wrong, affecting patient care and trust in technology.

Professional Considerations

Ethical AI practices and professional standards will be followed in this paper. The researcher will also share their findings openly with other researchers so they can help improve synthetic medical data for better healthcare solutions.

5. Background

Liver Cirrhosis and the Need for Data-Driven Research

Liver cirrhosis is a serious disease that causes permanent damage to the liver. It is a major health problem worldwide, especially in developing countries like Nigeria. The disease is mainly caused by chronic hepatitis B and C infections, alcohol use, and poor access to healthcare. According to the World Health Organization (WHO), liver cirrhosis is among the top causes of death globally, with over 1 million deaths recorded each year. In Nigeria, hepatitis B and C are

widespread, and many people do not get proper treatment due to a lack of awareness and medical facilities [1].

Early diagnosis and treatment can improve survival rates, but healthcare providers struggle to detect liver cirrhosis in time. Advanced AI models can help in diagnosing and predicting disease outcomes, but these models need large amounts of high-quality patient data [15]. Unfortunately, in Nigeria, medical records are poorly kept, and data privacy laws make it difficult for researchers to access real patient records. This creates a major challenge in using AI for liver cirrhosis research which hinder timely diagnosis and treatment [16].

The Role of Synthetic Data in Medical Research

Synthetic data is artificial data generated by computers to mimic real-world information. In medicine, synthetic data can help researchers train AI models without using real patient records. This is important because real medical data are often hard to get due to privacy laws, ethical concerns, and poor record-keeping. With synthetic data, researchers can create datasets that look like real patient records but do not contain any actual patient information [17].

However, synthetic medical data must be realistic and diverse to be useful. If the data do not reflect the real characteristics of patients, AI models trained on them will not be reliable. Many current synthetic datasets lack important variations in patient demographics, severity of illnesses, and treatment outcomes [17]. Therefore, this research aims to solve this problem by improving the quality of synthetic liver cirrhosis data using AI techniques like generative adversarial networks (GANs) and reinforcement learning.

Previous Research on Synthetic Medical Data

Many researchers have studied how synthetic data can be used in healthcare. One well-known study by Choi et al. (2017) used GANs to generate synthetic electronic health records (EHRs) [18]. Their model, called MedGAN, successfully created artificial patient records that preserved important medical patterns. However, the data lacked enough diversity, which made it less effective in real-world applications [17]. Another study by Beaulieu-Jones et al. (2019) showed that synthetic data could be used to train AI models for predicting diseases [19]. However, they also found that the synthetic datasets needed further improvement to match real-world complexity [17].

Meanwhile, in liver disease research, few studies have explored the use of synthetic data. Most AI models for liver cirrhosis rely on limited real-world datasets. Thus, their approach faced challenges due to missing data and small sample sizes [17]. This shows the need for better synthetic datasets to support AI research in this field.

The Relationship Between This Project and Current Research

This project builds on previous research by improving synthetic liver cirrhosis data using advanced AI methods. While studies like MedGAN have shown that synthetic medical data can be useful, they have not fully addressed the issue of data accuracy and diversity. The researcher will use reinforcement learning to enhance synthetic datasets, making them more realistic and useful for AI training. This will help bridge the gap between real-world medical data and synthetic data, making AI models more reliable for liver disease diagnosis.

How Well Established is the Area of Study?

The use of synthetic data in medicine is a growing field. Many researchers are now exploring how AI-generated data can support medical research without violating patient privacy. In 2020, Google's DeepMind used synthetic data to train AI models for predicting kidney disease [20]. The results showed that AI could learn from synthetic data and still make accurate predictions.

In the case of liver cirrhosis, research on synthetic data is still limited. Most studies rely on real patient records, which are hard to access [17]. Therefore, this research will be one of the first to focus on generating high-quality synthetic data specifically for liver cirrhosis. By improving synthetic data using reinforcement learning, this work will contribute to a new and developing area of study.

Techniques and Theories Applied in This Research

This project will use two main AI techniques:

1. **Generative Adversarial Networks (GANs)** – GANs are AI models that generate synthetic data by learning from real-world samples. They consist of two neural networks: a generator and a discriminator. The generator creates fake data, while the discriminator tries to distinguish between real and fake data. Over time, the generator improves until it produces realistic synthetic data. GANs have been used successfully in other medical research areas, such as diabetes prediction and heart disease diagnosis [16].
2. **Reinforcement Learning** – This technique allows AI models to improve their performance by learning from past mistakes. In this study, reinforcement learning will be used to refine the synthetic liver cirrhosis data. This means the AI model will adjust its

data generation process based on feedback, ensuring that the synthetic data becomes more accurate and diverse.

By combining these techniques, this project aims to create synthetic datasets that can train AI models effectively, even when real medical data are unavailable.

Relevance to Industry and Society

The results of this study could have significant applications beyond the university. AI models trained on high-quality synthetic data can help doctors diagnose liver cirrhosis earlier, leading to better patient outcomes. In Nigeria, where medical data are often scarce, synthetic data could support AI-driven research without the need for large real-world datasets [16].

Tech companies and hospitals could also benefit from this research. AI-based diagnostic tools are becoming more popular, and synthetic data can help improve these technologies. Companies developing AI medical software could use my findings to create better prediction models for liver cirrhosis and other diseases.

Challenges and Future Directions

There are some challenges in using synthetic data for medical research. One major issue is ensuring that synthetic data truly reflects real-world patient characteristics. If the AI model generates inaccurate data, the resulting research could be misleading [17]. To address this, the study will validate its synthetic data by comparing AI models trained on synthetic data with those trained on real data.

Another challenge is ethical concerns. While synthetic data avoids privacy risks, researchers must still be careful in how they use AI-generated medical records. If synthetic data are not properly validated, they could lead to incorrect medical decisions [17]. The study will follow strict validation steps to ensure that the synthetic data are useful for real-world applications.

In the future, this research could be extended to other diseases. The same AI techniques used in this study could be applied to generate synthetic data for conditions like diabetes, kidney disease, and cancer. By improving synthetic medical data, researchers can continue to develop AI-driven solutions for healthcare challenges, especially in low-resource countries like Nigeria.

Conclusion

Liver cirrhosis is a major health issue, and AI can help in its early detection and treatment. However, real-world medical data are often limited due to privacy rules and poor record-keeping. Synthetic data offer a promising solution, but current datasets lack diversity and realism. This study aims to improve synthetic liver cirrhosis data using GANs and reinforcement learning, making them more useful for AI research.

This research builds on previous studies but introduces new techniques to enhance data accuracy and diversity. The results could benefit researchers, hospitals, and tech companies working on AI-driven healthcare solutions. If successful, this study will contribute to better AI models for diagnosing liver cirrhosis and could be extended to other diseases in the future.

References

[1] "World Hepatitis Day-In Nigeria, an estimated 20 million people are chronically infected | WHO | Regional Office for Africa," *WHO | Regional Office for Africa*, Feb. 03, 2025. https://www.afro.who.int/news/world-hepatitis-day-nigeria-estimated-20-million-people-are-chronically-infected

[2] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLOS Digital Health*, vol. 2, no. 1, p. e0000082, Jan. 2023, doi: 10.1371/journal.pdig.0000082.

[3] V. Thambawita *et al.*, "DeepFake electrocardiograms: the key for open science for artificial intelligence in medicine," *medRxiv (Cold Spring Harbor Laboratory)*, Apr. 2021, doi: 10.1101/2021.04.27.21256189.

[4] K. Bhanot, M. Qi, J. S. Erickson, I. Guyon, and K. P. Bennett, "The problem of fairness in synthetic healthcare data," *Entropy*, vol. 23, no. 9, p. 1165, Sep. 2021, doi: 10.3390/e23091165.

[5] I. Diouf *et al.*, "An approach for generating realistic Australian synthetic healthcare data," *Studies in Health Technology and Informatics*, Jan. 2024, doi: 10.3233/shti231079.

[6] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," *Npj Digital Medicine*, vol. 3, no. 1, Nov. 2020, doi: 10.1038/s41746-020-00353-9.

[7] C. Alloza *et al.*, "A case for synthetic data in regulatory Decision‐Making in Europe," *Clinical Pharmacology & Therapeutics*, vol. 114, no. 4, pp. 795–801, Jul. 2023, doi: 10.1002/cpt.3001.

[8] S. Bing, A. Dittadi, S. Bauer, and P. Schwab, "Conditional generation of medical time series for extrapolation to underrepresented populations," *arXiv (Cornell University)*, Jan. 2022, doi: 10.48550/arxiv.2201.08186.

[9] S. Suh *et al.*, "Supervised Segmentation with Domain Adaptation for Small Sampled Orbital CT Images," *arXiv (Cornell University)*, Jan. 2021, doi: 10.48550/arxiv.2107.00418.

[10] G. Nikolentzos, M. Vazirgiannis, C. Xypolopoulos, M. Lingman, and E. G. Brandt, "Synthetic electronic health records generated with variational graph autoencoders," *Npj Digital Medicine*, vol. 6, no. 1, Apr. 2023, doi: 10.1038/s41746-023-00822-x.

[11] Y. Bhomia, S. K. Mishra, and P. Tiwari, "Innovations in generative adversarial networks (GANs) for synthetic data generation in medical imaging: A review," *The Pharma Innovation*, vol. 8, no. 4S, pp. 22–25, Jan. 2019, doi: 10.22271/tpi.2019.v8.i4sa.25254.

[12] C. Yan *et al.*, "A Multifaceted benchmarking of synthetic electronic health record generation models," *Nature Communications*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/s41467-022-35295-1.

[13] E. B. Tapper, J. B. Henderson, N. D. Parikh, G. N. Ioannou, and A. S. Lok, "Incidence of and risk factors for hepatic encephalopathy in a Population‐Based cohort of Americans with cirrhosis," *Hepatology Communications*, vol. 3, no. 11, pp. 1510–1519, Sep. 2019, doi: 10.1002/hep4.1425.

[14] K. E. Emam, L. Mosquera, E. Jonker, and H. Sood, “Evaluating the utility of synthetic COVID-19 case data,” *JAMIA Open*, vol. 4, no. 1, Jan. 2021, doi: 10.1093/jamiaopen/ooab012.

[15] P. Decharatanachart, R. Chaiteerakij, T. Tiyarattanachai, and S. Treeprasertsuk, “Application of artificial intelligence in chronic liver diseases: a systematic review and meta-analysis,” *BMC Gastroenterology*, vol. 21, no. 1, Jan. 2021, doi: 10.1186/s12876-020-01585-5.

[16] S. L. Popa *et al.*, “Diagnosis of liver fibrosis using Artificial Intelligence: A Systematic review,” *Medicina*, vol. 59, no. 5, p. 992, May 2023, doi: 10.3390/medicina59050992.

[17] M. Di Giammarco, A. Santone, M. Cesarelli, F. Martinelli, and F. Mercaldo, “Evaluating Deep Learning Resilience in Retinal Fundus Classification with Generative Adversarial Networks Generated Images,” *Electronics*, vol. 13, no. 13, p. 2631, Jul. 2024, doi: 10.3390/electronics13132631.

[18] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, “Generating Multi-label Discrete Patient Records using Generative Adversarial Networks,” arXiv (Cornell University), Jan. 2017, doi: 10.48550/arxiv.1703.06490.

[19] B. Beaulieu-Jones et al., “Trends and focus of Machine learning applications for health Research,” JAMA Network Open, vol. 2, no. 10, p. e1914051, Oct. 2019, doi: 10.1001/jamanetworkopen.2019.14051.

[20] D. J. Sexton and C. Judge, “Assessments of Generative AI as Clinical Decision Support Ought to be Incorporated into Randomised Controlled Trials of Electronic Alerts for Acute Kidney Injury,” Mayo Clinic Proceedings Digital Health, vol. 2, no. 4, pp. 606–610, Oct. 2024, doi: 10.1016/j.mcpdig.2024.09.004.

Student and First Supervisor Project Sign-off			
	Name	**Signature**	**Date**
STUDENT: I agree to complete this project:	Michael Vincent Udousoro	M.V.U.	14th February 2025
SUPERVISOR: I approve this project proposal:			
Supervisor Comments/Feedback			