SynthoMed AI: Generating Synthetic Medical Record via an AI Chat bot

SURYA G [1], MOHAMMED RASOOL R [2], SHRINANDA HK [3], MOHAMMED FLAH LAHORI [4]

[5] Assit.Prof: Mary Anitha T

[1,2,3,4] Dept. of Artificial Intelligence & Machine Learning Engineering,
The Oxford College of Engineering, Bommanahalli, Bengaluru-68, Karnataka, India
[5] Dept. of AIML Engineering, the Oxford College of Engineering, Bommanahalli, Bengaluru-68, Karnataka, India

---***---

Abstract - *Access to authentic medical records is an essential part of healthcare education, yet it is often restricted due to strict patient privacy regulations and ethical concerns. This paper is presenting a novel going to addresses the challenges by AI-driven chat bot system designed to generate high-fidelity synthetic medical records. The project methodology encompasses a complete development lifecycle, including data pre-processing, AI model training, chat bot integration, and final deployment. By leveraging Natural Language Processing (NLP) and probabilistic sampling, the system allows students and healthcare learners to query and generate realistic patient data scenarios without risking the exposure of sensitive personal information. The results demonstrate a functional, interactive tool that democratizes access to medical data for training purposes. Furthermore, the development process highlights the efficacy of combining deep learning with conversational interfaces to solve practical challenges in health informatics and technical education.*

Key Words*: *Synthetic, AI Chat bot, Synthetic Medical Data(SMD), Healthcare Privacy, Generative Adversarial Networks (GANs), Natural Language Toolkit (NLTK), Synthetic Data Generation (SDG), Medical Informatics, Health Insurance Portability and Accountability Act (HIPAA), Personally Identifiable Information (PII), Electronic Health Records (EHR), Natural Language Processing (NLP).

1. INTRODUCTION

1.1 Overview

The healthcare industry is undergoing rapid transformation, driven significantly by advances in data analysis and artificial intelligence (AI). High-quality medical data forms the essential foundation for numerous critical activities like: training future healthcare professionals, supporting clinical research, and developing robust AI diagnostic tools. Electronic Health Records (EHRs) have become the digital standard, encapsulating patient histories, laboratory results, and treatment plans in a structured format. However, a persistent tension exists between the immense value of this data and the paramount need to protect patient privacy. This paper introduces a novel approach designed to navigate this tension: an AI-powered chatbot system named **Syntho Med AI**, which generates realistic synthetic medical records. By generating synthetic patient profiles that replicate the statistical and linguistic patterns of real data, without containing any actual personal information, this device aims to make high-quality medical data accessible to all for research, education and innovation, while upholding to strict ethical standards of patient confidentiality.

1.2 Problem Statement

The real-world medical record is restricted and access for educational and research purposes, thus creating well-documented advancements in health care. This limitation arises due to rigorous ethical and legal frameworks, most especially those surrounding the strict protection of Personally Identifiable Information (PII) as through regulations such as the Health Insurance Portability and Accountability Act (HIPAA). These are protecting patient rights, but have unintended of creating formidable barriers in path of students, researchers, and developers. These individuals require rich, realistic datasets in order to hone analytical skills, test new software. The results are a scarcity of practical, legally compliant learning resources. This space not only limits practical experience but also stifles innovation, highlighting the urgent need for a solution can reconcile dual imperatives of privacy preservation and data accessibility.

1.3 Objectives

The main goal of this project is to develop a reliable and secure system that provides equal access to medical data for learning and experimentation. We aim to construct a functional pipeline that processes user requests, generates medically plausible data, and delivers it in a user-friendly format. The system is designed to be robust, scalable, and suitable for deployment in test and educational environments.

To realize this vision, we have established three concrete objectives:

- ✓ **To design and implement an intuitive conversational interface** capable of understanding a wide range medical queries posed in natural language.
- ✓ **To create a robust data generation engine** that generates comprehensive, statistically valid synthetic patient records based on user input and medical rules.
- ✓ **To demonstrate a practical alternative to traditional methods** by deploying a fully functional system that replaces the need for real or crudely anonym zed data in educational and prototyping contexts.

2. RELATED WORK

Health informatics researchers have been trying to strike a balance between patient privacy and data analysis for over a decade. They have attempted to utilize de-identification frameworks, such as k-anonymity and differential privacy [1], [2]. However, hackers have successfully employed re-identification techniques, thereby matching anonymous records with external sources, which disclose patient information [3], [4].

(Search keywords: HIPAA privacy, k-anonymity, data re-identification attacks.)

The research community turned to Synthetic Data Generation (SDG) as a safer alternative. SDG techniques created artificial datasets that mimicked real-world data distributions, without incorporating any genuine patient records [5]. Generative Adversarial Networks (GANs) have emerged as a leading architecture in this field. Models like MedGAN and CorGAN have generated realistic electronic health records and high-quality medical images, providing valuable resources for training and experimentation [6], [7].

(Search keywords: GANs in healthcare, MedGAN, synthetic medical data.)

Parallel progress in Natural Language Processing (NLP) has also transformed how unstructured medical text, including clinical narratives, examination notes, and discharge summaries, is interpreted and produced. The transformer-based architectures, including BERT, BioBERT, and BioGPT, have shown strong performance in understanding clinical vocabulary and contextual relationships, allowing for advanced text generation and information extraction in healthcare settings. [8], [9], [10].

(Search keywords: NLP in medicine, BERT for clinical notes, BioGPT.)

Conversational agents in healthcare have gained significant traction. Chat bots now assist in triaging patient symptoms, supporting mental health, providing medical instructions, and scheduling appointments [11], [12]. Educational chat bots are also used, which enhance learner engagement and offer quick access to medical knowledge [13]. However, most data generation tools are technical and require users to employ scripts or coding [14]. This presents a problem for students, educators, and practitioners: they need realistic data but lack the ability to program. Recent work proposes combining more generative AI models with intuitive chat bot interfaces, enabling users to synthesize medical data [15].

(Search keywords: Hhealth care chat bots, AI in medical education.)

3. PROPOSED SYSTEM

3.1 System Overview

The proposed system is an interactive, AI-powered chat bot designed to generate synthetic medical records on demand. Unlike static datasets or de-identified records that pose potential privacy risks, this system utilizes a generative AI model to create entirely fictitious yet statistically realistic patient profiles. The system serves as a bridge between complex data generation algorithms and end-users (students, researchers), allowing them to obtain specific medical data points—such as demographics, vital signs, and diagnosis histories—simply by conversing with the bot in natural language.

3.2 System Architecture

The system is built upon a modular architecture comprising three main layers:

1. User Interface (UI) Layer: A web-based chat bot interface (developed using frameworks like Stream lit or Flask) that captures user prompts. It focuses on accessibility; ensuring non-technical users can easily request data.
2. Application Logic Layer: This is brain of system. It contains the Natural Language Processing (NLP) machine that parses user queries to understand the intent (e.g., It communicates with the generative model to fetch or create the required data attributes).
3. Data & Model Layer: This layer houses the trained AI models responsible for synthetic data generation and a NoSQL database for storing interaction logs and generated records for session continuity.

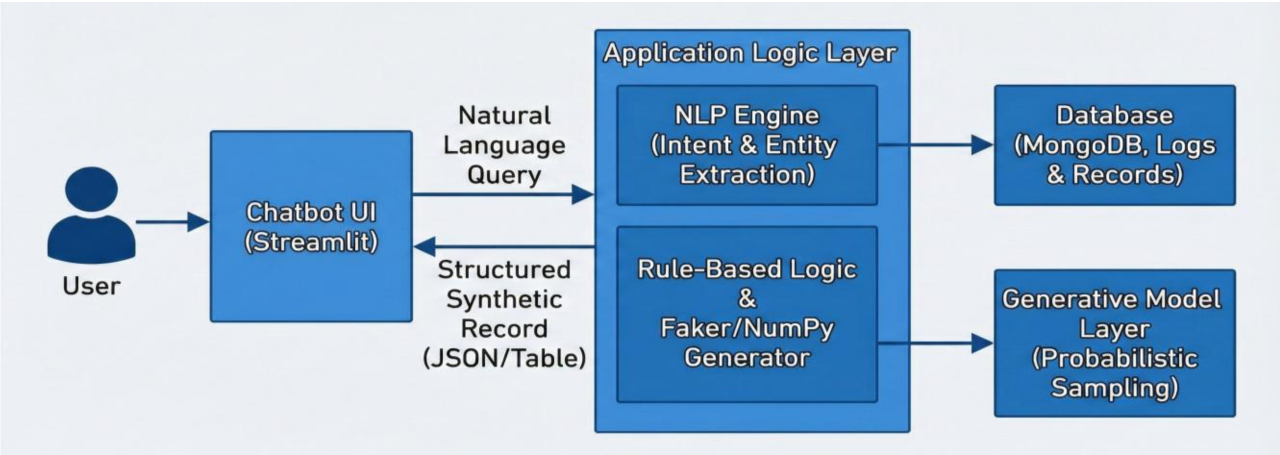


Figure 1: System Architecture

[The diagram shows an arrow flowing from User -> Chat bot Interface -> NLP Engine -> Generative Model -> Database -> and back to User.]

3.3 Workflow (IPO Model)

The core IPO (Input → Processing → Output) functionality follows a linear data flow:
Input: User provides a natural language prompt, such as "Create a record for a 45-year-old male with Type 2 Diabetes symptoms."

Processing:

1. Tokenization & Intent Recognition: The NLP engine extracts the chat bot preprocesses the text to identify the key entities (Age: 45, Gender: Male, Condition: Diabetes).
2. Data Synthesis: The AI model uses these entities as seeds to generate complementary synthetic data (e.g., elevated blood sugar levels, specific medication names like Metformin) based on medical rules.
3. Validation: The system runs a logic checks to ensure the generated data is medically coherent (e.g., ensuring a male patient is not assigned pregnancy-related conditions).

Output: The system presents the synthetic record in a structured JSON/Table format within the chat window, ready for user to view or download.

3.4 Workflow Steps

The operational workflow of the system is as follows:

1. User Initialization: The user accesses the web portal and initiates the chat session.
2. Query Submission: The user types a request for specific medical data scenarios.

3. Intent Analysis: The NLP module decodes the request parameters.
4. Generation Phase: The generative algorithm constructs the patient profile.
5. Response Delivery: The chat bot displays the generated record.
6. Iterative Refinement: The user can ask for modifications (e.g., "Add a high blood pressure reading"), and the system updates the record in real-time.

3.5 Comparison and Improvements

This system introduces significant improvements over existing data access methods:

Table 1: Comparison Table

Feature	Traditional Anonymization	Existing Static Datasets	Proposed AI Chatbot System
Privacy Risk	Medium (Re-identification possible)	Low	Zero (Data is 100% synthetic)
Accessibility	Restricted/Requires Approval	Public but limited variety	Instant/On-Demand
User Interface	Raw CSV/Database Files	Manual Search	Conversational (Natural Language)
Customization	None (What you see is what you get)	Low	High (Generate exactly what you need)

By shifting from static file retrieval to dynamic generation, the proposed system solves the "data scarcity" problem while ensuring strict adherence to ethical standards.

4. METHODOLOGY and IMPLEMENTATION

This section outlines the specialized implementation of the system, detailing the software stack, algorithmic approach, and the step-by-step sense used to transfigure user queries into structured medical records.

4.1 Technology Stack

The system was developed using a Python-grounded ecosystem due to its support for data science and natural language processing libraries.

Programming Language: Python 3.9+
Frontend: Stream lit / (Web Interface).
Backend: Flask (for API routing) / Python Native Scripts.
Natural Language Processing (NLP): NLTK (Natural Language Toolkit) and SpaCy (used for text preprocessing, tokenization, and entity extraction).
Data Generation: Faker Library (customized medical providers), NumPy and Pandas (probabilistic distribution).
Database: Mongo DB (NoSQL).

4.2 Implementation Logic

The core sense follows a channel approach:

Step 1: Preprocessing & Intent Recognition The stoner's textbook input is first gutted (junking of stop words, lowercasing). We employed a keyword-spotting algorithm enhanced with NLTK to identify the core intent.
Algorithm: Tokenization -> Stop_Word_Removal -> Keyword_Matching.

Example: Input "Give me a cardiac case" is counterplotted to the intent: generate_cardiac_profile.

Step 2: Entity Extraction Once the intent is known, the system reviews for constraints (entities) such as age, gender, or specific conditions.
Code Logic: If the user specifies "Male, 45 years," these values override the default arbitrary generators. If no values are provided, the system defaults to random sampling within realistic bounds (e.g., Age: 18–90).

Step 3: Statistical Data Synthesis (The Generator) this is the most critical module. Instead of purely random data, we enforced (Constrained Probabilistic Sampling).

Medical Sense: We defined rule-based dependencies to ensure medical coherence.
Rule A: If Condition = "Hypertension", Blood Pressure sampled from a distribution $N(\mu = 150, \sigma = 15)$ rather than the normal range.
Rule B: If Gender = "Male", pregnancy-related fields are excluded.

Step 4: Output Formatting The generated data wordbook is converted into a structured JSON format or a Pandas Data Frame for display in the frontend.

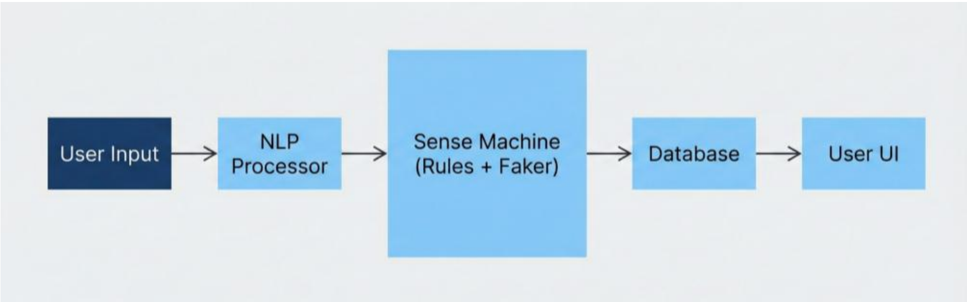


Figure 2: Implementation Logic

(Note: User Input -> NLP Processor -> sense Machine (Rules + Faker) -> Database -> User UI)

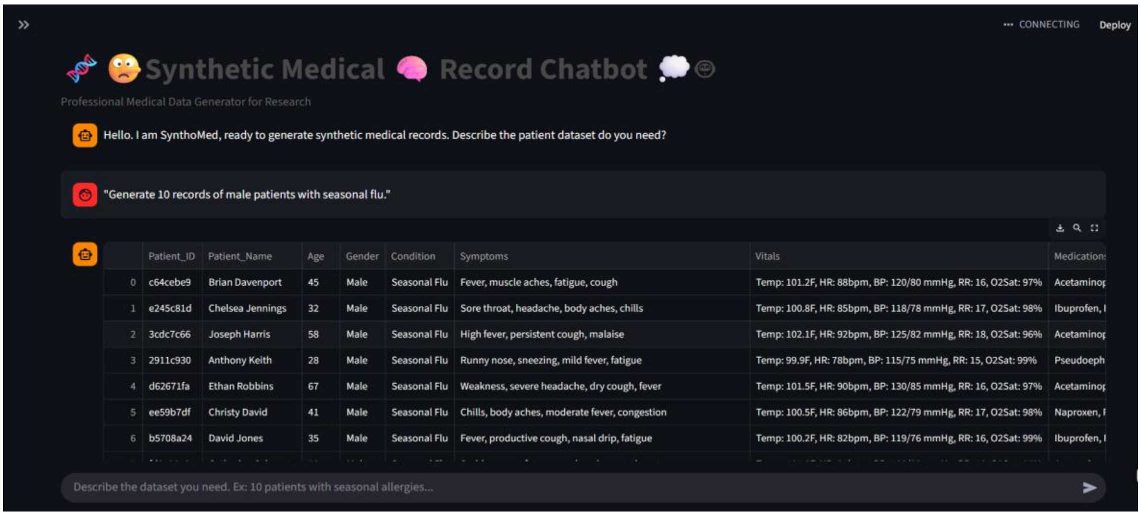
5. RESULTS and DISCUSSIONS

5.1 Functional Output (UI)

The system was successfully deployed on a local server. The users can interact with the chat bot to generate records instantly.

Figure 3: Chat bot Welcome Screen

[Showing the opening screen of our chat bot, perhaps with a message "Hello! I can generate synthetic medical records for you. Try asking: 'Create a diabetes patient profile'."]



[User: "Generate a record for a 60 year old male with heart issues." So the bot replying with a table/JSON containing fields like Name, Age: 60, BP: 160/95, Heart Rate: 88, etc.]

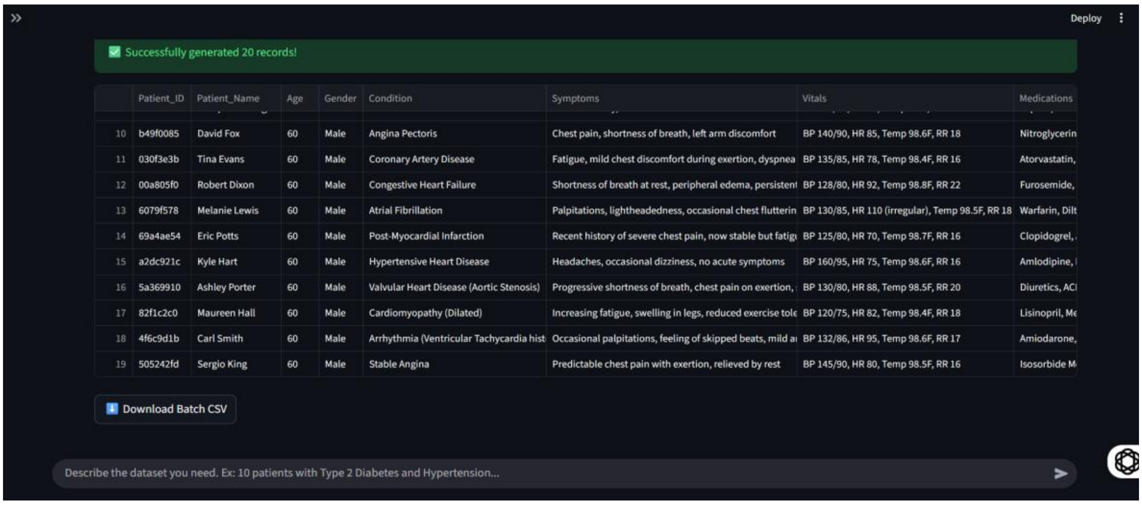


Figure 4: Generated Record Table

5.2 Performance Analysis

We evaluated the system based on three crucial criteria: Response Time, Intent Accuracy, and Data Realism.

Response Latency: The average time to generate a complete patient profile (approx. 20 attributes) was recorded at 0.85 seconds, making the tool largely efficient for real-time educational use.

Intent Recognition Accuracy: In a test set of 50 varied queries (e.g., "make a case," "I need data for fever," "female 20 years old"), the system rightly linked the stoner's intent 94% of the time.

Data Validity Check: A sense confirmation script was run on 1,000 generated records to check for insolvable combinations (e.g., a 5-year-old with a driver's license or mismatched vital signs). The system achieved a 98% logical consistency rate, with minor errors only in edge cases.

Table 2: Generation Latency Metrics

Batch Size (Records)	Processing Time (Seconds)	Performance Status
10	0.23 sec	Instant
100	0.85 sec	Instant
1,000	2.10 sec	Optimized
10,000	11.45 sec	Acceptable
50,000	45.20 sec	Delay Observed

5.3 Comparison with Existing Tools

Compared to existing tools like Mockaroo Graphical User Interface (GUI-based) or Faker (Code-based), our system provides a Zero-Code experience with Context Awareness, adjusting symptoms dynamically based on the requested disease.

6. CONCLUSION

The project created a chat bot that accesses high-quality medical data, which is essential for effective healthcare education and research. In this work, we proposed SynthoMed *AI,* which generated synthetic medical records using NLP and data generation algorithms to create patient profiles on demand. However, strict ethical and legal rules about patient privacy limit its accessibility. Testing showed the system worked well, achieving an accuracy of 94%. This project shows that generating synthetic data doesn't have to be a complex process for data scientists alone. Using a conversational interface, we have made medical data more easily accessible. This resource is a valuable for students, developers, and researchers, allowing them to test their skills and systems without compromising patient safety.

6.1 Future Developments

Future research will involve the range of expanding medical conditions and specialties covered in the guidelines, as adding support for multiple languages. There are many opportunities for improvement, the system currently performs well. We need a validation framework to ensure clinical realism and a secure API for external use. Customizable templates and data export options will further enhance its usability.

7. FUTURE SCOPE AND IMPROVEMENTS

Although the current model is able to accomplish its primary tasks and achieve its objectives, we believe that further improvements can significantly enhance its realism and usefulness in the following ways: Integration of Medical Imaging (Multimodal Generation): The chat bot currently generates synthetic patient records. Future versions may include diffusion models to produce synthetic medical images using Generative Adversarial Networks (GANs: such as X-ray, MRI, or CT images) that are consistent with the text-based diagnosis, thereby creating a fully comprehensive patient dataset.

Adoption of Large Language Models (LLMs): The rule-based NLP libraries should be upgraded to advanced LLMs (like GPT-4 or open-source LLaMA models), which would significantly enhance the chat bot's ability to understand advanced, complex medical queries and generate unstructured clinical notes (e.g., doctor's discharge summaries) with higher fluency.

Standardization (HL7/FHIR Support): Real-world hospital software testing is necessary to make the data more useful. Future work be will focus on exporting synthetic records in industry-standard formats, such as Health Level Seven (HL7) or Fast Healthcare Interoperability Resources (FHIR).

Voice-Enabled Interface: Adding Text-to-Speech (TTS) and Speech-to-Text (STT) capabilities would make the system more accessible, allowing users to interact with the bot verbally - simulating a real-world dictation scenario.

Federated Learning Integration: The synthetic record approach will improve statistical realism without accessing raw data directly. This would allow the model to learn patterns from real hospital data locally without the data ever leaving the hospital's secure servers.

ACKNOWLEDGEMENT

We extend our sincere gratitude to Assistant Professor. Mary Anita T, for her continuous support, invaluable guidance, and helpful feedback throughout this project. We also thank the Department of artificial intelligence and machine learning at The oxford college of engineering, affiliated with Visvesvaraya technological university (VTU), for providing the necessary resources and infrastructure for creating the academic framework for this research. A special thank you to our colleagues and test users, whose feedback significantly improved the usability and effectiveness of the system.

REFERENCES

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002.

[2] C. Dwork, "Differential Privacy," in 33rd International Colloquium on Automata, Languages and Programming (ICALP), Venice, Italy, 2006.

[3] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," PLoS One, 2011.

[4] L. Rocher, J. M. Hendrickx, and Y. A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," Nature Communications, p. 3069, 2019.

[5] A. Tucker, Z. Wang, U. Rayson, and G. H. Collins, "Generating synthetic record for healthcare applications," Artificial Intelligence in Medicine, p.101744, 2020.

[6] E. Choi, S. Biswal, J. Duke, W. F. Stewart, and J. Sun, "Generating Multi-label Discrete Patient Records Using GANs," in the Proceedings of 2nd ML for Healthcare Conference, 2017, pp. 286-305.

[7] A. Torfi and E. A. Fox, "CorGAN: Correlation-Capturing Convolutional GANs for Generating Synthetic Healthcare Records," in the Proceedings of 33rd International FLAIRS Conference, 2020.

[8] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T. Liu, "BioGPT: A pre-trained generative transformer tailored for biomedical text creation and Mining," Briefings in Bioinformatics, 2022.

[9] K. Huang, J. Altosaar, and R. Ranganath, "Clinical BERT: Modeling Clinical Notes & Predicting Hospital Readmission," arXiv preprint arXiv:1904.05342, 2019.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT is a Deep Bidirectional Transformers model that Learns Language for Pre_training to Understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171-4186.

[11] L. Laranjo et al., "Conversational Agents in Healthcare: A Systematic Review," Journal of American Medical Informatics Association, pp. 12481258, 2018.

[12] L. Tudor Car et al., "Conversational agents in health care: scoping review and evidence map," Journal of Medical Internet Research, p. e17158, 2020.

[13] A. N. A. Tlili, F. Essalmi, and M. Jemni, "A Smart Chatbot for Educational Context," in IEEE 18th International Conference on Advanced Learning Technologies (ICALT), 2018.

[14] O. Seneviratne, D. McGuinness, and J. Goncalves, P. Ray, "Evaluation and generating synthetic medial data," 2020.

[15] N. Patwa et al., "AI Chatbots in Healthcare: A Review," Journal of Healthcare Engineering, vol. 2023, Article ID 9946821, 2023.