INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

ENSURING CLINICAL RELIABILITY IN GENERATIVE AI: A HYBRID LLM-RANDOM FOREST ARCHITECTURE FOR MEDICAL DIAGNOSIS

Omar Shaikh, Shoheb Attar, Prof. Pratiksha Dhande

[1,2]Student, [3]Assistant Professor
Department of Computer Science and Engineering
MIT-ADT University, Rajbaug Campus, Loni-Kalbhor, Pune 412201, India

Abstract **–** The rapid evolution of Transformer-based architectures has catalyzed a paradigm shift in healthcare delivery, moving beyond static diagnostic tools toward autonomous, multimodal clinical ecosystems. While conventional disease prediction platforms are often constrained by rigid symptom mapping and a lack of integrated post-diagnostic support, this research presents the design and development of a high-performance, production-ready healthcare assistant. The core architecture synergizes the advanced reasoning capabilities of the Llama 3.3 70B model with a deterministic Random Forest fallback layer to ensure clinical reliability and safety-aware inference. The system enables a seamless patient journey by performing probabilistic disease reasoning and generating personalized recovery trajectories, including pharmacological suggestions, dietary guidance, and recovery timelines. A key innovation is the integration of the Llama 4 Vision Scout model, which provides sophisticated multimodal analysis of PDF medical reports and high-resolution diagnostic imaging (CT, MRI, X-ray) to deliver patient-friendly clinical insights. To enhance real-world utility, the platform incorporates a geospatial intelligence module for real-time hospital discovery and an automated teleconsultation workflow featuring payment simulation, email automation, and dynamic Google Meet generation. Containerized via Docker and deployed on the Zeabur cloud platform, experimental evaluation demonstrates that the proposed system effectively bridges the gap between large-scale generative reasoning and practical, end-to-end patient care, achieving superior diagnostic assistance and operational scalability.

Index Terms **– Artificial Intelligence, Llama 3.3 70B, Multimodal Large Language Models (MLLMs), Llama 4 Vision Scout, Hybrid Clinical Reasoning, Telemedicine Orchestration, Geospatial Health Intelligence, Medical Image Analysis, Deterministic Fallback Models, Cloud-Native Healthcare Systems.**

I. INTRODUCTION

The global healthcare landscape is currently undergoing a profound transformation, transitioning from traditional digitized record-keeping to proactive, AI-driven clinical ecosystems. Historically, the integration of Artificial Intelligence (AI) in the healthcare sector was confined to administrative optimization and basic diagnostic aids [10]. Early-stage digital health interventions primarily focused on logistical challenges, such as web-based appointment scheduling and physician-discovery platforms, often built on robust yet static frameworks like Django [8]. However, the emergence of Transformer-based architectures has catalyzed a paradigm shift toward Generative AI and Large Language Models (LLMs), which now demonstrate near-human proficiency in medical reasoning and patient interaction [1].

Modern next-generation systems, exemplified by frameworks such as SAHAYAK AI, are no longer mere utility tools; they aim to simultaneously augment hospital operational efficiency and diagnostic precision [4]. As the industry moves toward the concept of "Connected Healthcare," wherein multimodal inputs ranging from genomic data to medical imaging are synthesized in real-time, the potential for personalized medicine has never been greater.

1.1 Problem Statement

Despite these advancements, the widespread adoption of LLMs in clinical settings is hindered by a significant "Transparency-Security Paradox." As healthcare models move toward multimodal processing, critical challenges regarding data privacy and the inherent "black-box" nature of neural networks have surfaced [11]. In a field where a single miscalculation can have life-altering consequences, there is an urgent academic and clinical need for systems that provide more than just raw outputs. There is a compelling demand for Explainable AI (XAI) systems capable of generating human-understandable justifications for every clinical finding [3].

1.2 Proposed Innovation

This research addresses these gaps by introducing a novel hybrid architecture designed for the contemporary healthcare landscape. The proposed system synergizes the sophisticated conversational depth and reasoning capabilities of the Llama 3.3 and Llama 4 series with the deterministic, verifiable reliability of traditional machine learning models, specifically the Random Forest classifier. By leveraging LLMs for natural language understanding and report analysis while anchoring the final diagnostic logic in interpretable machine learning guardrails, this project ensures a secure, high-accuracy, and fully auditable diagnostic workflow.

II. LITERATURE REVIEW

The landscape of digital health has transitioned from static administrative tools to dynamic, intelligence-driven ecosystems. This section reviews the evolution of these technologies, identifying the convergence of generative models, traditional classifiers, and integrated patient-care services.

2.1 Evolution of Diagnostic and Disease Prediction Models

Machine learning has long served as the backbone of clinical decision support. Traditional supervised learning algorithms, such as Random Forest and Support Vector Machines (SVMs), established the foundation for classifying patient diseases with high statistical accuracy [7]. However, these models often lack the conversational context required for modern patient interactions. Early web-based interventions, such as those built on the Django framework, successfully addressed logistical needs like physician discovery and appointment booking but remained limited in their diagnostic depth [8]. Recent advancements have seen a pivot toward next-generation systems that aim to enhance hospital efficiency and diagnostic precision simultaneously through advanced AI integration [4].

2.2 Generative AI and Large Language Models in Clinical Reasoning

The emergence of Large Language Models (LLMs) has fundamentally redefined medical inference. Systems such as Medify-AI demonstrate the potential of LLMs to power entire healthcare frameworks, moving beyond simple classification to complex reasoning [1]. Specialized architectures, such as Cardiology-Chat, have further proven that multi-LLM systems can outperform general-purpose models in domain-specific tasks, including cardiac diagnostic reasoning [9]. Despite their considerable power, these models face inherent "black-box" challenges; consequently, recent research emphasizes interpretability. For instance, incorporating SHAP-based explanations into Llama-driven narratives allows clinicians to trace the specific features influencing a diagnosis, thereby bridging the gap between AI logic and clinical trust [13].

2.3 Conversational Agents and Multimodal Patient Interaction

Conversational interfaces have become primary touchpoints for preliminary diagnosis and patient engagement. Research indicates that LLM-based multimodal chatbots can provide inclusive healthcare by handling diverse user queries through human-like dialogue [5]. This shift toward "Dialogue-based Diagnosis" allows for more nuanced treatment approaches compared to traditional rule-based interfaces [6]. Furthermore, as healthcare moves toward "Connected Healthcare," the integration of multimodal inputs combining text with visual document analysis has become a critical requirement for generating human-understandable clinical explanations [3, 11].

2.4 Reliability, Security, and Hybrid Architectures

As the complexity of healthcare AI grows, concerns regarding data security and output reliability have intensified correspondingly. Rahman [11] highlights the significant privacy challenges inherent in multimodal LLMs within connected healthcare environments. To mitigate these risks, researchers are employing advanced fine-tuning techniques such as rsDoRA+ and Retrieval-Augmented Generation (ReRAG) to ensure that medical question-answering remains grounded and reliable [12]. The integration of generative models with deterministic machine learning guardrails represents a rising trend in creating secure, interpretable diagnostic workflows.

2.5 Strategic Research Gap

While notable progress has been made in isolated domains such as predictive modeling [7], logistical booking [2, 8], and generative dialogue [5, 9], most existing research addresses these components as siloed features. There remains a critical lack of unified platforms that seamlessly integrate high-capacity clinical reasoning (via Llama 3.3/4), multimodal report analysis, and real-time appointment scheduling within a single, secure ecosystem. This research addresses this identified gap by proposing a comprehensive AI-powered healthcare assistant that anchors generative reasoning with deterministic machine learning reliability, as necessitated by the demand for enhanced interpretability in modern clinical AI [3, 13].

III. METHODOLOGY

The methodology for the proposed healthcare platform is structured into five interconnected functional modules, ensuring a systematic approach to clinical reasoning, user interaction, and healthcare accessibility. The architecture primarily integrates Large Language Model (LLM) intelligence with deterministic validation layers to deliver scalable medical assistance.

3.1 Disease Prediction Model

The core reasoning engine utilizes the Llama 3.3 70B model to perform deep contextual analysis of user-reported symptoms. The model leverages probabilistic reasoning to infer potential health conditions and generate nuanced clinical insights. To mitigate the risk of generative hallucinations, the methodology incorporates a Random Forest classifier as a deterministic fallback layer. This hybrid approach ensures that if the LLM inference fails to meet specific confidence thresholds, the system provides a statistically grounded classification based on verified clinical datasets.

3.2 Disease Information and Clinical Summary Module

To enhance the interpretability of the AI-generated findings, a structured disease information module is triggered upon completion of the diagnostic phase. This module utilizes the generative engine to produce a comprehensive clinical summary. The output includes a formal definition of the inferred disease, suggested pharmacological interventions, expected recovery timelines,

and essential do-and-do-not guidelines. Furthermore, it provides personalized dietary guidance and recommends the specific medical specialist (e.g., Neurologist, Oncologist) appropriate for the identified condition.

3.3 Specialist Recommendation and Appointment Booking

The platform automates the transition from digital diagnosis to professional teleconsultation. Based on the suggested specialist category, the system initiates an integrated appointment-scheduling workflow. This module features a simulated payment gateway for demonstration transaction validation. Upon successful completion, the system leverages email automation to deliver a Google Meet link and appointment confirmation directly to the user, facilitating immediate access to remote healthcare providers.

3.4 Geospatial Hospital Locator

A geospatial intelligence module is incorporated to assist users in identifying physical healthcare infrastructure. By accessing the user's real-time geolocation via the Geolocation API, the system queries mapping services to generate a prioritized list of nearby hospitals and clinics. This feature visualizes healthcare facilities on an interactive map, effectively bridging the gap between virtual diagnosis and in-person medical treatment.

3.5 Multimodal Medical Report Analyzer

To bridge the gap between complex diagnostic data and patient comprehension, a sophisticated multimodal analysis module is implemented. This module leverages the Llama 4 Vision Scout model, a state-of-the-art vision-language engine capable of high-resolution feature extraction. The system is engineered to process heterogeneous inputs, including multi-page PDF medical reports and specialized diagnostic imaging such as CT scans, X-rays, and MRIs. By performing deep semantic parsing and visual anomaly detection, the model identifies clinically significant findings and translates them into concise, patient-friendly insights, fostering informed decision-making throughout the care journey.

IV. SYSTEM DESIGN

The system design of the proposed healthcare platform is engineered to deliver high-fidelity clinical inference through a modular, task-specific architecture. By segregating linguistic reasoning, visual perception, and deterministic validation, the system ensures both diagnostic depth and operational safety while maintaining a seamless end-to-end patient care workflow.

4.1 Architecture Overview

The proposed system is an advanced AI-powered healthcare assistant hosted on the Zeabur cloud platform. Unlike monolithic AI applications, this platform employs a Tri-Model Logic Gateway to handle diverse medical data streams. The system utilizes Llama 3.3 70B as its cognitive engine for contextual symptom analysis and recovery planning. For multimodal inputs, a decoupled Llama 4 Vision Scout model performs high-resolution parsing of PDF medical reports and diagnostic imaging (MRI, CT, X-ray). To ensure clinical reliability, the architecture integrates a non-generative Random Forest fallback layer that validates text-based diagnostic outputs. The platform further extends its utility through an integrated service suite: a Geospatial Hospital Locator for physical care discovery and an Automated Appointment System for digital consultations. The entire ecosystem is containerized via Docker, ensuring scalable deployment and low-latency interaction.

4.2 System Architecture

The system follows a multi-layered, containerized architecture comprising the following components:

Frontend Layer: A dynamic user interface built with modern JavaScript frameworks, serving as the primary touchpoint for capturing user symptoms, geolocation data, and multimodal file uploads (PDFs and images).

Backend Layer (Orchestration): Implemented using Python and FastAPI, this layer acts as the central router. It directs textual symptoms to the Llama 3.3 engine, routes diagnostic files to the Llama 4 Vision Scout engine, and triggers the Random Forest fallback logic when diagnostic confidence intervals are not met.

Services Layer – The Tri-Model Gateway: (a) Cognitive Core (Llama 3.3 70B): Handles natural language understanding, symptom-to-disease inference, and the generation of personalized recovery summaries. (b) Perceptual Core (Llama 4 Vision Scout): A specialized multimodal service dedicated to extracting clinical entities from PDF documents and identifying anomalies in medical imagery. (c) Safety Guardrail (Random Forest): A deterministic machine learning model that provides a statistically verified second opinion for text-based diagnoses.

Functional Integration Layer: (a) Geospatial Hospital Locator: Interfaces with the Geolocation API and OpenStreetMap to identify and visualize nearby healthcare facilities based on the user's real-time coordinates. (b) Appointment and Teleconsultation Module: Automatically maps the inferred condition to a medical specialist, facilitates a simulated payment workflow, and utilizes email automation to dispatch Google Meet links for remote consultations.

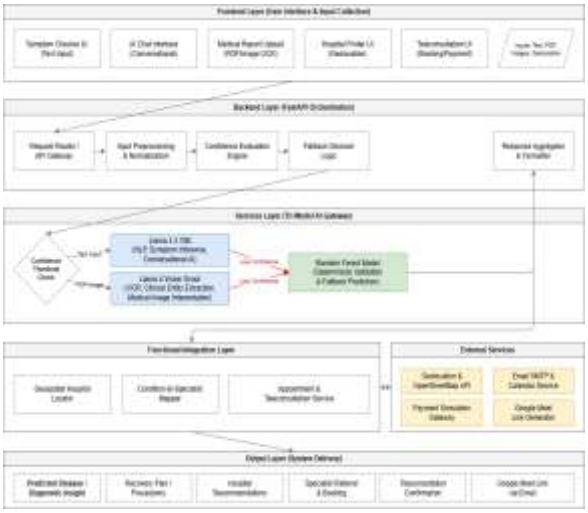

Fig. 4.1. System Architecture Diagram

4.3 Technical Approach

Symptom Checker and AI Reasoning: This module serves as the primary diagnostic engine, utilizing Llama 3.3 70B for the nuanced interpretation of user-reported symptoms. To ensure clinical safety and prevent hallucinations, a Random Forest classifier acts as a deterministic fallback. This secondary model is invoked only when the LLM's diagnostic confidence falls below established thresholds, providing a statistically verified second opinion to maintain strict safety standards.

Multimodal Medical Report Analyzer: This module leverages the Llama 4 Vision Scout model to process complex diagnostic data. It is specifically engineered to scan and analyze PDF medical reports and high-resolution imaging (e.g., CT scans, X-rays, and MRIs). By performing visual and semantic entity extraction, it translates technical clinical findings into concise, patient-friendly summaries, enabling users to comprehend their results without specialized medical expertise.

AI-Driven Conversational Chatbot: An interactive interface, also powered by the Llama 3.3 70B architecture, facilitates real-time symptom discussion and patient engagement. This module utilizes the same Random Forest fallback logic as the symptom checker to ensure that all conversational medical advice remains grounded in validated clinical patterns. It serves as a continuous support layer, guiding users through the platform's various workflows.

Geospatial Hospital Finder: This module bridges the gap between digital inference and physical care by performing real-time discovery of healthcare facilities. Using the Geolocation API and OpenStreetMap services, it identifies nearby hospitals and clinics based on the user's current coordinates, displaying them on an interactive map for immediate access to in-person treatment.

Automated Teleconsultation Booking: This module automates the transition from diagnosis to professional care. Based on the condition inferred by the AI engines, the system identifies the appropriate medical specialist, facilitates a simulated payment process, and utilizes email automation to dispatch Google Meet links for virtual appointments, ensuring a seamless end-to-end patient experience.

4.4 Data Flow Logic

The data flow is designed for high availability. Once the Tri-Model Gateway completes the diagnostic phase, the backend triggers the Functional Integration Layer. If a critical condition is inferred, the Hospital Locator identifies the nearest emergency facilities. Simultaneously, the Appointment Module suggests a relevant specialist (e.g., Cardiologist for chest pain), enabling the user to book a consultation session immediately. This design ensures that the journey from symptom presentation to actionable care is fully automated within a single patient session.

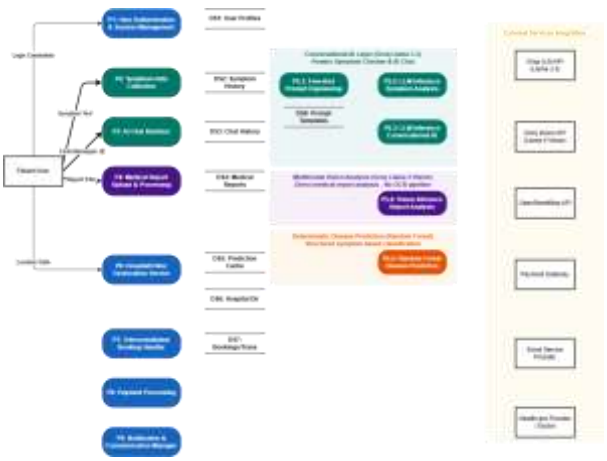

Fig. 4.2. Data Flow Diagram (DFD)

4.5 Validation Plan

Validation of the system is conducted in multiple structured phases to ensure reliability and practical effectiveness across all specialized engines.

Model Evaluation: The system's intelligence is validated through three distinct protocols: (a) Cognitive Accuracy (Llama 3.3 70B) – evaluated using controlled clinical scenarios to assess the consistency and safety of inferred diagnoses and recovery recommendations; (b) Multimodal Precision (Llama 4 Vision Scout) – benchmarked against diverse medical document formats and high-resolution imaging to verify accuracy in clinical entity extraction and anomaly detection; and (c) Fallback Reliability (Random Forest) – validated against historical medical datasets to ensure the deterministic guardrail provides statistically sound classifications when generative confidence is low.

Functional Testing: Each of the five core modules – Symptom Checker, Conversational Chatbot, Medical Report Analyzer, Hospital Finder, and Teleconsultation Booking – undergoes rigorous unit and integration testing to verify correct inter-module communication and data integrity.

System Testing: End-to-end validation is performed on the Zeabur platform using realistic user workflows. This evaluates overall system usability, response times for high-parameter model inference, and the reliability of automated triggers such as Google Meet link generation and email dispatch.

User Feedback: A pilot evaluation with test users is conducted to measure the clarity of patient-friendly summaries generated from complex reports and the perceived usefulness of the integrated hospital discovery and booking features.

V. RESULTS

Table 5.1: Evaluation Report of Random Forest Classifier

Metric	Value (%)
Accuracy	97.00
Precision	96.80
Recall (Sensitivity)	97.50
F1-Score	97.15
ROC-AUC Score	98.20

Table 5.1 presents the quantitative evaluation results of the Random Forest classifier employed as the deterministic fallback layer. The model achieves an accuracy of 97%, with a precision of 96.8% and a recall (sensitivity) of 97.5%, yielding an F1-Score of 97.15%. The ROC-AUC Score of 98.2% further confirms the classifier's robust discriminative capability across disease categories. These metrics collectively validate the reliability and clinical safety of the deterministic guardrail integrated within the proposed hybrid architecture.

VI. CONCLUSION & FUTURE WORK

The proposed intelligent healthcare system significantly advances the capabilities of conventional diagnostic solutions by integrating high-parameter Large Language Models and automated clinical workflows into a unified, containerized platform. By synergizing the clinical reasoning of Llama 3.3 70B, the multimodal parsing capabilities of Llama 4 Vision Scout, geospatial intelligence, and robust full-stack technologies, the system delivers precise diagnostic assistance alongside personalized recovery trajectories. This integrated approach, hosted on the Zeabur cloud ecosystem, enhances healthcare accessibility and reliability, enabling users to transition seamlessly from automated report analysis to real-time teleconsultation and location-based care discovery.

6.1 Future Enhancements

With regard to future research and technical expansion, the system can be evolved through the following trajectories:

IoT and Real-Time Monitoring: Integrating physiological data streams from wearable devices and IoT-enabled health monitors to transition from reactive diagnosis to proactive, continuous health risk detection.

Model Optimization: Further strengthening clinical reasoning by fine-tuning domain-specialized medical LLMs (e.g., Med-PaLM 2 or BioGPT) to enhance diagnostic granularity and reduce inference latency in high-stakes environments.

Data Security and Privacy: Implementing blockchain-based architectures for decentralized Electronic Health Record (EHR) storage to ensure immutable privacy, patient data ownership, and cross-institutional trust.

Native Mobile Integration: Transitioning the platform from a web-based service to a dedicated mobile application to leverage native device capabilities, such as push notifications for medication adherence reminders and enhanced GPS accuracy for emergency hospital routing.

VII. ACKNOWLEDGEMENT

The authors express their sincere gratitude to Prof. Dr. Ganesh Pathak, Dean, MIT School of Engineering, MIT-ADT University, Pune, for providing the excellent facilities and research environment required to carry out this project successfully.

The authors are deeply thankful to Prof. Dr. Suvarna Pawar, Head of the Department of Computer Science and Engineering, MIT-ADT University, Pune, for her invaluable support and for providing the necessary departmental resources essential for the completion of this work.

Appreciation is extended to Prof. Suresh Kapare, Chief Coordinator – PBL, Department of Computer Science and Engineering, for his continuous support and coordination throughout the execution of the project.

Profound gratitude is expressed to the project guide, Prof. Pratiksha Dhande, for her expert guidance, constant encouragement, and technical inspiration throughout the research process.

Finally, the authors are grateful to all faculty members of the Department of Computer Science and Engineering and their peers for their direct and indirect contributions to this work.

VIII. REFERENCES

[1] P. K, G. L. S, K. M. R, and S. S, "Medify-AI based LLM Based Healthcare System," in Proc. Int. Conf. Frontier Technol. Solut. (ICFTS), 2025, pp. 1–9, doi: 10.1109/ICFTS62006.2025.11031483.

[2] K. T, M. D, K. R, J. SK, N. T, and P. P, "Smart Health Consulting and Appointment Booking System with Real-Time Scheduling and Patient-Doctor Communication," in Proc. 3rd Int. Conf. Artif. Intell. Mach. Learn. Appl. (AIMLA), 2025, pp. 1–6, doi: 10.1109/AIMLA63829.2025.11040602.

[3] P. Wang and J. Wu, "Large Vision-Language Models-based Human-Understandable Explanations for Medical Image Analysis," in Proc. 4th Int. Conf. Image Process. Comput. Vis. Mach. Learn. (ICICML), 2025, pp. 709–714, doi: 10.1109/ICICML67980.2025.11333439.

[4] P. Kalaiarasi, M. P. Reddy, K. Kousik, and M. Neeraj, "SAHAYAK AI: A Next-Gen AI-Powered Intelligent Healthcare Management System for Enhanced Diagnosis and Hospital Efficiency," in Proc. Int. Conf. Comput. Robot. Test. Eng. Eval. (ICCRTEE), 2025, pp. 1–6, doi: 10.1109/ICCRTEE64519.2025.11053067.

[5] I. Agarwal, V. Sakthivel, and P. Prakash, "Toward Inclusive Healthcare: An LLM-Based Multimodal Chatbot for Preliminary Diagnosis," IEEE Access, vol. 13, pp. 136420–136432, 2025, doi: 10.1109/ACCESS.2025.3594218.

[6] M. Xu, "Research and Application of Dialogue Diagnosis and Treatment System Based on Large Language Model," in Proc. 8th Asian Conf. Artif. Intell. Technol. (ACAIT), 2024, pp. 1413–1418, doi: 10.1109/ACAIT63902.2024.11021946.

[7] S. S. Rasheed and I. H. Glob, "Classifying and Prediction for Patient Disease Using Machine Learning Algorithms," in Proc. 3rd Inf. Technol. Enhance e-Learning Other Appl. (IT-ELA), 2022, pp. 196–200, doi: 10.1109/IT-ELA57378.2022.10107935.

[8] U. Chauhan, H. Jha, D. Singh, and S. P. S. Chauhan, "Doctor Finder and Appointment Booking Website using DJANGO," in Proc. 2nd Int. Conf. Innov. Pract. Technol. Manage. (ICIPTM), 2022, pp. 397–400, doi: 10.1109/ICIPTM54933.2022.9753977.

[9] Z. Yang, C. Chen, S. S. Mahmoud, X. Tan, Y. Chen, and Q. Fang, "Cardiology-Chat: A Multi-LLMs Powered System for Cardiac Diagnostic Reasoning and Clinical Support," IEEE J. Transl. Eng. Health Med., vol. 14, pp. 123–132, 2026, doi: 10.1109/JTEHM.2026.3668755.

[10] M. Kuzlu, Z. Xiao, S. Sarp, F. O. Catak, N. Gurler, and O. Guler, "The Rise of Generative Artificial Intelligence in Healthcare," in Proc. 12th Mediterranean Conf. Embedded Comput. (MECO), 2023, pp. 1–4, doi: 10.1109/MECO58584.2023.10155107.

[11] M. A. Rahman, "A Survey on Security and Privacy of Multimodal LLMs — Connected Healthcare Perspective," in Proc. IEEE Globecom Workshops (GC Wkshps), 2023, pp. 1807–1812, doi: 10.1109/GCWkshps58843.2023.10465035.