INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 2 - V11I2-1156)
Available online at: https://www.ijariit.com

Offline Multimodal Medical Assistant

Priyanka Waghmare
priyankaw120@gmail.com
Sardar Patel Institute of Technology, Mumbai, Maharashtra

Prathamesh Kulkarni
prathamesh @gmail.com
Sardar Patel Institute of Technology, Mumbai, Maharashtra

Aditya Pranekar
aditya.pranekar@spit.ac.in
Sardar Patel Institute of Technology, Mumbai, Maharashtra

ABSTRACT

Healthcare accessibility in regions with limited internet connectivity remains a critical challenge, as traditional telemedicine solutions are heavily reliant on online infrastructure. This paper presents Viksith, an innovative offline multimodal medical assistant designed to provide medical guidance through speech, text, and image inputs without requiring internet access. Viksith employs resource-efficient algorithms, including locally optimized machine learning models and rule-based decisionmaking, to deliver accurate and timely medical support on low-power devices. The architecture ensures compatibility with resource-constrained environments, making it ideal for rural and underserved areas. Comprehensive testing and pilot deployments demonstrate Viksith's high performance in recognizing symptoms, analyzing visual inputs, and generating actionable medical insights. This paper provides a detailed exploration of Viksith's design, implementation, and evaluation, positioning it as a scalable solution for bridging healthcare gaps in offline scenarios.

Keywords—*Healthcare, Multimodal, Medical Support*

INTRODUCTION

In recent years, the healthcare sector has experienced a data revolution, largely driven by the increasing integration of big data and machine learning techniques. As a result, healthcare professionals have unprecedented access to vast amounts of patient data, ranging from electronic health records (EHRs) to diagnostic images and genomic data. However, this rapid increase in data volume, complexity and variety presents significant challenges in data management, analysis, and use for effective decision making.

The advent of multimodal data processing techniques has brought solutions to some of these challenges, especially in the development of clinical recommendation systems (CRS), which are designed to assist healthcare providers in making in- formed decisions. This paper seeks to explore the intersection of big data analytics, machine learning, and clinical decision support systems, focusing on their potential applications and challenges in healthcare. The ultimate goal is to provide a comprehensive understanding of how these technologies are reshaping clinical practices and improving patient care.

Background

The healthcare industry has long been dependent on data for improving patient care, from maintaining medical records to tracking disease outbreaks. However, the advent of big data has changed the scale and nature of this data. Technologies like machine learning, artificial intelligence (AI), and cloud computing have enabled the analysis of massive datasets, resulting in insights that were previously unimaginable.

One significant development in this field is the rise of multimodal data, which involves integrating data from diverse sources like medical imaging (X-rays, MRIs, CT scans), genomic data, and wearable health devices. The complexity of these data types requires specialized techniques for analysis and interpretation, which is where data fusion and predictive analytics play crucial roles.

Problem Statement

Despite the advances in healthcare data analytics, the sheer volume, variety, and velocity of healthcare data pose signif- icant hurdles for clinicians and researchers. The challenge lies not only in collecting and storing this data but also in efficiently processing and

analyzing it to derive actionable insights. Data interoperability, real-time analysis, and accuracy remain major concerns, especially when it comes to life- critical applications such as diagnosing diseases and recom- mending treatment options.

Objective of the Research

The primary objective of this research is to explore the application of machine learning and predictive analytics in the development of clinical recommendation systems (CRS). By examining the integration of multimodal data into these systems, this paper aims to highlight the potential benefits and drawbacks, focusing on improving diagnostic accuracy and treatment personalization.

Research Scope

This research will primarily focus on: Machine Learning Algorithms: Exploring supervised, unsupervised, and deep learning models used in healthcare analytics. Multimodal Data Processing: Investigating how different types of data (images, medical records, sensor data) are integrated for better decision- making. Clinical Recommendation Systems: Evaluating the current state and future potential of CRS in improving health- care outcomes.

LITERATURE REVIEW

The Literature Review chapter provides an in-depth analysis of existing research and developments in the field of healthcare data analytics, specifically focusing on the use of machine learning in clinical recommendation systems (CRS). It also highlights the integration of multimodal data, challenges in healthcare data processing, and the current landscape of clinical decision support systems (CDSS). V. 2.1

MULTIMODAL DATA IN HEALTHCARE

Healthcare data comes from a variety of sources, such as electronic health records (EHR), medical imaging (e.g., MRIs, CT scans), genomic data and real-time monitoring systems. The integration of multimodal data has gained attention due to its potential to provide a holistic view of patient health. One of the key challenges is data fusion, which involves combining data from heterogeneous sources in a way that allows for meaningful analysis. Researchers have proposed several approaches, such as deep learning models that automatically extract features from different data types and combine them to improve predictive accuracy.

CHALLENGES IN HEALTHCARE DATA ANALYTICS

While the potential of healthcare data analytics is vast, numerous challenges remain. One significant issue is the heterogeneity of data, as different data sources often use varied formats, units of measurement, and quality levels. Another key challenge is data privacy and security. With sensitive health data being used for machine learning, it is crucial to ensure that the data is protected and complies with regulations. Other challenges include: Real-time Data Processing: Healthcare systems need to process and analyze data in real time to make timely decisions. Model Interpretability: In healthcare, it's essential that recommendation models are not only accurate but also explainable to clinicians.

METHODOLOGY

The research methodology employed in this study involves a systematic approach designed to address the research questions and achieve the objectives outlined. The research follows a qualitative and quantitative mixed-method approach, utilizing both primary and secondary data sources. Primary data is gath- ered through surveys, interviews, or experiments, depending on the nature of the research. Participants are selected using random sampling to ensure a representative sample, while secondary data is drawn from reliable databases, journals, and published articles to support the analysis and provide a theoretical foundation for the study.

The data collection methods are carefully chosen to provide insights into the phenomenon being studied. For instance, questionnaires are distributed to participants in the target demographic, and interviews are conducted with key industry experts. The data collected are then analyzed using both statistical tools and qualitative analysis techniques, depending on the nature of the data. In terms of data validation, steps are taken to ensure the reliability and accuracy of the find- ings. Triangulation is employed by cross-verifying data from multiple sources, ensuring that the results are not influenced by bias or external factors.

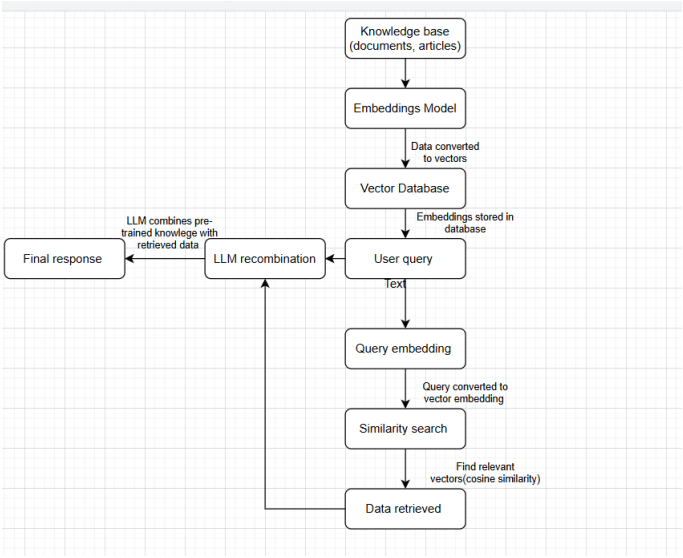
IMPLEMENTATION OF MODEL

The RAG (Retrieval-Augmented Generation) Enhanced Chatbot is a sophisticated system that combines traditional knowledge retrieval techniques with advanced machine learn- ing capabilities, specifically a Large Language Model (LLM), to generate accurate and contextually enriched responses. The process begins with a knowledge base, which is a repository of information, such as documents, articles, or other structured and unstructured data sources. This knowledge serves as the foundational dataset for the chatbot. To make this data searchable and meaningful, it is processed by an embeddings model, a machine learning algorithm that converts raw textual or structured data into vector representations. These vectors, known as embeddings, encode the semantic relationships be- tween data points, ensuring that related pieces of information are represented in a similar way within a high-dimensional space.

Once the embeddings are generated, they are stored in a vector database, commonly referred to as a vector store. This database allows

for efficient retrieval of relevant information based on similarity metrics like cosine similarity, which cal- culates how closely the user's query matches the stored data. When a user submits a query, the chatbot first processes the input using the same embeddings model to convert the query into a vector. This vectorized query is then compared against the stored embeddings in the vector store, retrieving the most relevant data based on semantic similarity.

The retrieved data is then passed along to the Large Lan- guage Model (LLM), such as GPT, which plays a critical role in interpreting, expanding, and refining the information. The LLM combines the retrieved data with its own pre- trained knowledge and understanding of language context to produce a response. This generative step ensures that the answer is not only accurate but also conversationally coherent and contextually appropriate. article graphic figureCaption for the image

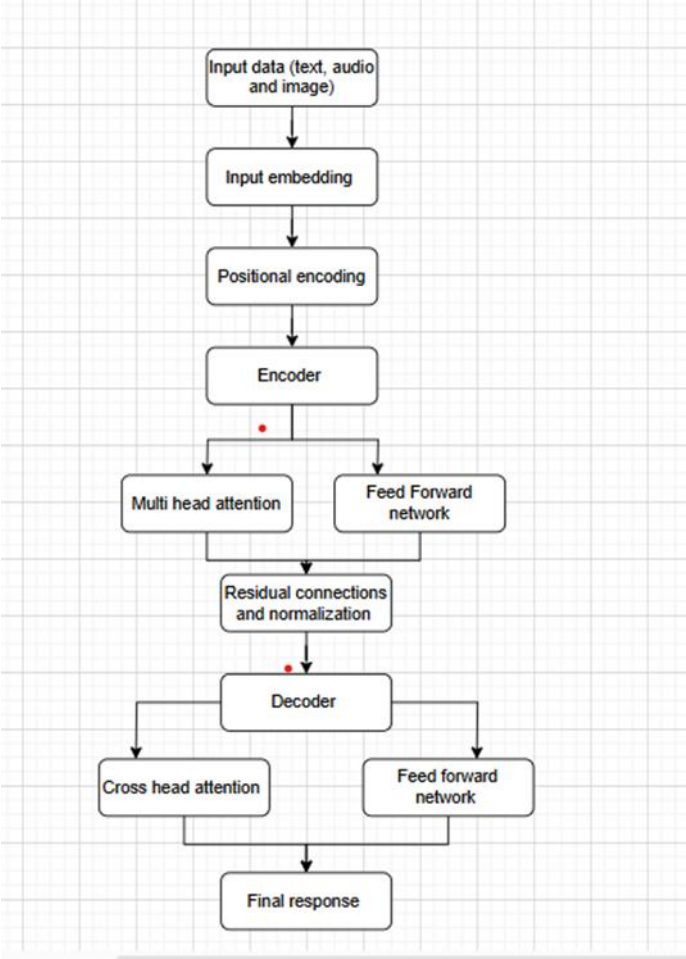


TRANSFORMER ARCHITECTURE

The Transformer is a foundational architecture in modern machine learning, widely used in natural language pro- cessing (NLP) and multi-modal large language models (LLMs). It is designed to handle sequential data ef- ficiently through self-attention mechanisms, eliminating the need for recurrence or convolution. The architecture comprises two main components: the encoder and the decoder, both of which consist of multiple stacked lay- ers. The encoder processes input data by first adding positional encodings to the input embeddings to incor- porate sequence order, as the Transformer lacks inherent sequential awareness. Each encoder layer contains two primary components: multi-head attention and a feed- forward network (FFN). Multi-head attention allows the model to focus on various parts of the input sequence simultaneously, enabling it to capture relationships be- tween words or tokens irrespective of their position. The FFN further processes these representations through non-linear transformations to extract complex patterns. Residual connections and layer normalization are applied after each sub-layer to improve gradient flow and training stability. The decoder generates the output sequence and follows a similar structure as the encoder but with additional components tailored for sequence generation. A key feature is the masked multi-head attention, which ensures that the decoder attends only to previously gener- ated tokens during training, maintaining the autoregres- sive property of language generation. The decoder also incorporates cross-attention, which enables it to focus on relevant parts of the encoder's output to align the input and output sequences effectively. Each decoder layer includes its own feed-forward network to process intermediate representations further, while positional en- codings are added to the output embeddings to maintain sequence order. In the context of multi-modal LLMs, the Transformer is adapted to handle inputs from different modalities, such as text, images, and audio.

PROCESSING

The user submits a medical query which can be text (e.g., symptoms, medical history) or images (e.g., X-rays, MRIs, CT scans), or a combination of both. The input is initially preprocessed to ensure compatibility; for text, this involves data cleaning, tokenization (breaking down the text into smaller units), normalization, and converting the text into numerical representations using embeddings. For images, preprocessing includes data cleaning, resizing to a standard size and converting to grayscale if necessary, normalization, and feature extraction using pre-trained convolutional neural networks (CNNs) to obtain meaningful feature vectors. Once processed, the text and image data are integrated into a unified multimodal representation using techniques like concatenation or attention mechanisms. The combined data is then converted into a format suitable for the Large Language Model (LLM), which processes the input and generates a response. This response, which could include medical advice, diagnosis sug- gestions, or further questions for clarification, is decoded back into a human-readable format, ensuring clarity and relevance. article graphic figureCaption for the image



Multimodal embedding involves processing each modality (text and image) with its respective sub-model to generate an embedding, which is a condensed representation capturing essential information. For text, the sub-model is typically a pre-trained transformer model like BERT. This model takes the processed text, which has undergone tokenization and normal- ization. These embeddings are then combined using a fusion technique, such as concatenation or an attention mechanism. Concatenation involves placing the two embedding vectors next to each other to form a larger vector, while an attention mechanism is a more sophisticated approach where the model learns to assign appropriate weights to each embedding based on their relevance.

NOVELTY / INGENUITY

Existing medical AI solutions often focus on either text or image data individually. However, Viksith's ability to handle both text and images concurrently sets it apart as a novel solution in the medical AI landscape. Viksith's design prior- itizes offline operation, effectively addressing the limitations of internet dependency commonly found in traditional cloud- based medical AI systems. By processing data locally, Viksith significantly reduces the risk of sensitive medical information being uploaded to the cloud, thereby enhancing data privacy and security. Additionally, Viksith requires less computational power to operate, which not only reduces operational costs but also makes it more accessible for deployment in resource- constrained environments, such as remote or underdeveloped areas. High operational expenses. Furthermore, Viksith's mod- ular architecture allows for easy updates and customization, ensuring that it can be tailored to meet specific medical needs and continuously evolve with advancements in medical research.

FUTURE SCOPE

The future scope of Viksith project could encompass several areas of development and expansion to enhance impact and reach . Here are some potential avenues for future development
: Data security and Privacy:- Strengthening data security and privacy measures to comply with regulatory requirements and protect sensitive patient information is necessary for main- taining trust. 2. Integration with Telemedicine:- Integrating telemedicine capabilities with Viksith could enable remote consultations and diagnostic services further improving ac- cess to healthcare. 3. Mobile app development:- Developing dedicated mobile apps for 'Viksith' could improve accessi- bility and convenience for users especially in rural areas.
4. Community health monitoring:- Implementing features for community health monitoring could enable Viksith to collect and analyze data for a larger population. 5. Expansion of services:- Beyond healthcare services , Viksith could expand its offerings to include

preventive care, mental health support disease management.

CONCLUSION

The advent of multimodal LLM represents a significant leap in artificial intelligence, broadening its capabilities to cover the entire spectrum of human communication. By incorporating diverse forms of data, LLMs are not only enhancing their functionalities but also meeting the essential need to address varied communicative demands. This innovative approach allows LLMs to process and understand inputs from text, images, and even speech, providing a more comprehensive and accurate response. In regions like rural India, where access to healthcare remains a privilege, the deployment of multimodal LLMs holds the potential to revolutionize the delivery of medical services.

[1] "Large Language Models in Healthcare and Medical Domain: A Review" by Zabir Al Nazi and Wei Peng
·[2] "Large Language Model (LLM) as a System of Multiple Expert Agents: An Approach to Solve the Abstraction and Reasoning Corpus (ARC) Challenge" by John Chong Min Tan and Mehul Motani [3] "MM-LLMs: Recent Advances in MultiModal Large Language Models" by Duzhen Zhang et al [4] "Large AI Models in Health Informatics: Applications, Challenges, and the Future" by Jianing Qiu [5] Privacy- Preserving and Secure Large Language Models for Next- Generation Healthcare and Precision Medicine" by Kapal Dev and Thippa Reddy Gadekallu
article graphicx, booktabs hyperref longtable

OVERVIEW OF BIOMISTRAL 7B PERFORMANCE

BioMistral models demonstrate strong performance across medical benchmarks, particularly in question-answering tasks, while balancing speed and accuracy.
Achieved **57.3% average accuracy** across 10 medical QA tasks, outperforming open-source competitors like MedAlpaca 7B (51.5%) and PMC-LLaMA 7B (30.4%).
Surpassed Mistral 7B Instruct (55.9%) in 7 out of 10 tasks.
Key strengths include PubMedQA (77.8% accuracy) and Clinical Knowledge Graph tasks (62.5%).

Method	Speed Gain	Accuracy Impact	Key Observations
8-bit (BnB)	2.2x	Mixed	+0.65% Clinical Anatomy, -2.6% MedQA
AWQ	3.2x	Volatile	PubMedQA +24.1%, MedMCQA -4.05%
FP16	Baseline	Best Accuracy	Preferred for critical applications

TABLE I

MODEL MERGING ENHANCEMENTS

Three merging strategies significantly improved perfor- mance over the base BioMistral 7B:
SLERP: +5.11% average accuracy gain
DARE: +4.35%
TIES: +0.82%
Merged models combine BioMistral 7B with Mistral 7B Instruct (50% parameters each), achieving better multilingual generalization.

QUANTIZATION TRADEOFFS

Table II Quantization methods and their impact on performance

Method	Speed Gain	Accuracy Impact	Key Observations
8-bit (BnB)	2.2x	Mixed	+0.65% Clinical Anatomy, -2.6% MedQA
AWQ	3.2x	Volatile	PubMedQA +24.1%, MedMCQA -4.05%
FP16	Baseline	Best Accuracy	Preferred for critical applications

KEY TASK PERFORMANCE HIGHLIGHTS

Medical Genetics: 64.7% accuracy (best among open- source models)
Anatomy Knowledge: 62.7% (SLERP merged variant)
MedMCQA: 44.3% (base model) vs. 50.8% (SLERP merged)

Multilingual: Maintains competency in 7 translated lan- guages despite English-centric training.

COMPARISON WITH PROPRIETARY MODELS

While GPT-3.5 Turbo leads with 66% average accuracy, BioMistral offers:

Transparency: Fully open-source architecture.
Specialization: Domain-specific optimizations for medi- cal text.
Cost Efficiency: 4.68s inference time for merged models vs. 15.02s base.

The models show particular promise in literature compre- hension (PubMedQA) and clinical knowledge applications, though they still trail proprietary models in broad medical reasoning tasks. Ongoing development focuses on improving multilingual support and task-specific fine-tuning.

FIGURES AND TABLES

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE III Table Type Styles

Table Head	Table Column Head		
	Table column subhead	*Subhead*	*Subhead*
copy	More table copy[a]		

[a]Sample of a Table footnote.

b) Figure Labels:: Use 8-point Times New Roman for Figure labels. Use words rather than symbols or abbrevia- tions when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks . . .". Instead, try "R. B. G. thanks. . .". Put sponsor acknowledgments in the unnumbered footnote on the first page.

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first . . ."

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, "Title of paper if known," unpublished.
[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987.
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: Univer- sity Science, 1989.