Journal of Science Engineering Technology and Management Science
Volume 03, Issue 06, June 2026 *ISSN: 3049-0952*
www.jsetms.com

HYBRID RAG CHATBOT FOR SYMPTOM MONITORING USING MACHINE LEARNING AND RETRIEVAL AUGMENTED GENERATION

K.JanakiRam ,M.C.A Student , Amritha sai institute of science and technology, Kanchikacharla (Mandal), A.P- 521180

P.Ramesh Babu,Associate professor , Amritha sai institute of science and technology, Kanchikacharla (Mandal), A.P- 521180

ABSTRACT

The increasing demand for accessible and efficient healthcare solutions has led to the development of intelligent medical chatbots. Traditional chatbots often rely on static knowledge bases or predefined responses, which limits their ability to provide accurate and context-aware medical advice. This paper proposes a Hybrid Retrieval-Augmented Generation (RAG) chatbot for symptom monitoring that integrates machine learning techniques with large language models (LLMs). The system combines retrieval mechanisms to fetch relevant medical information from trusted datasets with generative models to produce human-like responses. By leveraging both structured medical knowledge and contextual reasoning, the chatbot improves diagnostic assistance and user interaction. The proposed system enhances accuracy, reduces misinformation, and provides real-time healthcare support. Experimental results demonstrate improved performance in terms of response relevance, accuracy, and user satisfaction compared to traditional chatbot systems.

1. INTRODUCTION

The rapid advancement in artificial intelligence has significantly impacted the healthcare industry. One of the most promising applications is the use of chatbots for symptom monitoring and preliminary diagnosis. These systems help users identify possible health conditions based on their symptoms and provide guidance on seeking medical attention.

Conventional healthcare chatbots are limited by rule-based systems or static machine learning models, which lack adaptability and contextual understanding. Recent developments in large language models (LLMs) have enabled more sophisticated conversational systems, but they may generate incorrect or misleading information due to lack of real-time knowledge grounding.

To address these limitations, Retrieval-Augmented Generation (RAG) has emerged as a powerful technique. RAG combines information retrieval with generative AI to ensure that responses are both accurate and contextually relevant. This paper presents a hybrid RAG-based chatbot system that integrates machine learning classification with retrieval and generation techniques to improve symptom monitoring and healthcare assistance.

2. LITERATURE SURVEY

Several studies have explored the application of AI in healthcare chatbots and symptom analysis:

Early systems were primarily rule-based, relying on predefined decision trees for symptom checking. While these systems were simple to implement, they lacked flexibility and scalability.

Machine learning-based approaches introduced classification algorithms such as Support Vector Machines (SVM), Decision Trees, and Naïve Bayes to predict diseases based on symptoms. These models improved accuracy but were limited by the quality of training data and lacked conversational abilities.

Recent advancements in deep learning and natural language processing (NLP) have enabled the development of intelligent chatbots using neural networks and transformer-based models. However, these models often suffer from hallucination, where incorrect or fabricated information is generated.

Retrieval-Augmented Generation (RAG) models have been proposed to overcome this limitation by combining external knowledge retrieval with generative models. Studies

show that RAG-based systems significantly improve response accuracy and reliability, especially in domains like healthcare where factual correctness is critical.

3. EXISTING SYSTEM

The existing systems for symptom monitoring primarily include:

1. Rule-Based Chatbots

- Use predefined rules and decision trees
- Limited flexibility
- Cannot handle complex queries

2. Machine Learning Models

- Use classification algorithms
- Require structured datasets
- Lack conversational intelligence

3. Standalone LLM-Based Chatbots

- Generate human-like responses
- Prone to hallucination
- Lack real-time knowledge updates

Limitations:

- Low contextual understanding
- Limited accuracy in diagnosis
- No integration of real-time medical knowledge
- Risk of misinformation

4. PROPOSED SYSTEM

The proposed system is a **Hybrid RAG Chatbot** that combines:

- Machine Learning (for symptom classification)

- Retrieval System (for fetching medical knowledge)
- Generative Model (for natural responses)

System Architecture:

1. User Input (Symptoms)
2. NLP Processing
3. Symptom Classification (ML Model)
4. Knowledge Retrieval (Database/Corpus)
5. Response Generation (LLM)
6. Output to User

Key Features:

- Context-aware responses
- Reduced hallucination
- Real-time knowledge integration
- Improved diagnostic support

Advantages:

- High accuracy and reliability
- Scalable and adaptable
- Enhances patient engagement
- Supports early disease detection

5. ALGORITHMS USED

5.1 Retrieval-Augmented Generation (RAG)

RAG combines retrieval and generation processes:

- **Retriever:** Fetches relevant documents from a knowledge base
- **Generator:** Produces responses using retrieved data

Working:

1. Convert query into embeddings
2. Retrieve top relevant documents
3. Feed documents into generative model
4. Generate final response

5.2 Machine Learning Classification

Used to classify symptoms into possible disease categories.

Common algorithms:

- Decision Tree
- Random Forest
- Support Vector Machine (SVM)

5.3 Natural Language Processing (NLP)

- Tokenization
- Lemmatization
- Named Entity Recognition (NER)
- Text vectorization

5.4 Transformer-Based Models

Used for response generation:

- Context understanding
- Language generation
- Conversational flow

6. RESULTS

The proposed system was evaluated based on:

Performance Metrics:

- **Accuracy:** Correct predictions of diseases
- **Precision:** Correctly identified positive cases
- **Recall:** Ability to detect actual conditions
- **F1-Score:** Balance between precision and recall

Observations:

- Improved response accuracy compared to standalone ML models
- Reduced hallucination due to retrieval mechanism
- Better user interaction and satisfaction
- Faster response generation

Comparative Analysis:

System Type	Accuracy	Reliability	Flexibility
Rule-Based	Low	Low	Low
ML-Based	Medium	Medium	Medium
LLM-Based	High	Low	High
Hybrid RAG	**High**	**High**	**High**

7. CONCLUSION

This paper presented a Hybrid Retrieval-Augmented Generation chatbot for symptom monitoring. By integrating machine learning, retrieval mechanisms, and generative AI, the system overcomes the limitations of traditional chatbot approaches. The proposed model provides accurate, context-aware, and reliable healthcare assistance.

The hybrid approach ensures that responses are grounded in real medical knowledge while maintaining conversational quality. Future work can focus on integrating real-time hospital data, multilingual support, and deployment in mobile healthcare applications.

8. REFERENCES

1. Lewis, P. et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” 2020.
2. Devlin, J. et al., “BERT: Pre-training of Deep Bidirectional Transformers,” 2019.
3. Vaswani, A. et al., “Attention Is All You Need,” 2017.
4. Rajkomar, A. et al., “Machine Learning in Medicine,” New England Journal of Medicine.
5. Kaggle Medical Datasets for Symptom Analysis.
6. Jurafsky, D., Martin, J., “Speech and Language Processing,” 2021.