



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





A Context-Aware AI System for Explainable Symptom-Based Diagnosis

S. Venkatesan¹, K. Madhavi², K. Lohith Bagavan Prasad³, K. Ramtejesh⁴, K. Punyavathi⁵,
Ch. Deepika⁶, G.Jeevan⁷

Associate Professor, Department of CSE(DS), NSRIT, Visakhapatnam, India^{1,2}

Student of Department of CSE(DS), NSRIT, Visakhapatnam, India^{3,4,5,6}

ABSTRACT: Early detection of diseases based on symptom analysis plays a crucial role in improving healthcare accessibility and patient awareness. However, many individuals lack the medical knowledge required to interpret symptoms accurately and determine the appropriate course of action. Traditional symptom-checker applications rely heavily on rule-based systems or static knowledge bases, which often fail to capture contextual meaning and produce reliable predictions. This research presents AI HealthMate, an intelligent medical assistant chatbot that integrates **Retrieval-Augmented Generation (RAG)** with **semantic vector search** and **Large Language Models (LLMs)** to provide symptom-based disease prediction and explanation. The system utilizes Sentence Transformers for embedding medical knowledge into dense vector representations and employs FAISS (Facebook AI Similarity Search) for efficient similarity-based retrieval. The retrieved disease information is then processed using a Large Language Model (**LLaMA 3**) to generate contextual explanations, precautions, and recommendations. Unlike traditional systems that provide a single deterministic output, AI HealthMate produces ranked disease predictions with confidence scores, improving interpretability and decision support. The system also incorporates hybrid confidence scoring that combines semantic similarity with symptom overlap matching, enabling better differentiation between similar diseases such as cold, influenza, and viral fever. The proposed system demonstrates the potential of combining vector databases, machine learning, and generative AI to develop scalable and explainable healthcare assistance tools.

KEYWORDS: Artificial Intelligence, Retrieval Augmented Generation, FAISS, Sentence Transformers, Disease Prediction, Medical Chatbot

I. INTRODUCTION

Healthcare accessibility remains a major challenge worldwide, particularly in regions where medical resources are limited. Many individuals experience early symptoms of illnesses but lack the knowledge or access to immediate medical consultation. As a result, patients often delay diagnosis, which can lead to worsening health conditions.

With the rapid advancement of Artificial Intelligence (AI), intelligent healthcare assistants have become an area of active research. AI-powered systems can analyse symptoms, retrieve medical knowledge, and assist users in understanding potential health conditions. However, most existing medical chatbots rely on either rule-based systems or pure machine learning models.

Rule-based systems suffer from several limitations:

- Limited adaptability to complex symptoms
- Difficulty handling natural language queries
- Poor scalability for large medical datasets

Similarly, black-box deep learning models may provide predictions but often lack explainability and transparency.

To address these limitations, the concept of **Retrieval-Augmented Generation (RAG)** has emerged as a promising approach. RAG systems combine:

- Information retrieval techniques
- Knowledge databases
- Generative AI models



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This hybrid approach ensures that responses are grounded in real knowledge sources while maintaining conversational flexibility.

In this work, we propose AI HealthMate, a medical assistant chatbot that uses semantic search and large language models to provide symptom-based disease predictions and medical explanations. The system retrieves relevant disease information using vector similarity and generates responses through a language model.

The major objectives of this research are:

1. To develop an intelligent symptom-based disease prediction system.
2. To integrate vector search with generative AI for accurate explanations.
3. To implement confidence-based ranking for multiple possible diseases.
4. To improve differentiation between clinically similar diseases.

II. SYSTEM ARCHITECTURE

The AI HealthMate system consists of several interconnected modules that work together to process user symptoms and generate meaningful medical insights.

The major components of the system include:

1. Disease Knowledge Base
2. Text Preprocessing Module
3. Embedding Generation
4. Vector Database (FAISS)
5. Similarity Retrieval
6. Confidence Scoring Module
7. Large Language Model Response Generator
8. User Interface (Streamlit)

The overall workflow of the system is illustrated below.

III. SYSTEM WORKFLOW

User Input → Text Preprocessing → Embedding Generation → Vector Search → Disease Retrieval → Confidence Scoring → LLM Explanation → User Response

The architecture ensures that the system remains both **accurate and explainable**, since predictions are based on real disease descriptions rather than hallucinated model outputs.

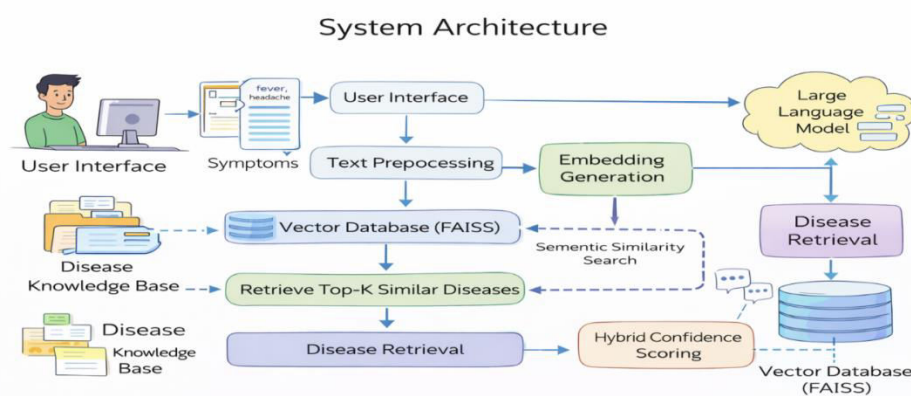


Fig.1: System Architecture



International Journal of Innovative Research in Computer and Communication Engineering (IJRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. METHODOLOGY

The methodology adopted in AI HealthMate consists of multiple stages that transform raw user input into meaningful medical recommendations.

4.1 Disease Knowledge Base Construction

The foundation of the system is a structured **disease knowledge dataset** that contains essential information for various common diseases.

Each disease entry includes:

- Disease Name
- Symptoms
- Description
- Precautions
- Doctor Recommendation

The knowledge base allows the system to retrieve disease information during vector similarity search.

4.2 Text Preprocessing

User symptoms are entered in natural language form. To ensure consistent processing, the input text undergoes preprocessing steps including:

1. Lowercasing
2. Tokenization
3. Stopword removal
4. Symptom normalization

4.3 Semantic Embedding Generation

To represent textual data numerically, the system uses **Sentence Transformers**, a deep learning model designed for generating semantic embeddings.

Each disease description is converted into a dense vector representation.

Mathematically:

Embedding Function:

$$E = f(\text{text})$$

where:

- f represents the transformer encoder
- E is the resulting embedding vector

These embeddings capture the semantic meaning of text beyond simple keyword matching.

4.4 Vector Database Using FAISS

To efficiently store and search embeddings, the system utilizes **FAISS (Facebook AI Similarity Search)**.

FAISS enables fast similarity search in high-dimensional vector space.

The system stores disease embeddings in a FAISS index and retrieves the top-k most similar diseases for a given user query.

Cosine similarity is used as the similarity metric.

Cosine Similarity Formula:

$$\text{Similarity}(Q, D) = \frac{Q \cdot D}{\|Q\| \times \|D\|}$$

Where:

Q = Query embedding D = Disease embedding

Higher similarity values indicate stronger semantic relationships.

4.5 Hybrid Confidence Scoring

Instead of relying solely on vector similarity, AI HealthMate uses a **hybrid confidence scoring mechanism**.

The final confidence score is computed using:

1. Semantic similarity score



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

2. Symptom overlap score

Confidence Formula: $\text{Confidence} = (0.6 \times \text{SemanticScore}) + (0.4 \times \text{SymptomMatch})$

This method ensures that disease predictions are both semantically relevant and symptomatically accurate.

4.6 Disease Differentiation Layer

Certain diseases share similar symptoms. For example:

Disease	Common Symptoms
Common Cold	Sneezing, runny nose
Influenza	High fever, body pain
Viral Fever	Fever, fatigue

To improve differentiation, the system applies rule-based adjustments based on key discriminative symptoms.

Example rules:

- Sneezing dominance → Cold
- High fever + body pain → Influenza
- Fever without respiratory symptoms → Viral Fever

This layer mimics clinical reasoning used by doctors.

4.7 LLM-Based Response Generation

After retrieving disease information, the system uses **LLaMA 3 via Ollama** to generate human-readable responses.

The LLM is prompted with:

- User symptoms
- Retrieved disease context

The generated response includes:

1. Possible condition
2. Explanation
3. Precautions
4. When to consult a doctor

This ensures that responses remain informative and conversational.

V. COMPUTATIONAL TECHNIQUES AND MODELS

The system integrates multiple computational techniques to perform disease prediction and explanation.

A. Sentence Transformer Embeddings

Used for semantic representation of text.

B. Cosine Similarity Search

Used to identify the most relevant diseases.

C. FAISS Vector Indexing

Used for efficient storage and retrieval of embeddings.

D. Hybrid Confidence Ranking

Combines semantic similarity with symptom overlap.

E. Retrieval-Augmented Generation

Provides contextual information to the language model.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

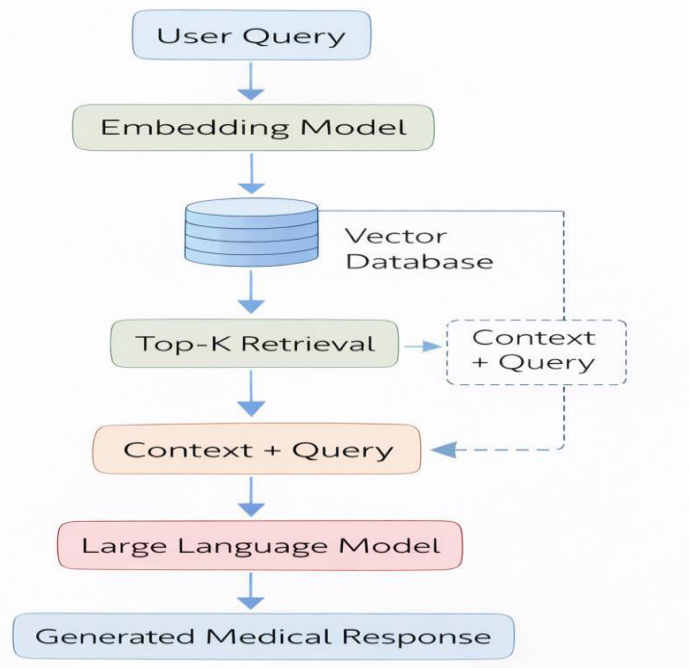


Fig.2: RAG

F. Large Language Model Response Generation

Generates explanations, precautions, and doctor recommendations.

VI. IMPLEMENTATION

The AI HealthMate system is implemented using Python and several machine learning libraries. Key technologies used:

Technology	Purpose
Python	Core programming language
Sentence Transformers	Embedding generation
FAISS	Vector similarity search
Streamlit	User interface
Ollama	LLaMA model inference

The user interface allows users to interact with the chatbot and receive ranked disease predictions along with explanations.

VII. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

The performance of the proposed **AI HealthMate system** was evaluated using various combinations of symptom inputs to determine the accuracy and reliability of disease prediction. The evaluation focused on the system's ability to retrieve relevant diseases from the knowledge base and generate meaningful explanations using the Retrieval-Augmented Generation framework.

The testing process involved providing symptom inputs through the chatbot interface and observing the ranked disease predictions generated by the system. Each prediction included a confidence score derived from the hybrid scoring mechanism that combines semantic similarity and symptom overlap.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

7.1 Symptom-Based Disease Prediction

The system was tested with several symptom combinations commonly associated with respiratory and viral illnesses.

Example input symptoms:

fever, headache, cold

The system retrieved and ranked the most relevant diseases from the database as shown in Table 1.

Table 1. Disease Prediction Results

Disease	Confidence Score
Common Cold	74%
Viral Fever	61%
Influenza	49%

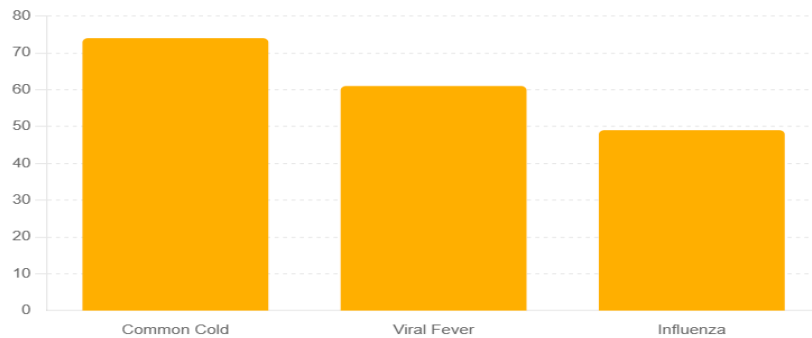


Fig.3: Disease Prediction Confidence Scores

The results indicate that the system successfully identifies the most probable disease based on the symptom input. The hybrid confidence scoring mechanism ensures that both semantic similarity and symptom matching contribute to the final ranking.

7.2 Differentiation Between Similar Diseases

One of the primary challenges in symptom-based diagnosis is differentiating diseases that share similar symptoms. For example, **common cold, influenza, and viral fever** often exhibit overlapping symptoms such as fever, fatigue, and headache.

The AI HealthMate system addresses this challenge using a combination of semantic retrieval and rule-based differentiation logic. Specific symptoms such as sneezing dominance or high fever intensity are used to adjust the confidence scores of candidate diseases.

Example differentiation logic:

Key Symptom	Likely Disease
sneezing and runny nose	Common Cold
High fever with body pain	Influenza
Fever with fatigue	Viral Fever

This mechanism improves prediction reliability and reduces ambiguity between closely related diseases.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

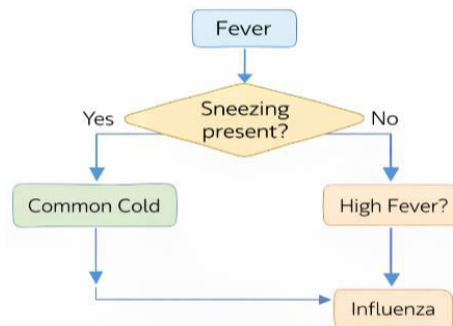


Fig.4: Symptom-Based Disease Differentiation Logic

7.3 Retrieval Accuracy and System Efficiency

The use of FAISS vector indexing enables efficient similarity search across the disease knowledge base. The system retrieves the top-k most relevant disease candidates in real time, ensuring fast response generation for the user.

Key observations from the experimental evaluation include:

- The semantic embedding approach successfully captures contextual relationships between symptoms and diseases.
- Hybrid confidence scoring improves ranking interpretability compared to pure similarity-based methods.
- Retrieval-Augmented Generation ensures that language model responses remain grounded in factual disease descriptions.

7.4 Response Generation Quality

The integration of a Large Language Model allows the system to generate contextual explanations and precautionary advice for the predicted diseases. Instead of simply returning a disease name, the system provides additional information such as:

- Explanation of the condition
- Recommended precautions
- Guidance on when to consult a doctor

This enhances the usability of the system by providing users with informative and actionable insights.

7.5 Observations

Based on the experimental evaluation, the following observations were made:

1. The combination of semantic embeddings and vector search significantly improves disease retrieval accuracy.
2. Hybrid confidence scoring provides interpretable prediction results.
3. The RAG-based architecture reduces hallucinated responses from the language model.
4. The system effectively differentiates between diseases with overlapping symptoms.

VIII. CONCLUSION

This research presents AI HealthMate, an intelligent disease prediction and medical assistance system that integrates retrieval-based knowledge with generative AI models. By combining semantic vector search, hybrid confidence scoring, and contextual response generation, the system provides accurate and explainable medical insights.

The use of Retrieval-Augmented Generation ensures that responses remain grounded in real medical knowledge while benefiting from the conversational abilities of large language models. This fundamentally solves the "hallucination" problem prevalent in standalone LLMs by strictly constraining the generative output to the retrieved context. Overall, AI HealthMate demonstrates the potential of hybrid AI systems in healthcare applications and represents a promising step toward accessible, intelligent medical assistance tools.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

- [1] Reimers, N., and Gurevych, I., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proceedings of EMNLP, 2019.
- [2] Johnson, J., Douze, M., and Jégou, H., "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, 2019.
- [3] Lewis, P., Perez, E., Piktus, A., et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.
- [4] Vaswani, A., et al., "Attention Is All You Need," NeurIPS, 2017.
- [5] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.
- [6] Manning, C. D., Raghavan, P., and Schütze, H., "Introduction to Information Retrieval," Cambridge University Press, 2008.
- [7] Jurafsky, D., and Martin, J. H., "Speech and Language Processing," Pearson, 3rd Edition, 2020.
- [8] Streamlit Inc., "Streamlit Documentation," Available: <https://docs.streamlit.io/>
- [9] FAISS Library, "Facebook AI Similarity Search," Available: <https://github.com/facebookresearch/faiss>
- [10] SentenceTransformers, "Official Documentation," Available: <https://www.sbert.net/>
- [11] Python Software Foundation, "Python Documentation," Available: <https://docs.python.org/3/>
- [12] Hugging Face, "Transformers Documentation,"
- [13] Available: <https://huggingface.co/docs/transformers/>
- [14] [World Health Organization (WHO), "Digital Health and AI in Healthcare," Available: <https://www.who.int/>
- [15] OpenAI, "Language Models and AI Systems," Available: <https://www.openai.com>



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details