International Journal of Computing and Digital Systems
2025, VOL. XX, NO. 1, 1–10

http://dx.doi.org/10.12785/ijcds/XXXXXX

# Semantic Search System For Ayurvedic Medicine Using Knowledge Graph and Retrieval-Augmented Generation

**Durairaj Thenmozhi [1] and J. Giftson Jeba Selva Raja[2]**

[1]*Sri Sivasubramaniya Nadar College of Engineering, India*
[2]*Sri Sivasubramaniya Nadar College of Engineering, India*

*Received XX 2025, Revised XX 2025, Accepted XX 2025*

**Abstract:** Ayurvedic medicine, one of the oldest holistic healing systems, contains vast textual knowledge that is often unstructured and stored in PDFs, making information extraction and retrieval challenging. Traditional keyword-based search systems do not capture the semantic meaning of user queries, leading to inaccurate or incomplete results. This research work addresses this gap by proposing a Semantic Search System for Ayurvedic Medicine that integrates retrieved generation with a Knowledge Graph. The system leverages NLP, vector embeddings, and structured knowledge representation to provide accurate and contextually relevant answers. The RAG component extracts text from Ayurvedic documents, fragments it and embeds it for semantic search, while the knowledge graph (KG) component constructs a graph with entities such as disease, symptoms, and treatment, enabling hybrid retrieval by combining vector and keyword search. Several queries are dynamically generated for an efficient graph-based search. The extracted contexts from both components are combined and passed to a large language model for the final generation of the answer. To evaluate system performance, a curated data set comprising 15 Ayurvedic diseases was used. The evaluation was carried out using 10 domain-specific user queries and the system responses were assessed against expert-verified reference answers using the metrics namely ROUGE-1 and ROUGE-L scores. Comparative analysis showed that the RAG pipeline provided fluent, context-rich answers, while the KG pipeline delivered structured, factually grounded responses. The hybrid RAG+KG model demonstrated improved overall accuracy, outperforming standalone models, with notable gains in handling complex queries. This integrated approach enhances semantic understanding and retrieval accuracy, enabling precise and meaningful responses grounded in Ayurvedic principles. The system provides a scalable and robust solution for semantic search in Ayurveda medicine.

**Keywords:** Semantic Search, Ayurveda Medicine, Ontology, Natural Language Processing, Knowledge Graph, Retrieval-Augmented Generation, Large Language Model

## 1. INTRODUCTION

Ayurvedic medicine, an ancient system of holistic healing, has been practiced forthousands of years and remains widely used today. With the growing globalinterest in alternative medicine and natural remedies, Ayurveda has gainedsignificant recognition. However, the vast body of Ayurvedic knowledge is stilllargely inaccessible due to its traditional textual representation, often stored inunstructured formats such as books and PDFs. Extracting relevant and accurateinformation from these sources poses a significant challenge, especially forresearchers, healthcare practitioners, and individuals seeking Ayurvedictreatments.

In recent years, researchers have explored the application of artificial intelligence,particularly Natural Language Processing and Knowledge Graphs, to improveinformation retrieval in traditional healthcare domains like Ayurveda. Machinelearning models have been used to generate personalized treatmentrecommendations based on a patient's prakriti [1][2], while ontology-basedframeworks like the Ontology-based Concept Extraction and Classificationalgorithm help extract structured concepts from Ayurvedic texts. KnowledgeGraphs have enabled semantic querying of key entities such as doshas, herbs, andtreatments [3]. However, most of these systems focus only on structured dataand struggle to handle the vast unstructured content found in classical Ayurvedicliterature. Additionally, keyword-based search methods often fail to capture thecontextual and personalized nature of Ayurvedic knowledge.Conventional search engines rely on keyword matching, which typically does notreflect the deeper semantic meaning of user queries. This leads to incompleteor irrelevant search results, reducing the effectiveness of information retrieval forAyurvedic diseases, remedies, and treatments. Furthermore, the lack of structuredknowledge representation limits the ability to perform advanced queries and derivemeaningful in-

*E-mail address: theni_d@ssn.edu.in, giftsonjebaselvaraja2320049@ssn.edu.in*

sights.Recent advancements in AI, particularly in NLP and knowledge representation,offer promising solutions to these challenges. Techniques likeRetrieval-Augmented Generation (RAG) and vector-based semantic searchmodels enable context-aware retrieval, making them ideal for improving access toAyurvedic knowledge [4]. By integrating structured knowledge representationwith semantic search, we can significantly enhance the accuracy, contextualrelevance, and accessibility of Ayurvedic information.

## 2. LITERATURE REVIEW

Ayurveda, as an ancient system of medicine, is gaining global recognition, and researchers are increasingly exploring modern technologies like artificial intelligence (AI) and machine learning (ML) to enhance its accessibility and application. Hrishikesh Terdalka et al. [5] highlight the potential of Ayurveda to integrate with contemporary scientific methods, particularly ML, to advance personalized treatments based on dosha imbalances. Unlike conventional treatments that target symptoms, Ayurveda focuses on identifying the root cause of diseases through prakriti (body constitution) analysis. Machine learning techniques such as decision trees, cosine similarity, and random forests can optimize Ayurvedic diagnosis by analyzing vast patient datasets to predict treatment outcomes with higher precision. These techniques facilitate the organization of large volumes of patient information, enabling the identification of symptom patterns and their relationships with dosha imbalances, ultimately improving treatment recommendations.

Ligandro Singh Yumnam et al.,[6] presented a system for automatically building and searching knowledge graphs for user comprehension of structured datasets.However, conventional techniques in processing unstructured data are often time consuming and hard to scale leading to difficulties in extracting relevant knowledge in timely and accurate manner. However, in automating knowledge graph how it can be beneficial is that the system can take unstructured data which are in the form of text or picture and organizes the data in such a way that there is an outline of entities and relationships thus giving a framework for knowledge. Adzhept Lumin's KGC and KQ engine enhances the process of straining information from large volumes of complex data and also facilitates data organization. The user issues queries, and the system explores the graph and finds the relevant insights to address questions, thus enabling efficient analysis of the available data and making sound decisions. Test results showed that this automated solution helps in understanding and drawing sense out of overwhelming amounts of data and more importantly spotting invisible trends and relationships. This advances the approach to existing propositions because it not only speeds up the process of arriving at decisions but also increases the accuracy of data retrieval and therefore can be seen as advantageous for firms that generate or handle vast amounts of data in need of swift and accurate information.

M. Gayathri et al.,[2] created a framework is aimed at achieving ontologically grounded semantic search in order to realize the retrieval of much more precise and pertinent information from texts related to elements of Ayurvedic literature. Query-based systems relying on keywords tend to fail to provide adequate results when dealing with vast amounts of raw data, and more so, within specific domains such as Ayurveda. Ayurveda is a domain with a rich vocabulary and complex linkages between a number of properties including, the doshas, the ailments, the cures, and the herbs. For this reason, the proposed framework utilizes a knowledge structure in the form of ontology as a core component for storing hierarchical relations within Ayurvedic information. Therefore, this ontology alleviates the problem of complexity of the data in Ayurveda and makes it possible for the search engine to 'understand' the user's query with an intelligence that grasps how different ayurvedic concepts are inter-related.

An integral part of the suggested framework is the Ontology-based Concept Extraction and Classification (OCEC) algorithm, which supports the processes of concept its recognition, extraction, and classification in the field of Ayurveda. OCEC advanced algorithm combines text mining abilities in order to fill up the prepared ontology with the relevant concepts and their relations drawn from a piece of text comprising Ayurvedic literature. This includes all terms describing the diseases, the treatments, the symptoms, the medicinal herbs and also how they are all interconnected in terms of which herbs relieve which symptoms and which symptoms are related to which dosha. The OCEC algorithm in performing both functions - structuring the Ayurvedic material and effectively indexing it so that the information presented to the user is as per his or her expectation. For example, if a user types ”Ayurvedic treatments for asthma”, the ontology-based search engine fetches results based on the indexing done by OCEC and presents users with relevant results, such as related herbs, (for example, Tulsi, and Licorice), treatment procedures such as Panchakarma and their indications such as respiratory dosha or symptoms. In such cases, the model includes an additional semantic layer that helps the system not to provide answers based only on key search words to the user but to provide more content and understanding of the Ayurvedic text in general.

Aaditi Narkhede et al.,[7] investigated the use of machine learning methods in order to develop and enhance existing Ayurvedic medicine recommendation systems. Ayurveda approaches health on an individualized basis, customizing all therapies to match the patient's specific body type, or prakriti, and the specific balance conditions residing in the physical and mental aspects of the person. Conventional techniques, however, greatly depend on the practitioners' subjectivity and the number of the available trained personnel. To solve these problems, Narkhede et al. outlined a treatment recommendation system, where machine learning algorithms such as decision trees and neural networks, were used to improve the treatment accuracy and reduce the treatment recommendation restrictions.

The approach begins with the training of different types of machine learning models using some historical patient information related to the demographics, health records, symptoms and treatment response to Ayurveda. Decision trees are applied in stratifying the patients and suggesting the appropriate treatment pathways based on the existing data patterns, and for more intricate databases, neural networks are utilized. Such neural networks can capture how various parameters of a patient relate to the treatments and their outcomes allowing the system to provide more holistic recommendations as prescribed in Ayurveda.

Simran Khan et al.,[8] presented a Ayurvedic medicines by implementing a knowledge graph together with the Hyperlink Induced Topic Search (HITS) algorithm to facilitate targeted treatment recommendations. This approach solves a problem that is significant in the context of Ayurvedic medicine practised today – recommending treatments that suit a patient's health status, that is not only current symptoms but medical history and other demographic factors. The knowledge graph is a database of Ayurveda lexicon in which various entities like herbs, therapies, doshas, symptoms, and illnesses have their associations. This helps the treatment system to understand treatment hierarchies and associations in Ayurveda as the concepts interact with one another creating this three-dimensional visio-graphic construal of their in-patient's profile. The system, therefore, improves its relevance within the context by making use of these user's health and demographic data in order to be able to better target the individual's treatment to be most appropriate for their body and mind. The incorporation of the HITS algorithm is a key aspect of the system as it impacts the decision of which treatments to offer so as to maximize patient outcomes. HITS allows for each treatment to be assigned a certain authority rating and a treatment position made as per its position in the knowledge graph. Those treatments that gain high authority treatment scores are likely to be perceived to be more general in practice and appropriate for use within Ayurveda

Building knowledge bases from Sanskrit literature is not without hurdles. More so in the case of Ayurveda since Sanskrit is a highly intricate language and a lot of technical information is present in these ancient medical texts. M. Harshini et al.,[1] took this problem into account with an emphasis on the Ayurvedic text Bhāvaprakāśanighantu: which is an ancient treatise rich in the description of medicinal plants and their uses. This text was annotated manually by the researchers to identify the entities in the text such as herbs, diseases, dosha, treatments and the relations among the entities and consequently a knowledge graph about Ayuvedic concepts was generated.In order to operationalize and simplify the access to this knowledge graph, the authors put together a detailed ontology which is a structural arrangement of elements with respect to the entities and their relationships in the domain of Ayurveda. The ontology serves as data standardization for the knowledge graph in order to depict relevant data as per Ayurvedic concepts. As a result, query templates were also created for the purpose of allowing the users to search the knowledge graph efficiently.

Expanding on the role of AI in Ayurveda, Sharma et al.[9] introduce AyUR-bot, an AI-driven chatbot designed to provide 24/7 guidance on Ayurvedic nutrition, remedies, and wellness. The chatbot integrates the Llama 2 Large Language Model (LLM) with Retrieval-Augmented Generation (RAG) to generate accurate and context-aware responses. By leveraging a specialized Ayurvedic document repository, AyUR-bot enhances accessibility and efficiency in delivering Ayurvedic knowledge to users. The system's accuracy is evaluated using BERTScore, demonstrating its ability to provide insightful and reliable responses. This chatbot represents a significant step forward in making Ayurvedic knowledge more accessible to the general public, bridging the gap between traditional wisdom and modern AI-driven healthcare solutions.

Kumar et al. [10] take a similar approach by developing a hybrid Ayurvedic drug recommendation system that centralizes diverse Ayurvedic knowledge sources to improve medication suggestions. Ayurveda's treatment methods involve a vast array of medicinal plants and formulations, making it challenging to determine the most suitable drug for a given condition. To address this, the proposed system integrates structured and historical Ayurvedic literature with generative AI models. By processing a centralized Ayurvedic data repository, the system refines drug recommendations, improving both accuracy and scalability. This AI-driven approach enhances clinical decision-making and supports practitioners in selecting the most effective Ayurvedic formulations for their patients.

Rahman et al. [11] further explore the integration of domain-specific knowledge into large language models (LLMs) for medical herb recommendations. Traditional AI chatbots often rely on generic internet-based knowledge, which may not be suitable for specialized domains like Ayurveda. To overcome this limitation, the proposed model fine-tunes Mistral 7B, a 7-billion parameter LLM, using academic journals on Indonesian medicinal herbs. The dataset is processed using LangChain, sentence transformers, and FAISS indexing, enabling efficient retrieval and similarity matching. This fine-tuned model ensures that the chatbot delivers Ayurvedic recommendations based on scientifically validated sources rather than general web-based information.

Overall, the literature demonstrates significant advancements in integrating AI, ML, knowledge graphs, and ontologies with Ayurveda. These technologies facilitate structured knowledge representation, improve information retrieval, and enhance personalized treatment recommendations. The combination of automated knowledge extraction, semantic search frameworks, AI-driven chatbots, and LLM-based recommendations is transforming the accessibility and accuracy of Ayurvedic knowledge.

## 3. PROPOSED SYSTEM

The proposed system combines Retrieval-Augmented Generation and a Knowledge Graph for semantic search and structured knowledge retrieval in Ayurvedic medicine, integrating flexible text retrieval with precise structured querying to address challenges in accessing diverse knowledge sources.

The system architecture Figure 1 comprises two parallel pipelines, each playing a crucial role in enhancing the accuracy, contextual relevance, and interpretability of the responses generated by the system:

1) **RAG Pipeline**: This pipeline is responsible for extracting and retrieving unstructured text from Ayurvedic documents, such as classical texts, research papers, and online repositories. It employs a Retrieval-Augmented Generation (RAG) approach, which first retrieves relevant textual information from the corpus using advanced search techniques and then integrates this information into a Large Language Model (LLM) to generate precise and context-aware responses[1].
2) **KG Pipeline**: The Knowledge Graph (KG) pipeline focuses on constructing and querying a structured knowledge representation of Ayurvedic concepts, including diseases, symptoms, treatments, herbs, and causal relationships. This structured ontology, built using Neo4j Aura, enables efficient semantic reasoning and relationship inference. Queries to the KG leverage Cypher queries, allowing the system to retrieve explicit and implicit knowledge based on structured Ayurvedic principles.

The outputs of both pipelines are synergistically combined to produce responses that are accurate, comprehensive, and contextually rich. The RAG pipeline provides broad textual context, capturing nuanced details from unstructured sources, while the KG pipeline ensures factual precision and structured insights. This hybrid approach empowers the Large Language Model (LLM) to generate responses that are not only factually accurate but also semantically consistent with the holistic and interconnected nature of Ayurvedic knowledge. Additionally, the system incorporates mechanisms for iterative refinement, allowing it to adapt to user feedback and evolving knowledge bases, thereby ensuring long-term relevance and scalability.

## 4. METHODOLOGY

### A. Dataset Description

The dataset used to develop and evaluate the system was derived from a carefully selected Ayurvedic text. From this book, fifteen chapters were chosen, each focusing on a distinct disease. These chapters provided a comprehensive range of conditions that encompass a variety of symptoms, treatments, and herbal remedies within Ayurvedic medicine.

This research work comprises unstructured Ayurvedic texts extracted from PDF documents, which include information on diseases, symptoms, causes, treatments, and medicinal herbs. The texts were preprocessed to remove noise and formatted for consistency. Key entities such as ”Disease,” ”Symptom,” ”Cause,” ”Treatment,” and ”Herb” were identified and structured into a Knowledge Graph. The dataset was further enriched with metadata, including page numbers, character counts, and token counts, to facilitate efficient retrieval and analysis

### B. Retrieval-Augmented Generation Pipeline

The Data Ingestion and Preprocessing phase begins with acquiring Ayurvedic texts, primarily in PDF format, which serve as the foundational knowledge source for the system. To extract text efficiently, the system utilizes PyMuPDF (fitz), a robust library designed for high-precision text retrieval from PDF documents. Once extracted, the raw text undergoes a series of preprocessing steps to enhance its usability.

This includes metadata extraction, such as identifying page numbers, counting words, and capturing other structural details that contribute to better document organization. Additionally, custom text formatting techniques are applied to eliminate noise, such as headers, footers, and other redundant elements, ensuring that the extracted content is clean and structured. The preprocessing pipeline also standardizes fonts, removes special characters, and ensures consistent spacing to improve downstream processing. Furthermore, domain-specific text cleaning is performed to filter out irrelevant content while preserving Ayurvedic terminology and contextual meanings.

Text chunking and embedding play a crucial role in processing and retrieving relevant information efficiently. Chunking involves splitting text into logical units, such as 8-sentence chunks, using the RecursiveCharacterTextSplitter to ensure coherent segmentation while preserving contextual meaning. Once the text is chunked, the embedding process converts these chunks into numerical vector representations using the all-mpnet-base-v2 sentence transformer model. These embeddings enable efficient similarity searches, allowing the system to retrieve the most relevant information based on user queries.

Semantic Search and Retrieval: The system processes user queries by embedding them into the same vector space as the document embeddings, ensuring a consistent representation for comparison. Using similarity matching techniques, it computes the cosine similarity between the query embedding and the precomputed chunk embeddings, identifying the most relevant information [2]. Finally, a Top-k retrieval mechanism selects and returns the most relevant

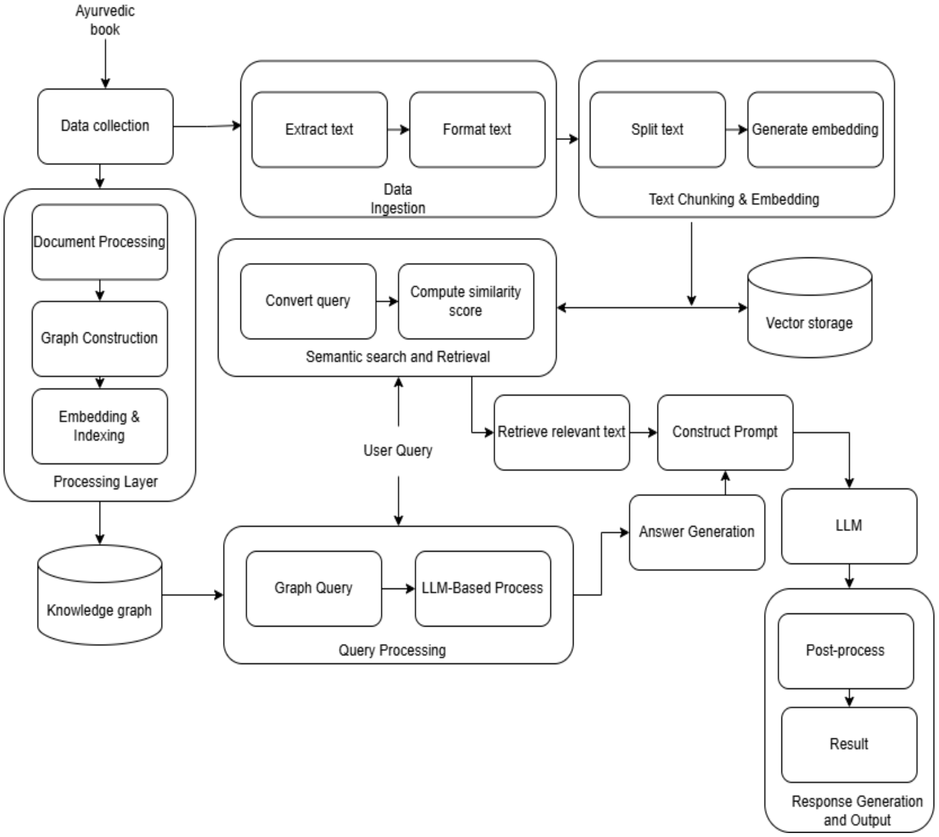

Figure 1. System architecture

text chunks, which are then used for context augmentation in generating accurate and meaningful responses.

To retrieve the relevant context, the algorithm follows the steps outlined in Algorithm 1. Additionally, a ranking and filtering mechanism is applied to refine the retrieved results, ensuring that high-confidence responses are prioritized. Future enhancements could include adaptive weighting techniques to dynamically balance semantic similarity with domain-specific constraints.

### C. *Knowledge Graph (KG) Pipeline*

Graph construction process involves defining key entities and their relationships to model Ayurvedic knowledge effectively. The primary entities include Disease, Symptom, Herb, Treatment, Cause, and Prohibited, each representing a crucial aspect of medical understanding. Relationships such as HAS SYMPTOM, TREATED BY, and CAUSED BY establish meaningful connections between these entities, enabling efficient querying and inference. For implementation, the system utilizes the Neo4j graph database for structured storage and retrieval of knowledge. Additionally, PyPDFLoader and RecursiveCharacterTextSplitter are employed for document processing, ensuring efficient extraction and segmentation of relevant information from Ayurvedic texts [12].

**Algorithm 1 Retrieval-Augmented Generation (RAG)**

```
procedure RAGPipeline( extitinput)
    input: extitAyurvedic texts (PDF format)
    output: extitAugmented responses using relevant extracted knowledge
    extittext ← extract_text( extitinput)
    extitmetadata ← extract_metadata( extitinput)
    extitclean_text ← format_text( extittext)
    for each extitchunk C in extitclean_text do
        extitchunks ← chunk_text( extitC, size=8 sentences)
        extitembeddings ← embed_text( extitchunks, model=all-mpnet-base-v2)
    end for
    extitquery_embedding ← embed_text( extituser_query, model=all-mpnet-base-v2)
    extitrelevant_chunks ← cosine_similarity( extitquery_embedding, extitembeddings)
    extittop_k_chunks ← select_top_k( extitrelevant_chunks, k)
    extitcontext ← aggregate( extittop_k_chunks)
    extitresponse ← generate_response( extitcontext, extituser_query)
    return extitresponse
end procedure
```

Hybrid Search mechanism enhances information retrieval by leveraging both structured and unstructured data sources. Embeddings are generated for 18 Knowledge Graph (KG) entities using HuggingFaceEmbeddings, enabling semantic similarity searches that go beyond keyword-based retrieval. Additionally, Query Translation plays a vital role in bridging natural language input with graph-based queries. This is achieved through GraphCypherQAChain, which converts user queries into Cypher queries, allowing efficient traversal and retrieval of relevant Ayurvedic knowledge from the KG.

To further enhance accuracy and contextual relevance, the system utilizes Groq Llama3-70B, a powerful language model capable of interpreting complex natural language inputs and generating precise Cypher queries for structured data retrieval. To construct the knowledge graph, the algorithm follows the steps outlined in Algorithm 2.

**Algorithm 2 Knowledge Graph Construction**

```
procedure GraphConstruction( extitinput)
  input: extitAyurvedic texts (PDF format)
  output: extitKnowledge graph with structured Ayurvedic information
  extittext ← extract_text( extitinput)
  extitclean_text ← format_text( extittext)
  extitchunks ← chunk_text( extitclean_text)
  extitgraph ← initialize_graph(Neo4j)
  for each extitentity E in {Disease, Symptom, Herb, Treatment, Cause, Prohibited} do
    add_node( extitgraph, extitE)
  end for
  for each extitrelationship R in {HAS_SYMPTOM, TREATED_BY, CAUSED_BY} do
    add_relationship( extitgraph, extitR)
  end for
  extitembeddings ← embed_entities( extitgraph, model=HuggingFaceEmbeddings)
  procedure HybridSearch( extituser_query)
    extitquery_embedding ← embed_text( extituser_query, model=HuggingFaceEmbeddings)
    extitrelevant_entities ← cosine_similarity( extitquery_embedding, extitembeddings)
    extitcypher_query ← translate_query( extituser_query, model=GraphCypherQAChain)
    extitresults ← execute_cypher( extitgraph, extitcypher_query)
    if extitresults are insufficient then
      extitenhanced_query ← refine_query( extituser_query, model=Groq Llama3-70B)
      extitresults ← execute_cypher( extitgraph, extitenhanced_query)
    end if
    return extitresults
  end procedure
end procedure
```

*D. Integration and Response Generation*

Prompt Augmentation: The system enhances the generated responses by combining retrieved text chunks from the RAG pipeline with structured query results from the Knowledge Graph (KG) pipeline, forming a comprehensive and contextually enriched prompt. This process ensures that the responses are both data-driven and semantically accurate. The augmented prompt follows a structured format where an instruction explicitly guides the model to generate answers based on the provided context [10]. The context section integrates relevant text chunks retrieved via the RAG approach alongside structured Ayurvedic knowledge obtained from the KG pipeline. Finally, the query section contains the user's question, ensuring that the response remains relevant and aligned with the original intent. The format follows: Instruction: Answer the Ayurvedic query using the context below. Context: [RAG chunks] + [KG results]. Query: [User question]. This structured approach enhances the LLM's ability to generate precise, context-aware, and semantically meaningful responses.

LLMInference: The system utilizes the google/gemma-2b-it model to generate responses by leveraging both retrieved unstructured text and structured knowledge from the KG. This ensures that the outputs are contextually relevant and semantically accurate. To enhance clarity and coherence, a post-processing step is applied, which removes redundant tokens, refines sentence structures, and ensures that the final response is concise, readable, and aligned with Ayurvedic principles [6].

By combining knowledge graph data with retrieval-augmented generation, the model delivers precise, context-aware, and semantically accurate responses grounded in Ayurvedic principles.

## 5. RESULTS

The system was evaluated using a carefully curated set of user queries pertaining to various aspects of Ayurvedic medicine, with a particular focus on comparing the performance of two distinct approaches: a Knowledge Graph (KG)-based Question Answering (QA) system and a Retrieval-Augmented Generation (RAG) system. This evaluation was designed to measure not only the lexical relevance but also the semantic accuracy and factual alignment of the generated responses, thereby assessing the practical utility of each system in addressing domain-specific informational needs.

The evaluation dataset comprised representative queries across multiple Ayurvedic categories, including diagnostics, etiology, pathogenesis, and treatment. Example queries included "What are the causes of varicose veins?", "What are the symptoms of hay fever?", "How can Pittavrita Vyana be prevented?", and "What is the treatment for dry hemorrhoid?". These queries were selected to span both modern medical terminology and classical Ayurvedic concepts, thus testing the systems' ability to handle translational and integrative knowledge tasks.

For each query, the QA pipeline was executed independently using both the KG based and RAG-based systems. The KG system generated answers by traversing curated graph triples derived from classical Ayurvedic texts, while the RAG system retrieved semantically relevant passages from the corpus and generated answers through an LLM.

To quantitatively assess the performance of each system, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric suite was employed—specifically ROUGE-1, ROUGE-2, and ROUGE-L. These metrics evaluate the n-gram overlap and longest common subsequence between the generated and reference answers, serving as proxies for relevance, fluency, and informativeness. Higher ROUGE scores indicate a closer match to the reference in terms of content and structure, providing an empirical basis for comparing the two approaches.

Detailed results for selected queries are presented in Tables I and II, which report ROUGE-1 and ROUGE-L scores, respectively. Table I presents the ROUGE-I scores of our 3 variations namely RAG-based, KG-based and RAG+KG (hybrid method). Similarly, Table II shows the ROUGE-L scores for our methods.

| **Medical Query** | **RAG** | **KG** | **Hybrid** |
|---|---|---|---|
| Causes of Varicose Veins | 0.6325 | 0.4719 | 0.4390 |
| Symptoms for Hay Fever | 0.2200 | 0.1905 | 0.3238 |
| Prevention of Pittavrita Vyana | 0.0847 | 0.5324 | 0.0992 |
| Treatment for Dry Haemorrhoid | 0.2353 | 0.1500 | 0.3516 |
| Prevention of Pramehaja Sarsapi | 0.0559 | 0.1333 | 0.1412 |
| Treatment for Allergic Asthma | 0.3409 | 0.1818 | 0.2051 |
| Causes of Pittaja-Granthi | 0.0968 | 0.3913 | 0.4056 |
| Recommended Diet for Dhatuksayaja Madhumeha | 0.1008 | 0.4122 | 0.2008 |
| Treatment for Migraine | 0.2388 | 0.5366 | 0.2500 |

TABLE I. ROUGE-1-Based Performance Analysis for KG, RAG, and Hybrid (RAG+KG) Systems

Key Takeaways: The RAG system excels in queries requiring fluent, context-rich answers, such as "Causes of Varicose Veins" and "Treatment for Allergic Asthma." The KG system provides structured, accurate responses and performs best in ontology-aligned queries like "Prevention of Pittavrita Vyana" and "Treatment for Migraine." The Hybrid system shows potential in balancing fluency and precision, outperforming in queries like "Causes of Pittaja-Granthi" and "Treatment for Dry Haemorrhoid," though its performance is inconsistent due to missing data in some cases (e.g., "Recommended Diet for Dhatuksayaja Madhumeha"). A refined hybrid approach, addressing data gaps and optimizing integration, could yield optimal results across a broader range of queries, particularly in specialized domains like Ayurveda.

| **Medical Query** | **RAG** | **KG** | **Hybrid** |
|---|---|---|---|
| Causes of Varicose Veins | 0.4786 | 0.4045 | 0.3293 |
| Symptoms for Hay Fever | 0.1600 | 0.1190 | 0.2286 |
| Prevention of Pittavrita Vyana | 0.0508 | 0.3597 | 0.0661 |
| Treatment for Dry Haemorrhoid | 0.1961 | 0.0500 | 0.2857 |
| Prevention of Pramehaja Sarsapi | 0.0559 | 0.0848 | 0.0941 |
| Treatment for Allergic Asthma | 0.2955 | 0.1212 | 0.1026 |
| Causes of Pittaja-Granthi | 0.0645 | 0.2029 | 0.2238 |
| Recommended Diet for Dhatuksayaja Madhumeha | 0.0672 | 0.3053 | 0.0672 |
| Treatment for Migraine | 0.1791 | 0.4878 | 0.2083 |

TABLE II. ROUGE-L-Based Performance Analysis for KG, RAG, and Hybrid (RAG+KG) Systems

### *A. Hybrid QA System Evaluation*

To leverage the strengths of both the Knowledge Graph (KG) and Retrieval-Augmented Generation (RAG) approaches, we experimented with a hybrid Question-Answering (QA) system.

In this setup, the primary answer is generated through the RAG model, which benefits from the flexibility and fluency of large language models, while the knowledge graph is used to refine or supplement the generated response with relevant triples, ensuring factual accuracy and domain-specific precision. This hybrid strategy aims to combine the best aspects of both approaches: the generative capabilities of RAG, which excels at understanding and creating coherent responses, and the structured, fact-based retrieval from the KG, which ensures that the information is contextually correct and relevant to Ayurvedic knowledge.

We tested this hybrid system using several representative queries from the Ayurvedic domain, focusing on questions involving both conceptual and factual complexities. The results, as measured by ROUGE scores, demonstrated modest improvements in cases where the accuracy of specific facts and domain knowledge was essential. For example, queries regarding Ayurvedic pathogenesis or treatment principles showed clear improvements, indicating that the hybrid system can better handle the nuanced, context-dependent nature of Ayurvedic medicine.

In one case, a query regarding the causes of varicose veins yielded promising results, with the hybrid system achieving a ROUGE-1 score of 0.439 and a ROUGE-L score of 0.329, outperforming the individual RAG and KG models when assessed for overall response quality and coherence. These results suggest that combining the generative strengths of RAG with the precise, domain-specific data from the knowledge graph offers a significant advantage in terms of response balance and depth. However, the improvements were more pronounced in complex queries and less so for simpler or more general questions. This finding underscores the potential of hybrid systems in specialized domains like Ayurveda, where domain-specific knowledge and context are critical, but also highlights the need for further fine-tuning to optimize performance across a broader range of queries.

The observed performance improvements support the hypothesis that hybrid systems, which combine generative and retrieval-based techniques, can provide more accurate and contextually rich responses in complex, domain-specific fields. While the current results are promising, further refinement and evaluation with a broader set of queries will be necessary to fully realize the potential of hybrid QA systems in areas such as Ayurveda, where both precision and contextual relevance are paramount.

### B. Limitations

While evaluation metrics such as ROUGE provide useful insights into the syntactic overlap between generated and reference responses, they do not fully capture semantic accuracy, contextual appropriateness, or the clinical validity of the answers—especially in a nuanced domain like Ayurveda. Therefore, relying solely on such metrics can overlook subtle but important aspects of response quality.

Another significant limitation lies in the system's reliance on the quality of the underlying embeddings and the structure of the Knowledge Graph. Embeddings trained on general language models may lack the domain-specific sensitivity required for Ayurvedic terminology, metaphors, and contextual usage. Similarly, the effectiveness of the Knowledge Graph depends heavily on the accuracy and completeness of the ontology design, entity relationships, and the granularity of the stored data.

Furthermore, the system does not yet incorporate temporal, geographical, or personalized factors, which can be crucial in Ayurvedic diagnosis and treatment. For example, seasonal changes, user constitution (*Prakriti*), and local environmental factors are often key to accurate recommendations in Ayurveda but are currently not accounted for.Lastly, user interaction is still relatively static, with limited support for clarification, dialogue-based follow-ups, or handling of ambiguous queries. These factors collectively suggest that while the system shows potential, significant scope remains for improvement in terms of coverage, adaptability, and intelligent interaction.

### C. Advantages

The integration of symbolic and neural reasoning enables the hybrid QA system to address complex, domain-specific queries effectively. By combining a structured knowledge graph with neural retrieval and generation capabilities, the system is capable of bridging the gap between formalized Ayurvedic knowledge and the variability of human language.

The knowledge graph ensures that the system's responses are grounded in authoritative Ayurvedic knowledge, preserving the integrity and authenticity of traditional medical concepts. It allows for precise semantic reasoning, where relationships among diseases, symptoms, treatments, and causes are explicitly represented and exploited for accurate query resolution.

Simultaneously, the RAG (Retrieval-Augmented Generation) pipeline enhances the system's flexibility in understanding and interpreting diverse natural language queries. By leveraging a large language model (LLM) with retrieval mechanisms, the system can contextualize and synthesize information from multiple sources, even when queries are vague, implicit, or colloquially phrased.The architecture is inherently scalable, allowing for the easy incorporation of new texts, treatments, and ontological extensions without requiring major structural changes. This extensibility supports continuous evolution of the system in line with new research findings, user feedback, or updates to Ayurvedic literature.At the same time, the LLM ensures that the final responses are not only factually accurate but also natural, coherent, and semantically rich. This makes the system accessible to a broader audience, from practitioners and researchers to students and general users, by translating complex knowledge into clear and engaging language.

Furthermore, the semantic search capabilities enabled by the knowledge graph provide more than just direct answers—they support exploratory and inferential interactions. Users can navigate the interconnected concepts within the domain, uncover hidden relationships, and pursue deeper understanding through guided reasoning paths. This positions the system as a powerful tool for educational enrichment, clinical decision support, and advanced Ayurvedic research.

### D. Workflow Summary

When a user inputs a natural language query, such as "causes of Varicose Veins," the system initiates a dual-pipeline architecture to derive a comprehensive, semantically enriched response. This architecture integrates the flexibility of Retrieval-Augmented Generation (RAG) with the precision of a Knowledge Graph (KG), combining unstructured textual evidence with structured domain knowledge in Ayurveda.

The RAG pipeline begins by converting the query into a dense vector representation using a transformer-based embedding model. This embedding is used to perform a semantic similarity search over a corpus of curated Ayurvedic documents, extracting contextually relevant text chunks—even in cases where the user's wording does not directly match document phrasing. This mechanism ensures high recall and contextual alignment, leveraging the rich language and subtle nuances found in classical Ayurvedic literature.

In parallel, the KG pipeline processes the query through a structured reasoning layer. The natural language input is parsed and transformed into a Cypher query using intent classification and entity-relation mapping, tailored to the Ayurvedic ontology. For instance, the query "What are the causes of Varicose Veins?" is mapped to entities (e.g., *Varicose Veins*) and relationships (e.g., *hasCause*) and translated into a Cypher statement. This query is then executed against a Neo4j-based knowledge graph, extracting concise, fact-based results rooted in domain-specific hierarchies and interconnections.

As an example, the KG might identify causes such as Weak Nervous System, Blood Loss, Blocked Circulation, Weak Digestion and Malnutrition, Pregnancy, Menopause, Heavy Menstruation, Prolonged Hanging of Legs, Old Age, and Emotional Factors like Anxiety and Fear. Each of these causes is annotated with domain-informed descriptions, citing sources or expert validation where available. This structured representation provides transparency, traceability, and depth, acting as a curated backbone to support decision-making or further exploration.

The outputs from the RAG and KG pipelines converge in the synthesis module powered by a Large Language Model (LLM). This model harmonizes the broad, expressive content from unstructured text with the precise, authoritative data from the knowledge graph. The LLM integrates and contextualizes both sources, ensuring that the final output is coherent, semantically grounded, and aligned with traditional Ayurvedic principles.

This fusion of symbolic and sub-symbolic AI enables the system to respond not just with relevant information, but with interpretative depth—bridging the gap between ancient textual wisdom and modern information retrieval technologies. The end result is a response that is not only accurate and comprehensive but also explainable and culturally resonant. It supports a diverse user base, including Ayurvedic practitioners, researchers, and patients, by enhancing both usability and reliability of the information presented.

## 6. CONCLUSIONS

This research work proposed a semantic search system for Ayurvedic Medicine by integrating Retrieval Augmented Generation (RAG) and Knowledge Graph (KG) technologies to enhance information retrieval and accessibility. Traditional keyword-based search methods often fail to capture the deeper semantic meaning of Ayurvedic texts, leading to inefficient and inaccurate search results. By leveraging RAG and KG, the system overcomes these limitations by enabling semantic understanding, structured knowledge retrieval, and context-aware responses. It effectively bridges the gap between unstructured Ayurvedic texts and structured knowledge, ensuring that users— including practitioners, researchers, and the general public—can access precise and relevant information with ease. This approach not only enhances the discoverability of Ayurvedic insights but also supports evidence-based decision-making by providing well-structured and contextually enriched responses.

While the current system demonstrates promising results, several enhancements can further improve its capabilities. One key area for improvement is Advanced AI Integration, where fine-tuning domain-specific large language models (LLMs), such as Mistral 7B, on Ayurvedic datasets can enhance response quality and contextual understanding. Another critical enhancement is Enhanced Query Processing, which involves implementing multi-hop reasoning in the knowledge graph to handle complex queries, such as "Which herb treats cough caused by cold weather?" This would significantly improve the depth and accuracy of responses. Additionally, Real-World Deployment and Validation is essential to ensure the system's accuracy and usability. Collaborating with Ayurvedic practitioners for validation and conducting user studies to measure retrieval performance and response relevance would provide valuable feedback for further refinement. Finally, ensuring Sustainability and Scalability is crucial for long-term efficiency. Optimizing computational performance through techniques like LLM quantization can reduce resource consumption while maintaining high-quality outputs, making the system more scalable for widespread adoption.

## References

[1] M. Pavithra, M. Harshini, P. S., R. Deepalakshmi, and R. Vijayalakshmi, "Ayurvedic elixirs for digital generation using machine learning," *Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, pp. 1–7, 2024.

[2] M. Gayathri, "A framework for ontology based semantic search system in ayurvedic medicine," *Computational Intelligence and Machine Learning*, vol. 2, pp. 1–5, 2021.

[3] M.Gayathri, "Ontology based concept extraction and classification of ayurvedic documents," *Procedia Computer Science*, pp. 511–516, 2020.

[4] Paneru, Biplov, B. Thapa, and B. Paneru, "Leveraging ai in ayurvedic agriculture: A rag chatbot for comprehensive medicinal plant insights using hybrid deep learning approaches," *Telematics and Informatics Reports*, pp. 1–4, 2024.

[5] H. Terdalkar, A. Bhattacharya, M. Dubey, S. Ramamurthy, and B. N. Singh, "Semantic annotation and querying framework based on semi-structured ayurvedic text," *World Sanskrit Conference*, pp. 155–173, 2023.

[6] L. S. Yumnam, A. Jain, C. Usha G, and P. D. Cyril, "Exploring ayurvedic medicine recommendation using machine learning techniques," *IJFMR*, vol. 6, pp. 1–6, 2024.

[7] A. Narkhede, “Knowledge graph and similarity based retrieval method for query answering system,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 9, pp. 1–9, 2022.

[8] S. Khan, A. Saify, S. Gosaliya, D. Jain, and J. Zalte, “Medimatch: Ai-driven drug recommendation,” *2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, pp. 1342–1349, 2024.

[9] R. Patil, S. Yeolekar, O. Khade, Y. Kadam, P. Ingole, and S. Patil, “Ayur bot: A revolutionary ayurvedic chatbot empowered by generative ai,” *8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pp. 1–6, 2024.

[10] N. Singh, A. Sah, S. M. Vartika, V. Malik, R. Sobti, and B. Brahma, “A hybrid ayurvedic drug recommendation system with generative AI,” *Innovative Computing and Communications (ICICC 2024)*, vol. 1020, pp. 1–3, 2024.

[11] D. Firdaus, I. Sumardi, and Y. Kulsum, “Integrating retrieval-augmented generation with large language model mistral 7b for indonesian medical herb,” *JISKA (Jurnal Informatika Sunan Kalijaga)*, pp. 230–243, 2023.

[12] L. Li, P. Wang, J. Yan, Y. Wang, S. Li, J. Jiang, Z. Sun, B. Tang, T.-H. Chang, S. Wang, and Y. Liu, “Real world data medical knowledge graph: construction and applications,” *Artificial Intelligence in Medicine*, vol. 103, pp. 1–5, 2020.