

A Hybrid Retrieval Augmented Generation Approach for Explainable Clinical Decision Support Using Vector Databases and Best Matching 25

Bodireddy Mahalakshmi

mahalakshmi.bodireddy@pccoepune.org

Pimpri Chinchwad College of Engineering

Smita Khairnar

Pimpri Chinchwad College of Engineering

Shilpa Gite

Dr. D. Y. Patil University

Anurag Lengure

Pimpri Chinchwad College of Engineering

Anuj Loharkar

Pimpri Chinchwad College of Engineering

Swaraj Mahadik

Pimpri Chinchwad College of Engineering

Soham Mahajan

Pimpri Chinchwad College of Engineering

Research Article

Keywords: BioBERT, Clinical Decision Support System, Healthcare AI, Hybrid Search, Retrieval-Augmented Generation, Vector Database

Posted Date: June 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9787855/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Hybrid Retrieval Augmented Generation Approach for Explainable Clinical Decision Support Using Vector Databases and Best Matching 25

Bodireddy Mahalakshmi^{1*}, Smita Khairnar¹, Shilpa Gite²,
Anurag Lengure¹, Anuj Loharkar¹, Swaraj Mahadik¹,
Soham Mahajan¹

^{1*}Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India.

^{2*}School of Science and Technology, Dr. D. Y. Patil University, Pune, India.

*Corresponding author(s). E-mail(s):

mahalakshmi.bodireddy@pccoepune.org;

Contributing authors: smita.khairnar@pccoepune.org;

shilpa.gite@dpu.edu.in; rajesh.anurag22@pccoepune.org;

anuj.loharkar22@pccoepune.org; swaraj.mahadik22@pccoepune.org;

soham.mahajan22@pccoepune.org;

Abstract

Clinical decision making in modern healthcare industry faces problems due to the growing amount of medical knowledge, scattered data sources, and the lack of patient centered information. Traditional Clinical Decision Support Systems heavily depend on rules and keyword searches, which always lack an understanding of context and personalization. This paper puts forth a Clinical Decision Support System that is assisted by Retrieval Augmented Generation (RAG) method. It fuses Electronic Health Records with unstructured medical literature. The system uses a hybrid search approach that fuses BM25 lexical search with BioBERT based semantic search. The results are combined using Reciprocal Rank Fusion. Additionally, the system has risk prediction for early disease diagnosis. Experimental evaluation of the retrieval system on roughly 182,000 medical records achieved a Precision@5 of 0.81, a Recall@10 of 0.85, and an MRR

of 0.83, surpassing BM25 independent and dense baselines by up to 9.5%. This framework boosts quality of evidence and improves reasoning based on individual patients. It helps make timely decisions and increase reliability in complex, high risk situations.

Keywords: BioBERT, Clinical Decision Support System, Healthcare AI, Hybrid Search, Retrieval-Augmented Generation, Vector Database

1 Introduction

The healthcare industry is generating an exponentially increasing volume of medical literature, research results, and guidelines, causing cognitive overload for medical practitioners in searching for relevant information during decision making [1]. Traditional keyword dependent search tools are not quite good in identifying the semantic relationships between medical terms and may produce useless results or miss critical context.

Clinical Decision Support Systems (CDSS) are basically used to help medical practitioners in making diagnoses and treatment decisions. Current systems heavily depend on static rule based algorithm or keyword based searches. Although these approaches provide structured results, they lack the ability to recognize context or handle complex medical queries effectively.[2] This makes them less useful. Consequently, today’s CDSS systems face difficulties due to the fragmented nature of data sources from research databases, clinical guidelines, and electronic health records (EHRs). They are also faced with a lack of understanding of meaning, rapidly evolving medical knowledge, and a lack of insights into specific patients.

To overcome these drawbacks, this paper proposes a Retrieval Augmented Generation (RAG) [3] based Clinical Decision Support System that combines hybrid information retrieval and large language model generation for providing accurate, interpretable, and contextually aware medical responses.

The main contributions of this work are summarized as follows:

1. A hybrid retrieval framework integrating BioBERT based dense semantic search with BM25 lexical retrieval using Reciprocal Rank Fusion (RRF).
2. A dual interface architecture supporting both general medical research and patient specific therapeutic optimization.
3. An automated risk classification and alerting mechanism for early identification of critical clinical conditions.
4. Integration of large medical knowledge sources, including United States Medical Licensing Examination textbooks, PubMed and ArXiv publications, and WHO,FDA clinical guidelines.

Following this introduction, the arrangement of the research paper is explained below. Section 2 gives a look at existing studies regarding retrieval augmented generation and clinical decision support systems. Because the research needs clarity, Section 3 shows the suggested structure, the design of the system, information origins, and the mixed search process. Many people believe that the test arrangement and the outcomes of the assessment are explained in Section 4. In Section 5, the primary observations and useful thoughts are examined. Section 6 finishes the research paper and points to the path for later investigations. Declaration Section states all the required declarations of our work followed by the data availability statement.

2 Related Work

Recently, several recent studies involving RAG, multimodal learning [4], and AI-based predictive approaches have been conducted on Clinical Decision Support

Systems (CDSS) for improving access to and use of clinical information. Specifically, a multimodal retrieval augmented generation approach for accessing textual and visual biomedical data was proposed by Lahiri and Hu (2024). Zero-shot Retrieval-Augmented Generation has proved efficient in increasing factual accuracy and reducing hallucinations when summarizing unstructured electronic health records in aged-care settings, but Alkhalaf et al. [4] approach is confined to disease-specific literature without patient-level integration. In this case, the model achieved a reduction in documentation burden but lacked diagnostic reasoning and integration with external medical knowledge [5]. Furthermore, the study by Alkan et al. [6] involved an analysis of AI-driven clinical decision support systems with respect to ethical, explainability and validation issues, but it lacks an effective operational framework. On the other hand, Thwal et al. [7] designed a federated learning-based personalized CDSS to solve the privacy problem in clinical settings; however, this model did not include evidence-based reasoning and semantic retrieval of biomedical knowledge. Additionally, models developed using deep learning and multimodal physiological signals have proved effective in predicting outcomes in clinical applications; however, the models work as standalone prediction approaches in such cases without any integration in CDSS workflows [8, 9].

Several studies have explored retrieval, generation, personalization, and prediction as distinct modules of Clinical Decision Support Systems (CDSS). Recent developments in Retrieval-Augmented Generation (RAG) [6], multimodal learning, and AI-based predictive approaches indicate some positive trends. Nonetheless, many of the existing approaches are limited to disease-specific literature or unstructured clinical notes, without much patient-level information incorporated into them. Although privacy preserving models like federated learning have played an important role in addressing the issue of data security, they do not incorporate any semantic retrieval or evidence-based reasoning approaches [7]. All of these limitations reflect a significant research gap in building an effective clinical decision support solution that incorporates patient data, biomedical knowledge, and explainable reasoning.

To address the limitations mentioned above, this study suggests the design of a unified hybrid architecture based on retrieval and RAG. It will utilize semantic/lexical search, patient-level modeling, and multimodal learning. Moreover, the suggested solution will make use of both patient-specific data as well as external biomedical literature for providing personalized clinical recommendations. Through the combination of predictive modeling and generation with knowledge-awareness and explainability approaches, it will provide an end-to-end solution for CDSS.

3 Data and Methodology

The upcoming few subsections provide a description of the dataset used along with an explanation of the designed methodology, including data preprocessing, design, and evaluation procedures.

3.1 Architecture

The proposed design can be divided into three-tier modular architecture, as shown in Fig. 1, for scalable medical knowledge retrieval and patient-specific decision-making. It is comprised of three primary layers: Data Ingestion, Hybrid Retrieval, and AI Generation.

Data Ingestion Layer consolidates heterogeneous medical information sources including research literature, clinical guidelines, pharmaceutical database, medical textbooks, and Electronic Health Records (EHR). The ingested texts are standardized into one single format for semantic access across structured and unstructured data. Hybrid Retrieval Layer uses a combination of semantic similarity search with lexical

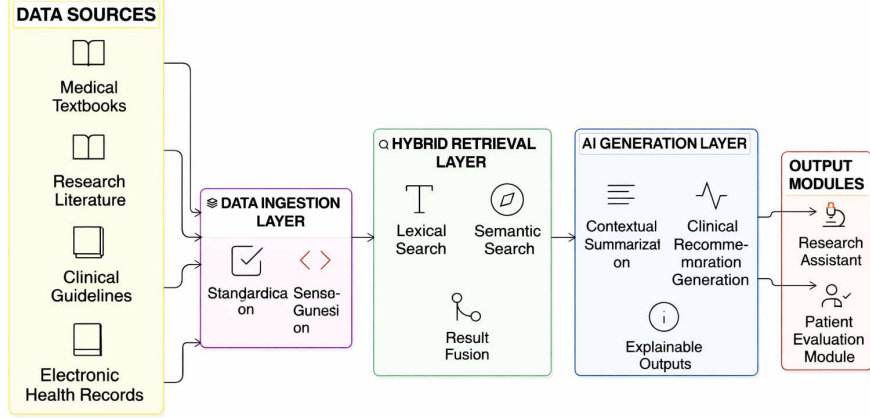


Fig. 1 High level 3 tier workflow of the designed system.

matching that enables robust information discovery in different clinical queries.

AI Generation Layer leverages retrieved texts for producing context-aware summaries and clinical suggestions; the output is generated from the source materials to ensure factual correctness and transparency.

The architecture facilitates two key workflows: (i) research assistant module to explore literature evidence, and (ii) patient evaluation module to fuse clinical records with retrieved knowledge and generate personalized treatment guidance, risk assessment, and monitoring recommendations. Individual component implementation details are elaborated in the following sections.

Figure 2 shows the secured personalization pipeline created in the above-described architecture, enabling patient-specific inference. The process starts with the clinician authentication through OAuth-based login, followed by the validation of JSON Web Token (JWT) and session creation.

A dedicated Security Layer enforces session management, token expiration, and response boundaries on the basis of roles to control access to the system. After being authenticated, clinicians are sent to the Profiling Engine where their structured profiles are generated. These profile attributes and session data are stored in MongoDB.

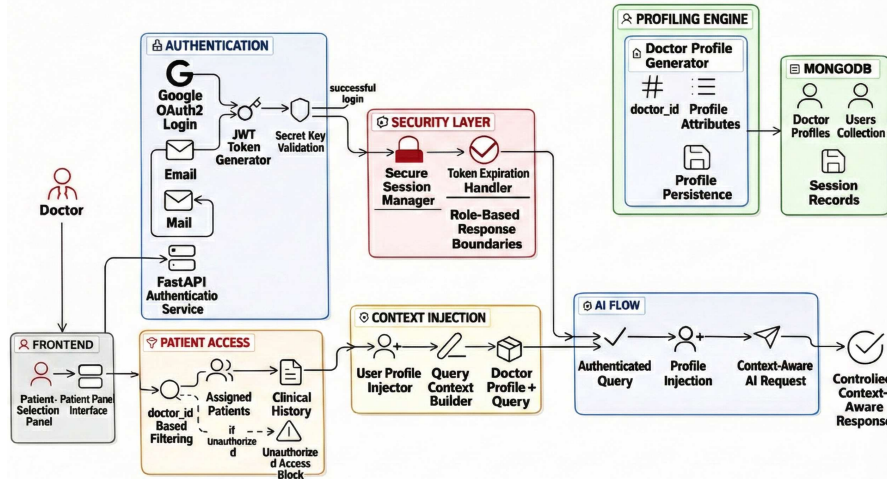


Fig. 2 Architecture of the secure personalization and AI reasoning pipeline.

Patient Access Control filters the input request on the basis of clinician Id to restrict unauthorized access and allows doctors to fetch clinical history information only for the assigned patients.

Context Injection Module integrates clinician profiles, authorized patient data, and the query to create contextualized final queries. Final queries are then passed to AI Model Inference.

3.2 Data Sources and Vector Database Infrastructure

In this section, we outline the various data sources used and the vector database infrastructure developed to perform scalable semantic retrieval. The objective is to create an integrated medical knowledge repository by consolidating the knowledge from diverse trusted sources and providing access based on semantic similarity.

Table 1 Summary of integrated data sources

Data source category	Data points	Origins
Medical textbooks	~126,000	[10]MedRAG (HuggingFace)
Research literature	~3,000	PubMed[11], arXiv[12], bioRxiv[13]
Guidelines & drug data	~1,000	WHO[14], FDA[15], RxNorm[16]
StatPearls[17] NCBI Bookshelf	~52,000	Clinical Reviews
Total vector records	~182,000	—

The medical knowledge sources used in our pipeline are provided in Table 1. In contrast to randomly collecting data, the process of semantic search follows a disease-based methodology. First, a list of clinically relevant diseases such as diabetes, cancer,

hypertension, etc., is considered, and information pertaining to them is gathered from different sources. In addition to literature and research articles, other important clinical knowledge repositories including essential medicine list are also collected. PubMed data is obtained via NCBI E-utilities API e-util. For each disease, the corresponding PubMed article ids are first retrieved, and then their complete data is obtained. Similar techniques are used to obtain papers from arXiv and bioRxiv. For these articles, critical fields like titles and summaries are retrieved via XML/JSON parsing. In addition to literature data, clinical information is also taken into consideration. Data about drugs is acquired from the FDA API and RxNorm, whereas the WHO web site is scraped to obtain clinical guidelines and the essential medicine list. Embeddings are created using BioBERT [18] through a SentenceTransformer implementation. It is worth noting that BioBERT [18] is fine tuned for semantic similarity purposes and thus is a good choice for creating embeddings for biomedicine. Every article is converted into a 768-dimensional vector that captures its meaning. In contrast to other methods which use chunked versions of text data, in our implementation, embeddings are created using full document texts. The embeddings thus generated are stored in a vector database named Qdrant. Cosine similarity is adopted as the distance measure in the Qdrant vector database. An entry in the vector database consists of an Id, a vector, and the payload consisting of the text of the document. The Qdrant database internally employs the approximate nearest neighbor (ANN) search using Hierarchical Navigable Small World (HNSW) graph. Processing and storage of documents happen in batches in order to prevent reaching the limits imposed by the API and reduce overhead during indexing. The query submitted by the user is embedded using BioBERT [18] model and the vector is compared against the entries in the database. In addition, a combined keyword/semantic search is performed using BM25 as the scoring function. This overall pipeline allows efficient and context aware information retrieval and fuses easily with components that assist us ahead, as shown in Fig. 3.

3.3 Methodology

Description of the complete pipeline of the proposed framework is included in the subsection. The proposed pipeline aims at reducing information loss and keeping explanations as the data moves from one step to another.

Different tasks have been assigned to each of these steps. Semantic encoding detects contextual information; hybrid retrieval achieves high accuracy and precision, whereas generation generates justified structured output.

3.3.1 Query Processing Pipeline

Query processing consists of a set of processing operations that convert input data into dense semantic vectors employed for retrieval purpose. To achieve this objective, the input to our model is first tokenized and embedded using BioBERT [18], whose output embeddings (context embeddings) are pooled through averaging technique to produce the following dense query:

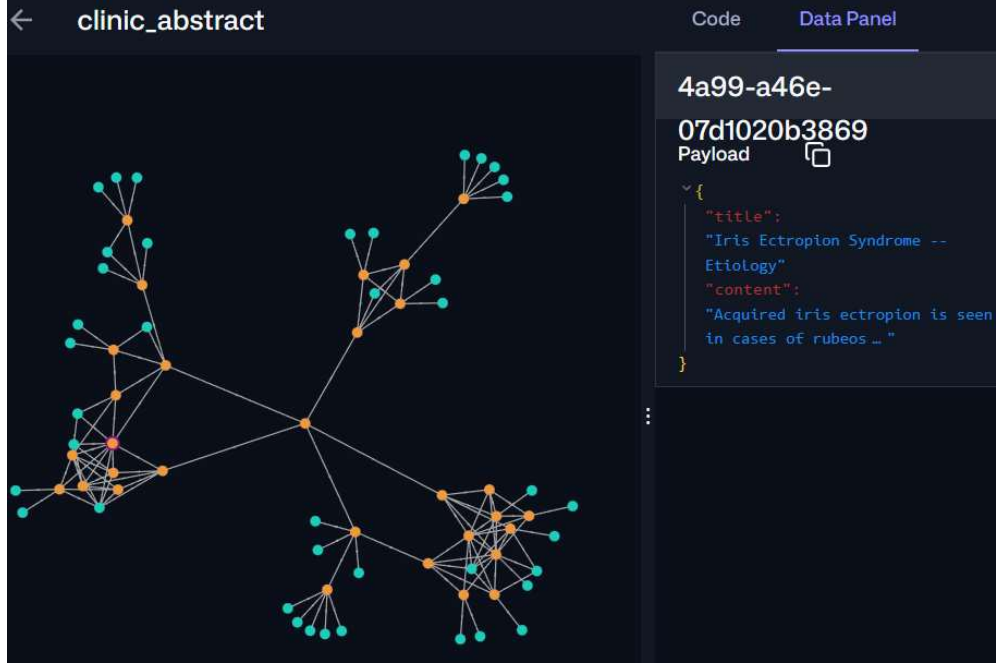


Fig. 3 Dense Vector Representation.

$$E_q = \frac{1}{n} \sum_{i=1}^n h_i \quad (1)$$

Here, h_i is the context embedding of i^{th} token and n is the number of tokens in the query. This process provides us with 768-dimensional vectors which are further used for query representation in retrieval task.

As compared to keyword representations, the major benefit of utilizing the BioBERT embeddings for representing our query is that it allows preserving domain-specific semantic relationships like synonyms (e.g., “myocardial infarction” vs. “heart attack”).

Moreover, query preprocessing such as token conversion to lower case, denoising, and medical term normalization further helps us improve embedding generation quality and subsequent retrieval process.

3.3.2 Hybrid Retrieval with Reciprocal Rank Fusion

A hybrid strategy combining both approaches is adopted to provide a solid retrieval context. The user query is represented using BioBERT and compared against the vector embeddings stored in Qdrant using cosine similarity metric, depicted in Fig. 3 below:

$$CosSim(x, y) = \frac{x \cdot y}{||x||, ||y||} \quad (2)$$

While cosine similarity performs the retrieval by leveraging semantic proximity, instead of exact lexical match, which is suitable in case of clinical queries where terminology can be different. Furthermore, since we used Qdrant, it guarantees Approximate Nearest Neighbor (ANN) Search for scalability in large-scale medical data.

On the other side, BM25 keyword-based retrieval helps to find an exact medical terminology, which is important in cases when there is a necessity to retrieve exact keywords, which should not be semantically generalized. The resulting relevance scores are combined in the following manner:

$$Score(d) = \alpha \cdot BM25(d, q) + (1 - \alpha) \cdot CosSim(E_q, E_d) \quad (3)$$

where α is the contribution ratio of sparse/dense retrieval. It enables tuning the importance of precise keyword matches (BM25) versus semantics similarity (cosine), according to the nature of the query.

Furthermore, in order to increase the quality of the ranking, fusion is performed using Reciprocal Rank Fusion (RRF) [19], given as:

$$RRF(d) = \sum_{i=1}^N \frac{1}{k + r_i(d)} \quad (4)$$

where $r_i(d)$ is a rank of document d obtained with each method and k is a smoothing constant ($k = 60$). If the document appears on both list, its combined scores will be summed, and the ultimate ranking will be computed on the combined score.

Reciprocal Rank Fusion technique is highly useful in solving inconsistency among ranks produced with different retrieval models, as it favors those documents that have a consistent rank across all retrievers rather than relying on absolute score.

3.3.3 Response Generation and Patient Specific Optimization

As illustrated in Fig. 4, the retrieved documents are injected into a language large model (LLaMA-3.3-70B) running on the Groq Language Processing Unit (LPU) to output a context-aware response in the predefined format.

$$Response = LLM(q, D_{retrieved}) \quad (5)$$

Here q refers to the user’s query while $D_{retrieved}$ refers to the top- k medical documents that were retrieved.

The context-aware response generation is aided by the retrieved documents in order to minimize hallucination while ensuring the factual accuracy of the answers. The retrieved documents are gathered together in the context window in order to assist the language model to generate answers based on reliable medical resources. This way, the proposed method is significantly more reliable than independent generative models.

In the case of patient-specific questions, the patient profile, treatment history, and medical guidelines are injected into the generative model based on which the patient will receive recommendations regarding his/her medical needs. These recommendations include suggestions regarding the type and amount of medication to be taken, duration of treatment, monitoring parameters, and risk level (minor, medium, or critical).

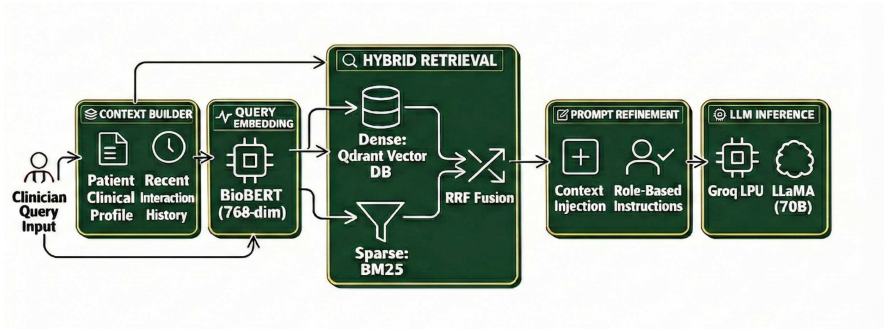


Fig. 4 Context aware response generation flow in the proposed RAG framework.

Context injection enables patient-specific attributes to be added into the prompt in a structured way.

3.3.4 Risk Classification and Alert Mechanism

The platform has an automated risk classification mechanism, which will evaluate the generated response to detect risk levels. Indicators for critical risk levels would be terms like “life-threatening”, “severe”, “urgent”, and “emergency”; whereas “moderate”, “monitor closely”, and “follow up required” would indicate medium-level risks.

The minor, medium, and critical risk levels will serve as the decision-making trigger for notifications to patients and clinicians. The risk detection is done through keywords recognition and contextual understanding of the generated response.

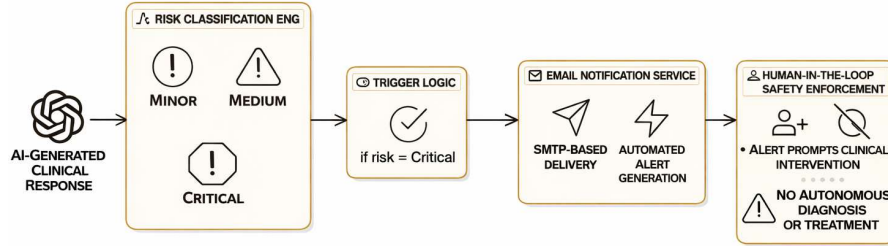


Fig. 5 Critical Risk Mailing Logic.

System allows automated alert generation using severity evaluation, allowing early identification of high risk clinical scenarios. As illustrated in Fig. 5 by adding risk signals with generated recommendations, the framework helps carry out timely intervention.

4 Experimental Evaluation

The system provides mechanisms for generating alerts through severity detection of patient conditions. As depicted in Fig. 5, the inclusion of risk and recommendation generation will help facilitate early interventions.

4.1 Retrieval Performance

The suggested hybrid retrieval method has been evaluated using Precision@5, Recall@10, and Mean Reciprocal Rank measures against the sparse (BM25) and dense baselines. The Reciprocal Rank Fusion (RRF) [19] based method has been observed to achieve better results than the individual retrieval methods using all the measures, as shown in Table 2.

Compared to the best performing baseline model, our hybrid model achieves statistically significant improvements ($p < 0.05$), thus validating the use of both lexical and semantic retrieval cues for knowledge intensive medical IR tasks.

Table 2 Retrieval performance comparison

Method	P@5	R@10	MRR
BM25 only	0.68	0.72	0.71
Dense only	0.74	0.79	0.76
Hybrid (RRF)	0.81	0.85	0.83
Improvement	+9.5%	+7.6%	+9.2%

Table 2 compares the retrieval performance of different methods.

4.2 Response Quality Evaluation and Confidence Scoring

Generated responses were evaluated across qualitative dimensions: accuracy, relevance, medical correctness, alignment to user intent, coverage of key clinical aspects.

A confidence score is calculated by combining normalized relevance scores from dense semantic retrieval (Qdrant) and sparse lexical search (BM25), as shown in 6:

$$Confidence(d) = \alpha \cdot Qdrant_{norm}(d) + (1 - \alpha) \cdot BM25_{norm}(d) \quad (6)$$

where $Qdrant_{norm}(d)$ denotes normalized cosine similarity, $BM25_{norm}(d)$ represents normalized lexical relevance, and α is empirically set to 0.7. The final scores are used to rank documents before output generation and this provides clinicians a robust and clear reliability indicator to support evidence based verification.

5 Discussion

The results show that the fusion of domain-specific embeddings (from word2vec) into a retrieval-based model significantly improve reliability and contextual understanding of clinical information. In addition, the incorporation of BioBERT as opposed to standard BERT-based models; and, the aggregation of several medically-related data sources, result in improved ability to understand both the semantic meaning and the domain-specific relevance when compared to traditional keyword-based methods.

One of the most significant advantages of this project is its inclusion of diverse data sources within a singular knowledge-base. By integrating textbooks, research articles,

clinical practice guidelines, and drug databases into one comprehensive knowledge base, the system provides an extremely well-rounded and functional knowledge base. Additionally, the focus on collecting data from a disease-centric perspective ensures that the resultant dataset remains clinically relevant and focused.

The use of full-document embeddings in this system is another advantage. As opposed to many other systems, which utilize small blocks of embedded chunks of content to represent documents, this method allows for the preservation of full context within each piece of medical text.

In terms of scalability, the utilization of a vector database combined with ANN-assisted indexing allow for efficient scaling of the system, while maintaining minimal performance degradation.

However, this architecture does have several trade-offs. The use of full-document embeddings can lead to less precise retrieval details; and, as such, dense retrieval may fail to retrieve certain cases in which exact keyword matching is necessary. Furthermore, the generation of embeddings via BioBERT can be computationally expensive, which could impact real-time performance unless some form of hardware acceleration exists.

6 Conclusion

This paper describes a clinical decision support system (CDSS) that employs a retrieval augmented generation (RAG) technique to tackle the challenges in clinical reasoning and medical information retrieval. The proposed system offers precise and interpretable responses to clinical queries. It achieves this by BioBERT embeddings, hybrid search(BM25), semantic search with Reciprocal Rank Fusion, and LLM.

The proposed system integrates over 82,000 points of medical information from different sources. An automated risk categorization and notification system enhances clinical safety. Moreover, hybrid search offers a 9.5% boost in precision over single strategy.

The primary contribution of this study comes from its capacity to integrate lexical and semantic search strategies, which leads to enhanced evidence coverage and reduced information loss, a critical requirement in clinical aspects that deal with life threatening scenarios. The architecture demonstrates that RAG and hybrid retrieval systems, when combined with patient context, create a scalable AI decision support system, which provides clear explanations and practical applications for medical use, establishing a foundation for responsible deployment in real medical environments.

Future Scope. Although the proposed model shows great promise in hybrid retrieval-based clinical decision support, there are a number of possible improvements that could further improve its performance and deployment.

One direction that should be explored is introducing multimodality. By doing so, the model will also leverage other types of input like medical imaging (X-ray, MRI, CT scan) in addition to the knowledge in the text form. This would make our current model a multimodal RAG system and it will allow us to jointly reason about visual and textual information.

Additionally, a useful feature to introduce into the system is automated understanding of pictograms, where graphical visual aids (charts, graphs, and annotated medical images) are interpreted by the machine using vision language models.

Lastly, achieving cross-institutional interoperability and demonstrating the performance of the system in the real-world clinical setting is essential for success. In order to do so, we need to optimize latency and enhance explainability of the results.

As you see above, each of those improvements would bring us closer to making the system scalable, multimodal, and self-learning platform for intelligent diagnosis.

Acknowledgements. We the authors of this work acknowledge the contributions and insights of the healthcare professionals who provided domain expertise during requirements design, system design and final evaluation of response generated.

Declarations

Funding -Not applicable

The authors received no external funding for this work.

Ethical approval -Not applicable

This study does not involve any clinical trial, human participant study, collection of personal patient information or any intervention requiring ethical review. Therefore, ethical approval was not applicable.

Consent to participate -Not applicable

The study did not involve the recruitment or participation of human subjects. Therefore, consent to participate was not applicable.

Consent to publish -Not applicable

No individual participant data, personal information, or identifiable records were included in this study. Therefore, consent to publish was not applicable.

Competing Interests

The authors declare that they have no financial or non-financial interests that could have influenced the work reported in this paper.

Data Availability

The knowledge base used for this work was created based on publicly available resources relating to medicine and healthcare, including MedRAG, PubMed, arXiv, bioRxiv, WHO guidelines, FDA resources, RxNorm, and StatPearls. All source materials are available from their respective providers and are cited in the References section of this manuscript.

The vector database created as a result of this study was hosted on a Qdrant cloud free tier platform to serve experimental purposes and thus cannot be guaranteed to be preserved forever. Processed data sets and implementation details can be made available by contacting the corresponding author under reasonable requests and data source licenses.

References

- [1] D. W. Bates, M. Cohen, L. L. Leape, J. M. Overhage, M. M. Shabot, and T. Sheridan, “Reducing the frequency of errors in medicine using information technology,” *J. Amer. Med. Inform. Assoc.*, vol. 8, no. 4, pp. 299–308, 2001.
- [2] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, “Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success,” *BMJ*, vol. 330, no. 7494, p. 765, 2005.
- [3] P. Lewis et al., “Retrieval augmented generation for knowledge-intensive NLP tasks,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.
- [4] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, “Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records,” *Journal of Biomedical Informatics*, vol. 156, 104662, 2024.
- [5] E. H. Shortliffe and B. G. Buchanan, “A model of inexact reasoning in medicine,” *Math. Biosci.*, vol. 23, no. 3–4, pp. 351–379, 1975.
- [6] M. Alkan, A. Yildirim, and H. Demir, “Artificial intelligence-driven clinical decision support systems: ethical, explainability, and validation challenges,” *Artificial Intelligence in Medicine*
- [7] C. M. Thwal, K. Thar, Y. L. Tun, and C. S. Hong, “Attention on Personalized Clinical Decision Support System: Federated Learning Approach,” *arXiv preprint arXiv:2401.11736*, 2024.
- [8] A. Choi, K. Lee, H. Hyun, K. J. Kim, B. Ahn, K. H. Lee, S. Hahn, S. Choi, and J. H. Kim, “A novel deep learning algorithm for real-time prediction of clinical deterioration in the emergency department for a multimodal clinical decision support system,” *Scientific Reports*, vol. 14, no. 1, 39627310, 2024.
- [9] K. Gasmi, O. Hrizi, N. Ben Aoun, I. Alrashdi, A. Alqazzaz, O. Hamid, M. O. Altaieb, A. E. M. Abdalrahman, L. Ben Ammar, M. Mrabet, and O. Necibi, “Enhanced multimodal physiological signal analysis for pain assessment using optimized ensemble deep learning,” *Computer Modeling in Engineering & Sciences*, vol. 143, no. 2, pp. 2459–2489, 2025.

- [10] MedRAG Dataset, “Hugging Face Repository.” Available: <https://huggingface.co/datasets/MedRAG>.
- [11] National Center for Biotechnology Information (NCBI), “PubMed,” Available: <https://pubmed.ncbi.nlm.nih.gov>.
- [12] arXiv, “arXiv e-Print Archive,” Available: <https://arxiv.org>.
- [13] bioRxiv, “bioRxiv Preprint Server,” Available: <https://www.biorxiv.org>.
- [14] World Health Organization, “WHO Clinical Guidelines,” Available: <https://www.who.int/home>.
- [15] U.S. Food and Drug Administration, “FDA Drug Databases and Regulatory Information,” Available: <https://www.fda.gov/drugs/development-approval-process-drugs/drug-approvals-and-databases>.
- [16] U.S. National Library of Medicine, “RxNorm: Normalized Names for Clinical Drugs,” Available: <https://www.nlm.nih.gov/research/umls/rxnorm>.
- [17] StatPearls Publishing, “StatPearls [Internet],” Treasure Island (FL): StatPearls Publishing; 2024. Available: <https://www.ncbi.nlm.nih.gov/books/>.
- [18] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [19] G. V. Cormack, C. L. Clarke, and S. Büttcher, “Reciprocal rank fusion outperforms Condorcet and individual rank learning methods,” *Proc. SIGIR*, pp. 758–759, 2009.
- [20] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, “Benchmarking Retrieval-Augmented Generation for Medicine,” *arXiv preprint arXiv:2402.13178*, 2024.