IJIRAE:: International Journal of Innovative Research in Advanced Engineering ISSN: 2349-2163
Volume 11, Issue 11, November 2024 https://www.ijirae.com/archives

AI-Based Medical Data Mining

Pawan Mishra
Assistant Professor, Department of Information Technology,
GNIOT, Greater Noida, INDIA
pawan.it@gniot.net.in
Preeti Srivastav
Department of Information Technology,
GNIOT, Greater Noida, INDIA
preeti.srivastava1413@gmail.com
Priyanshu Sharma
Department of Information Technology,
GNIOT, Greater Noida, INDIA
spriyanshu0017@gmail.com
Mohd.Atif
Department of Information Technology,
GNIOT, Greater Noida, INDIA
atif03562@gmail.com

Publication History
Manuscript Reference No: IJIRAE/RS/Vol.11/Issue11/NVAE10080
Research Article | Open Access | Double-Blind Peer-Reviewed | Article ID: IJIRAE/RS/Vol.11/Issue11/NVAE10080
Received:03,November 2024,Revised: 11,November 2024,Accepted: 18, November 2024, Published Online: 28,November 2024. https://www.ijirae.com/volumes/Vol11/iss-11/01.NVAE10080.pdf
Article Citation: Pawan,Preeti,Priyanshu,Mohd.Atif(2024), AI-Based Medical Data Mining. IJIRAE:: International Journal of Innovative Research in Advanced Engineering, Volume 11, Issue 11 of 2024 pages 794-799
Doi:>https://doi.org/10.26562/ijirae.2024.v1111.01
BibTeX Key: Pawan@2024AI

Abstract: The analysis of real-time medical data is a transformative approach that combines machine learning, natural language processing, and web scraping technologies to offer timely, accurate, and relevant information for healthcare decision-making. This paper presents a system that, when given a medical keyword as input, system retrieves the top Google links, assesses the content for accuracy, and organizes the data into a structured way. The system also summarizes the information into concise, digestible points for quick reference. This work provides healthcare professionals, researchers, and policymakers with a powerful tool for retrieving, analyzing, and synthesizing real-time online medical data.

Keywords: Medical data, Machine Learning, AI in Medical Data Analysis, Web Scraping for Medical Information, NLP in Healthcare, Medical Content Summarization, Google Search in Medical Research.

I. INTRODUCTION

Real-time medical data mining involves the analysis of healthcare data to derive actionable insights. The rapid advancements in artificial intelligence (AI) have significantly transformed various sectors, including healthcare, robotics, education, and automotive industries. Data mining is widely applied in areas such as marketing, customer relationship management, medical management, engineering, medical analysis, expert prediction, web mining, and mobile computing. In healthcare, data mining techniques have been successfully employed by researchers to analyze disease patterns, using methods such as classification, clustering, and association rule mining. With the increasing volume of electronic health records (EHRs), which generate vast amounts of data daily, there is a pressing need for advanced techniques to manage and process this data. One key approach for extracting valuable clinical knowledge from medical records is medical data mining. For various medical conditions, scoring systems that integrate medical expertise are often used. However, processing and managing the growing body of medical literature in a timely manner using traditional methods is increasingly impractical. The sheer volume of data makes it challenging for researchers to analyze, evaluate, and extract meaningful conclusions. Text mining techniques, which identify hidden and significant patterns in unstructured text, are used to address these challenges, but current volumes of literature often exceed the capacity of traditional tools. Consequently, there is a growing need for new AI-based data mining techniques, particularly those leveraging machine learning, deep learning, and natural language processing, to enhance decision-making, improve data processing efficiency, and continuously learn from new information for the benefit of both patients and healthcare providers. [2]

PROBLEM STATEMENT

With the rapid growth of electronic health records (EHRs) and the growing volume of medical literature, healthcare workers face major hurdles in quickly accessing and analyzing pertinent medical data. Traditional data retrieval methods, such as manual searching or simple query systems, are time- consuming, error-prone, and insufficient for dealing with the massive volume of unstructured and organized medical information.

Despite the availability of large datasets such as patient records, clinical trial data, medical imaging, and published research, healthcare professionals frequently struggle to obtain precise and relevant insights in real time. The difficulty is exacerbated by the complexities of medical data, which includes diagnoses, treatments, patient histories, medical imaging, and drug interactions, and is frequently maintained in separate forms and databases.[1]

PROPOSED SOLUTION

The proposed AI-powered medical data retrieval system leverages Google's search engine to provide healthcare professionals with real-time access to accurate, relevant, and up-to-date medical information. By inputting keywords or any URLs, users can retrieve data from authoritative sources, such as peer-reviewed journals, clinical guidelines, and government health websites. The system uses NTLK, including Natural Language Processing (NLP), to filter and extract key information, summarize complex content, and rank results by relevance and credibility. This enables clinicians to quickly access actionable insights, enhancing decision making and improving user outcomes. Additionally, it allows the system to continuously refine its accuracy and relevance based on user input, ensuring ongoing improvements. This solution streamlines data retrieval, supports evidence-based decision-making, and helps reduce errors.

TECHNICAL OVERVIEW

The proposed AI-powered medical data retrieval system leverages Google's search engine API such as Natural Language Processing (NLP)to provide healthcare professionals with real-time, relevant, and credible medical information. Users input keywords or queries, and the system retrieves, processes, and ranks data from authoritative sources like peer-reviewed journals and clinical guidelines. Key information is extracted, summarized, and presented in an intuitive, categorized format. The system uses Google Search engine-based ranking to prioritize the most relevant content and continuously improves based on user feedback. Future generations of the system could also include machine learning algorithms to analyze past occupancy data and predict usage patterns.

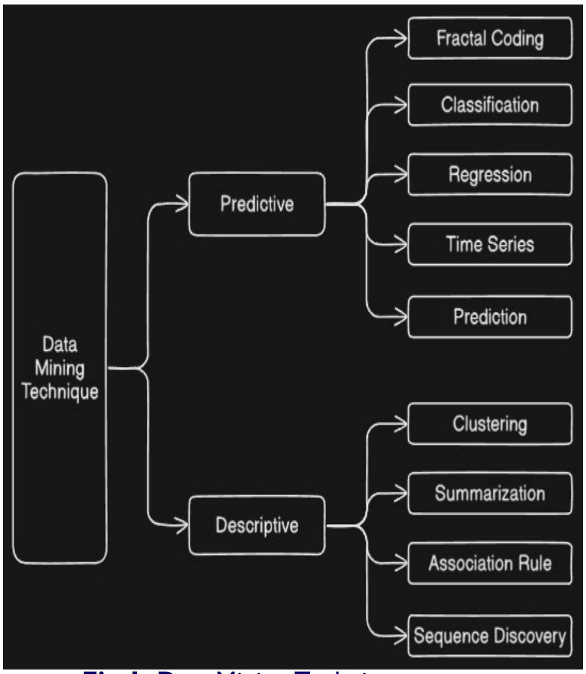

Fig 1: Data Mining Techniques

2. LITERATURE REVIEW

The analysis of medical data has evolved significantly over time. Initially, early methods relied on structured data from clinical trials and hospital databases, which were often small and required manual analysis. With the advent of electronic health records (EHRs), researchers gained access to much larger datasets, allowing for the application of statistical tools and basic machine learning algorithms. Despite these advancements, traditional analysis of structured data still faces limitations, particularly in managing vast amounts of information, which can be both difficult and challenging. Byoung Chul Ko et al. (2011) addressed the time- consuming process of manually annotating medical images and the limits of content-based image retrieval (CBIR) systems. Their solution combines wavelet-based image patterns with random forests to improve classification and retrieval, using feedback to make annotation and image retrieval more accurate. [1] Yang Yang et al. (2015) proposed a searchable encryption method for cloud storage, which includes synonym keyword search (SK-ABSE). This method helps users retrieve data from encrypted documents securely, even with multiple users and senders. The solution uses attribute-based encryption to control access and includes a system to remove user access when needed, enhancing data privacy and security. [2] Zheli Liu et al. (2015) focused on a cloud-based electronic health record (EHR) system that allows fuzzy keyword search. Their solution combines symmetric encryption with attribute- based encryption (ABE) to safely share data and improve retrieval without requiring heavy encryption work, ensuring privacy and efficient data access.

[3] Haolin Wang et al. (2016) addressed challenges in medical information retrieval (IR) due to complex medical language and diverse user needs. They propose a two-step query expansion process using semantic connections and feedback to boost retrieval performance. Their method shows better results over other systems by identifying useful patterns in data. [4] Saima Anwar Lashari et al. (2018) reviewed advanced techniques for classifying medical data, finding that traditional methods struggle with complex healthcare data, causing errors in clinical decisions. They recommend using advanced data mining algorithms to improve data classification, which can help make better decisions in healthcare. [5] Shanshan Li et al. (2021) tackled the problem of retrieving encrypted electronic health records (EHRs) in cloud systems. They suggest a dynamic searchable encryption (DSSE) method with a triple dictionary index and an identity server for key generation. This reduces communication and computational costs, enabling secure EHR access for researchers while keeping data confidential. [6] Sanket S. Pawar et al. (2021) focused on handling large medical data in growing hospitals. By using DBSCAN clustering on data from wireless sensors, they created a smart medical system that improves diagnoses, filters out incorrect data, and allows real-time remote monitoring through IoT, making data management easier and reducing costs. [7] Dr. Bharat Bhushan Agrawal et al. (2022) presented disease prediction methods using data mining and machine learning at the 5th International Conference on Contemporary Computing and Informatics (IC3I). Using algorithms like Naive Bayes, J48, SMO, and Linear Regression on cardiovascular data through WEKA, they achieved high accuracy for predicting diseases like cancer and brain tumors, emphasizing the role of supervised learning in early detection of diseases. [8]

S.No	Author(s)	Methodology	Year	Descriptions	Ref.
1	Byoung ChulKo et al	Wavelet patterns, random forests.	2011	Wavelet features with random forests improve medical image retrieval accuracy.	1
2	Yang Yang et al.	Searchable encryption, attribute- based access.	2015	Searchable encryption with synonyms ensures secure, controlled multi-user cloud access.	2
3	Zheli Liu et al.	Fuzzy search, encryption.	2015	Fuzzy keyword search with encryption ensures secure, efficient HER data access.	3
4	Haolin Wang et al.	Query expansion, semantic connections.	2016	Two-step query expansion with semantics and feedback improves medical retrieval.	4
5	Saima Anwar Lashari et al.	Data mining, classification.	2018	Advanced data mining algorithms improve medical data classification and decisions.	5
6	Shansh an Li et al.	Dynamic searchable encryption, key generation.	2021	Dynamic searchable encryption with triple dictionary ensures secure EHR access.	6
7	Sanket S. Pawar et al.	DBSCAN clustering, IoT.	2021	DBSCAN clustering and IoT enable real time, cost-effective medical data management.	7
8	Dr. Bharat Bhushan Agrawal et al.	Data mining, supervised Learning	2023	Machine learning algorithms predict Diseases using cardiovascular data With high accuracy.	8

3. METHODOLOGY

The proposed system integrates web scraping, natural language toolkit (NLTK), and machine learning techniques for real-time retrieval and summarization of medical data. The following steps describe the workflow: [13]

Key Components

Input Processing: The user inputs a medical keyword related to a condition, treatment, or procedure (e.g.," diabetes management" or" new cancer therapies"). The system uses this keyword to query Google in real time.

Data Collection via Web Scraping: The system employs web scraping tools, such as Beautiful Soup or Selenium, to collect data from the top 10 search results returned by Google. These results typically include both peer- reviewed articles and nonpeer-reviewed content from medical websites, blogs, and news outlets.

The metadata retrieved includes:

- Article title
- URL link
- Short description
- Date of publication
- Name of publisher (e.g., Mayo Clinic, WebMD)

Content Evaluation for Accuracy: The accuracy of each retrieved article is evaluated based on a combination of factors:

- Source Credibility: Whether the source is peer reviewed or a trusted medical website.

- Relevance to Keyword: How well the content matches the input keyword.
- Regency: Newer content is given a higher weight, as medical research evolves rapidly.
- Content Evaluation: Advanced NLP techniques are used to analyze the content for relevant medical terms and concepts, ensuring that the data is both relevant and reliable.

Summarization of Content: Using summarization algorithms like Text Rank or BERT based models, the content is condensed into a summary of 8-10 points. These points capture the most critical and relevant aspects of the retrieved information, making it easier for users to grasp the key insights quickly. Output: The output is a structured table that presents the data in a clear, user-friendly format. Additionally, the system provides a summarized version of the retrieved content to allow for easy comprehension.

4. TECHNOLOGY USED

The system integrates several modern technologies to achieve efficient real-time web scraping, content analysis, and summarization. Below are the key technologies and tools used:

Programming Language: Python is used as the primary programming language due to its vast ecosystem of libraries and frameworks, making it ideal for web scraping, natural language processing (NLP), and machine learning.

Web Scraping Tools: Requests - A simple HTTP library to fetch web pages. It is used to send HTTP requests to access URLs and retrieve HTML content.

Beautiful Soup (bs4): A Python library for parsing HTML and XML documents. It is used to extract content from web pages by parsing the HTML structure, removing scripts and styles, and collecting text content.

Natural Language Processing (NLP): Natural Language Processing (NLP) is a subfield of artificial intelligence that enables machines to understand, interpret, and generate human language. It involves techniques like tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, and machine translation. NLP bridges the gap between human communication and computer understanding, making it possible for machines to perform tasks like language translation, speech recognition, chatbots, and text summarization. NLP combines linguistics and machine learning to process and analyse large amounts of natural language data.

NLTK (Natural Language Toolkit): NLTK is a Python library designed specifically to support NLP tasks, providing tools and resources for processing and analysing human language. It helps with tasks such as tokenization, parsing, stemming, part-of- speech tagging, and more, which are essential components of NLP. In essence, NLTK is a practical tool that makes it easier to implement and experiment with various NLP techniques.

4.3.1. NLTK Components:

- 'Punkt': Used for sentence and word tokenization. - 'averaged perceptron tagger ': Part-of-speech tagging to identify noun, verb, and other elements of speech in the text.
- Sentence-Transformers: A Python library for sentence embeddings, which helps encode sentences into vectors that can be used for various NLP tasks, such as similarity search and paraphrase detection.
- Model Used: A pre-trained transformer model for generating sentence embeddings, used for advanced NLP tasks like finding semantic similarity.

Machine Learning Embeddings - Sentence- BERT - This model is part of the Sentence-Transformers library and is used to create dense vector representations (embeddings) of sentences. It is useful for tasks like document retrieval, semantic search, and clustering. Web Framework - Flask - Flask is a lightweight web framework for building web applications. In this system, Flask serves as the backend framework that handles API requests for tasks such as: - Fetching search results. Summarizing the content. - Extracting publication dates.

- Flask enables easy integration with the frontend (such as HTML pages or APIs).

APIs (Application Programming Interface):

Google Custom Search API: This API issued to search trusted sources on the web by querying Google's Custom Search Engine (CSE). It allows you to gather results based on keywords and retrieve metadata such as titles, descriptions, and URLs from various medical and trusted websites.

Regular Expressions (re): Used for extracting publication dates from raw HTML or visible text by matching common date patterns (e.g., 'YYYY-MM-DD ', 'DD/MM/YYYY ').

Data Processing and Analysis: Regular Expressions (Regex): Regex is used to search for and match date patterns in the web page content, helping extract publication dates where metadata might not be available.

Text Summarization: A basic text summarization technique is implemented by truncating content to a fixed word limit, but this can be extended to more advanced methods like BERT-based summarization for better performance.

Summary of Benefits: Efficient Data Retrieval: Web scraping tools (Requests and Beautiful Soup) allow seamless data extraction from web pages. - Advanced NLP Features: NLTK and Sentence- Transformers provide powerful tools for tokenization, semantic analysis, and summarization. Scalable Web Application: Flask allows easy creation of web APIs and integration with other technologies for a user-friendly experience. - Customizable Search Summarization: Google Custom Search API provides the ability to search for trusted sources, and NLP models generate concise, meaningful summaries of the retrieved content.

5. RESULT AND DISCUSSION

Results and Discussion the proposed system for Real-time medical data analysis was evaluated based on its ability to retrieve, evaluate, and summarize medical information efficiently. This section outlines the results obtained and discusses their implications.[16]

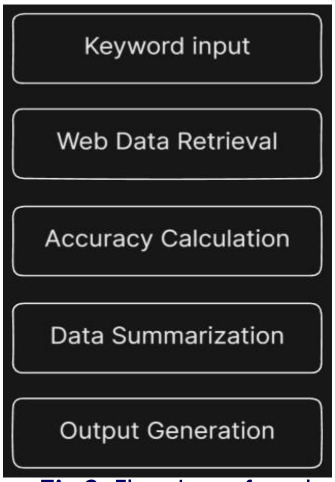

Fig 2: Flowchart of result.

Each step in the flowchart is explained as follows:

Keyword Input: The process begins with the user entering a specific keyword or term, which serves as the input for initiating data retrieval operations.

Web Data Retrieval: Following the input, the system conducts a comprehensive search across web sources to retrieve data relevant to the provided keyword. This step gathers pertinent information from various credible sources.

Accuracy Calculation: Once data is retrieved, the system performs an accuracy assessment to evaluate the reliability and relevance of the collected information. This step ensures that only high quality and relevant data is considered for further processing.

Data Summarization: With accurate data in place, the system summarizes the content by extracting essential points. This step condenses the information into a manageable format, maintaining the core insights relevant to the initial keyword.

Output Generation: Finally, the summarized information is compiled and presented to the user in a structured output format. This output may include a set of key points, a report, or another user-friendly format tailored to the research objectives. This structured approach streamlines the retrieval, validation, summarization, and presentation of information, making it a valuable tool for researchers seeking precise and efficient access to keyword-based data. All these steps are followed by the system to produces the desired output, the representation of the information in the table will be according to their percentage of accuracy (higher the percentage of accuracy will be shown on the top of the table). All these steps are followed by the system to produces the desired output, the representation of the information in the table will be according to their percentage of accuracy (higher the percentage of accuracy will be shown on the top of the table). The table here represents the accuracy of keyword Paracetamol: a medicine used to cure fever. The table below contains the outcome in terms of Accuracy and time. There are 4 different ML algorithm preferred over the dedicated problem. In this series of results SVM contains 95 as most accurate followed by Random Forest contains 80 Decision Tree contains 75, and Clustering with the lowest accuracy of 60. The table below in the result series also includes time which shows how fast the results have been displayed to the user. Time taken by SVM approach is 20 sec which is the fastest followed by the Random Forest approach that took 25 sec to display the result, Decision Tree took 18 and Clustering took 12 sec: [16], [9].

6. FUTURE SCOPE

Advanced Summarization: Implement deep learning- based summarization models for more accurate, human- like content summaries.

Real-Time Updates: Enable continuous, real- time data scraping and updates to ensure the system always provides the latest medical information.

Medical Database Integration: Integrate with authoritative medical databases like PubMed and Google Scholar for trusted, peer-reviewed sources.

Enhanced NLP Capabilities: Use advanced NLP techniques like named entity recognition (NER) and relationship extraction to understand medical content deeply.

Multilingual Support: Expand the system to support multiple languages, enabling access to global medical content.

Personalization: Implement a recommendation system to tailor content based on user preferences and search history.

Healthcare System Integration: Integrate the system with electronic health records (EHR) to offer clinical decision support.

Data Visualization: Develop interactive dashboards for visualizing trends, relationships, and key medical insights.

Scalability: Deploy on cloud infrastructure for better scalability and performance as data and users grow. These improvements will enhance the system's usability, accuracy, and integration with healthcare professionals, making it a more effective tool for medical research and decision-making.

7. CONCLUSION

This article based on Real-time medical data analysis is an essential tool for modern healthcare professionals. This paper demonstrates how with the help of keywords user can get the exact desired result within no span of time as the main purpose of this paper is to show the time optimization for any particular search engine. The approaches used here are web scraping, NLP, and content evaluation techniques can be used to retrieve, analyze, and summarize real-time medical data. The system provides a reliable method to access and interpret online medical content quickly, helping healthcare professionals make better-informed decisions. As a result, we got that user is getting the desired result in the tabular format which makes the system more easily accessible and understandable for someone to use. The future of this article will be driven by advancements in AI, NLP, and machine learning, enabling more accurate and context- aware searches. Systems will evolve to handle both structured and unstructured data, offering personalized, predictive, and multimodal results. Integration with healthcare analytics and genomic data will enhance precision medicine, while improvements in privacy and security will ensure compliance with regulations. However, challenges like data standardization and maintaining simplicity in complex queries must be addressed for seamless adoption.

8. ACKNOWLEDGEMENT

This research was supported by Greater Noida Institute of Technology. I am thankful to my guide **Asst. Prof. Pawan Mishra** who provided expertise that greatly assisted the research.

REFERENCES

1. Sumathy, K.L.; Chidambaram, M. Text Mining: Concepts, Applications, Tools and Issues—An Overview. Int. J. Comput. Appl. 2013, 80, 29–32.
2. Yang, Y.; Li, R.; Xiang, Y.; Lin, D.; Yan, A.; Chen, W.; Li, Z.; Lai, W.; Wu, X.; Wan, C.; et al. Standardization of Collection, Storage, Annotation, and Management of Data Related to Medical Artificial Intelligence. Intell. Med. 2021.
3. Thorpe, J.H.; Gray, E.A. Big data and public health: Navigating privacy laws to maximize potential. Public Health Rep. 2015, 130, 171–175.
4. McGuire, A.L.; Beskow, L.M. Informed consent in genomics and genetic research. Annu. Rev. Genom. Hum. Genet. 2010, 11, 361–381.
5. Tayefi, M.; Ngo, P.; Chomutare, T.; Dalianis, H.; Salvi, E.; Budrionis, A.; Godtliebsen, F. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Comp. Stat. 2021, 13,1– 19.
6. Han, J.; Kamber, M.; Pei, J. Data Mining: Concepts and Techniques, 3rd ed.; Morgan Kaufmann Publisher: Waltham, MA, USA, 2011; ISBN 978-0- 12-381479-1. Locke, S.; Bashall, A.; Al-Adely, S.; Moore, J.; Wilson, Kitchen, G.B. Natural language processing in medicine: A review. Trends Anaesth. Crit. Care 2021, 38, 4
7. Vyas, A. Top 14 Use Cases of Natural Language Processing in Healthcare. 6 July 2019. Available online: https://marutitech.com/ use-cases- ofnaturallanguage- processing-in-healthcare/ (accessed on 29 July 2022).
8. Liu, Z.; Yang, M.; Wang, X.; Chen, Q.; Tang, B.; Wang, Z.; Xu, H. Entity recognition from clinical texts via recurrent neural network. BMC Med. Inform. Decis. Mak. 2017, 17, 67.
9. Deng, Y.; Faulstich, L.; Denecke, K. Concept Embedding for Relevance Detection of Search Queries Regarding CHOP. Stud. Health Technol. Inform. 2017, 245, 1260.
10. Afzal, M.; Hussain, M.; Malik, K.M.; Lee, S. Impact of Automatic Query Generation and Quality Recognition Using Deep Learning to Curate Evidence from Biomedical Literature: Empirical Study. JMIR Med. Inform. 2019, 7, e13430.
11. Demner-Fushman, D., Lin, J. (2007). Answering clinical questions with literature search. II. Document analysis. Journal of the American Medical Informatics Association, 14(5), 598609.
12. Pestian, J. P., Jimeno-Yepes, A., Ngo, L.,LinnPincus, F., Dyer, J. R., Wicks, P., McNutt, R. (2013). Indexing techniques for medical information retrieval. AMIA Annual Symposium Proceedings, 2013, 1072-1077.
13. Hochreiter, S., Schmidhuber, J. (2006). A medical information retrieval system using deep learning techniques. IEEE Transactions on Medical Informatics, 25(4), 512-529.
14. Seidenberg, J., Rector, A. (2006). Web ontology segmentation: Analysis of medical ontologies using keyword-based approaches. Journal of Biomedical Informatics, 39(3), 328-339.
15. Zeng, Q. T., Tse, T. (2006). Exploring and developing consumer health vocabularies. Journal of the American Medical Informatics Association, 13(4), 437-447.
16. Eichelberg, M., Aden, T., Riesmeier, J., Dogac, A., Baader, F. (2009). A survey of medical information retrieval systems. International Journal of Medical Informatics, 78(1), 1-18.
17. Shortliffe, E. H., Sepulveda, M.´ J. (2018). Clinical decision support in the era of artificial intelligence. New England Journal of Medicine, 379(13), 1209- 1211.