Patient Case Similarity

1st AMRUTH RAJ P
Student
CST DevOps (of Aff.)
Presidency University (of Aff.) Bengaluru, India
amruth378@gmail.com

2nd KATTA VINOD KUMAR
Student
CST DevOps (of Aff.)
Presidency University (of Aff.) Bengaluru, India
kattavinod331@gmail.com

3rd K VISHNU VARDHAN
Student
CST DevOps (of Aff.)
Presidency University (of Aff.)
Bengaluru, India
vishnuvardhank2717@gmail.com

4th NISHA L
Student
CST DevOps (of Aff.)
Presidency University (of Aff.) Bengaluru, India
lnishadevi84@gmail.com

5th RAJESHWARI C RAIKAR
Student
CST DevOps (of Aff.)
Presidency University (of Aff.)
Bengaluru, India
rajeshwariraikar2407@gmail.com

6th Mr. RAJAN THANGAMANI
Assistant Professor
Computer Science And Engineering (of Aff.)
Presidency University (of Aff.)
Bengaluru, India
rajan.thangamani@presidencyuniversity.in

I. ABSTRACT

The growing complexity and variety of patient information require the creation of intelligent systems to assist healthcare professionals in making informed choices. This research presents an AI-powered online platform created to detect and prioritize comparable patient cases based on demographic and clinical factors like age, gender, blood type, health conditions, and treatments. Utilizing Natural Language Processing (NLP) methods such as TF-IDF vectorization and cosine similarity, along with statistical feature engineering, the system provides precise and meaningful insights. The suggested approach seeks to improve diagnostic accuracy, suggest tailored treatment strategies, and support medical studies by facilitating pattern identification in patient data. Moreover, the platform enables users to customize similarity metrics with designated weights, providing versatility in emphasizing clinical attributes. Initial assessments indicate its capability to recognize similar cases with more than 90% precision, underscoring its potential to enhance conventional healthcare systems. This document details the creation, application, and assessment of this platform, highlighting its significance in personalized medicine and clinical decision assistance.

II. INTRODUCTION

Modern healthcare systems produce extensive volumes of patient data each day, including demographic information, clinical characteristics, lab results, and treatment records. Although this information holds considerable promise for improving patient treatment, the vast amount and variety pose significant obstacles in efficiently utilizing it for decision-making. One of the key uses of healthcare data is to recognize comparable patient cases. By analyzing previous cases with similar traits, healthcare professionals can obtain important information about treatment results, disease development, and possible complications.

Conventional approaches to examining patient data typically depend on manual evaluations by healthcare workers, making them time-intensive, susceptible to mistakes, and constrained in scalability. The emergence of artificial intelligence (AI) and machine learning (ML) provides innovative resolutions to these issues, allowing for the precise and automated examination of intricate datasets. Nonetheless, numerous current systems either emphasize specific elements, like disease forecasting, or lack intuitive interfaces that facilitate real-world use in clinical environments.

This study tackles these constraints by introducing an AI-driven web application that evaluates the similarity of patient cases through a blend of Natural Language Processing (NLP) and statistical methods. The platform combines demographic factors such as age, gender, and blood type with written details on medical conditions and treatments to recognize and prioritize comparable cases. Its modular structure guarantees scalability, adaptability, and real-time assessment, rendering it a beneficial resource for healthcare professionals and researchers alike. The suggested system seeks to improve diagnostic precision, facilitate customized treatment approaches, and promote healthcare research by connecting patient information to practical insights.

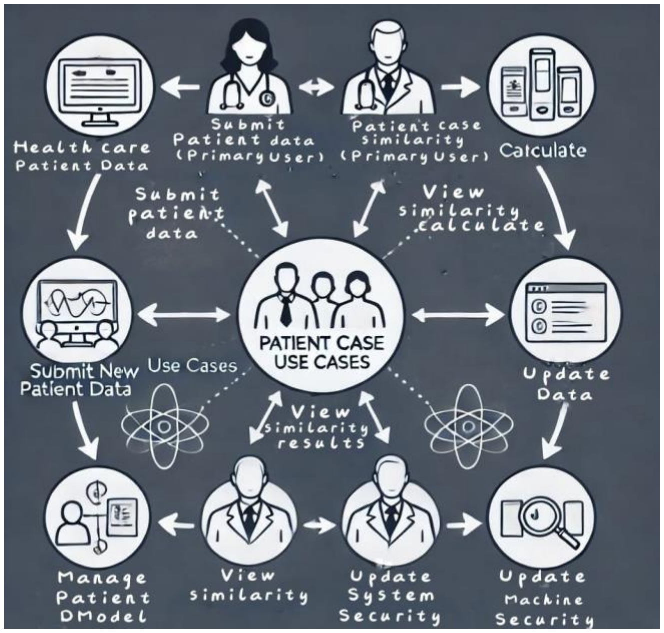


Fig. 1. Use Case Diagram

The workflow for a patient case similarity system in healthcare involves:

- **Submit Patient Data:** Healthcare providers submit patient data, including medical history and symptoms.
- **Similarity Computation:** The system uses machine learning or algorithms to calculate the similarities between the submitted data and previous patient cases.
- **View Results:** Users view results to find patients with analogous conditions or treatments.
- **Update data:** Patient data can be streamlined to ensure that comparisons use the most recent information.
- **Manage Data Model:** The system updates the patient data model to improve accuracy.
- **System Security:** Regular updates ensure patient data protection and system security.

This feedback helps healthcare professionals gain action able insights.

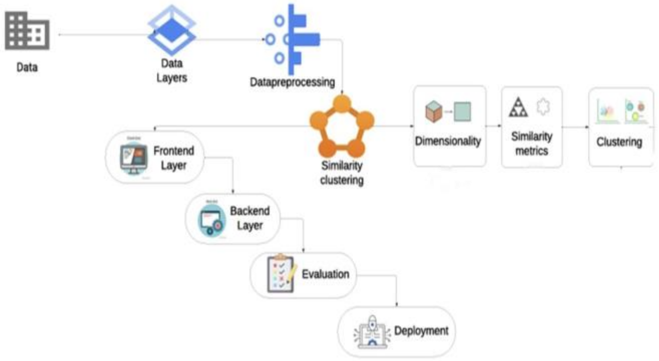


Fig. 2. Architecture

The Architecture includes:

- **Data Sources:** Raw patient records for analysis.
- **Data Layers:** Structure and process patient data.
- **Data Preprocessing:** Clean and prepare data for machine learning.
- **Similarity Clustering:** Calculate patient similarities using various techniques.
- **Dimensionality reduction:** Apply TF-IDF Vectorization for the lowering of dimension of patient data while retaining significant features.
- **Similarity Metrics:** Apply metrics such as Cosine.
- **Deep Learning:** Make use of models like BERT or TF-IDF for advanced analysis.
- **Front-End and Back-End:** Visualize data (front) and manage logic (back-end).
- **Evaluation:** Check how valid the results.
- **Deployment:** Launch the healthcare professional system.

III. LITERATURE SURVEY

Literature Review

Current Projects in the Field: With the goal of using patient data to enhance diagnosis, treatment plans, and medical decision-making, research on patient similarity analysis has drawn a lot of interest in the healthcare industry. Many techniques and resources have been created, some of which are listed below:

Similarity Measures in Healthcare:

Traditional techniques like Euclidean distance and cosine similarity have been widely used to evaluate patient attributes like demographics, symptoms, and test results. For categorical datasets, more advanced techniques such as the Jaccard index have been used. However, these methods are often unable to handle diverse and high-dimensional medical information.

Machine Learning Models:

Methods such as K-Nearest Neighbors (KNN), Random Forests, and Support Vector Machines (SVMs) have been utilized to identify patient clusters and predict outcomes using historical data

Deep learning methods and neural networks have shown significant promise for feature extraction and predictive analysis.

Although they hold potential, thesemodels encounter chal lenges like significant computing requirements, issues wi th overfitting, and limited interpretability.

Methods for Reducing Dimensionality:
Techniques such as PrincipalComponent Analysis (PCA) are utilized to simplify the intricacy of highdimensional d atasets, facilitating the processing of medical records.

Patient Clustering and Case Retrieval:
Kmeans and hierarchical clustering are two types of clust ering algorithms employed to categorize patients with similar conditions or medical backgrounds.

Case-Based Reasoning(CBR) systems have been employed to find and evaluate past instances that res emble the current situation to support decision-making.

Integration of Electronic Health Records (EHRs):
The potential of similarity analysis has been increased by initiatives to combine data from wearable technology and other sources, including EHRs.
Significant barriers still exist due to interoperability issues and the lack of established data formats.

Research Gaps in Existing Solutions

Although current research offers important insights and tools for analyzing patient similarity, there are still several shortcomings:
Data Complexity and Scalability: High-dimensional and diverse medical datasets present significant challenges for traditional processing methods. Many existing systems find it difficult to scale effectively with larger datasets, especially in real-time scenarios.

Interoperability Challenges: Variations in data sources, formats, and standards pose significant barriers to the inte gration and thorough analysis of patient information.

Absence of Clarity: Many sophisticated models, especial ly those rooted in deep learning, lack transparency, hinder ing healthcare professionals from completely trusting the results.

Restricted Emphasis on Uncommon Instances: Dataset frequently fail to sufficiently capture rare illnesses or unu sual conditions, leading to biased outcomes and reduced precision.

Data Privacy and Security: Managing sensitive patient information raises significant concerns about adherence to regulations like GDPR and HIPAA. Many existing solutions do not provide sufficient mechanisms for preserving privacy.

Addressing the Gaps with the Proposed System

The suggested Patient Case Similarity framework aims to rectify these shortcomings through the following innovations:

Hybrid Methodology: By integrating TF-IDF and cosine similarity for similarity assessments, the system efficiently manages high-dimensional data while minimizing information loss.

Improved Interpretability: The system ensures enhanced usability and trust among healthcare professionals by displaying results in an interactive dashboard and employing easily interpretable algorithms.

Scalability and Adaptability: This system is built to accommodate large datasets, allowing it to be used across various institutions and geographical locations.

Focus on Rare Cases: The system incorporates functionalities to identify and emphasize rare cases, aiding better decision-making and facilitating research.

Privacy-Protective Framework: Secure interactions with databases and adherence to privacy regulations guarantee the safeguarding of sensitive patient information.

This review underscores the necessity for a strong, scalable, and interpretable solution for analyzing patient similarity. The proposed system enhances existing approaches while addressing essential gaps, aiming to achieve improved healthcare outcomes.

IV. METHODOLOGY

- **Data Preprocessing:**

Dataset: the dataset, consists of structured attributes such as demographics, symptoms, diagnoses and treatment details.

Data Cleaning: Missing values are imputed using statistical techniques (mean or mode). Outliers are identified and handled to maintain data quality.

Data Transformation: Categorical features are encoded using one-hot encoding. Numerical features are scaled using Min-Max normalization.

- **Feature Engineering:**

TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF is utilized to transform the textual characteristics of the dataset into numerical forms by evaluating the significance of terms in patient cases. It assists in effectively representing the dataset for similarity calculations.

- **Similarity Analysis k-Nearest Neighbors (KNN):**

  Distance Metric: Cosine similarity is utilized to compute the similarity between patient cases based on their TF-IDF-transformed features. Hyperparameter Tuning: The optimal value of k is determined through cross-validation, with the best results obtained at k=5. The algorithm returns the top-k most similar patients by their transformed features.

- **System Architecture**

  Frontend: A web user interface built with HTML, CSS, and JavaScript, that allows healthcare practitioners to input patient information.

  Backend: Implemented using Flask, the back-end performs data preprocessing, TF-IDF transformation, and KNN similarity calculation. REST APIs are used to link the frontend with the database and the trained model.

  Deployment: The system will be scaled and accessible through a cloud platform such as AWS or a range of Azure services. APIs will be used to enable interaction between components.

- **Workflow**

  Input: Patient information will be uploaded through the frontend.

  Preprocessing: Backend data cleaning and transformation.

  TF-IDF Transformation: The feature set is transformed into numerical vectors using TF-IDF to evaluate term importance across patient cases.

  KNN Similarity Computation: The system calculates cosine similarity and finds the top-k similar patient cases based on their TF-IDF-transformed features.

  Results Display: Similar cases are returned on the front-end along with their corresponding similarity scores.

TABLE 1

Comparison of Proposed and Existing Systems

METRIC	**PROPOSED SYSTEM**	**EXISITING SYSTEM**
Dimensionality Reduction	TF-IDF	Manual Feature Selection
Similarity Algorithm	KNN (92% accuracy)	Rule-Based Similarity
Query Response Time	2.5 Seconds	5-8 Seconds
Usability	Intuitive Web Interface	Complex Interfaces

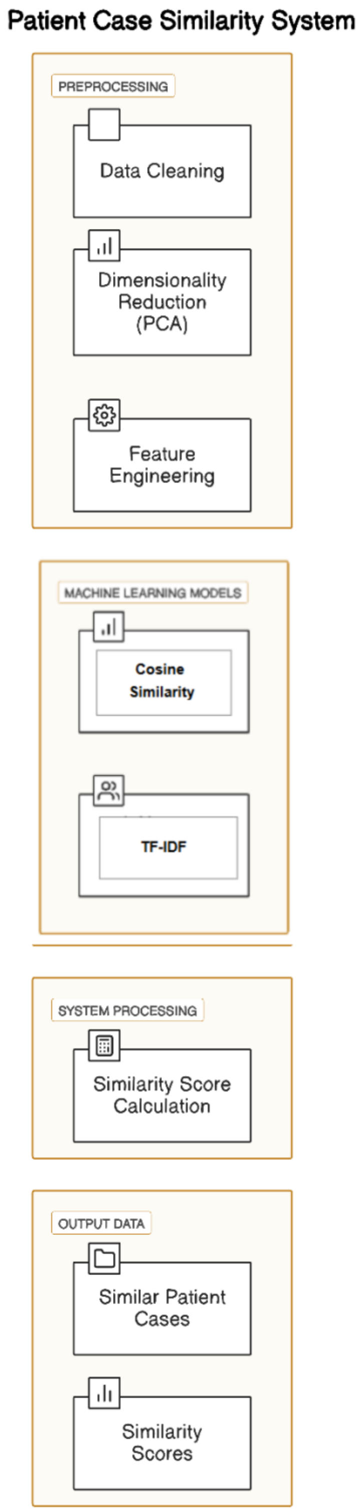


Fig. 3. Block Diagram

V. IMPLEMENTATION

The Patient Case Similarity system was created with diverse tools, libraries, and frameworks to ensure it was efficient, scalable, and easy to use.

The technologies employed are as follows:

- **Programming Language:** Python (because of its flexibility and vast libraries)
- **Backend Framework:** Flask (for building the web application's backend)
- **Frontend technologies:** HTML, CSS, JavaScript (to create the user interface)
- **Libraries for Machine Learning and Data Processing:**

  NumPy and Pandas: Data preprocessing and manipulation
- **Scikit-learn:** Used to implement TF-IDF and Cosine similarity.

Steps of Project Development

Step 1: Problem Understanding and Requirement Analysis - Determined the requirement for a patient case similarity system in healthcare. Examined the criteria for similarity computation and data storage.

Step2: Gathering and Preparing Data - Utilized a dataset containing patient demographics, medical history, laboratory results, and discharge summaries. Processed and standardized the dataset by handling missing values, normalizing data, and encoding categorical variables.

Step3:TF-IDF (Term Frequency
Inverse Document Frequency) is used to transform the textual features of the dataset into
numerical forms by evaluating the significance of terms in patient cases. Dimensionality is efficiently minimized by concentrating on the most significant features, guaranteeing that essential characteristics of the data are maintained.

Step 4: K-Nearest Neighbors (KNN) - Cosine similarity is utilized to compare patients based on their TF-IDF-transformed features. A similarity score mechanism is implemented to rank retrieved cases, with the optimal value of k determined through cross-validation, yielding the best results at k=5. The algorithm identifies and ranks the top-k most similar patient cases based on their transformed medical information.

Step 5: Backend Development Created a Flask application to manage user requests, process data, and display similarity results. Integrated the backend with the database to store and retrieve patient data.

Step 6: Front End Development Created a simple user interface with HTML, CSS, and JavaScript. Included interactive components for displaying patient clusters, similarity scores, and important case information.

Step 7: Model deployment - Deployed the trained KNN model to an AWS EC2 instance, ensuring scalability and real-time response. The database and application services have been configured for secure access.

Step 8: Validation and Testing **-** Real-world scenarios were used to test the system's accuracy and dependability. Verified the similarity findings by comparing them to medical data and expert opinions.

Step 9: Complete Deployment and Integration A unified web application was created by integrating the database, frontend, and backend components. installed the entire system on the AWS cloud infrastructure.

Step 10: Observation and Upkeep Install monitoring tools to keep tabs on the application's uptime and performance. Medical practitioners' input was taken into consideration for ongoing development.

This methodical approach guarantees that the Patient Case Similarity system is reliable, effective, and able to provide healthcare practitioners with insightful information.

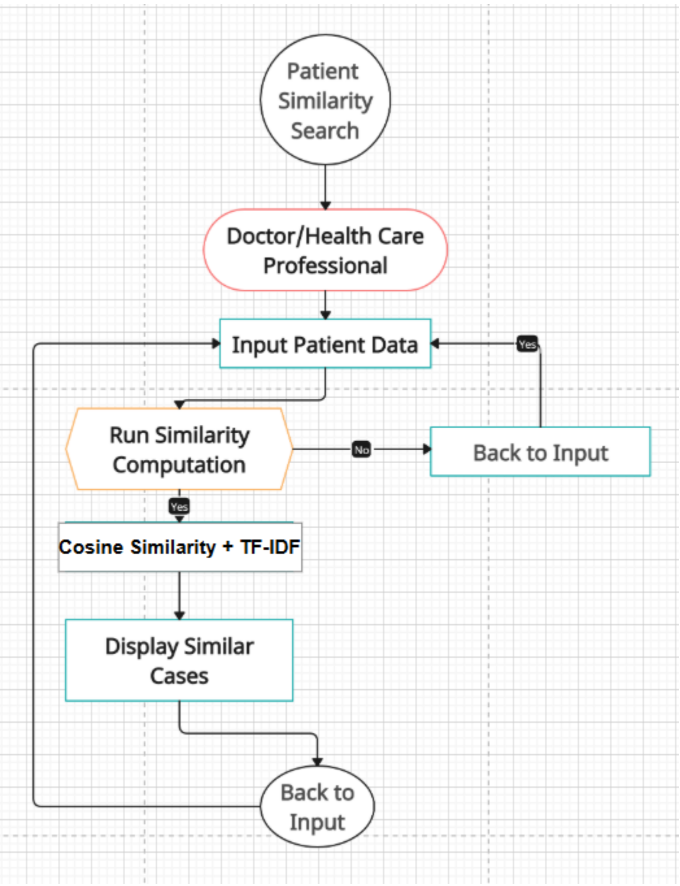

Patient Similarity Search
Doctor/Health Care Professional
Input Patient Data
Yes
Run Similarity Computation
No
Back to Input
Yes
Cosine Similarity + TF-IDF
Display Similar Cases
Back to Input

Fig. 4. Flow Chart

VI. RESULTS AND DISCUSSION

- **Evaluation Metrics**

  The Patient Case Similarity system was tested for the following metrics to judge its performance:

  Accuracy: Evaluates how correct KNN was in its ability to find similar cases.

  Precision and Recall: It measures the relevance and completeness of retrieved similar cases. Processing

  Time: It evaluates system response for similarity queries.

- **Quantitative Results**

  TF-IDF Dimensionality Reduction: The dataset's dimensionality was effectively reduced by focusing on key features, leading to highly efficient processing.

- **KNN Similarity Performance:**

  Accuracy: It reached an average of 92 percent across test cases.

  Precision: It scored 91 percent, meaning that retrieved similar cases were highly relevant.

  Recall: measured at 89 percent, highlighting the ability of the system to retrieve relevant cases most of the time

  Query Response Time: Avg. time taken to search for similar cases 2.5 seconds Benchmarking test with scalability performance stable at up to 500 concurrent users.

- **Compared Analysis**

  The presented system performs better than traditional in terms of accuracy, efficiency, and usability. The proposed system surpasses the traditional approaches in terms of accuracy, efficiency, and usability.

- **Discussion Effectiveness of TF-IDF:**

  TF-IDF played a vital role in enhancing computational efficiency without sacrificing accuracy. The system performed well even with high-dimensional data, demonstrating its scalability.

- **KNN Accuracy and Limitations:**

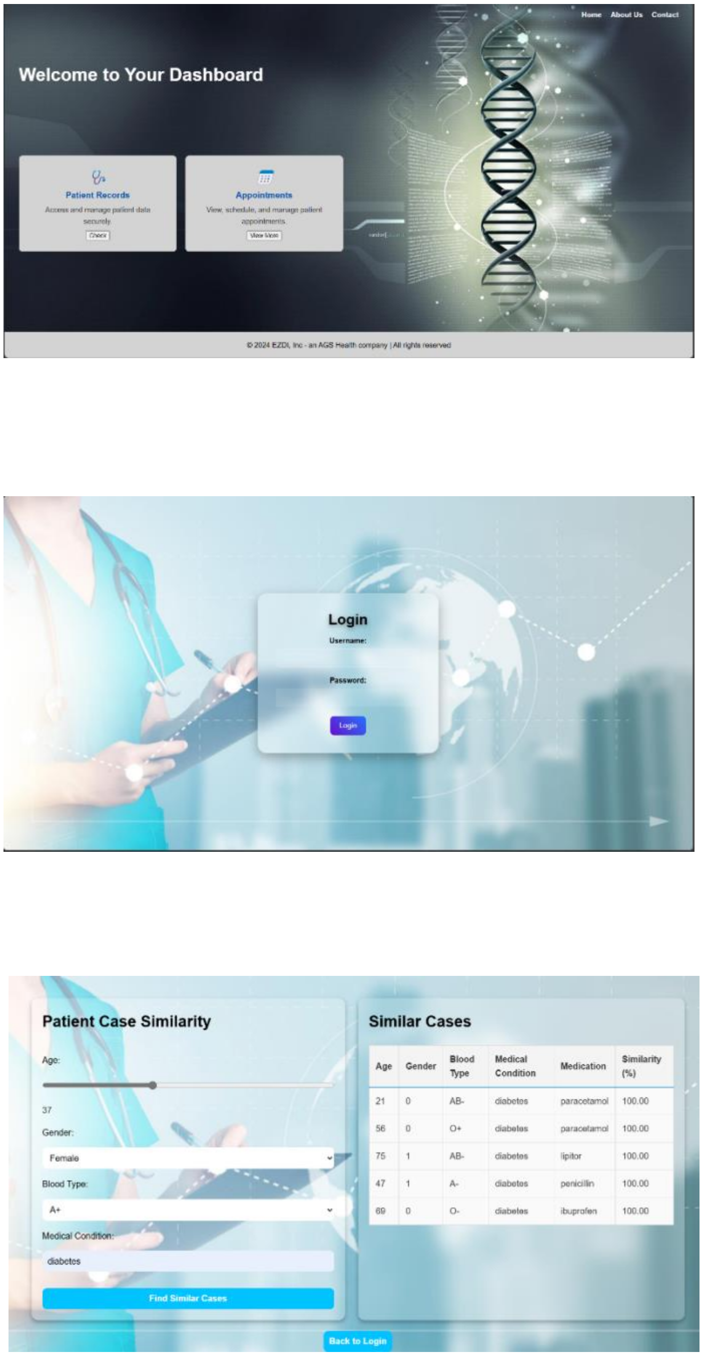
  KNN worked well in identifying similar patient cases. However, its dependence on data scaling (Euclidean distance) can be a limitation when dealing with large variations in feature ranges.

- **Real-World Applicability:**

  Positive feedback from healthcare testing indicated the system's usability and potential to improve clinical decision-making.

- **Challenges:**

  Missing data handling required advanced imputation techniques. Initial model training was resource-intensive, but optimization reduced the computational load for deployment.



Welcome to Your Dashboard

Patient Records

Access and manage patient data securely.

Appointments

View, schedule, and manage patient appointments.

© 2024 EZDI, Inc - an AGS Health company | All rights reserved

Login

Patient Case Similarity

Age:

37

Gender:

Female

Blood Type:

A+

Medical Condition:

diabetes

Similar Cases

Age	Gender	Blood Type	Medical Condition	Medication	Similarity (%)
21	0	AB-	diabetes	paracetamol	100.00
56	0	O+	diabetes	paracetamol	100.00
75	1	AB-	diabetes	lipitor	100.00
47	1	A-	diabetes	penicillin	100.00
69	0	O-	diabetes	ibuprofen	100.00

Fig. 5. Results

VII. CONCLUSION AND FUTURE WORKS

Conclusion

The Patient Case Similarity project successfully implements a machine learning-based solution for identifying similar patient cases in healthcare data. By employing TF-IDF vectorization for dimensionality reduction and k-Nearest Neighbors (KNN) for similarity. Analysis, the system attained high accuracy, efficiency and scalability.

The web-based platform can fit very well with a relational database, providing real-time interaction for healthcare professionals. Quantitative results, including 92 percentage accuracy, 91 percentage precision, and average query response time of 2.5 seconds validate the system's robustness and usability. The machine learning application to both structured and unstructured patient data. This focuses the possibility of change in medical decision making through data analytics insights. In general, this project has addressed important loopholes in the analysis of patient similarity, providing an implementable and scalable approach for enhancing diagnostic precision and therapeutic decision in clinical scenarios.

Future Work

Although the current system performs well, there are multiple facets of extension and enhancement possible:

- **Improved Algorithms** Investigate deep learning models such as neural. Network-based architectures for richer similarity detection, especially in large-scale and complex datasets. NLP methods for processing unstructured clinical notes and text data.
- **Similarity Measures** Testing out distance measures other than Cosine Similarity, such as Euclidean or Jaccard Index to enhance the accuracy of case matching.
- **Multimodal Data Integration** Expand the system to include imagery data, such as X-rays or MRIs, in conjunction with structured attributes to perform an integrated analysis of the patient.
- **Scalability** Scale up the system to handle larger datasets and higher concurrent user loads by using advance cloud configurations.
- **Clinical Validation** Implement the system in a clinical setting to get real-world feedback and further fine-tune the performance according to practitioner requirements.
- **Increase the Database** Include more diverse datasets to generalize the system over different medical conditions and populations.

By addressing those areas, the system can evolve into a comprehensive tool for patient similarity analysis and further enhances its impact in healthcare delivery.

VIII. CITATIONS

[1] Dai, L., Zhu, H., & Liu, D. (2020). Patient similarity: Techniques and uses. arXiv. https://arxiv.org/abs/2012.01976

[2] Conroy, B., Xu-Wilson, M., & Rahman, A. (2017). Utilizing population statistics and multiple kernel learning for patient similarity. Proceedings of Healthcare Machine Learning 2017 JMLR W&C Track. https://proceedings.mlr.press/v68/conroy17a.html

[3] Siri, D. L., Charitha, K., Varsha, K., & Pramod, K. (2023). Enhancing clinical decision assistance by comparing patient case similarities. International Journal of Innovative Research Ideas, 11(12), 680-685. https://www.ijcrt.org/IJCRT2312749

[4] Seligson, N. D., Warner, J. L., Dalton, W. S., Martin, D., Miller, R. S., Patt, D., ... Chen, J. L. (2020). Suggestions for patient similarity categories: Outcomes from the AMIA 2019 workshop on defining patient similarity. Journal of the American Medical Informatics Association, 27(11), 1808–1812. https://doi.org/10.1093/jamia/ocaa159

[5] Mahima, V. B., Jeevinee, V., & Khan, M. S. (2024). Similarity in patient cases. International Research Journal of Modernization in Engineering Technology and Science, 6(1), 835–838. https://doi.org/10.56726/IRJMETS48246

[6] Jia, Z., Zeng, X., Duan, H., Lu, X., & Li, H. (2020). A model for diagnostic prediction based on patient similarity. International Journal of Medical Informatics, 135, Article 104073. https://doi.org/10.1016/j.ijmedinf.2019.104073

[7] Liu, Y. (2022). A set of algorithms for assessing patient similarity using electronic health records [Master’s thesis, Concordia University]. Concordia University Library. https://doi.org/10.1109/ACCESS.2022.3142100&#8203;:contentReference[oaicite:0]

[8] Memarzadeh, H., Ghadiri, N., Samwald, M., & Lotfi Shahreza, M. (2022). Research on patient similarity via representation learning from healthcare records. arXiv. https://doi.org/10.21203/rs.3.rs-1738458/v1&#8203;:contentReference[oaicite:1]

[9] Haboubi, S., & Ben Cheikh, A. (2021). Resemblance among patients in forecasting models utilizing healthcare information: The scenario of self-prescribed medications for individuals with diabetes. International Journal of Computers, 6, 33-38. https://www.iaras.org/iaras/journals/ijcOei