

Project Report

**“Patient Case Similarity”**

Batch Details

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Roll Number** | **Student Name** |
| 1 | 20211CDV0033 | Rajeshwari C Raikar |
| 2 | 20211CDV0034 | Nisha L |
| 3 | 20211CDV0041 | Katta Vinod Kumar |
| 4 | 20211CDV0055 | Amruth Raj P |
| 5 | 20211CDV0056 | K Vishnu Vardhan |

**School of Engineering,**

**Presidency University, Bengaluru.**

Under the guidance of,

Mr. RAJAN THANGAMANI,

Assistant Professor,

School of Computer Science,

Presidency University, Bengaluru.

**INTRODUCTION:**

In healthcare, data comes in many forms, including numerical values (such as lab results), categorical information (such as diagnoses), and unstructured text (like clinical notes). Analysing this diverse and complex data to find patterns, such as identifying similar patient cases, is vital for improving medical care but can be challenging.

Identifying similar patient cases can help healthcare providers make more informed and personalized decisions regarding diagnoses, treatments, and outcomes. Traditionally, this process has been manual, requiring time and subject to human bias. However, with advancements in machine learning and natural language processing (NLP), there is now the potential to automate this task.

This project focuses on developing a system that combines machine learning and NLP techniques to analyze both structured and unstructured patient data. By using algorithms that group patients based on shared characteristics and applying similarity measures, the system can compare patient cases effectively. A user-friendly web application will allow clinicians to quickly retrieve these similarity scores, helping them make data-driven decisions that enhance patient care.

**LITERATURE SURVEY:**

The *Patient Case Similarity Project* explores how machine learning and natural language processing (NLP) can be used to compare patient cases. This review covers key studies in machine learning in healthcare, similarity measures, NLP applications, and case studies, as well as challenges in implementing these systems.

1. **Machine Learning in Healthcare**

Machine learning is widely used in healthcare to analyse large datasets and make predictions. *Rajkumar et al. (2019)* studied deep learning models to predict patient outcomes from electronic health records (EHRs), emphasizing the importance of selecting the right features and reducing data complexity to improve accuracy. *K-Means clustering*, a popular method for grouping patients with similar traits, was shown by *Hernández et al. (2020)* to help in patient segmentation, which can improve personalized treatment. Another clustering technique, *DBSCAN*, has been useful in identifying patterns and outliers in patient data, which is important for understanding unusual cases.

1. **Similarity Measures in Medical Data**

Measuring the similarity between patient cases is crucial for comparing them. *Zhang et al. (2018)* demonstrated how *Cosine Similarity* can be used to compare clinical notes and medical records, helping find similar cases. Other methods like *Euclidean distance* (for numerical data) and the *Jaccard Index* (for categorical data) are also widely used, with the choice of method depending on the type of data being analysed.

1. **Natural Language Processing (NLP) in healthcare**

NLP is a key tool for analysing unstructured medical text, like clinical notes. *Kawasaki et al. (2020)* highlighted how NLP can extract useful information from these texts, making it easier to analyse patient data. Techniques such as *Named Entity Recognition (NER)* and *topic modelling* help identify important medical terms and relationships in the text.

Modern NLP models like *BERT* have been shown to improve text analysis by creating detailed

numerical representations of medical language. *Rashid et al. (2021)* showed how BERT improves the accuracy of patient similarity analysis by better understanding medical texts.

1. **Applications and Case Studies**

Many studies show the benefits of patient similarity analysis in real-world healthcare. *Chen et al. (2019)* developed a system to help doctors diagnose rare diseases by comparing patient data to historical cases. Similarly, *Mansoor et al. (2021)* used patient similarity to predict treatment outcomes, showing that it can improve personalized care by matching new cases with past similar cases.

1. **Challenges and Future Directions**

Despite its potential, there are challenges in using machine learning and NLP in patient case analysis. *Data privacy* is a major concern, as medical data is sensitive. Ensuring patient confidentiality while analysing data is crucial. Another challenge is *integrating different types of data* (numerical, text, etc.), which requires advanced frameworks.

There is also the issue of *model interpretability*—healthcare providers need to understand how these models work to trust their results. Future research should focus on making models easier to understand, while ensuring patient data remains secure and private.

**OBJECTIVES:**

1. **Clustering Algorithms for Patient Similarity Objective:** Implement clustering techniques to group similar patient cases.

Tasks:

* K-means: Use K-means clustering for grouping patients based on medical data, fine-tuning the number of clusters (k) for optimal results.
* Hierarchical Clustering: Apply hierarchical clustering to visualize patient case relationships through dendrograms.
* DBSCAN: Use DBSCAN to identify clusters based on density, useful for irregular or unknown cluster shapes.
* Dimensionality Reduction: PCA: Apply Principal Component Analysis (PCA) to reduce the dimensionality of high-dimensional patient data for efficient clustering.
* T-SNE: Use t-SNE for visualizing clusters in 2D/3D space, helping with interpretation.

1. **Similarity Metrics for Various Data Types Objective:** Apply appropriate similarity measures for different types of data.

Tasks:

* Cosine Similarity: Use for text data, such as patient notes or diagnoses (from NLP-extracted features).
* Euclidean Distance: Apply to numerical data such as age, lab results, or vital signs for comparing patient cases.
* Jaccard Index: Employ for binary categorical data (e.g., presence or absence of symptoms or conditions).
* Deep Learning Embeddings: BERT/GPT Embeddings: Use NLP-based models to generate vector representations of patient text data (e.g., medical reports, case notes).

1. **Implementation in Python Objective:** Use Python libraries for clustering, similarity, and deep learning.

Tasks:

* Scikit-learn: Implement clustering algorithms (K-means, Hierarchical, DBSCAN), and similarity measures (cosine, Euclidean, Jaccard) using the scikit-learn library.

1. **NLP Libraries:** Use libraries like transformers (Hugging Face) or spaCy to implement deep learning embeddings (BERT, GPT) for text similarity.
2. **Web Development Objective:** Build an interactive web app for visualizing and interacting with patient similarity results.

Tasks:

* Front-end (HTML/CSS/JS): Create an intuitive user interface to visualize patient similarity results, allow users to query and search for similar cases.
* Back-end (Python/Flask): Develop the backend using Flask to handle data input, clustering, similarity calculations, and serve results to the front-end.
* Database (MySQL): Set up a MySQL database for storing patient data and similarity clusters. Ensure efficient query handling for similarity searches.

1. **Evaluation Objective:** Evaluate the effectiveness of clustering and similarity analysis.

Tasks:

* Cluster Validation: Use metrics like Silhouette Score or Davies-Bouldin Index to assess the quality of clusters in high-dimensional data. Address the challenge of high dimensionality and complexity by optimizing clustering and similarity measures.
* Domain-Specific Evaluation: Validate clusters and similarity results with domain experts (e.g., healthcare professionals). Compare system generated similarities with real-world patient case studies.

1. **Addressing Limitations and Challenges Objective:** Handle potential issues such as limited generalization, high dimensionality, and scalability.

Tasks:

* High Dimensionality: Implement dimensionality reduction techniques (PCA, t-SNE) to overcome challenges in clustering and similarity in high-dimensional space.
* Limited Generalization: Regularly retrain models using new patient data and improve generalization through cross-validation and domain-specific tuning. By breaking down the objectives in this way, you can ensure each aspect of the project—from clustering to web development and evaluation—will be addressed methodically. Let me know if you'd like to refine or expand any of these objectives!

## **EXPERIMENTAL DETAILS:**

## **Hardware and Software Used:**

## Hardware: Personal computer or cloud computing resources (if needed for large datasets).

## Software: Python with libraries such as Pandas, Scikit-learn, Matplotlib, Seaborn, Flask (for the web app).

## **METHODOLOGY:**

## **Design Procedure:**

## Step 1: Load and clean the dataset

## Remove duplicate or irrelevant observations.

## Step 2: Perform exploratory data analysis to understand the distribution of features.

## Analyze the distribution and relationships of features through statistical summaries and visualizations.

## Step 3: Apply dimensionality reduction techniques (PCA) if necessary.

* Reduce feature complexity to simplify the data while retaining important variance.

## Step 4: Apply clustering algorithms (e.g., K-Means, DBSCAN) to group similar patients.

## Use methods like K-Means or DBSCAN to group patients based on their feature similarities.

## Step 5: Evaluate the results and identify key attributes defining each cluster.

## Assess the quality of clusters and determine the most significant features differentiating each cluster.

## Step 6: Build a model that can predict cluster membership for new patients.

## Develop a model to classify new patients into one of the identified clusters based on their features.

## Step 7: Visualize the clusters and important features.

## Create visualizations such as scatter plots, cluster maps, or feature importance charts to illustrate the clusters.

## Step 8: Interpret the results and make actionable recommendations.

## Draw insights from the clusters and suggest actionable recommendations for patient care or treatment strategies based on the findings.

## 

## **Timeline for Execution of Project:**

## 

## **Expected Outcomes:**

**1. Improved Diagnostic Accuracy and Treatment Planning:**

* By analysing both structured (numerical, categorical) and unstructured (text) patient data, the system enhances the accuracy of diagnosis and treatment decisions. Clinicians can access similar historical cases, helping them make more evidence-based, personalized decisions, which leads to better patient outcomes.

**2. Automation of Patient Case Analysis:**

* The project automates the traditionally manual process of reviewing and comparing patient cases. This automation saves time for healthcare professionals and reduces the risk of human error or subjective bias, allowing for quicker and more objective comparisons of cases.

**3. Integration of Diverse Patient Data:**

* The system successfully integrates structured data (such as patient vitals and lab results) with unstructured data (such as clinical notes). This unified patient representation enables a more holistic view of patient health, which is particularly useful for personalized medicine and case-based reasoning.

**4. Development of a User-Friendly Web Application:**

* A web-based interface built with Flask allows clinicians to input new patient data and retrieve real-time similarity scores. This easy-to-use application provides seamless access to valuable insights, supporting clinicians in their everyday decision-making processes.

**5. Application of Advanced Machine Learning and NLP Techniques:**

* The project demonstrates the effective application of **K-Means** and **DBSCAN** clustering algorithms for patient segmentation. Additionally, it incorporates **Cosine Similarity** and NLP models like **BERT** to process unstructured text (e.g., clinical notes) and convert it into numerical embeddings. These advanced techniques improve the system’s ability to identify relevant patient cases.

**6. Scalable and Adaptable Solution:**

* The system is designed to be scalable, capable of handling diverse types of medical data across different healthcare settings. This adaptability ensures that the solution can be applied to various clinical scenarios and is not limited to specific datasets or conditions.

**7. Contributions to Global Healthcare Goals (SDGs):**

* By contributing to **SDG 3 (Good Health and Well-Being)** and **SDG 10 (Reduced Inequalities)**, the project promotes better healthcare outcomes, reduces the gap in healthcare quality across regions, and fosters innovation in medical technology. The system supports personalized care while enhancing access to digital health solutions.

**8. Improved Clinical Workflows:**

* The automation of patient similarity analysis and case comparison streamlines clinical workflows, allowing healthcare professionals to make faster and more informed decisions. This efficiency improves patient care by reducing the time needed for manual review and enabling more effective use of healthcare resources.

**Conclusion:**

The Patient Case Similarity Project has successfully demonstrated the potential of machine learning and natural language processing (NLP) in improving healthcare decision-making. By clustering patients based on their medical data and identifying the key characteristics within these clusters, the system has provided actionable insights that can support personalized treatment and diagnosis. Additionally, the development of a predictive model allows for the classification of new patients into relevant clusters, offering a powerful tool for data-driven clinical decision-making.

The project’s integration of structured and unstructured patient data, combined with advanced clustering algorithms and similarity measures, has shown promise in enhancing diagnostic accuracy and treatment planning. The user-friendly web application makes this system accessible to healthcare professionals, enabling

real-time analysis and improving clinical workflows.

In conclusion, the Patient Case Similarity Project holds significant potential for real-world healthcare applications, offering a scalable, adaptable solution that can contribute to the advancement of personalized medicine. Through its focus on patient-centric care, the project paves the way for more accurate, efficient, and personalized healthcare delivery, ultimately improving patient outcomes.

**References:**

* **Dai, L., Zhu, H., & Liu, D.** (2020). *Patient similarity: Methods and applications*. arXiv. <https://arxiv.org/abs/2012.01976>
* **Conroy, B., Xu-Wilson, M., & Rahman, A.** (2017). *Patient similarity using population statistics and multiple kernel learning*. Proceedings of Machine Learning for Healthcare 2017 JMLR W&C Track. https://proceedings.mlr.press/v68/conroy17a.html
* **Siri, D. L., Charitha, K., Varsha, K., & Pramod, K.** (2023). *Improving clinical decision support through patient case similarity*. International Journal of Creative Research Thoughts, 11(12), 680-685. <https://www.ijcrt.org/IJCRT2312749>
* **Seligson, N. D., Warner, J. L., Dalton, W. S., Martin, D., Miller, R. S., Patt, D., ... Chen, J. L.** (2020). Recommendations for patient similarity classes: Results of 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>
* **Mahima, V. B., Jeevinee, V., & Khan, M. S.** (2024). Patient case similarity. International Research Journal of Modernization in Engineering Technology and Science, 6(1), 835–838. <https://doi.org/10.56726/IRJMETS48246>
* **Jia, Z., Zeng, X., Duan, H., Lu, X., & Li, H.** (2020). A patient-similarity-based model for diagnostic prediction. International Journal of Medical Informatics, 135, Article 104073. <https://doi.org/10.1016/j.ijmedinf.2019.104073>
* **Liu, Y.** (2022). A family of algorithms for patient similarity based on electronic health records [Master’s thesis, Concordia University]. Concordia University Library. [https://doi.org/10.1109/ACCESS.2022.3142100&#8203;:contentReference[oaicite:0](https://doi.org/10.1109/ACCESS.2022.3142100&)]
* **Memarzadeh, H., Ghadiri, N., Samwald, M., & Lotfi Shahreza, M.** (2022). A study into patient similarity through representation learning from medical records. arXiv. [https://doi.org/10.21203/rs.3.rs-1738458/v1&#8203;:contentReference[oaicite:1](https://doi.org/10.21203/rs.3.rs-1738458/v1&)]
* **Comito, C., Falcone, D., & Forestiero, A.** (2022). AI-driven clinical decision support: Enhancing disease diagnosis exploiting patients' similarity. IEEE Access, 10, 6878-6888. [https://doi.org/10.1109/ACCESS.2022.3142100&#8203;:contentReference[oaicite:0](https://doi.org/10.1109/ACCESS.2022.3142100&)]
* **Haboubi, S., & Ben Cheikh, A.** (2021). Similarity of patients in predictive models using medical data: Case of auto-prescription drugs for diabetic patients. International Journal of Computers, 6, 33-38. <https://www.iaras.org/iaras/journals/ijcOei>
* **R. W., Hsu, W., Lee, M. L., & Tan, N. C.** (2023). Using similar patients to predict complication in patients with diabetes, hypertension, and lipid disorder: A domain knowledge-infused convolutional neural network approach. Journal of the American Medical Informatics Association, 30(2), 273–281. [https://doi.org/10.1093/jamia/ocac212&#8203;:contentReference[oaicite:0](https://doi.org/10.1093/jamia/ocac212&)]
* **Wang, N., Huang, Y., Liu, H., Fei, X., Wei, L., Zhao, X., & Chen, H.** (2019). Measurement and application of patient similarity in personalized predictive modeling based on electronic medical records. BioMedical Engineering OnLine, 18(98), 1-15. <https://doi.org/10.1186/s12938-019-0718-2>