PATIENT CASE SIMILARITY

Batch Number: CAI-G24

Roll Number	Student Name
20211CAI0062	Pranav Ganesh
20211CAI0063	Prerna Kakade
20211CAI0069	Bhuvana V
20211CAI0085	Nida Aiyman

Under the Supervision of,

Dr. Mohammadi Akheela Khanum
Professor
School of Computer Science and Engineering
Presidency University

Name of the Program: **B.Tech**
Name of the HoD: **Dr. Zafar Ali Khan**
Name of the Program Project Coordinator: **Dr. Afroz Pasha**
Name of the School Project Coordinators: **Dr. Sampath A K / Dr. Abdul Khadar A / Mr. Md Ziaur Rahman**

Introduction

- Patient Case Similarity is used in healthcare systems, particularly in clinical decision support systems, to give similarity scores between the new and old patients.
- In this project, we are comparing a new patient's data with a historical patient. The data is gathered from electronic health records (EHRs) and various research papers.
- This is done to identify similar patterns and predict the disease. Further, we can also recommend treatment options based on the patterns.
- The main goal is to cluster patients based on heart diseases, diabetes and other diseases.
- After which we will improve the diagnostic accuracy, predictive models and optimize treatment plans for the patients.

Literature Review

Research Paper	Year	Advantages	Disadvantages
i) Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases	2008	• Improve case similarity measurement • Integrate natural language processing (NLP) for feature abstraction	The study acknowledges limitations, including the focus on only four feature types and the lack of contextual information in weighing features. Future research should explore the impact of additional features and contextual factors on similarity measures to enhance their applicability.
ii) Patient Similarity: Emerging Concepts in Systems and Precision Medicine	2016	• It aims to provide a foundation for the integration of computational tools and data analytics to enhance personalized healthcare. • The primary focus is on how patient similarity algorithms can transform medical decision-making by grouping patients.	• Data Challenges • Scalability Issues

Research Paper	Year	Advantages	Disadvantages
iii) Patient similarity for precision medicine: A systematic review	2018	• The primary aim of this line of research is to improve clinical outcomes for individual patients through more precise treatment targeting by leveraging on genetic, biomarker, phenotypic, or psychosocial characteristics. • That distinguish a given patient from others with similar clinical presentations	• Limited Database Scope • Lack of Real-World Application
iv) A patient-similarity-based model for diagnostic prediction	2019	• To simulate the clinical reasoning of doctors, retrieve analogous patients of an index patient automatically and predict diagnoses by the similar/dissimilar patients. • The main goal is to predict patient diagnoses by comparing the similarities between the clinical features of current patients and historical patient data.	• Limited Dataset Size • High Computational Costs

Research Paper	Year	Advantages	Disadvantages
v) Measurement and application of patient similarity in personalized predictive, modelling based on electronic medical records	2019	• The main goal of this research was to create a new way to measure how similar patients are, based on the data from electronic medical records (EMRs). • By measuring patient similarity, the study aimed to improve how we predict a patient's health outcomes, specifically focusing on diabetes.	• The study didn't fully use all the available data when calculating patient similarity • The models didn't include specific exclusion criteria when choosing patients for the study.
vi) Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding	2019	• Develop a framework to measure clinical similarities between patients based on EHRs. • Preserve temporal information in patient data, which is often lost in existing models.	• Loss of temporal information in existing methods: • High dimensionality and sparsity

Research Paper	Year	Advantages	Disadvantages
vii) Patient-Case Similarity	2020	• Develop a system to identify patients with similar medical histories. • Improve decision-making processes in clinical settings using patient data.	• Data Quality Dependency • Complex Medical Cases
viii) Patient similarity: methods and applications	2020	• Analyze and compute similarities between patients using electronic health records (EHRs), genetic, and other data. • Improve predictive models in healthcare by integrating patient-specific data from various sources.	• Information Loss • Complexity in Implementation

Research Paper	Year	Advantages	Disadvantages
ix)Patient similarity analytics for explainable clinical risk prediction	2021	• To develop an explainable and interpretable Clinical Risk Prediction Model (CRPM) by leveraging patient similarity analytics, specifically to improve explainability and interpretability. • To use real-world data from electronic medical records of patients with type-2 diabetes, hypertension, and dyslipidemia in Singapore to develop and validate the patient similarity model	• Incomplete Variable Set • Static Data Usage
x)A Novel Patient Similarity Network (PSN) Framework Based on Multi-Model Deep Learning for Precision Medicine	2022	• Utilizes multi-model deep learning to identify similarities among patients. • This patient similarity network (PSN) approach aims to enhance precision medicine by combining multiple data types, such as clinical records, genetic information, and imaging data.	• Data Heterogeneity and Dimensionality • Limited Availability of Open Datasets

Research Paper	Year	Advantages	Disadvantages
xi) Deep Dynamic Patient Similarity Analysis: Model Development and Validation in ICU	2022	• Develop a Novel Dynamic Patient Similarity Model • Validate Model Using Clinical Tasks	• Limited Clinical Application • Computational Complexity
xii) Patient Case Similarity	2024	• Improve healthcare analytics by leveraging data science techniques to enhance diagnostics, treatment recommendations, and patient care outcomes • The system utilizes machine learning (ML) and natural language processing (NLP) to identify similarities between patient cases, enabling more personalized, data-driven medical interventions.	• Prescription Recommendation Accuracy • Scalability Concerns

Existing method Drawback

i. Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases – 2008:-

The limitations include focus on only four feature types and the lack of contextual information in weighing features. Future research should explore the impact of additional features and contextual factors on similarity measures to enhance their applicability.

ii. Patient Similarity: Emerging Concepts in Systems and Precision Medicine – 2016:-

Complexity of algorithms – many of the proposed algorithms are not yet optimized for real-world clinical use due to their complexity and reliance on high-end computing infrastructure.

iii. Patient similarity for precision medicine: A systematic review – 2018:-

Lack of Deep Learning Exploration - This paper has minimal discussion on deep learning approaches, which are becoming increasingly important in handling high-dimensional medical data and patient similarity analysis. This could be a missed opportunity to explore advanced methods.

iv. A patient-similarity-based model for diagnostic prediction – 2019:-

Low Success Percentage - The model's success percentage (the percentage of patients for whom diagnoses were correctly predicted) is low (19%).

v. Measurement and application of patient similarity in personalized predictive, modelling based on electronic medical records – 2019 :-

The models didn't include specific exclusion criteria when choosing patients for the study. This could affect the accuracy of the predictions because not all patients may be equally relevant for the predictive task

vi. Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding – 2019:

Electronic Health Records are complex, and patient records contain sparse and high-dimensional data.

vii. Patient-Case Similarity – 2020:

The use of machine learning models leads to overfitting, where the model performs well on training data but fails to generalize new or unseen data.

viii. Patient similarity: methods and applications – 2020:-

During data transformation and integration, particularly in early integration strategies, there is a risk of losing valuable patient information.

ix. Patient similarity analytics for explainable clinical risk prediction – 2021:

Incomplete Variable Set - The model does not include all relevant variables like gender, race, diet, and lifestyle, which are associated with diabetes, hypertension, and hyperlipidemia (DHL) complications. The absence of such data limits the model's comprehensiveness.

x. A Novel Patient Similarity Network (PSN) Framework Based on Multi-Model Deep Learning for Precision Medicine – 2022

Data Heterogeneity and Dimensionality - The combination of diverse static and dynamic clinical data, including both structured and unstructured formats, presents difficulties in accurate modeling. Handling this high-dimensional, heterogeneous data is challenging and can lead to loss of information during data reduction processes such as those using autoencoders.

Information Loss During Encoding: The use of autoencoders for dimensionality reduction may result in accuracy loss.

xi. Deep Dynamic Patient Similarity Analysis: Model Development and Validation in ICU – 2022

Need for Clinical Protocol - The model requires a clinical protocol for practical implementation, which hasn't been discussed in this research

xii. Patient Case Similarity – 2024: Prescription Recommendation Accuracy: The system shows lower accuracy in prescription recommendations (61.27%), indicating room for improvement in this area.

Proposed Method

1. Data collection and Preprocessing
2. Feature Engineering
3. Model Development
4. Training and Evaluating the model
5. Clusters and Similarity Score
6. Web Development

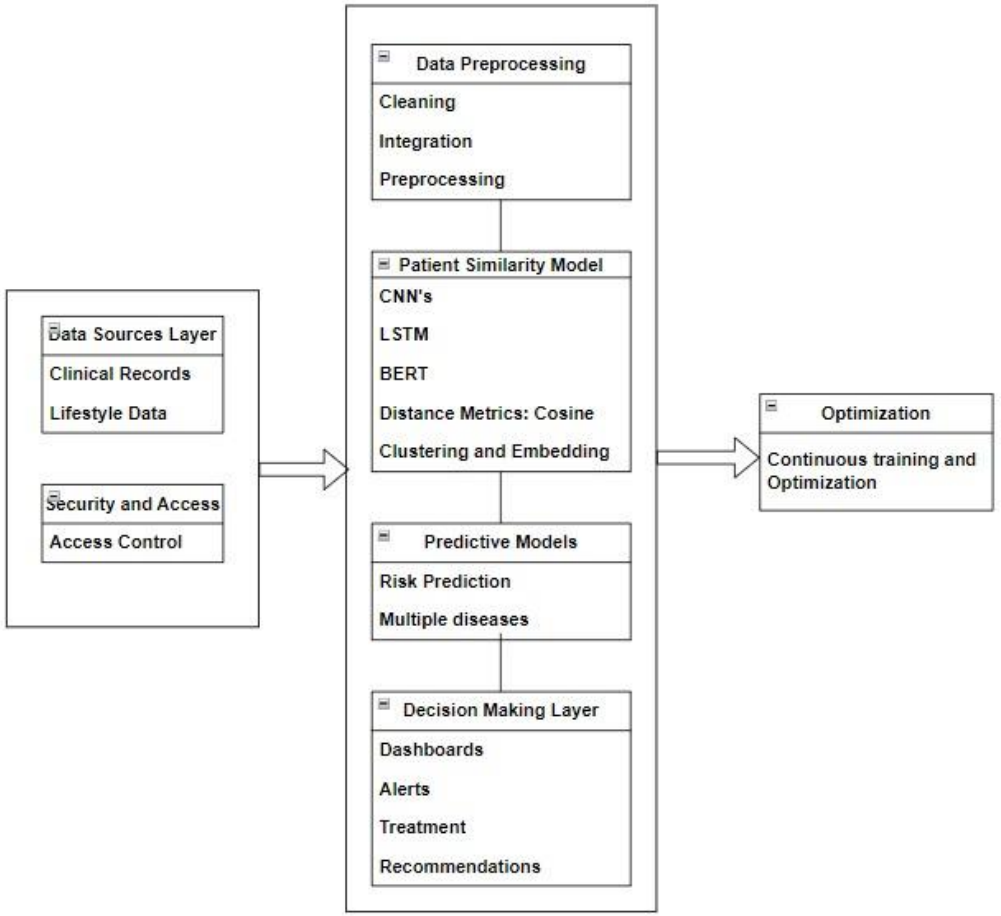
Objectives

- The goal is to leverage historical patient data to improve diagnosis, treatment planning, and outcomes.
- To provide a system which facilitates data driven clinical decision making. This gives insights on patients with similar symptoms and provides treatment accordingly.
- Enhancing risk predictions - Improving predictive models by analyzing patient similarities, allowing accurate diagnosis across a variety of diseases, not limited to heart diseases and diabetes.
- Contributing to reducing costs and improving the overall quality of healthcare.

Methodology/Modules

1. Data collection and Preprocessing: Using libraries such as pandas, numpy and scikit learn.
2. Feature Engineering: Using the transformer library and converting text to embeddings which can be used as features.
3. Model Development: Exploring models such as CNN and LSTM with the help of libraries such as tensorflow or keras.
4. Training and Evaluating the model: Training the model using test/train sets and for evaluation using metrics such as accuracy, precision, recall and F1 score.
5. Clusters and Similarity Score: Creating clusters by using clustering algorithms such as K-means and DBSCAN and finding the similarity scores based on the type of diseases.
6. Web Development: Frontend interface will help in visualizing results and allow user interactions.

Architecture



Timeline of Project

G24 PATIENT CASE SIMILARITY			
Project Start Date:	04-09-2024		Time allocated
Current Date:	14-10-2024		Progress
Weeks:	Wk 6		

Task	Start Date	End Date	Days	Progress
Review 0				
Title Selection & PPT	04-09-2024	13-09-2024	9	100%
GitHub Repository	05-09-2024	13-09-2024	8	100%
Finalizing Objectives	08-09-2024	13-09-2024	5	100%
Methodology	08-09-2024	13-09-2024	5	100%
Review 1				
Abstract	12-09-2024	15-10-2024	33	100%
Software Details	12-09-2024	15-10-2024	33	100%
Exploring Research papers	16-09-2024	15-10-2024	29	100%
Architecture Diagram	18-09-2024	15-10-2024	27	100%
Source Code Details	18-09-2024	15-10-2024	27	75%
Review 1 Report	18-09-2024	15-10-2024	27	100%
Review 2				
Algorithm	15-10-2024	19-11-2024	35	0%
50% implementation of Code	15-10-2024	19-11-2024	35	0%
50% Report Completion	15-10-2024	19-11-2024	35	0%
Review 3				
100% implementation of Code	19-11-2024	17-12-2024	28	0%
100% Report Completion	19-11-2024	17-12-2024	28	0%
Live Demo	19-11-2024	17-12-2024	28	0%
Final Viva				
Live Demo	19-11-2024	17-12-2024	28	0%
Plagiarism Check	19-11-2024	17-12-2024	28	0%
Hard Copy Report Submission	19-11-2024	17-12-2024	28	0%
Publication of Research Paper	19-11-2024	17-12-2024	28	0%

Hardware/Software components

Hardware:-

Laptop (11th Gen Intel(R) Core(TM) i5-11320H @ 3.20GHz 2.50 GHz, 16 GB RAM)

Software:-

i. Operating System:Windows

ii. Programming Language: Python 3.7 or higher - The primary language for machine learning and deep learning development.

iii. Deep Learning Libraries and Frameworks:TensorFlow with Keras API for ease of use and model building.

iv. Machine Learning Libraries:

- Scikit-learn: For classical machine learning algorithms, data preprocessing, and basic model evaluation.
- NumPy, Pandas: For data manipulation and analysis.
- Matplotlib, Seaborn: For data visualization.

v. Database Systems:SQL Database - for structured patient data storage.

vi. Jupyter Notebooks: For running interactive Python code and sharing results with the research team.

vii. Git: For version control and collaboration.

Expected Outcomes

- To diagnose diseases accurately by comparing old and new patient data.
- To have successfully created a web application that allows doctors to access patient data easily and view the similarity scores while patients have easy access to the treatments.

Conclusion

- Our project is mapped to SDG-3, that is, Sustainable Development Goals of Good Health and Well-Being.
- Our project combines advanced machine learning techniques and web applications to create a system for patient case similarity.
- Each area of healthcare can be iteratively improved based on performance metrics which helps achieve our goal of accurate similarity scores.

Github Link

Patient Case Similarity

References

i. Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases – 2008: https://www.sciencedirect.com/science/article/pii/S1532046408000440
ii. Patient Similarity: Emerging Concepts in Systems and Precision Medicine – 2016: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2016.00561/full
iii. Patient similarity for precision medicine: A systematic review – 2018: https://doi.org/10.1016/j.jbi.2018.06.001
iv. A patient-similarity-based model for diagnostic prediction – 2019: https://doi.org/10.1016/j.ijmedinf.2019.104073
v. Measurement and application of patient similarity in personalized predictive, modelling based on electronic medical records – 2019: https://doi.org/10.1186/s12938-019-0718-2
vi. Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding – 2019: https://arxiv.org/pdf/1902.03376
vii. Patient-Case Similarity – 2020: https://www.researchpublish.com/upload/book/Patient%20Case%20Similarity-8606.pdf
viii. Patient similarity: methods and applications – 2020: https://arxiv.org/pdf/2012.01976
ix. Patient similarity analytics for explainable clinical risk prediction – 2021: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01566-y
x. A Novel Patient Similarity Network (PSN) Framework Based on Multi-Model Deep Learning for Precision Medicine – 2022: https://doi.org/10.3390/jpm12050768
xi. Deep Dynamic Patient Similarity Analysis: Model Development and Validation in ICU – 2022: https://www.sciencedirect.com/science/article/pii/S0169260722004151
xii. Patient Case Similarity - 2024: https://www.doi.org/10.56726/IRJMETS48246

Project work mapping with SDG

- Our project is mapped to SDG-3,that is, Good health and Well Being.
- Our main goal is to cluster patients based on heart diseases, diabetes and other diseases and give similarity scores between old and new patients.
- This will help in improving the diagnostic accuracy, predictive models and optimize treatment plans for the patients.

Thank You