PATIENT CASE SIMILARITY

[1]Ms. Prerna Kakade, [2]Ms.Nida Aiyman, [3]Ms. Bhuvana V, [4]Mr. Pranav Ganesh, [5]Dr. Mohammadi Akheela Khanum
(student, student, student, student, professor)

Department of Computer Science and Engineering
Presidency University, Bengaluru, India

{prerna.20211cai0063, nida.20211cai0085, bhuvana.20211cai0069, pranav.20211cai0062, akheela.khanum}
@presidencyuniversity.in

Abstract: **The main aim of Patient Case Similarity web app is used to help doctors and researchers by comparing new patient data with past cases. It uses electronic health records and medical research to find patterns, and predict diseases. Our web app groups patients with similar conditions and gives a similarity score. This is used to spot trends and improve the accuracy of diagnosis. It also helps to build better prediction tools to improve patient care. We have applied machine learning algorithms which gives an innovative approach that enhances the modern day medical decision. This approach leads to better patient similarity outcomes.**

Keywords: **Similarity Score, Logistic Regression Model, K-Means Clustering, Silhouette Score**

1. INTRODUCTION

Patient Case Similarity is used in healthcare systems, mainly in clinical decision support systems, to give similarity scores between the new and old patients. In this project, we are developing a web application designed for doctors and researchers to enhance patient care and medical research by comparing a new patient's data with a historical patient. The data is gathered from electronic health records (EHRs) and various research papers. The main goal is to cluster patients based on heart diseases. After which we will improve the diagnostic accuracy and predictive models to give the similarity score between the patients.
First, we load the data into a program. Then, we look at the data to see patterns or connections. If there's missing information, we fix it by either filling it in or removing it. Next, we use charts to make the data easier to understand. We split the data into two parts: one to train the model and one to test it. We train the model with the first part and check how well it works with the second part. Lastly, we group similar data together to find patterns and give the similarity score.

2. LITERATURE REVIEW

The major work in the area of patient case similarity are:
Case-based reasoning (CBR) is a problem-solving paradigm used for improving case similarity measurement and integrating natural language processing for feature abstraction.[1]
The goal is to set the foundation for the integration of computational tools and data analytics to enhance personalized healthcare.[2]
The primary aim is to improve clinical outcomes for individual patients through more precise treatment targeting by leveraging on genetic, biomarker, phenotypic, or psychosocial characteristics.[3]
A patient-similarity-based framework is used to simulate the clinical reasoning of doctors, retrieve analogous patients of an index patient automatically and predict diagnoses by the similar/dissimilar patients.[4]
The main goal of individualized predictive modeling based on similar patients was to create a method to measure how similar patients are using data from electronic medical records (EMRs).[5]
Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding created a framework to measure clinical similarities between patients using EHRs and kept track of time-related information in patient data, which is often missed in other model.[6]
Patient similarity analysis is used to compute similarities between patients using electronic health records (EHRs), genetic, and other data.[7]

Development of an explainable and interpretable Clinical Risk Prediction Model (CRPM) by leveraging patient similarity analytics, specifically to improve explain ability and interpretability.[8]

The Patient Similarity Network (PSN) approach aims to improve precision medicine by using different types of data, like clinical records, genetic data, and imaging.[9]

A novel dynamic patient similarity model developed and validated model using clinical tasks.[10]

3. RESEARCH GAPS OF EXISTING METHODS

The study only looks at four types of features and doesn't consider the context when weighing them. Future research should look at how adding more features and context can improve similarity measures.[1]

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.[2]

Lack of Deep Learning Exploration - The paper talks very little about deep learning, which is now important for analyzing complex medical data and finding patient similarities. This might be a missed chance to use better methods.[3]

Low Success Percentage - The model's success percentage (the percentage of patients for whom diagnoses were correctly predicted) is low (19%).[4]

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.[5]

Electronic Health Records are complex, and patient records contain sparse and high-dimensional data.[6]

During data transformation and integration, particularly in early integration strategies, there is a risk of losing valuable patient information.[7]

The model doesn't include important factors like gender, race, diet, and lifestyle, which are linked to complications of diabetes, hypertension, and high cholesterol. Missing this data makes the model less complete.[8]

The mix of different types of clinical data, both structured and unstructured, makes it hard to create accurate models. Handling this complex data that can cause information to be lost during processes like auto encoders, which reduce the data's size. Using auto encoders for this can lead to a loss in accuracy.[9]

Need for Clinical Protocol - The model requires a clinical protocol for practical implementation, which hasn't been discussed in this research.[10]

4. PROPOSED METHODOLOGY

Our project involves various steps that have to be carried out in order to successfully achieve the goal. The first step of the process involves loading the datasets we have collected. This dataset will be used to train the model and we have loaded the data using the pandas library and its functions. Since we had multiple datasets we had to concatenate them to make a single dataset. The second step involves exploring the dataset, this step is essential because it gives an overview of the data which is instrumental in the process of building the model. While exploring the dataset we have used functions such as head(), tail(), shape(), info(), and describe(). The third step involves inspecting the data for missing values, it is an essential step as missing values and noise in the data can hamper the results and can give bad results. Using the function isnull() we can find the missing values in the data frame, and if there are any we can get rid of them by either dropping the unnecessary rows if they are less in number or filling the values using that column's mean or median values. If a column has multiple missing values and is not important, then we choose to drop those columns like is_smoking and education. The next step is preprocessing, this is a crucial step in data analysis which prepares raw data for further analysis and modeling. Once the preprocessing was done, we performed the next step of data visualization, this step will help us to understand the data better. We used Python libraries like matplotlib and Seaborn to perform visualization and obtained a few charts. A bar plot to compare the Age and Heart Rate of the patient. A Histogram to check the distribution of Heart Rates across various Frequencies. By viewing the Histogram, we can see whether the heart rate is normally distributed, skewed or bimodal. A pairplot to show the relationships between variables in the dataset.

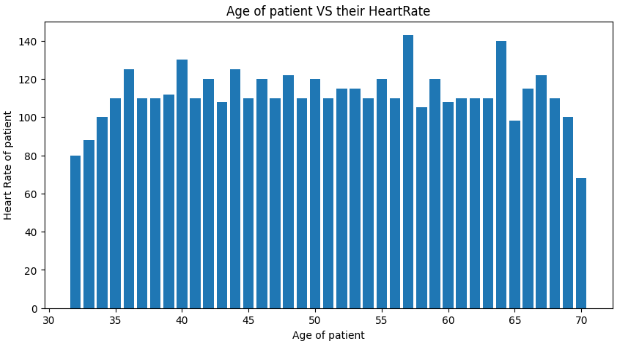


Fig 4.1: Bar Plot of Age VS. Heart Rate

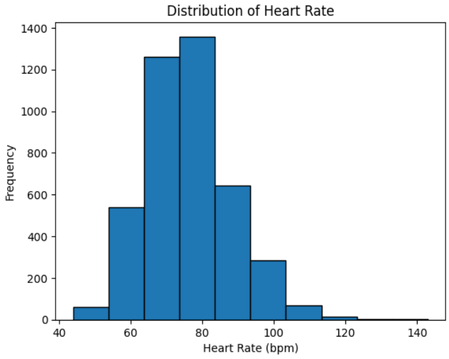


Fig 4.2: Histogram of Heart Rate

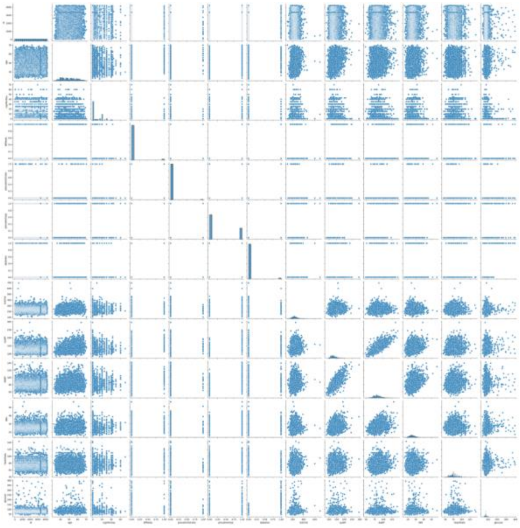


Fig 4.3: Pairplot of patient data

The next step involves splitting the dataset into training and testing. In most cases, about 70-80% of the data is used for training. And the remaining 20-30% is used for testing. In our project, we have utilized 70% of the data for training and 30% for testing. The next step involves training, in which the model is fed training data from which it learns patterns. The input training data contains features such as Age, cigsPerDay, BPMeds, prevalentStrok, prevalentHyp, Diabetes, totChol, sysBP, diaBP, BMI, heartrat, and Glucose. Once the training is done, the model is tested on new, unseen data. The model makes predictions and the results are compared to the actual outcomes to evaluate its performance. The testing phase is crucial because it makes sure that the model is not just memorizing the training data but also making accurate predictions on new unseen data. The models that we have used in our project are Logistic Regression, Decision Tree, and Random Forest. Logistic Regression is used to predict if the patient has heart disease or not. ‘1’ if the patient has heart disease and ‘0’ if the patient does not have heart disease based on features such as Age and cholesterol. Decision Tree is a flowchart that is used to make decisions based on different features. In our patient case similarity model, this model uses patient features to split the data into branches which leads to an accurate prediction. Random Forest is a collection of multiple decision trees. An ensemble method that combines multiple decision trees instead of just one to make an accurate and stable prediction. Clustering is a method in data analysis that groups similar objects or data points together, we have implemented K-Means clustering. Our code groups patient data into three clusters using K-Means Clustering and shows the results with a chart.

The average values of each feature for each cluster are calculated and shown. A scatter plot is made with patient age and cholesterol levels, and the points are colored based on their cluster.

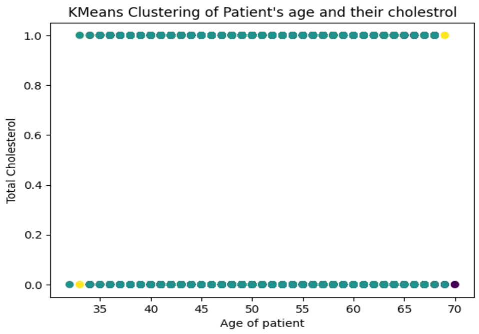


Fig 4.4: Scatter plot for K-Means Clustering

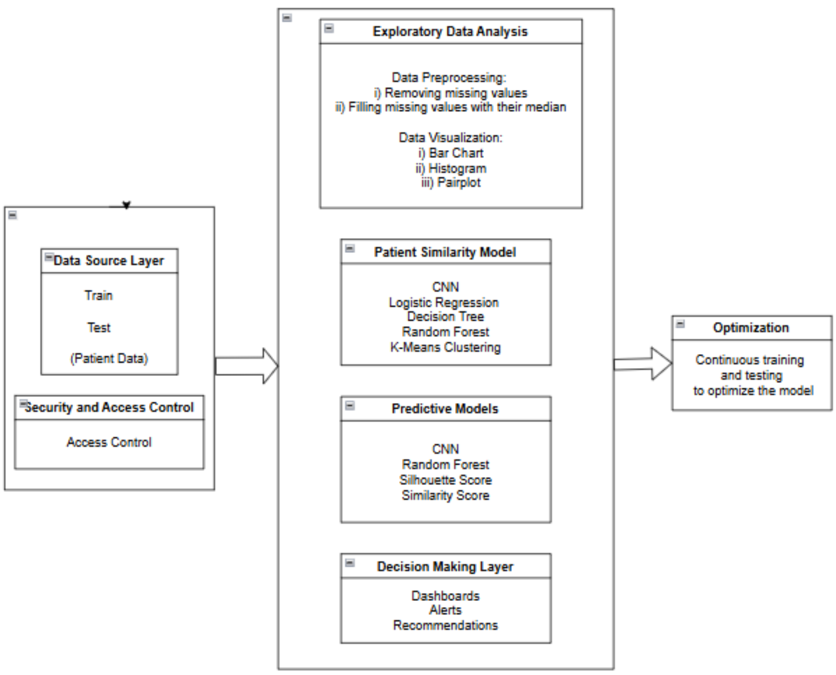
The next step involves calculating the Silhouette Score, it checks how well data points are grouped together by calculating the score for different numbers of clusters using K-Means. It compares how similar each point is to its own group compared to other groups. It gives the score for each number of clusters to help us choose the best one. A higher score testifies the groups are clear and better formed.

5. SYSTEM DESIGN AND IMPLEMENTATION

The diagram shows how the Patient Similarity System works. It starts with the data source layer, which stores patient data split into training and testing sets. This data is protected using Access Control. Next is Exploratory Data Analysis (EDA), where the data is cleaned by fixing the missing values with the median. Then, bar charts, histograms, and pair plots are used to visualize and understand the data.

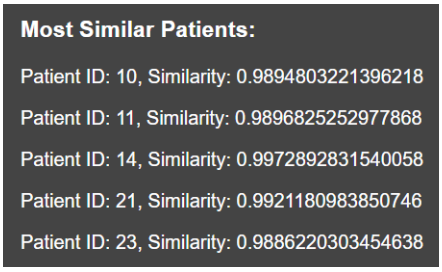
The Patient Similarity Model uses machine learning methods like CNN, logistic regression, decision trees, random forests, and K-means clustering to study patient data and find patients with similar characteristics. Then, the Predictive Models use techniques like CNN, random forests, and similarity scores to predict outcomes and assign scores.

The system's results feed into the Decision Making Layer, which generates dashboards, alerts, and recommendations to support medical decisions. Finally, the Optimization stage involves continuous training and testing of the models to improve their performance over time. This workflow ensures a seamless pipeline from data acquisition to actionable insights while focusing on security, analysis, prediction, and optimization.

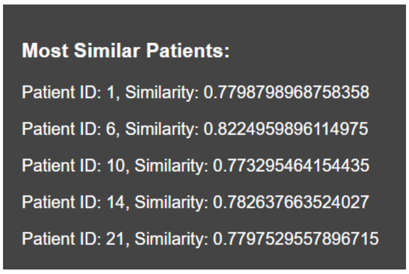


6. RESULT ANALYSIS

We merged the dataset “test.csv” and “train.csv” for data analysis and modeling. Missing values in columns such as cigsPerDay, BPMeds, totChol, BMI, heartRate, glucose, were filled with their median values.
Valuable insights were gained from visualization. A bar chart showed the relationship between age of patients and their heart rates. Histogram showed a normal distribution of heart rate values. Pairwise relationships between numeric variables were visualized using Pairplot.
Logistic Regression gave an accuracy of 85.84%, Decision Tree Classifier gave an accuracy of 75.42% and Random Forest Classifier gave an accuracy of 85.05%. The best model was determined by these accuracies, that is, Logistic Regression Model. K-Means clustering grouped patients together based on their features and for these clusters a Silhouette score was determined. Different ranges of k-values gave various silhouette scores, some even higher scores, which suggested better defined clusters.
We successfully determined the similarity score of a new patient in comparison to five patients from the historical data. The number of patients can be adjusted as desired and the model will still accurately compute the similarity scores.



```
Most Similar Patients:
Patient ID: 10, Similarity: 0.9894803221396218
Patient ID: 11, Similarity: 0.9896825252977868
Patient ID: 14, Similarity: 0.9972892831540058
Patient ID: 21, Similarity: 0.9921180983850746
Patient ID: 23, Similarity: 0.9886220303454638
```



```
Most Similar Patients:
Patient ID: 1, Similarity: 0.7798798968758358
Patient ID: 6, Similarity: 0.8224959896114975
Patient ID: 10, Similarity: 0.773295464154435
Patient ID: 14, Similarity: 0.782637663524027
Patient ID: 21, Similarity: 0.7797529557896715
```

7. CONCLUSION

The project "Patient Case Similarity" has successfully been used to demonstrate the power of Machine Learning in order to enhance clinical decision support systems. Early diagnoses of heart diseases can be made by making use of the historical patient data and identifying similarities among the patients. Machine Learning models such as Logistic Regression, Decision Trees and Random Forests have been used for reliable predictive performance. K-Means clustering has been used for grouping together of patients profiles based on similarity among them.

The project provides immense benefits because it not only reduces the overall healthcare costs by eliminating unnecessary tests but also promotes medical decision-making by improving efficiency. The user-friendly website promotes accessibility for medical professionals and patients improving user interaction and user satisfaction. It also reduces the gap between clinical practices and data driven insights by encouraging personalized medicine and better healthcare.

In the future, the project could be enhanced by expanding the dataset and combining Deep Learning for better predictive accuracy and clustering efficiency.

8. REFERENCES

[1] Cao, Hui, et al. "Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases." *Journal of biomedical informatics* 41.6 (2008): 882-888.
[2] Brown, Sherry-Ann. "Patient similarity: emerging concepts in systems and precision medicine." *Frontiers in physiology* 7 (2016): 561.
[3] Parimbelli, Enea, et al. "Patient similarity for precision medicine: A systematic review." *Journal of biomedical informatics* 83 (2018): 87-96.
[4] Jia, Zheng, et al. "A patient-similarity-based model for diagnostic prediction." *International journal of medical informatics* 135 (2020): 104073.
[5] Wang, Ni, et al. "Measurement and application of patient similarity in personalized predictive modeling based on electronic medical records." *Biomedical engineering online* 18 (2019): 1-15.
[6] Zhu, Zihao, et al. "Measuring patient similarities via a deep architecture with medical concept embedding." *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 2016.
[7] Dai, Leyu, He Zhu, and Dianbo Liu. "Patient similarity: methods and applications." *arXiv preprint arXiv:2012.01976* (2020).
[8] Fang, Hao Sen Andrew, et al. "Patient similarity analytics for explainable clinical risk prediction." *BMC*

medical informatics and decision making 21.1 (2021): 207.

[9] Navaz, Alramzana Nujum, et al. "A novel patient similarity network (PSN) framework based on multi-model deep learning for precision medicine." *Journal of Personalized Medicine* 12.5 (2022): 768.

[10] Sun, Zhaohong, et al. "Deep dynamic patient similarity analysis: model development and validation in ICU." *Computer Methods and Programs in Biomedicine* 225 (2022): 107033.