**Patient Case Similarity**

Perumalla Sai Surya1, Bhumpalli Vishnu Vardhan Reddy2, Sanjana R3, Lingamdhinne Akanksha4, Koyi Mithun5

*UG Students, Department of Computer Science and Engineering,*

*Under Guidance of* Dr. Sharmasth Vali

*Associate Professor, Department of Computer Science and Engineering,*

*Presidency University,*

*Bengaluru,*

*India.*

Email: saisuryaperumalla96092@gmail.com

***Abstract:*** *This is the approach to finding similar patients in terms of characteristics. It may shift how health care is going to develop itself, especially with the help of machine learning algorithms in finding patterns that may not be observable from the data given and in enhancing clinical decision-making. This project aims at developing a strong patient similarity analysis system based on decision trees.*

*The steps in the project include data collection and preprocessing, feature engineering, model training, evaluation, and finally, deployment. The quality and completeness of the data are necessary for any analysis. Feature engineering is actually the process of choosing and designing relevant features to describe patients. Decision trees learn decision rules that classify patients into similar groups. Some of the metrics used for determining the performance of the model are accuracy, precision, recall, and F1-score.*

*Data privacy, bias, and fairness must therefore be considered when applying the model practically. The model, therefore, must be explainable to the clinicians to gain their confidence and enhance uptake. Finally, there must be further learning for updates in the model toward achieving accuracy and relevance.*

*This will help us harness the similarity analysis of patients to enhance clinical decision-making, treatment planning, and accelerating medical research. It is a contribution toward the advancement of precision medicine, which improves patient outcomes in a broader sense.*

**Keywords –** *decision tree, similarities, preprocessing, visualization, prediction, accuracy*

1. **Introduction**

Machine learning in patience case similarity is a sub-part of artificial intelligence that studies the previous data of previous patients and tries to predict the current patient situation accurately using DECISION TREE algorithm. DECISION TREE algorithm is very suitable for this type of problem because it finds the relations between the patients by clustering the nearer values of the current patient data.

***1.1 Aims and Objectives:***

* Develop a strong and accurate decision tree-based model for patient similarity.
* To assess the performance of the model using appropriate values.
* To derive the most significant features responsible for patient similarity.
* To consider the possible uses of the model in clinical settings.

***1.2 Context and Motivation:***Data generation in the healthcare sector increased manifold in recent years with the advancements of EHR and wearable technologies. The large amount of data generated now can give an idea of providing insights, which can be transformed into benefits in the care of a patient. However, it is not easy to analyse or interpret such large amounts of data.

***1.3 Thesis Overview:***In this thesis, we attempt to investigate the application of decision trees for patient similarity analysis. We shall use the interpretability and efficiency of a decision tree in developing a model that can accurately identify similar patients with great potential in providing clinical decision-making insights. In this case, the student will explore data preprocessing, feature engineering, model training, evaluation, and interpretation. Finally, she will discuss the ethical implications of the use of patient data and suggest future possible research directions.

**2. Literature Review**

***2.1 How it will begin:***

* Data collection – Collecting different patient data with different diseases and situations.
* Data pre-processing – Identifying and removing the null values and inconsistent data in the data set for getting best accuracy.
* Data clustering – Grouping the patient data which have similarities.
* Training model
* Testing model accuracy

***2.2 Why machine learning:***

Algorithms that will present in the machine learning was so accurate We can upload more no of images in the form of dataset. And by using machine learning we even make the model for CSV files. As we all know that company like Amazon uses machine learning for the feature extraction and machine learning is used for determining the height and weight, it's dimensions where the feature extraction is very accurate Machine learning has more advantages. We will train the machine to identify and extract features according to our requirement. If we see the products in the Amazon, they can't describe the matter for every product. So by using machine learning we can compute its dimensions, We can determine precision, recall, accuracy and f1 score which was in machine learning.

***2.3 Types of problems solved using Machine Learning:***

Classification is a task that required the use of machine learning algorithms to learn how to assign a class label to a given data

* Let’s say that we are given to classifying a fruits and vegetables on basis of there category.
* Regression it help to investigate.
* The relationship between variables
* Means for example imagine if we collect a pack of apples on different stage of the year
* If we want to visualize the data x value of the each point is the day of the year It is sold and the y value is the price of the package. In this scenario, we can use to find a mathematical formula represents this data.
* This unable us to predict us the price of the apples give the day of year
* There are 3 types of regression

1. Linear regression
2. Polynomial regression
3. Logistic regression

***2.4 Types of Machine learning algorithm:***

* *Supervised Learning:* It is the method of teaching machine under the supervision and with structured data. It uses only labelled data. In this project we used supervised learning because it should learn the data with help of labels and previous condition particular diseases.
* *Reinforcement Learning:* It is the method of machine learning that learns on its own by feedback and experience. It will help this project to predict the medicine which to be used for the current patient based on old patient data. Then it checks for the changing environment and it will adapt the new environment accordingly.

***2.5 Why Python in ML for Patient case similarity:***

Python play a pivotal role in implementing ML models for patient case similarity due to its libraries and frameworks with the help of python data preprocessing is performed with libraries like pandas and numpy and scikit-learn these are commonly used for DECISION TREE and clustering and in python we use pytorch for deep learning approach and spancy for handling unstructured data and matplotlib or seaborn for visualization with the help of these libraries we can able to build scalable and accurate models.

***2.6 Breadth Context and Theory***

Literature review comprises patient similarity analysis in general and discusses their applications, challenges, and prevalence in healthcare. It will attempt to show how proper patient phenotyping is important and machine learning in healthcare plays a role.

***2.7 Work by Theme in Detail***

A concrete set of studies that have used decision trees in the similarity analysis of patients will be outlined. The methodologies applied, datasets utilized and metrics highlighted in terms of performance will be considered.

***2.8 Research Gap and Summary***

A list of gaps in current literature will be provided, including other rigorous analysis and further data sources into the decision tree models and the development of user-friendly interfaces for clinical applications.

**3. Methodology**

***3.1 Research Design***

This paper applies machine learning for creating a patient similarity model by means of decision trees. The proposed study design for the analysis of the existing patient history data is a retrospective design**.**

***3.2 Data Collection and Preprocessing***

• Sources: EHRs, clinical trials, biomedical literature

• Cleaning: Missing values, outliers, inconsistencies

• Feature Engineering: Relevant features such as demographics, medical history, lab results, genetic data

***3.3 Model Development and Training***

• Algorithm used:- Decision Tree, ID3, C4.5, CART

• Model Training-: Train the model on the pre-processed training data.

• Hyper parameter Optimization: Enhance the model by optimizing the hyperparameters succinctly

**3.4 *Model* *Evaluation***

• Evaluation Metrics: Make use of accuracy and precision, recall, F1-score, and also ROC curve for checking the performance of the model.

• Cross-validation: Test the model's generalization.

• Confusion Matrix: Also, consider taking the confusion matrix into consideration in order to know which patients are being misclassified.

***3.5 Ethics and Limitation***

• *Data Privacy:* Follow the norms of data privacy, that is, HIPAA norms.

• *Ethical Considerations:* Understand the possibilities of biasness of the model and keep the fairness of the model intact.

• *Limitations:* Discuss the limitations of the study.

1. **Analysis and Synthesis**

* *Data Analysis*: The preprocessed data is subjected to pattern and trend analysis.
* *Model Performance*: The performance of the decision tree model can be evaluated based on different metrics.
* *Feature Importance*: Identify which features are important, leading to the highest patient similarities.
* *Sensitivity Analysis*: Assess how different input parameters may be affecting the model's output.

**Flow of Project:**

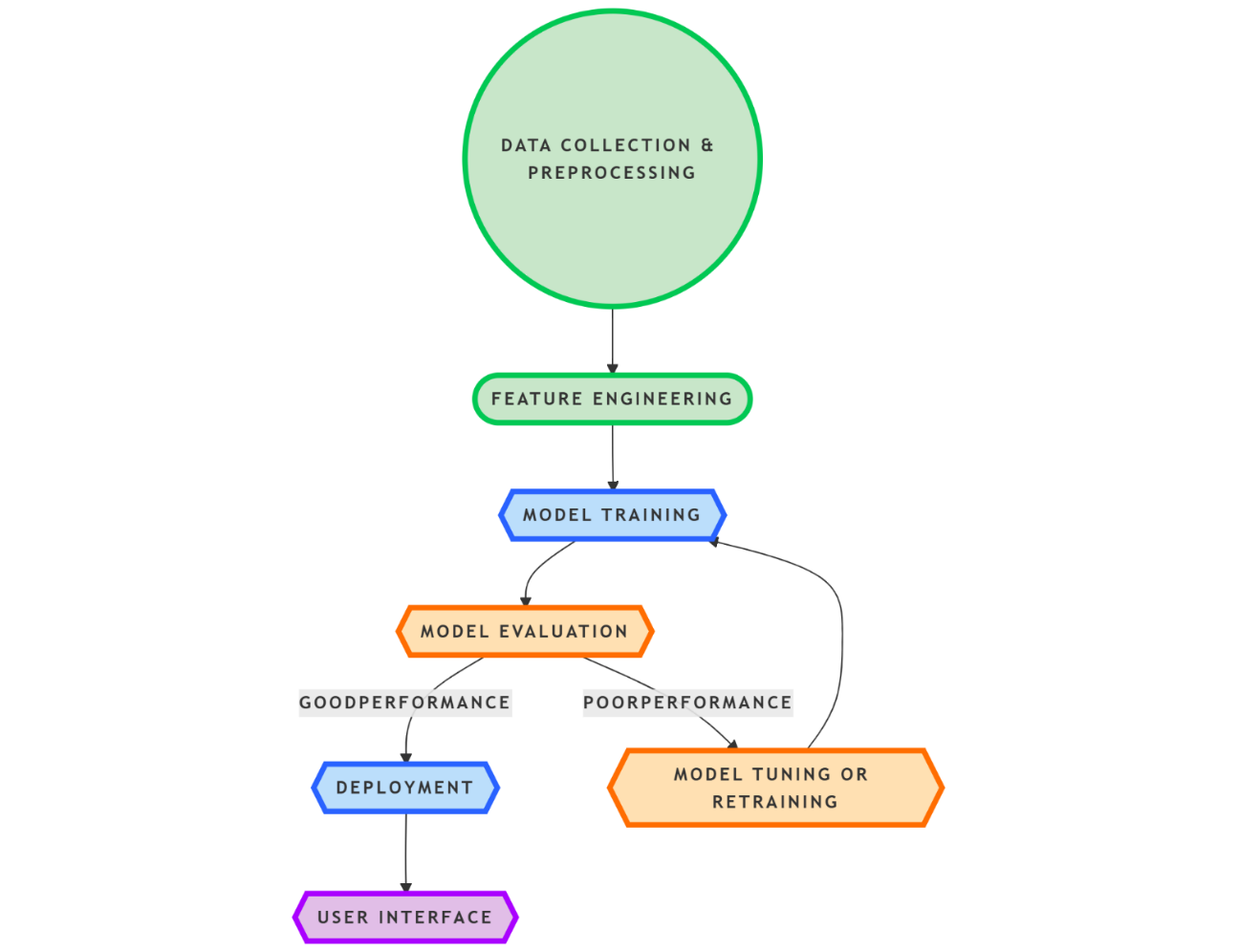


Figure-1: Work Flow of the Project

**1. Data Gathering and Preprocessing:**

* Obtain data about patients from various sources, which may include EHRs and clinical trials.
* Clean and preprocess the data, including checking for missing values, outliers, and inconsistencies
* Normalize/standardize the numerical data.
* Create feature engineering for relevant features.

**2. Feature Engineering:**

* Features that account for similarity between patients
* Feature selection techniques may also come in, like the filter methods, the wrapper method, and the embedded methods

**3. Model Selection and Training**

* Choose a suitable decision tree algorithm, such as ID3, C4.5, or CART
* Fit the decision tree model on the preprocessed and selected features.
* Tune parameters for optimal performance

**4. Model Evaluation**

* Critically evaluate the model's accuracy, precision, recall, F1-score, and ROC curve to assess the model's appropriateness.
* Test the ability of the model to generalize through cross-validation.
* Inspect the confusion matrix for the patients wrongly classified

**5. Model Deployment**

* Deploy the learned model into a clinical decision-making system or apply to relevant applications.
* While ensuring the model is optimally integrated into existing workflows.

**6. Tuning or Retraining Models:**

* In the event the model fails to perform satisfactorily, hyperparameters may be tuned, or the model may be retrained with additional data.
* Explore alternative algorithms or methods of ensemble to boost performance.

**7. Continue monitoring and enhancing the model**

* Continue to monitor the performance of your model. Retrain it periodically to maintain its accuracy.
* Interact with users to identify areas for improvement.
* Keep the model and the user interface updated with new knowledge and data.

**Implementing the Flowchart to an Agile Model:**

An Agile model like Scrum can be adopted in the development process of the patient similarity analysis.

***Scrum***

* *Sprint Planning:* Division of the project into an even more workable and manageable task set, which might be data collection, preprocessing of the data, training a model, evaluation, and finally deployment.
* *Sprint Execution:* Assign all these tasks to team members to execute iteratively.
* *Daily Scrum:* Keep daily stand-up meetings to track the progress and brainstorm any challenge.
* *Sprint Review:* Present the work to the stakeholders and collect the feedback for the work done.
* *Sprint Retrospective:* Review the sprint, identify lessons learned, and set goals for the next sprint.

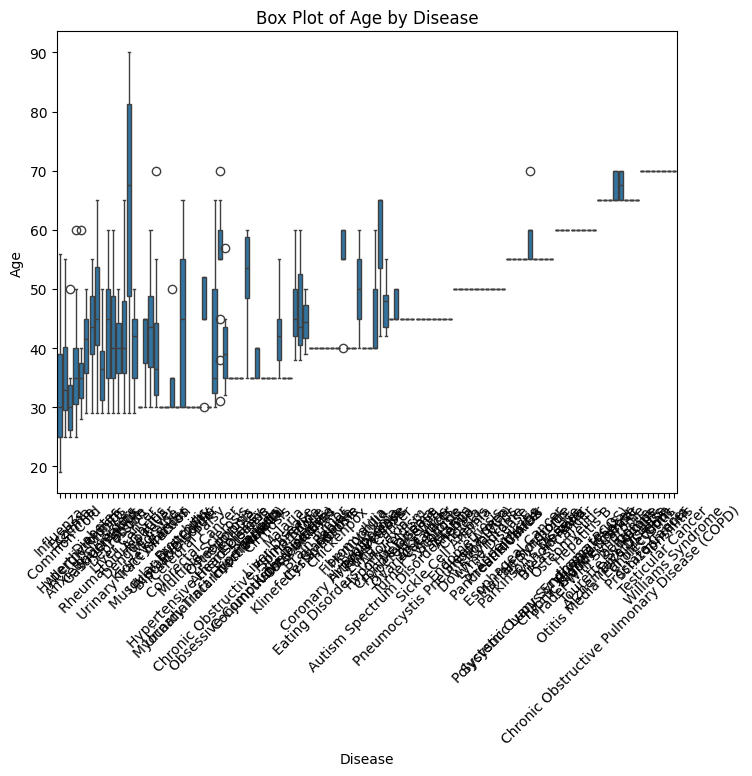
**Data Visualization for Patient Similarity Analysis**

Data Visualization is a very powerful tool in understanding and interpreting patient similarity. By visualizing the data as well as the results of the analysis, we are able to obtain valuable insights into what contributes to patients being similar to each other and into how good the decision tree model really is.

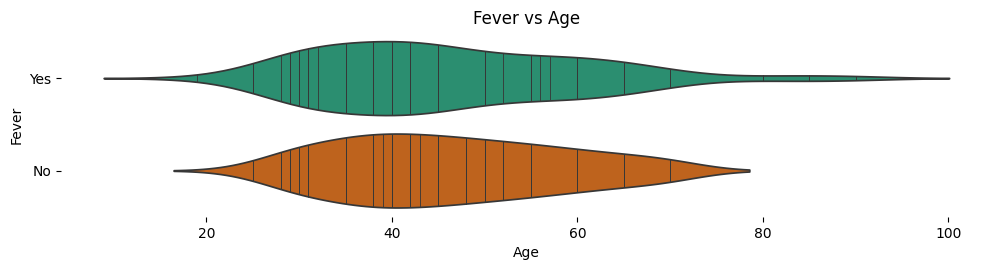
***Key Visualization Techniques:***

*Feature Importance Plots:*

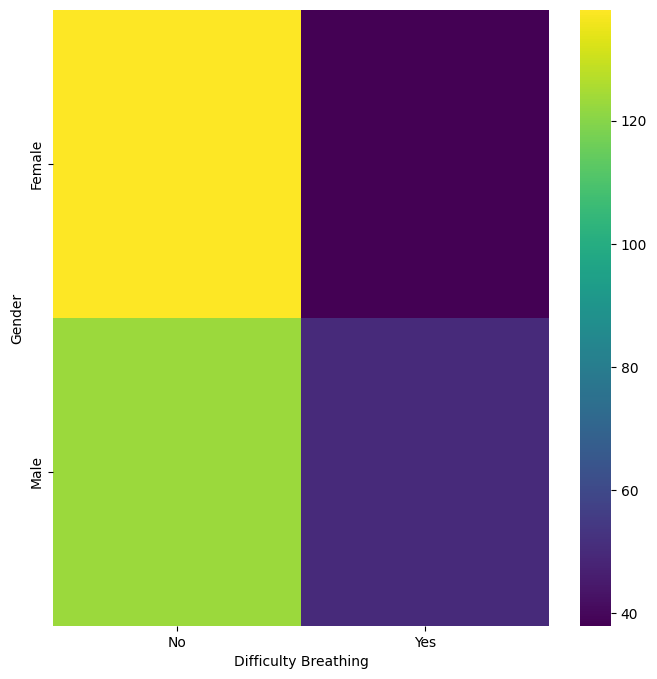
* Visualize the importance of different features in the decision tree.
* Find out which are the key factors that contributed to patient similarity.



**Figure-2:** Boxplot of Age by Disease - The boxplot provided visually shows the age spread across various diseases. It is important for patient similarity analyses, as it is important to show possible trends regarding the incidence of diseases with respect to age. Such understanding allows for clustering patients with similarities in age and diseases, making the similarity analysis more precise.



**Figure-3:** Violin plot of Fever vs Age - The violin plot would thus give a comprehensive overview of how age is distributed along various categories of the feature "Fever". It illustrates the density of and how ages are spread among cases with or without fever. In such a way, we can find what kind of pattern exists, or probably a relationship exists between age and fever by correlating the shapes and positioning of violins.



### Figure-4: Heat Map of Gender vs Difficulty in Breathing - It illustrates the relationship between gender and difficulty breathing. The frequency of occurrence of each combination is represented by the intensity of color.

### 

### Figure-5: Line plot: This line plot indicates the distribution of ages in the data set. It will represent the extent of ages, varied age groups' frequency, and outliers.

### 

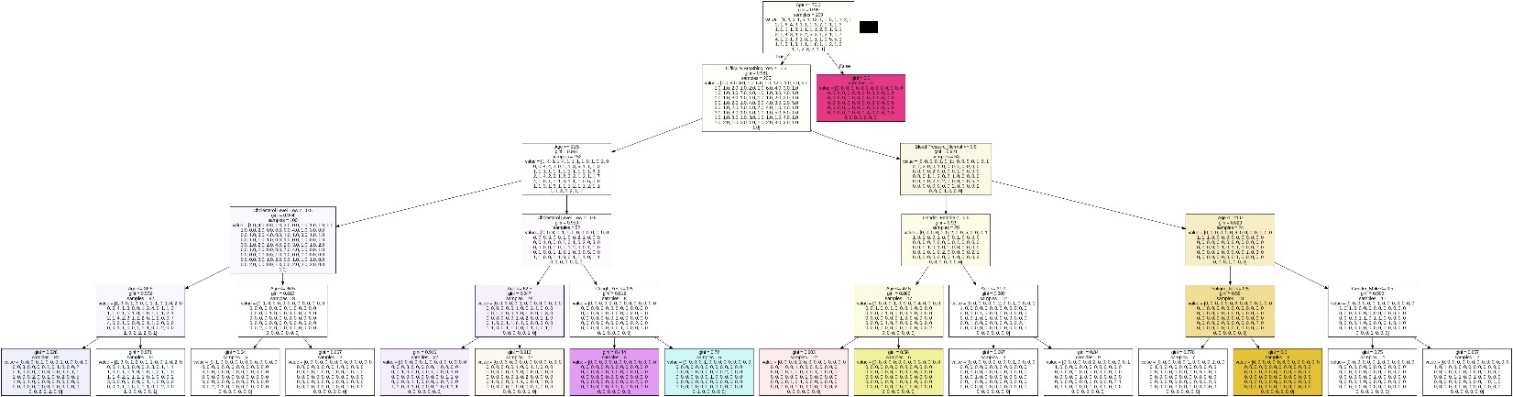
### Figure-6: Bar chart of count of the diseases according to gender - The bar chart is a distribution of diseases by gender. From the heights of the bars, we can infer potential gender disparities in the prevalence of disease. This information is very important in patient similarity analysis because it will enable us to group patients based on their gender and disease profile. This shows gender-specific patterns, by which understanding these patterns can improve the accuracy of similarity predictions through better analysis and modeling techniques.

### 

### Figure-7: Line plot of Fever vs Age - It will provide, by direct comparison, a graphic view of how the age distribution between patients who are feverish and those that are not compares. Trends between the two lines may allow some patterns or relationships between age and fever to be seen.

*Decision Tree Visualization:*

* Reflect on the structure of the decision tree.
* Describe the decision-making logic and rules used to classify patients.



### Figure-8: Decision Tree Visualization - This is an obvious and intuitive visualization of the decision tree in the model's decisions. Each node in this tree is a decision to the model based on its chosen feature, and different branches represent possible results. The leaves of the trees represent final classification or prediction.

*Patient Similarity Network:*

* Build a network graph where nodes indicate patients and edges indicate similarity of patients.
* Describe the clusters formed by patients who are similar and their characteristics.

*Time-Series Visualization:*

* Visualize patient trajectories over time for detecting patterns and trends.
* Assess comparative trajectories of similar patients to understand disease progression.

***Advantages of Data Visualization:***

1. *Better Understanding:* Visualization can enlighten complex patterns and relationships between the data.
2. *Improved Communication:* Visualization can represent insights in meaningful ways with clinicians and researchers.
3. *Informing Decisions:* Visualization can facilitate data-driven decision-making by presenting information directly and clearly.
4. *Identify Outliers:* Visualizations are used in identifying outliers and anomalies for the data.

**Role of One-Hot Encoding in Patient Similarity Analysis**

One-hot encoding is a very important technique that facilitates the conversion of categorical data into a numerical format that would be suitable for machine learning algorithms, such as decision trees. The reason one-hot encoding transposes categorical features into numerical ones is that it means that the decision tree clearly captures difference and admits better predictive results.

***Here's how one-hot encoding works:***

1. Identify categorical features from patient data, such as gender, race, diagnosis, or medication.

2. Encoding: for each categorical feature there would be a new binary feature created for each category.

3. Binary Representation: give the value of 1 to the relevant binary feature, in case it belongs to that category and otherwise 0. Suppose there's a categorical feature "Gender" with two categories, "Male" and "Female." One-hot encoding would result in two new binary features:

•  Is\_Male: 1 if the patient is male; else 0

•  Is\_Female: 1 if the patient is female; else 0

***Advantages of One-Hot Encoding:***

• *Categorical Information is Preserved:* One-hot encoding preserves the categorical nature of the data without inducing ordinal relationships between categories.

• *The model improves:* Numerical representation of categorical features yields better decision-making models.

•  *Better Interpretability:* One-hot encoding may lead to a decision tree that has more interpretations explicitly including the influence of each category.

*The following picture shows how it worked:*

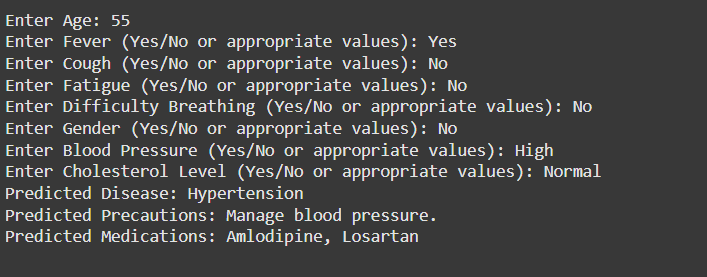


Figure-9: The Prediction of patient cases

1. **Disscussion**

***How Patient Similarity Analysis Can Be Helpful to Others***

Patient similarity analysis with decision trees and other machine learning algorithms can convert health into a better individualized and a more exact treatment approach. Some of the main benefits are as follows:

***Improvement for Patients***

Tailored Plans of Treatment Identifying similar patients would enable healthcare providers to have a chance to come up with specific plans suited to their needs hence giving good outcomes and adverse effects.

Early Onset Disease Detection Early on, it gives the chance for early identification of similar patients who may be suffering from the disease; hence it can provide the diagnosis and interventions much earlier.

Better Patient Experience Patient-centered needs and preferences are probably more familiar to provide empathetic or personalized care and services to patients.

***For Healthcare Providers:***

Enhancing Clinical Decision-Making Patient similarity analysis can be extremely useful for informed clinical decisions: treatment decisions, etc; regarding prognosis.

Optimal Resource Allocation: Patient group similarity can enable healthcare providers to make the best allocation of resources.

Research and Development: Patient similarity analysis accelerates drug development and discovery by detailing patient subgroups likely to respond well to a certain kind of treatment.

***For Researchers:***

*New Insight Discovery*: Patient similarity analysis may discover new disease subtypes and biomarkers.

*Finding of Novel Therapeutic Targets*: Mechanisms of disease will identify potential therapeutic targets.

*Advance Precision Medicine*: The similarity of patients is one of the most important factors of precision medicine, or tailored treatments for individual patients.

1. **Conclusion**

Patient similarity analysis by decision trees represents a powerful approach towards improved patient care. It matches patients who have similar characteristics, enabling clinicians to make more informed decisions related to diagnosis, treatment, and prognosis.

This project has demonstrated the classification of patients with similarities using decision trees. The developed model, having used a comprehensive dataset for its training, can predict outcomes of the patients and determine important factors that influence similarity.

In this regard, the current study faces limitations in regard to a decision tree, and more advanced methodologies need to be approached. Further inclusion of genomics and proteomics data can also improve the accuracy of the analysis with relevance to precision. In terms of adoptions, it is possible to have friendly interfaces for consumers in order to get affected by this model into real-world clinical workflows.

Further, patient similarity analysis may form the basis for future prognostications of personalized medicines directed at patients carrying out specific treatments based on individual characteristics.

**References**

1. **Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman, University of Maribor – FERI Smetanova 17, SI-2000 Maribor, Slovenia vili.podgorelec@uni-mb.si.**
2. **Han et al., “A Survey of Data Mining Techniques for Medical Diagnosis,” 2001.**
3. **"Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy," *2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*.**
4. **Azar, A.T., El-Metwally, S.M., “Decision tree classifiers for automated medical diagnosis,” *Neural Comput & Applic*, vol. 23, pp. 2387–2403, 2013.** [**https://doi.org/10.1007/s00521-012-1196-7**](https://doi.org/10.1007/s00521-012-1196-7)**.**
5. **"Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," *Information Fusion*, 2019.**
6. ***International Journal of Computer Science and Information Technology Research*, ISSN 2348-120X (online), vol. 8, issue 2, pp. 5–9, April–June 2020. Available at:** [**www.researchpublish.com**](http://www.researchpublish.com)**.**
7. **© Dillon Chrimes, Originally published in the *Interactive Journal of Medical Research*,** [**https://www.i-jmr.org/**](https://www.i-jmr.org/)**, 30 January 2023.**
8. **“How Healthcare Decision Trees Emerge and Function,” Published online by *Cambridge University Press*, 13 July 2023.**
9. **Brown, S.A., Chung, B.Y., Doshi, K. et al., “Patient similarity and other artificial intelligence machine learning algorithms in clinical decision aid for shared decision-making in the Prevention of Cardiovascular Toxicity (PACT): a feasibility trial design,” *Cardio-Oncology*, vol. 9, no. 7, 2023.** [**https://doi.org/10.1186/s40959-022-00151-0**](https://doi.org/10.1186/s40959-022-00151-0)**.**
10. **Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman, “Decision trees: an overview and their use in medicine.”**