SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V11I12P103

Volume 11 Issue 12, 16-22, December 2024

Original Article

Patient Case Similarity

Perumalla Sai Surya[1], Bhumpalli Vishnu Vardhan Reddy[2], Sanjana R[3], Lingamdhinne Akanksha[4], Koyi Mithun[5]

[1,2,3,4,5]*Computer Science & Engineering/Student, Presidency University, Bengaluru, India.*

[1]*Corresponidng Author : saisuryaperumalla96092@gmail.com*

Received: 13 October 2024 Revised: 17 November 2024 Accepted: 11 December 2024 Published: 30 December 2024

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 to develop 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, further learning is needed to update the model to achieve 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 patient 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 the DECISION TREE algorithm. THE 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. Use references to provide the most salient background rather than an exhaustive review. The last sentence should concisely state your purpose for carrying out the study or a summary of the results [2].

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 has increased manifold in recent years with the advancements of EHR and wearable technologies. The large amount of data generated now can provide insights that can be transformed into benefits in the care of a patient. However, it is not easy to analyze 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

This section should extend, not repeat, the information discussed in Introduction [4]. In contrast, a Calculation Section represents a practical development from a theoretical basis [5].

2.1. How it will Begin

- Data collection – Collecting different patient data with different diseases and situations.

- ta preprocessing – Identifying and removing the null values and inconsistent data in the data set to get the best accuracy.
- Data clustering – Grouping the patient data that have similarities.
- Training model
- Testing model accuracy

2.2. Why Machine Learning

The algorithms that will be present in the machine learning were so accurate. We can upload more images in the form of a dataset. Moreover, by using machine learning, we even made the model for CSV files. As we all know, a company like Amazon uses machine learning for feature extraction, and machine learning is used to determine height, weight, and dimensions, where feature extraction is very accurate. Machine learning has more advantages. We will train the machine to identify and extract features according to our requirements. If we see the products on Amazon, they cannot 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 in machine learning.

2.3. Types of Problems Solved using Machine Learning

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

- Let us say that we are given the ability to classify fruits and vegetables on the basis of their category.
- Regression helps to investigate.
- The relationship between variables
- For example, imagine if we collect a pack of apples at a different stage of the year.
- If we want to visualize the data, the x value of 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 that represents this data.
- This enables us to predict the price of the apples given the day of the year.
- There are 3 types of regression
  1. Linear regression
  2. Polynomial regression
  3. Logistic regression

2.4. Types of Machine Learning Algorithms

- ✓ *Supervised Learning:* It is the method of teaching machines under supervision and with structured data. It uses only labelled data. In this project, we used supervised learning because it should learn the data with the help of labels and previous conditions, particularly 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 to be used for the current patient based on old patient data. Then, it checks for changes in the environment and adapts to the new environment accordingly.

2.5. Why Python in ML for Patient Case Similarity

Python plays 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, 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

The 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 how 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 the current literature will be provided, including other rigorous analyses 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 to create a patient similarity model by means of decision trees. The proposed study design for the analysis of the existing patient history data is retrospective**.**

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 using the preprocessed training data.
- Hyperparameter 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 to check the performance of the model.
- Cross-validation: Test the model's generalization.
- Confusion Matrix: Also, consider considering the confusion matrix in order to know which patients are being misclassified.

3.5. Ethics and Limitations

- Data Privacy: Follow the norms of data privacy, that is, HIPAA norms.
- Ethical Considerations: Understand the possibilities of bias in the model and keep the fairness of the model intact.
- Limitations: Discuss the limitations of the study.

4. 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

All figures in the manuscript should be numbered sequentially using Arabic numerals (Example: Figures 1 and 2), and each figure should have a descriptive title. The figure number and title should be typed single-spaced and centered across the bottom of the Figure in 8-point Times New Roman, as shown below. The figure captions should be editable and be written below the figures.

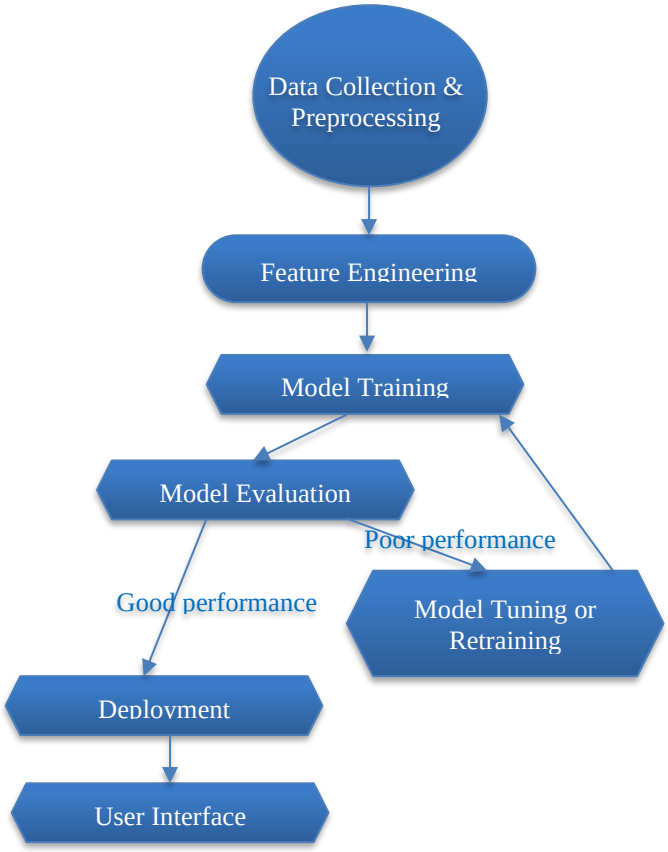

Fig. 1 Work flow of the project

4.1. 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.

4.2. 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

4.3. 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

4.4. 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.

4.5. Model Deployment

- Deploy the learned model into a clinical decision-making system or apply it to relevant applications.

- While ensuring the model is optimally integrated into existing workflows.

4.6. 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 an ensemble to boost performance.

4.7. 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.

4.8. Implementing the Flowchart to an Agile Model

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

4.8.1. 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 challenges.
- Sprint Review: Present the work to the stakeholders and collect feedback for the work done.
- Sprint Retrospective: Review the sprint, identify lessons learned, and set goals for the next sprint.

4.9. Data Visualization for Patient Similarity Analysis

Data Visualization is a very powerful tool for 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.

4.9.1. Key Visualization Techniques

Feature Importance Plots:

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

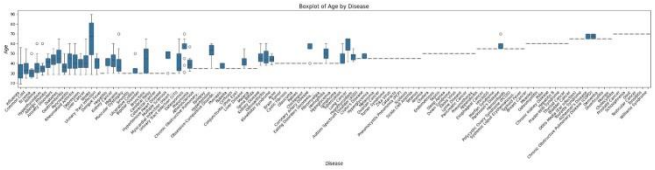

Fig. 2 Boxplot of age by disease

The boxplot provided visually shows the age spread across various diseases. Patient similarity analyses need 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.

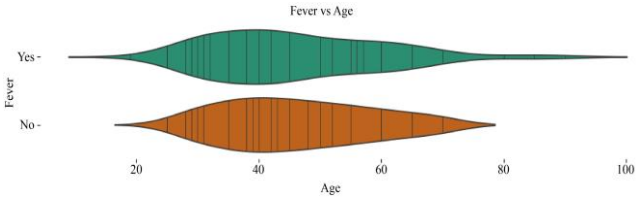

Fig. 3 Violin plot of fever vs. age

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.

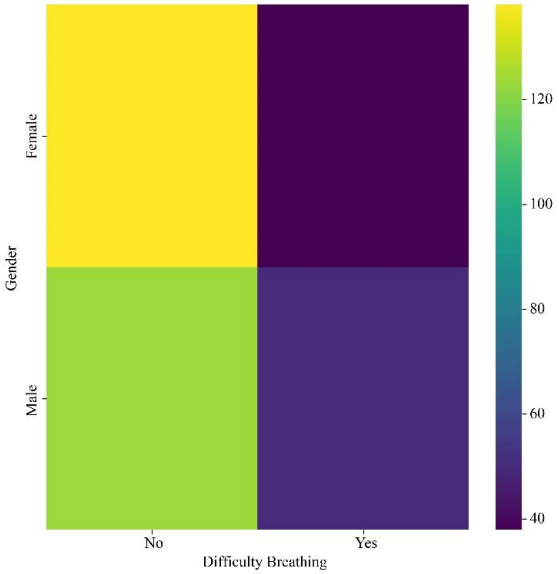

Fig. 4 Heat Map of Gender vs. Difficulty in Breathing

It illustrates the relationship between gender and difficulty breathing. The intensity of color represents the frequency of occurrence of each combination.

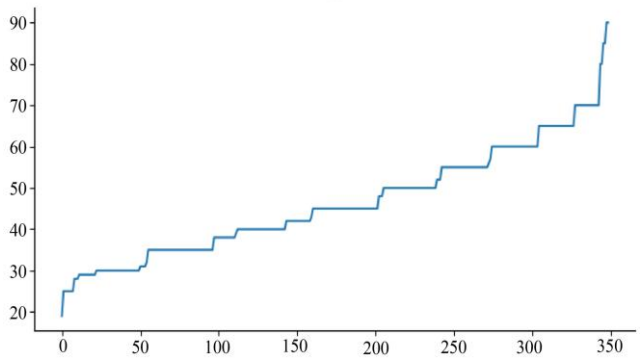

Fig. 5 Line plot

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.

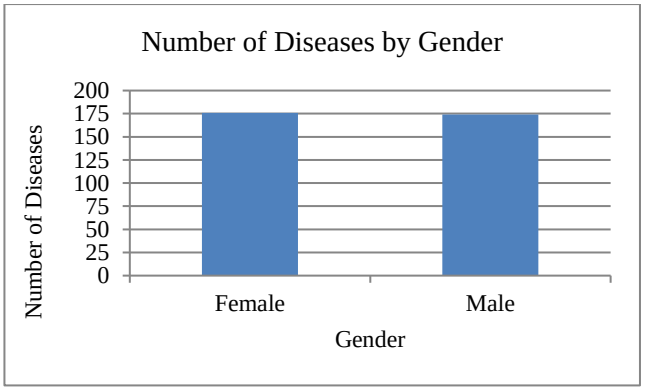

Fig. 6 Bar chart of count of the diseases according to gender

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, which can improve the accuracy of similarity predictions through better analysis and modelling techniques.

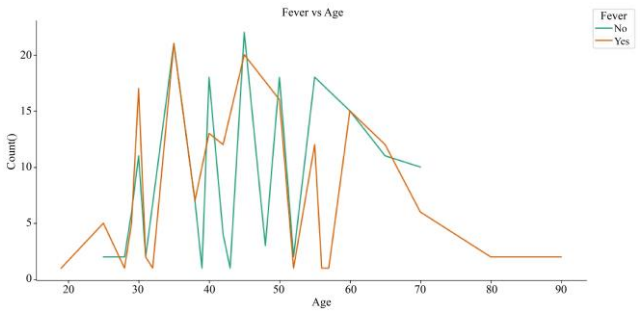

Fig. 7 Line plot of fever vs Age

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.

4.10. Decision Tree Visualization

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

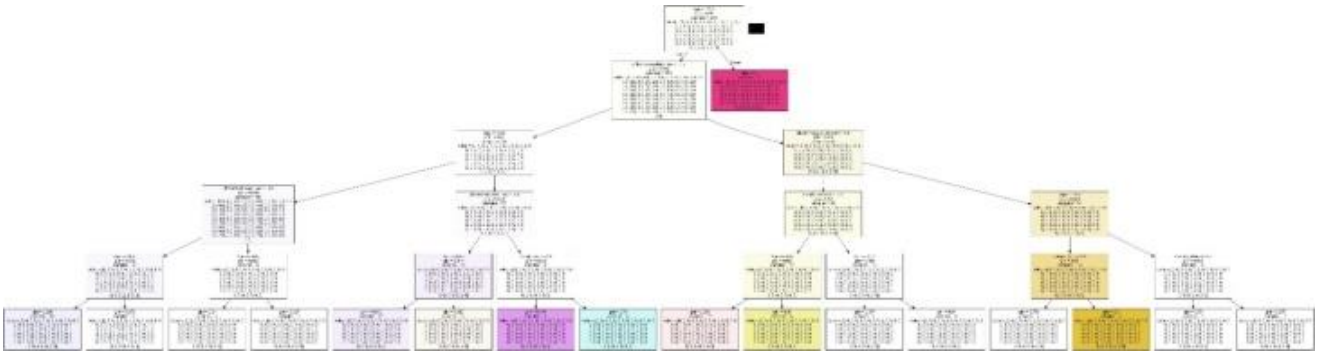

Fig. 8 Decision tree visualization

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 the final classification or prediction.

4.11. Patient Similarity Network

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

4.12. Time-Series Visualization

- Visualize patient trajectories over time to detect patterns and trends.
- Assess comparative trajectories of similar patients to understand disease progression.

4.13. 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 to identify outliers and anomalies in the data.

4.14. 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 differences and admits better predictive results.

Here's how one-hot encoding works

- Identify categorical features from patient data, such as gender, race, diagnosis, or medication.
- Encoding: For each categorical feature, a new binary feature would be created for each category.
- Binary Representation: give the value of 1 to the relevant binary feature, in case it belongs to that category and otherwise 0. Suppose there is 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

4.15. Advantages of One-Hot Encoding

- Categorical Information 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 explicit interpretations, including the influence of each category.

The following picture shows how it worked.

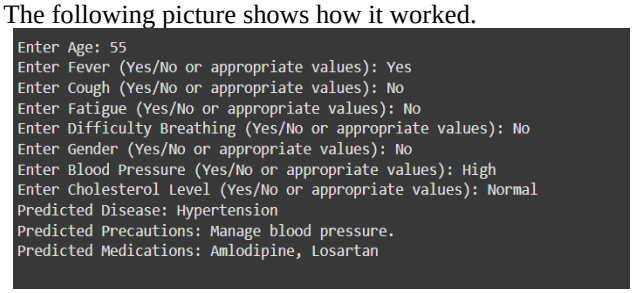

Fig. 9 The Prediction of Patient Cases

5. Discussion

5.1. 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 more exact treatment approach. Some of the main benefits are as follows:

5.2. 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 providing good outcomes and adverse effects.

Early Onset Disease Detection Early on, it gives a 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 providing empathetic or personalized care and services to patients.

5.3. 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.

5.4. 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.

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

6. 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 the 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 be affected by this model in real-world clinical workflows.

References

[1] Vili Podgorelec et al., “Decision Trees: An Overview and their Use in Medicine,” *Journal of Medical Systems*, vol. 26, no. 5, pp. 445-463, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[2] Han et al., “*A Survey of Data Mining Techniques for Medical Diagnosis*,” 2001.

[3] L.W.C. Chan et al., “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)*, Hong Kong, China, pp. 467-470, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ahmad Taher Azar, and Shereen M. El-Metwallym, “Decision Tree Classifiers for Automated Medical Diagnosis,” *Neural Computing and Applications*, vol. 23, pp. 2387-2403, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[5] Marinka Zitnik et al., “Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities,” *Information Fusion*, vol. 50, pp. 71-91, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Vishakha Sharma et al., "Patient-Case Similarity," *International Journal of Computer Science and Information Technology* Research, vol. 8, no. 2, pp. 5-9, 2020. [Publisher Link]

[7] Dillon Chrimes, "Using Decision Trees as an Expert System for Clinical Decision Support for COVID-19," *Interactive Journal of Medical Research*, vol. 12, no. 1, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Ramalingam Shanmugam, *Chapter 7 - How Healthcare Decision Trees Emerge and Function*, Data-Guided Healthcare Decision Making, Cambridge University Press, pp. 188-198, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sherry-Ann Brown 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, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]