

A Project Report

On

**“Patient Case Similarity”**

Batch Details

|  |  |  |
| --- | --- | --- |
| Sl. No. | Roll Number | Student Name |
| 1 | 20211CIT0047 | Lingamdhinne Akanksha |
| 2 | 20211CIT0043 | Sanjana R |
| 3 | 20211CIT0013 | Bhumpalli Vishnu Vardhan Reddy |
| 4 | 20211CIT0055 | Koyi Mithun |
| 5 | 20211CIT0079 | Perumalla Sai Surya |

**School of Computer Science,**

**Presidency University, Bengaluru.**

Under the guidance of,

Dr. Sharmasth Vali

School of Computer Science,

Presidency University, Bengaluru

**CONTENTS**

1. Introduction about Project
2. Literature Review
3. Objectives

## Methodology

## Literature Survey

## Dataset

1. Timeline for Execution of Project
2. Outcomes
3. Conclusion
4. References

**INTRODUCTION**

The objective of patient case similarity is to identify similar patients based on their medical reports. Identification of similar patient cases be useful for improving patient outcome, for treatment or drug recommendation to a new patient, prediction of clinical outcome, clinical decision support, research on those cases.  
  
Decision tree is a popular machine learning algorithm that can effectively applied to the patient case similarity task. They can create a tree like model where each internal node represent a test on attribute and each leaf node represents the class label.

**The process of using decision tree for patient case similarity involves following steps:**

**1. Data Preprocessing:** Clean the patient data, handles the missing values and inconsistencies.

**2. Feature Selection:** Chooses the relevant features that are used predictive for patient case similarity.

**3. Decision Tree Construction:** Builds the decision tree by recursive partitioning of the data based on most informative attributes.

**4. Similarity Assessment**: Use the constructed decision tree for classifying the new patients as similar or dissimilar to the existing patients.

**LITERATURE REVIEW**

**Advantages:**

**• Handles both numerical and categorical data:** Decision tree can handle the both types of data without requiring preprocessing.

**• Can handle missing data:** Decision tree can be handles missing values during the classification process.

**• Robust to outliers:** Decision tree is the relatively robust to the outliers, as they create partitions based on the attributes value.

**• Scalable:** Decision tree can handle the large datasets effectively.

**• Can be used for feature selection:** Decision tree can be helpful for identify the most important feature for predicting patient similarity.

**• Can be combined with other algorithms:** Decision tree can be used as part of ensemble method like random forest.

**• Can be used for both classification and regression tasks:** Decision tree can be adapted for the both tasks, making them versatile for different types of patient similarity problems.

**Limitations:**

** Overfitting:** Tree can overfit the data.

** Noise sensitive:** Tree is sensitive to noise**.**

** Biased**: Tree is biased towards feature with more levels.

** Unstable:** Small changes in the data can causes big changes in tree.

** Limited expressiveness:** Trees can’t express complex relationships between

features.

** Continuous variables:** Trees are less efficient with continuous variables.

** Imbalanced data:** Trees are biased towards the majority class.

** Computational cost:** Trees can be expensive for large data and deep trees.

** Don’t capture feature interactions:** Trees can’t capture complex feature

interactions.

**OBJECTIVES**

**Objective 1:** Develops the robust and scalable patient similarity algorithms that can be effectively identifies the similar patients across the diverse datasets and handles large-scale data.  
• Rationale: Many existing algorithms may be faces the challenges with scalability and generalizability across the different datasets. A robust and scalable algorithms can be improve the applicability and impacts of patient similarity analysis.  
 **Objective 2:** Explores the integration of the patient case similarity algorithms with clinical decision support systems to facilitate the personalized treatment recommendation and improves the patient outcomes.  
• Rationale: While patient case similarity algorithm have shown promise, their integration into the clinical workflow is crucial for the real world impact. Developing the effective integration strategy can enhances the adoption and the utilization of these algorithm.  
 **Objective 3:** Investigates the role of patient case similarity algorithm in identifying novel biomarkers and the therapeutic targets for some specific diseases.  
• Rationale: Patient case similarity analysis can be helpful for identify the subgroups of the patients with similar characteristics that can respond alternate to treatment. By analyzing the subgroups, researchers can be potentially discovers the novel biomarkers and the therapeutic targets.  
 **Objective 4:** Addresses the ethical and the privacy concerns that are associated with patient case similarity analysis, particularly in the relation with data sharing.  
• Rationale: Ensure the ethical and the responsible use of thew patient data is crucial. Developing some guidelines and frameworks to address the privacy concerns.

**METHODOLOGY-**

* **EXPERIMENTAL DETAILS**
* Microsoft Excel
* Python compiler
* **DESIGN PROCEDURE:**

**1. Problem Definition and Goal Setting:**

* Clearly defines the objective for the project, such as identifying similar patient cases for the treatment recommendations.

**2. Data Collection and Preprocessing:**

* It gathers the similar medical data, which includes patient demographics, symptoms, treatments, and outcomes.
* Cleans and preprocess the data to handle missing values, inconsistencies.
* Normalizes the numerical data to ensures comparability.
* Considers feature engineering to create the new features that can be more informative.

**3. Feature Selection:**

* Identifies the relevant features that are predictive to the patient case similarity.
* Use technique like correlation analysis, or wrapper methods to select the most effective features.

**4. Decision Tree Construction:**

* Chooses the decision tree algorithm based on characteristics of your data and the desired properties of model.
* Trains the decision tree model on pre-processed data, creating a tree-like structure where each node can represents a test on attribute and each leaf node represents class label.

**5. Model Evaluation:**

* Use the appropriate metrics to evaluate the performance of the decision tree model on the validation dataset.
* Considers the techniques like cross-validation to access the model's generalization ability.

**6. Model Optimization:**

* If initial model performance is not satisfactory, explore techniques like pruning, boosting, or bagging to improve the model accuracy and to prevent overfitting.
* Experiments with the different decision tree algorithms and hyperparameters to find the best configuration.

**7. Deployment and Use:**

* It deploys the trained decision tree model in production environment.
* It integrates the model into clinical workflow or research tool to the support decision-making and analysis.
* It continuously monitors the model performance and updates it as needed based on new data and change requirements.

**LITERATURE SURVEY:**

1. **Given the fact that applications of data mining increasingly find their way into medicine, which different techniques in data mining in terms of effectiveness and efficiency and interpretability best apply in relation to each of the following applications in medicine: diagnosis/prognosis, classification, risk factor analysis?**A. Data mining is the analysis that involved very big amounts of data in order to derive meaningful patterns and connections, thereby increasing possibilities to predict future trends and make the most suitable decisions. Data mining applications are significant in marketing, banking, medicine, etc. In this paper, we present an overview of data mining applications in medicine to provide a clear view of the challenges and previous works in this area for researchers. Data mining techniques such as Decision Tree, Random Forest, K-means Clustering, Support Vector Machine, Logistic Regression, Neural Network, Naive Bayes, and association rule mining are used for diagnosing, prognosis, classifying, constructing predictive models, and analysing risk factors of various diseases. The main objective of the paper is to analyse and compare different data mining techniques used in the medical applications. We present a summary of the results and provide comparison analysis of the data mining methods employed by the reviewed articles.  
  
  
2. **Given that COVID-19 complexity, and considering that no free publicly available chatbot could determine the severity of infection available, what factors and associations involving the body systems, viral infections, comorbidities, and the clinical presentations significantly contribute to the development of severe cases of COVID-19?**A. Viral Infection Severity: Viral load and existence of viral variants could be related to the severity of the viral infection.  
•Comorbidities- Cardiovascular disease, diabetes, and other chronic respiratory disorders that may predispose a patient. According to the introduction of this article reviewed, the following factors and interactions would significantly contribute to the development of severe cases of  
COVID-19  
Involvement of Body Systems: In this case, there could be involvement of specific body systems such as lungs and heart.t to severe COVID-19.  
Manifestations-Specific symptoms and manifestations such as shortness of breath, chest pains, and abnormal blood tests can be strong candidates for serious cases.  
The complex decision tree developed within this paper suggests that this phenomenon of developing the severity of disease with COVID cannot be put down to any one of these factors, but rather a combination of all. Further validation of these results along with more in-depth research into the additional factors which might lead to the development of disease severity is suggested.

**3.Considering the advantages of decision trees, particularly those about precision, interpretability, and validation, how have they been applied in medical decision-making, as well as what are the most important challenges and limitations that need to be addressed?**

A. They are used on an incredibly large scale in making medical decisions to diagnose, prognosticate, decide on the treatment, and even assess risk. Decision trees are appropriate for medical applications because they provide an interpretable model and can handle both numerical and categorical data.

•Some of the main challenges and limitations that characterize the use of decision trees in medical decision-making include:

• Overfitting: Decision trees can overfit over the training data and therefore will end up generalizing poorly.

• Noisy: A decision tree is highly sensitive to data containing noise in it. In other words, the accuracy is affected by noise.

**4.** **Given the advantages of decision trees for medical decision-making, what are some key challenges and limitations associated with their application in this domain, how have researchers addressed these challenges to make them better, and what can be recommended?**

**A**. Decision trees offer several advantages for medical decision making. However, there is a need to address the following limitations:

a. Overfitting: The training data may be overoptimized by decision trees, leading to poor generalization performance.

b. Sensitivity to noise: The decision tree models can be very noisy sensitive, which leads to error in predictions.

c. Biasness: Decision trees become biased if it is having more levels for features that causes biased decisions.

d. Limited expressiveness: It may not capture complex relationships between attributes.

e. Computational complexity: For large datasets and deep trees, it's computationally costly.

To overcome all these problems, researchers have approached many techniques such as:

o Pruning: shrinking the size of the decision tree to prevent overfitting. n Ensemble methods: combining more than one decision tree in order to improve the accuracy and, in some cases, also bias (the random forest, gradient boosting). n Feature selection: to reduce the dimensionality of the data some of the most relevant features are picked, which could improve model performance. n Advanced decision tree algorithms: more complex algorithms that could handle noisy and more complex relations between features.

o Hybrid approaches: Combining decision trees with other machine learning techniques to exploit their complementary strengths.

**5. On the general applicability of decision support systems in forming medical decisions, how do the different types of decision tree classifiers, one single decision tree, boosted decision trees, and decision tree forests, compare on performance and efficiency for tasks like the breast cancer detection?**

**A.** From the paper, the following conclusions can

Be drawn in regard to different performance of decision tree classifier performances compared against:

BDTs are generally superior than SDTs for general performance in terms of accuracy, sensitivity, specificity, and many other measures of performance.

In conjunction with DTFs, very good performance was also found in the validation phase and were also less overfitting prone.

 The classifier adopted might be based on some needs and limitations. For instance, SDT can be used wherever speed is of essence. BDT or DTF can be used anywhere where the tasks involved require high precision as well as dependability.

**6.** **Given that decision-making is critical in most fields, what is the correct approach to using decision trees such that value and probability are weighed upon differently for various results to illuminate the processes involved in decision-making?**

A. Decision trees provide a useful framework for bringing together two important elements: value and probability, to decision-making processes. Assigning values to different outcomes and considering their associated probabilities, decision trees can guide people and organizations to choose the expected value-maximizing alternatives.

•Value Quantification: Measurability of different outcomes is not possible-values are often intangible or subjective. Utility theory or multi-criteria decision analysis can be applied.

•Probability Estimation: Estimation of the probability of different outcomes through reliable data and appropriate statistical methods is essential. Bayesian networks or other probabilistic models can be applied along with prior knowledge and uncertainty.

• Calculating Expected Value: The expected value is found by multiplying each possible outcome by its probability and then adding up all the results. This gives you a figure that you can use to express the potential outcomes of a decision, in a quantitative manner.

• Sensitivity Analysis: The sensitivity analysis shall identify on which factors the impact occurs and which of the factors used in the problem would have an effect on the expected value such that the robustness of the decision-making process is established.

7. **Given the potential of patient similarity analysis in clinical outcome prediction, how well can various measures and algorithms of similarity identify similar patients for survival prediction in hepatocellular carcinoma?**

A. It is possible to predict the survival of HCC patients using locoregional chemotherapy by means of the similarity analysis of patients, as indicated by the study published. Results regarding promising accuracy, sensitivity, and specificity for this model were produced by utilization of a proposed algorithm called SimSVM along with its 14 clinical and imaging parameters similarity measures.

However, the external validity of these findings to larger and more diverse patient populations must be evaluated through future research. Additional similarity measures and algorithms are further explored with the aim of improving the performance of patient similarity analysis for survival prediction in HCC.

It is also important to note here the limitations of a patient similarity alone for

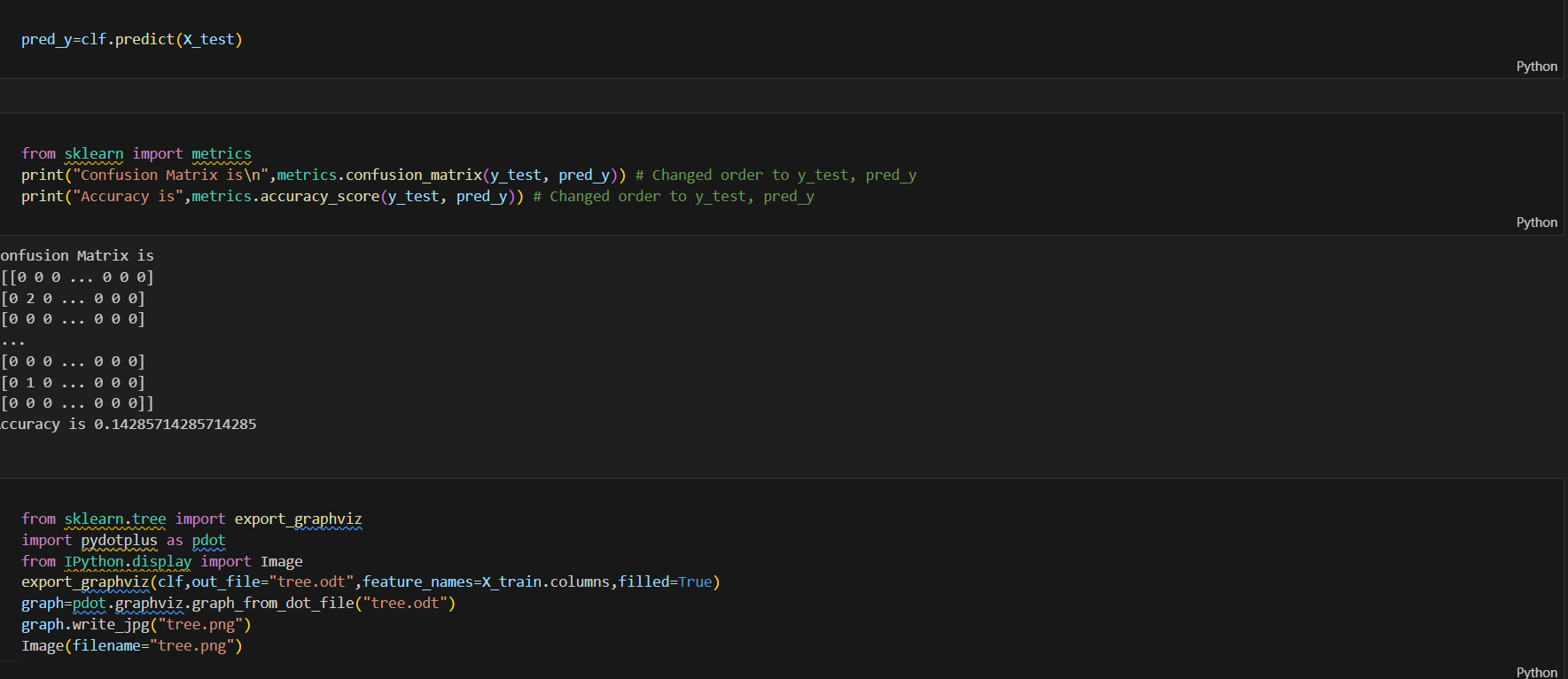
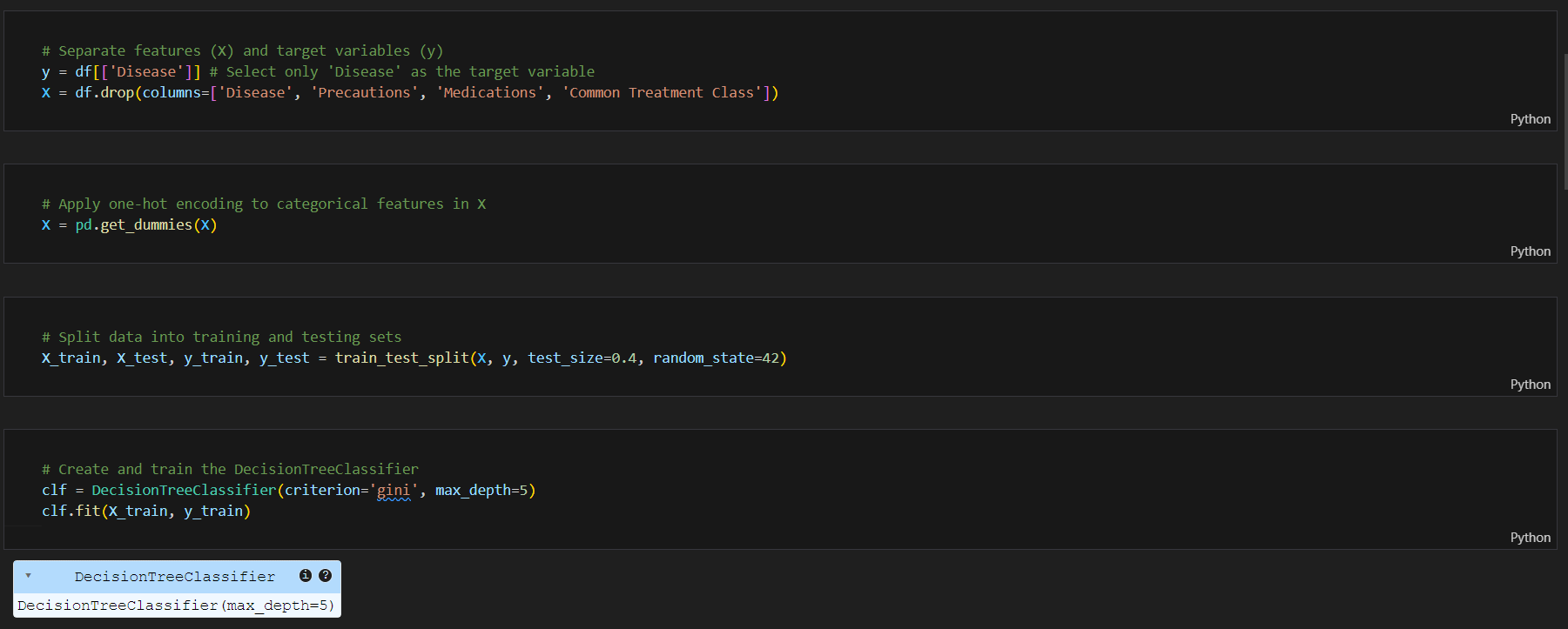
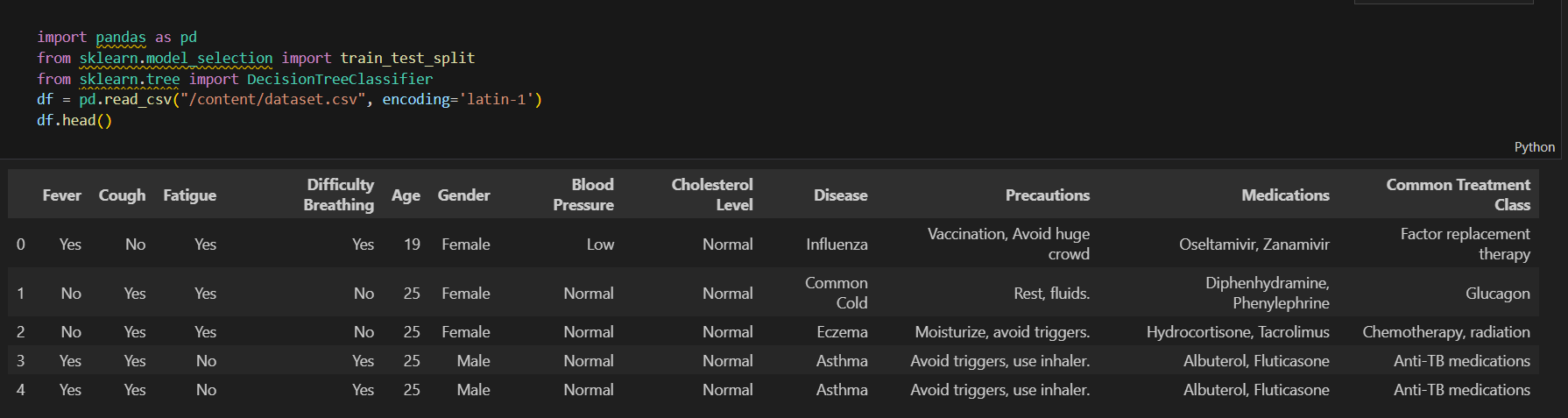
Probably the most interesting application is using the approach above to predict the outcome of clinical cases. However, there are other factors including a specific treatment protocol and the extent to which comorbidities are prevalent, that may alter survival; thus, complete analysis involving patient similarity and other predictive factors may be necessary to produce accurate and reliable predictions.

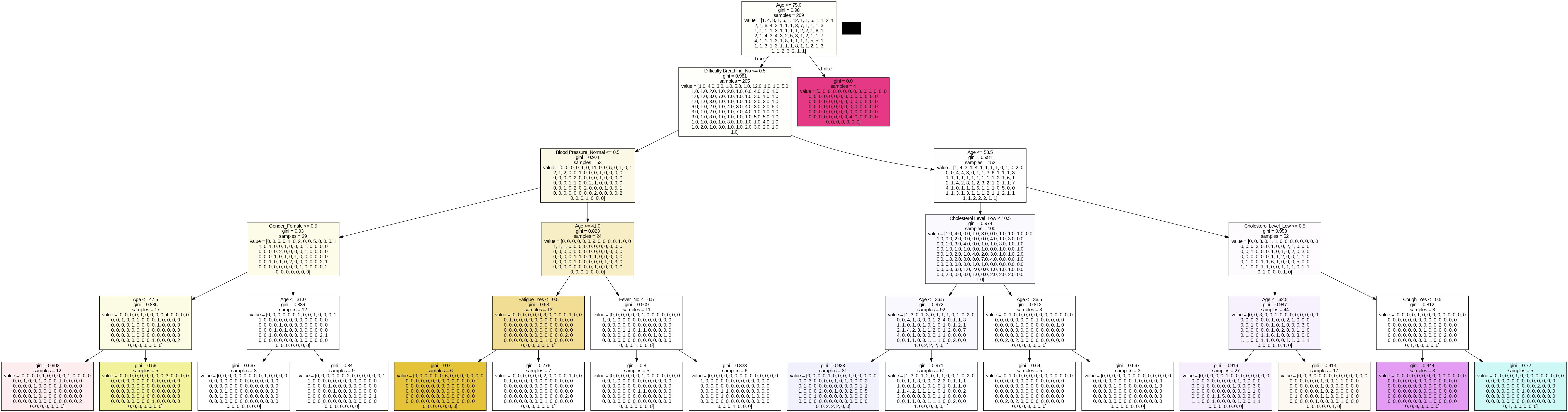
8**. With the importance of patient similarity algorithms on rising with precision medicine, what are the challenges and limitations associated with using them, and how to overcome these challenges to make it effective?**A. Patient similarity algorithms have an absolute benefit in making the clinical efficiency enhanced and provide a good amount of selection and recognition of similar patients, patient trajectory prediction, and the implementation of clinical decision-making support. In addition, however, there are a number of challenges as well.  
¬ Data Quality and Completeness: Patient data quality may affect very much the accuracy of similarity algorithms. The results would be biased if there is missing or inconsistent data.  
¬ Feature Selection: For patients' similarity, the most appropriate feature must be selected. Models with too many dimensions can suffer from overfitting; whereas, very simple models may fail to catch all the important interrelations.  
¬ Selection of Similarity Metric: The variation in similarity metric can also result in different findings. While some particular types of data and applications appear to be more appropriate for certain metrics than for others, some metrics are very broadly applied.  
¬ Computational Efficiency: Calculating patient similarity can be computationally very expensive, at least when big data and complex models are used, especially in practice. Algorithms and hardware fast enough make it possible.  
study discussed in the paper presents evidence for the promise of artificial  
in light of knowledge deficiencies in cardio-oncology and for the benefit of better patient outcomes. In theory, using machine learning algorithms trained on massive data sets, these can make accurate predictions about cardiovascular disease risk and make evidence-based recommendations about care.  
The design and implementation of such clinical decision aids must consider  
♣Data Quality and Quantity: The quality and quantity of the data used for training algorithms would determine¬  
Interpretability: Some similarity algorithms are clearly intuitive, and it is unclear which factors are actually contributing to scores; this could reduce their utility in the clinic.  
  
9. **What is the promise of AI-based clinical decision aids to advance assessment and care of the exponentially rising population of cancer survivors diagnosed with heart disease and ability to surmount barriers to attain the best possible care?**A. The study the accuracy of the prediction. More desirable large dataset and diverse set enveloping relevant clinical and imaging information is a prerequisite.  
♣Algorithm Selection: The appropriate machine learning algorithm that will be used to suit cardiovascular risk prediction is most critical. Since all the factors influencing the interpretability, computational efficiency, and performance are valid, then this needs to be taken into account.  
♣Integration with Clinical Practice: Making it a part of the electronic health record can facilitate both the healthcare provider and patient to utilize it.  
♣Ease of Use: It must be found easy to use by both the clinicians and patients in case they have to rely on it.  
♣Ethical Issues: The tool should align with ethical considerations into data privacy, bias, and how it may possibly impact clinical decision-making.  
  
10**. Indicate what would majorly challenge and how to work around it as such great potential with the approach of networks is considered in precision medicine.**A. These network-based approaches include issues such as data quality, network construction and analysis, integration with clinical data, and generalizability. The traditional techniques that have been used to address these challenges include data preprocessing, network inference methods, tools for visualization, integration frameworks, and rigorous evaluation to further advance the integration of network-based approaches into precision medicine.

**DATASET:** (sample)

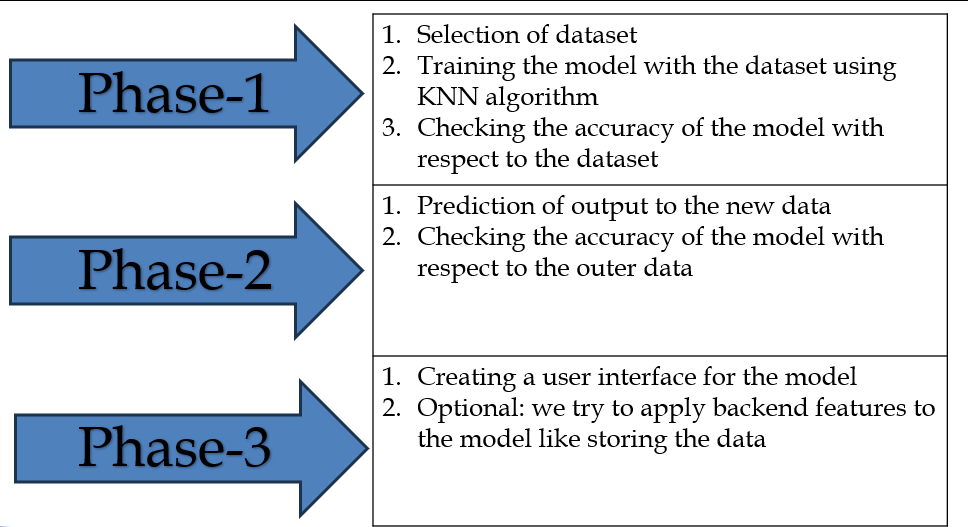
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Fever** | **Cough** | **Fatigue** | **Difficulty Breathing** | **Age** | **Gender** | **Blood Pressure** | **Cholesterol Level** | **Disease** |
| **Yes** | **No** | **Yes** | **Yes** | **19** | **Female** | **Low** | **Normal** | **Influenza** |
| **No** | **Yes** | **Yes** | **No** | **25** | **Female** | **Normal** | **Normal** | **Common Cold** |
| **No** | **Yes** | **Yes** | **No** | **25** | **Female** | **Normal** | **Normal** | **Eczema** |
| **Yes** | **Yes** | **No** | **Yes** | **25** | **Male** | **Normal** | **Normal** | **Asthma** |
| **Yes** | **Yes** | **No** | **Yes** | **25** | **Male** | **Normal** | **Normal** | **Asthma** |
| **Yes** | **No** | **No** | **No** | **25** | **Female** | **Normal** | **Normal** | **Eczema** |
| **Yes** | **Yes** | **Yes** | **Yes** | **25** | **Female** | **Normal** | **Normal** | **Influenza** |
| **Yes** | **Yes** | **Yes** | **Yes** | **25** | **Female** | **Normal** | **Normal** | **Influenza** |
| **No** | **Yes** | **No** | **No** | **28** | **Female** | **Normal** | **Normal** | **Hyperthyroidism** |
| **No** | **Yes** | **No** | **No** | **28** | **Female** | **Normal** | **Normal** | **Hyperthyroidism** |
| **Yes** | **No** | **No** | **Yes** | **28** | **Male** | **High** | **Normal** | **Asthma** |
| **No** | **Yes** | **Yes** | **No** | **29** | **Female** | **Normal** | **Low** | **Allergic Rhinitis** |
| **No** | **Yes** | **No** | **No** | **29** | **Female** | **Normal** | **High** | **Anxiety Disorders** |
| **No** | **No** | **No** | **No** | **29** | **Female** | **Low** | **Normal** | **Common Cold** |
| **No** | **No** | **No** | **No** | **29** | **Male** | **Low** | **Normal** | **Diabetes** |
| **No** | **Yes** | **No** | **No** | **29** | **Female** | **Normal** | **Normal** | **Gastroenteritis** |
| **Yes** | **No** | **No** | **No** | **29** | **Female** | **High** | **Normal** | **Pancreatitis** |
| **No** | **Yes** | **Yes** | **Yes** | **29** | **Female** | **High** | **High** | **Rheumatoid Arthritis** |
| **Yes** | **Yes** | **Yes** | **Yes** | **29** | **Male** | **High** | **Normal** | **Depression** |
| **Yes** | **Yes** | **Yes** | **Yes** | **29** | **Female** | **Normal** | **Normal** | **Liver Cancer** |
| **Yes** | **Yes** | **Yes** | **Yes** | **29** | **Female** | **Normal** | **Normal** | **Stroke** |
| **Yes** | **Yes** | **Yes** | **No** | **29** | **Male** | **High** | **High** | **Urinary Tract Infection** |
| **Yes** | **No** | **Yes** | **No** | **30** | **Female** | **Normal** | **Normal** | **Dengue Fever** |
| **Yes** | **No** | **Yes** | **No** | **30** | **Female** | **Normal** | **Normal** | **Dengue Fever** |
| **No** | **Yes** | **Yes** | **No** | **30** | **Male** | **High** | **High** | **Eczema** |
| **Yes** | **Yes** | **Yes** | **No** | **30** | **Male** | **High** | **High** | **Gastroenteritis** |
| **Yes** | **Yes** | **Yes** | **Yes** | **30** | **Male** | **High** | **Normal** | **Hepatitis** |
| **No** | **No** | **Yes** | **No** | **30** | **Male** | **Normal** | **Normal** | **Kidney Cancer** |
| **Yes** | **No** | **No** | **No** | **30** | **Female** | **Normal** | **Normal** | **Migraine** |
| **No** | **Yes** | **Yes** | **No** | **30** | **Female** | **Normal** | **Normal** | **Migraine** |
| **No** | **No** | **Yes** | **No** | **30** | **Male** | **High** | **High** | **Muscular Dystrophy** |
| **No** | **Yes** | **Yes** | **No** | **30** | **Male** | **Normal** | **Normal** | **Sinusitis** |
| **Yes** | **Yes** | **No** | **No** | **30** | **Female** | **Normal** | **Normal** | **Ulcerative Colitis** |
| **No** | **Yes** | **Yes** | **No** | **30** | **Female** | **Normal** | **Normal** | **Ulcerative Colitis** |
| **Yes** | **Yes** | **No** | **Yes** | **30** | **Female** | **Normal** | **Normal** | **Asthma** |
| **Yes** | **Yes** | **No** | **Yes** | **30** | **Female** | **Normal** | **Normal** | **Asthma** |
| **Yes** | **Yes** | **Yes** | **Yes** | **30** | **Female** | **Normal** | **Normal** | **Asthma** |
| **No** | **No** | **Yes** | **No** | **30** | **Female** | **High** | **High** | **Bipolar Disorder** |
| **Yes** | **Yes** | **No** | **Yes** | **30** | **Female** | **Low** | **Normal** | **Bronchitis** |
| **Yes** | **Yes** | **Yes** | **Yes** | **30** | **Male** | **High** | **High** | **Bronchitis** |
| **Yes** | **Yes** | **Yes** | **Yes** | **30** | **Male** | **High** | **High** | **Bronchitis** |
| **No** | **No** | **Yes** | **Yes** | **30** | **Female** | **Normal** | **Normal** | **Cerebral Palsy** |
| **No** | **No** | **Yes** | **No** | **30** | **Female** | **Normal** | **High** | **Colorectal Cancer** |
| **Yes** | **No** | **No** | **No** | **30** | **Male** | **High** | **High** | **Eczema** |
| **No** | **No** | **Yes** | **No** | **30** | **Female** | **High** | **High** | **Hypertensive Heart Disease** |
| **Yes** | **Yes** | **Yes** | **No** | **30** | **Male** | **High** | **Normal** | **Influenza** |
| **Yes** | **Yes** | **Yes** | **Yes** | **30** | **Female** | **Normal** | **Normal** | **Influenza** |
| **No** | **No** | **Yes** | **No** | **30** | **Female** | **High** | **High** | **Multiple Sclerosis** |
| **Yes** | **Yes** | **Yes** | **Yes** | **30** | **Female** | **High** | **High** | **Myocardial Infarction (Heart...** |
| **Yes** | **No** | **No** | **No** | **30** | **Female** | **Normal** | **Normal** | **Urinary Tract Infection (UTI)** |
| **No** | **No** | **No** | **Yes** | **31** | **Male** | **Normal** | **Low** | **Asthma** |
| **No** | **No** | **No** | **Yes** | **31** | **Male** | **Low** | **Normal** | **Osteoporosis** |
| **Yes** | **No** | **Yes** | **Yes** | **31** | **Male** | **High** | **High** | **Common Cold** |
| **Yes** | **No** | **No** | **No** | **31** | **Female** | **Normal** | **Normal** | **Migraine** |
| **Yes** | **No** | **Yes** | **Yes** | **32** | **Female** | **High** | **Normal** | **Pneumonia** |
| **No** | **No** | **Yes** | **No** | **35** | **Female** | **Normal** | **Low** | **Allergic Rhinitis** |
| **No** | **Yes** | **Yes** | **Yes** | **35** | **Female** | **Normal** | **High** | **Asthma** |
| **Yes** | **Yes** | **Yes** | **Yes** | **35** | **Female** | **Normal** | **Normal** | **Asthma** |
| **No** | **Yes** | **Yes** | **Yes** | **35** | **Female** | **High** | **Normal** | **Asthma** |
| **No** | **Yes** | **Yes** | **Yes** | **35** | **Female** | **High** | **Normal** | **Asthma** |
| **No** | **No** | **Yes** | **No** | **35** | **Male** | **Normal** | **Normal** | **Atherosclerosis** |
| **Yes** | **Yes** | **Yes** | **Yes** | **35** | **Male** | **Normal** | **Normal** | **Chronic Obstructive Pulmonary...** |
| **Yes** | **Yes** | **Yes** | **No** | **35** | **Male** | **High** | **Normal** | **Common Cold** |
| **No** | **Yes** | **No** | **No** | **35** | **Male** | **High** | **Normal** | **Eczema** |
| **No** | **No** | **Yes** | **No** | **35** | **Male** | **High** | **High** | **Epilepsy** |
| **Yes** | **Yes** | **Yes** | **No** | **35** | **Female** | **High** | **Normal** | **Hypertension** |
| **Yes** | **Yes** | **Yes** | **No** | **35** | **Female** | **Normal** | **Normal** | **Hyperthyroidism** |
| **No** | **No** | **Yes** | **No** | **35** | **Male** | **Normal** | **Normal** | **Obsessive-Compulsive Disorde...** |
| **Yes** | **Yes** | **Yes** | **Yes** | **35** | **Female** | **Normal** | **Normal** | **Pneumonia** |
| **Yes** | **Yes** | **Yes** | **Yes** | **35** | **Female** | **Normal** | **Normal** | **Pneumonia** |
| **Yes** | **No** | **No** | **No** | **35** | **Female** | **Normal** | **Low** | **Psoriasis** |
| **No** | **Yes** | **Yes** | **No** | **35** | **Female** | **Normal** | **Low** | **Psoriasis** |
| **Yes** | **No** | **Yes** | **No** | **35** | **Female** | **High** | **Normal** | **Rubella** |
| **Yes** | **No** | **Yes** | **No** | **35** | **Female** | **High** | **Normal** | **Rubella** |
| **No** | **Yes** | **Yes** | **No** | **35** | **Male** | **High** | **High** | **Urinary Tract Infection (UTI)** |
| **Yes** | **Yes** | **No** | **Yes** | **35** | **Male** | **Normal** | **Normal** | **Asthma** |
| **Yes** | **Yes** | **No** | **Yes** | **35** | **Male** | **Normal** | **Normal** | **Asthma** |
| **No** | **No** | **Yes** | **No** | **35** | **Female** | **Normal** | **High** | **Cirrhosis** |
| **No** | **Yes** | **No** | **No** | **35** | **Female** | **High** | **High** | **Conjunctivitis (Pink Eye)** |
| **No** | **No** | **Yes** | **No** | **35** | **Female** | **Normal** | **High** | **Depression** |
| **Yes** | **No** | **Yes** | **Yes** | **35** | **Male** | **Low** | **High** | **Gastroenteritis** |
| **Yes** | **Yes** | **Yes** | **No** | **35** | **Male** | **High** | **High** | **Hyperthyroidism** |
| **Yes** | **Yes** | **Yes** | **No** | **35** | **Male** | **High** | **High** | **Hyperthyroidism** |
| **No** | **No** | **Yes** | **No** | **35** | **Male** | **High** | **High** | **Kidney Cancer** |
| **No** | **No** | **Yes** | **No** | **35** | **Female** | **High** | **High** | **Liver Cancer** |
| **No** | **No** | **Yes** | **No** | **35** | **Male** | **High** | **High** | **Liver Disease** |
| **Yes** | **No** | **No** | **No** | **35** | **Male** | **High** | **High** | **Malaria** |
| **Yes** | **No** | **No** | **No** | **35** | **Male** | **High** | **High** | **Malaria** |
| **No** | **Yes** | **Yes** | **No** | **35** | **Male** | **High** | **High** | **Migraine** |
| **Yes** | **No** | **No** | **No** | **35** | **Male** | **High** | **High** | **Migraine** |
| **No** | **No** | **Yes** | **No** | **35** | **Male** | **Normal** | **High** | **Pancreatitis** |
| **Yes** | **Yes** | **Yes** | **No** | **35** | **Male** | **Normal** | **Low** | **Rheumatoid Arthritis** |
| **No** | **No** | **Yes** | **No** | **35** | **Female** | **Normal** | **Low** | **Rheumatoid Arthritis** |
| **No** | **No** | **Yes** | **No** | **35** | **Female** | **Normal** | **Normal** | **Spina Bifida** |
| **No** | **No** | **Yes** | **No** | **35** | **Male** | **High** | **High** | **Ulcerative Colitis** |
| **Yes** | **No** | **No** | **No** | **35** | **Male** | **High** | **High** | **Ulcerative Colitis** |
| **Yes** | **No** | **Yes** | **No** | **35** | **Female** | **Normal** | **High** | **Urinary Tract Infection** |
| **No** | **Yes** | **No** | **No** | **38** | **Female** | **Low** | **Normal** | **Allergic Rhinitis** |
| **No** | **No** | **No** | **No** | **38** | **Female** | **Normal** | **High** | **Depression** |
| **No** | **No** | **No** | **No** | **38** | **Male** | **Normal** | **Low** | **Gastroenteritis** |

**OUTCOMES**

****

****

**TIMELINE OF THE PROJECT**

****

**CONCLUSION**

The authors chose to apply decision trees in order to identify which patient cases presented similar information according their medical records. We successfully separated patients into similar and dissimilar groups by building a decision tree model. The transparency of the decision trees offered substantial insight into what was driving patient similarity. A decision tree, although it has its own limitations in that they can be sensitive to noise and overfit the data as well as bias variance trade-off issue, is a very useful way of such patient case similarity problem. For further work an ensemble method like Random Forests might improve the model and address some of those limiting factors. In summary, this work indicates the usefulness of decision trees in clinical setting with regard to patient outcomes.

**REFERENCES**

* 1. **A Survey of Data Mining Techniques for Medical Diagnosis" by Han et al. (2001)**
  2. **©Dillon Chrimes. Originally published in the Interactive Journal of Medical Research (https://www.i-jmr.org/), 30.01.2023.**
  3. **Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman**

**University of Maribor – FERI**

**Smetanova 17, SI-2000 Maribor, Slovenia**

[**vili.podgorelec@uni-mb.si**](mailto:vili.podgorelec@uni-mb.si)

* 1. **Decision trees: an overview and their use in medicine**

**Vili Podgorelec 1, Peter Kokol, Bruno Stiglic, Ivan Rozman**

* 1. **Azar, A.T., El-Metwally, S.M. Decision tree classifiers for automated medical diagnosis. *Neural Comput & Applic* 23, 2387–2403 (2013).** [**https://doi.org/10.1007/s00521-012-1196-7**](https://doi.org/10.1007/s00521-012-1196-7)
  2. **How Healthcare Decision Trees Emerge and Function**

**Published online by Cambridge University Press: 13 July 2023**

* 1. **Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy**

**Published in:**[**2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)**](https://ieeexplore.ieee.org/xpl/conhome/5695144/proceeding)

* 1. **International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online) Vol. 8, Issue 2, pp: (5-9), Month: April - June 2020, Available at:** [**www.researchpublish.com**](http://www.researchpublish.com)
  2. **Brown, SA., 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* 9, 7 (2023).** [**https://doi.org/10.1186/s40959-022-00151-0**](https://doi.org/10.1186/s40959-022-00151-0)

**10.Machine learning for integrating data in biology and medicine: Principles,**

**practice, and opportunities**

**2019, Information Fusion**