Final Project: A comparative analysis of the performance of Neural Network and Decision Tree ML Algorithms in detecting breast-cancer using a medical imaging dataset.

Xiaohong Deng

April 15, 2023

1 Introduction

Cancer of the breast is the second deadliest type of cancer. Early detection is critical in lowering mortality rates. However, the lack of a reliable early diagnostic approach has been a major challenge, leading to diagnostic inaccuracies. As a result, machine learning approaches are growing popular among biologists as a means of detecting breast cancer in a timely and effective manner [3]. In article [5], it is highlighted that the superior accuracy of machine learning compared to human radiologists, leads to reduced rates of false positives and false negatives in breast cancer detection. When it comes to detecting breast cancer, Neural Networks and Decision Trees are suitable methods as they have the ability to identify intricate patterns and relationships within large datasets. Decision Trees are capable of categorizing data into specific groups based on a set of rules or conditions, which can aid in identifying the most significant factors associated with breast cancer. Meanwhile, Neural Networks can learn from massive amounts of data and recognize complex patterns, making them well-suited for detecting breast cancer by analyzing multiple variables. [4]. This study aims to investigate and compare the effectiveness of Neural Networks and Decision Trees Algorithms in detecting breast cancer using medical imaging datasets. Our primary objective is to examine the null hypothesis(H_0) that there is no notable difference in the accuracy of breast-cancer detection between these two machine learning approaches.

2 Literature review

2.1 Decision Trees: Definition & Mechanics

The decision tree is a popular supervised machine-learning technique used for classification and regression analysis. The algorithm operates by dividing data into subsets based on a particular feature repeatedly until each subset becomes pure. The decision tree constructs a model of decisions and their corresponding outcomes in a tree-like structure, where internal nodes represent attribute tests, branches display test results, and leaf nodes depict class labels or numerical values. [2].

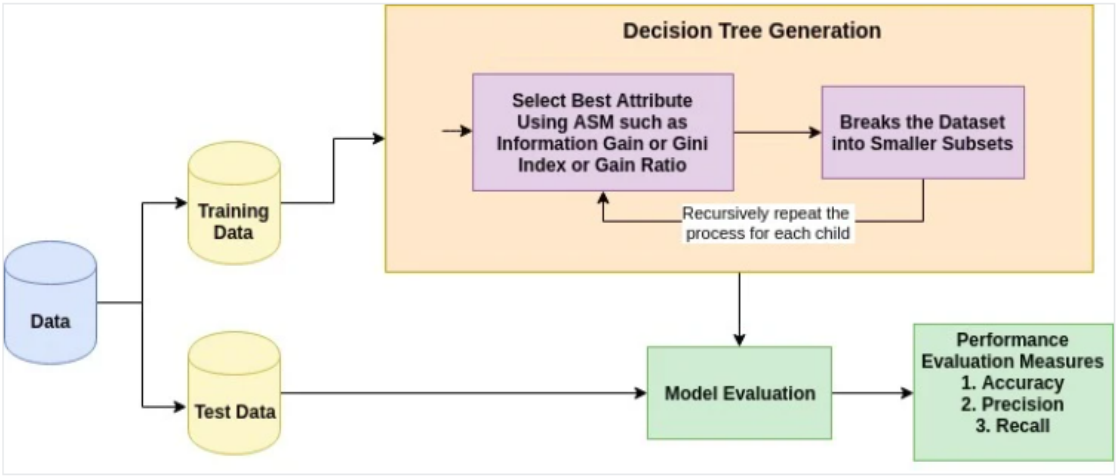

Figure 1: How Do Decision Trees Work?

2.2 Neural Networks: Definition & Mechanics

Neural Networks are machine learning models that emulate the structure and behavior of the human brain. These models are made up of interconnected neurons, which receive inputs, assign weights to them, and then utilize an activation function to yield output. In the phase of training, the neural network uses the backpropagation technique to adjust the weights of inter-neuron connections. This process seeks to lessen the discrepancy between the network's predicted output and the desired output [4].

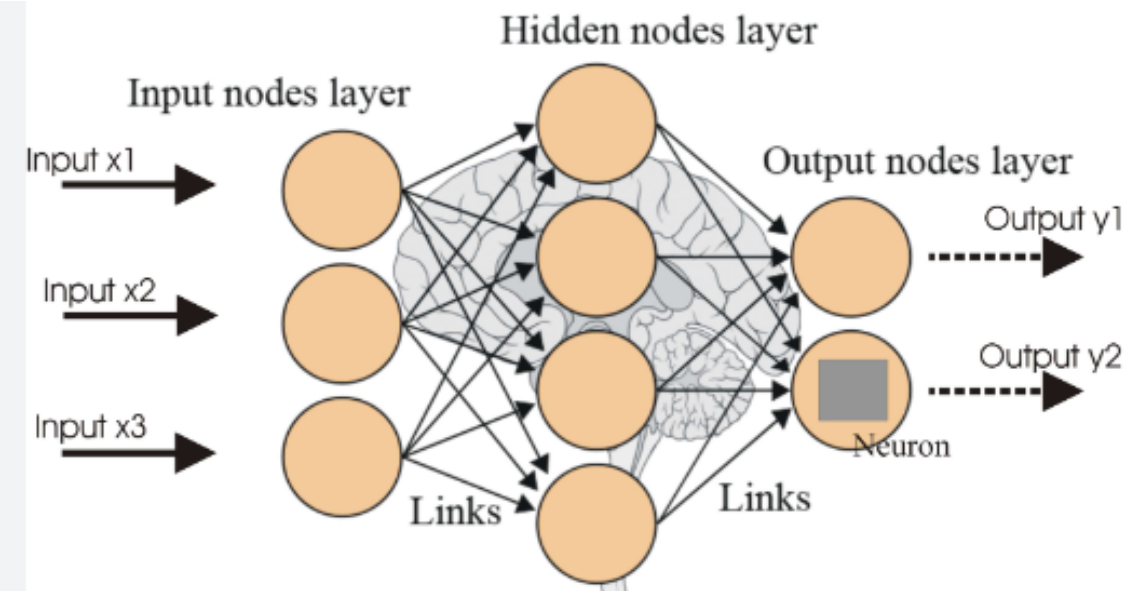

Figure 2: Simple Neural Network

2.3 Decision Tree for Breast Cancer Detection

To use Decision Trees to detect breast cancer from medical images, the first step is to preprocess the data by identifying relevant image features and converting them into numerical values. These numerical values are then used as inputs for the decision tree model, which learns to classify images as benign or malignant by identifying patterns in the data. The algorithm can be trained on a portion of the data and tested on a separate set to assess its performance. To optimize decision trees, we can adjust their parameters, such as pruning the tree to avoid overfitting or setting the minimal amount of samples required to divide a node. Additionally, ensemble methods like as boosting, bagging, and random forests can be utilized to improve the accuracy of Decision Trees by combining predictions from multiple trees. The use of these methods can help to reduce variance and bias, which enhances the overall performance of the model [1].

2.4 Neural Network for Breast Cancer Detection

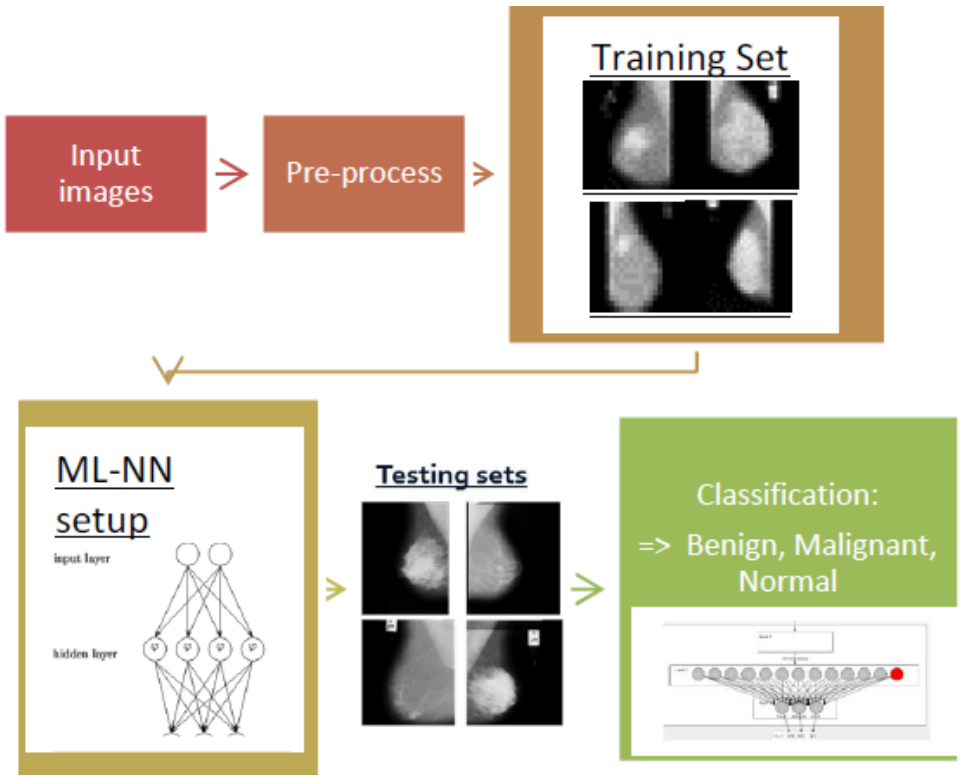

Figure 3: The Neural Network Flow for Detecting Breast-Cancer

To use the Multilayer Perceptron algorithm(a feedforward artificial neural network algorithm) to detect breast cancer from medical images, medical images are first filtered and labeled to improve their quality. These labeled images are then used as training data for the MLP, which is made up of multiple layers of perceptions - simple processing units. Each perceptron receives multiple inputs, applies weights to them, and uses an activation function to produce an output. Throughout the training

process, the weights of inter-neuron connections are modified via backpropagation to reduce the difference between the desired output and the actual output. After the MLP has been trained, it can be used to make predictions about new, unobserved data. [7].

3 Methods

3.1 Experiment Design

To compare the accuracy of Neural Networks and Decision Trees algorithms, the experiment was designed as follows:

- Acquire the medical imaging dataset from the Weka platform and divide it into two sets, one for training and one for testing, using a random split. Assign 66% of the dataset to training and 34% to testing.
- Create a Decision Tree model through the J48 algorithm and a Neural Networks model through the Multilayer Perceptron algorithm in Weka, a machine-learning platform.
- Assess the accuracy of both models on the testing set to evaluate their performance.
- Perform the above steps 30 times, each time using a different random sample of the dataset for training and testing.
- Calculate the accuracy of each model for each iteration and obtain the mean accuracy over the thirty iterations.
- Conduct a statistical analysis on the accuracy results to examine the hypothesis that there is a significant difference in the efficacy of Neural Networks and Decision Trees algorithms.

3.2 Data Source

For this project, we will use medical imaging data available on the Weka platform, consisting of images from cancer patients and healthy individuals. The dataset will be randomly divided into two sets: 66% to the training set and 34% to the testing set. To assess the performance of two machine learning algorithms (Decision Tree and Multilayer Perceptron), we will build 30 models for each algorithm using different random samples of the dataset. Statistical analysis will be performed to compare the accuracy of Neural Networks and Decision Trees algorithms and determine if there is a significant difference in their performance.

3.3 Metrics: Accuracy, precision, and recall

Commonly used metrics to evaluate the performance of a classification algorithm include accuracy, precision, and recall.
Accuracy is calculated as the percentage of true predictions produced by the algorithm out of all predictions, using Formula(1).

Precision measures the percentage of true positive predictions produced by the algorithm out of all positive predictions, using Formula (2).
Recall measures the percentage of true positive predictions made by the algorithm out of all actual positive instances in the data, using Formula (3).

Formula(1): Accuracy = (TP +TN) / (TP +TN +FP +FN)
Formula(2): Precision = TP / (TP +FP)
Formula(3): Recall = TP / (TP +FN)
Where TP refers to true positive; TN refers to true negative; FP refers to false positive; FN refers to false negative [6].
The following are the metrics reports for the Decision Tree and Multilayer Perceptron algorithms.

Row Labels	Accuracy	Precision	Recall
Decision Tree	71.1%	73.7%	73.7%
Neural network	67.6%	76.1%	76.1%
Grand Total	**69.3%**	**74.9%**	**74.9%**

Figure 4: Accuracy, Precision, and Recall of Decision Tree and Multilayer Perceptron

3.4 Data analysis – the student t-test

The dataset presented below is a result of an experiment conducted in Weka. It

Times	Decision Tree	Neural network
1	73.95833333	72.91666667
2	63.26530612	67.34693878
3	71.875	67.70833333
:	:	:
:	:	:
:	:	:
30	65.97938144	70.10309278

Figure 5: Accuracy Report of Multilayer Perceptron and Decision Tree

showcases the performance of both Neural networks and Decision Tree models on a

specific task. The dataset includes 30 observations for each model. The objective of the analysis is to examine if there is a statistically significant difference in the accuracy of the two models. To examine the hypothesis, we can use a two-sample t-test, assuming that the variances of the two populations are equal. Therefore, we should choose a heteroscedastic t-test.

To perform a heteroscedastic t-test in Excel, we can use a build-in function T.TEST

Figure 6: perform a heteroscedastic t-test

with the "Tails" argument set to 2 (two-tailed distribution) and the "Type" argument set to 2(two-sample equal variance). The resulting p-value will indicate the significance of the difference between the means of the two populations. If the p-value is below 0.05, we have sufficient evidence to reject the null hypothesis(H_0) and affirm that there exists a significant statistical difference between Decision Tree and Neural Network.

4 Results and Analysis

4.1 Hypothesis Test to Analyze the Experiment Result

Goal: The hypothesis test is to determine whether there is a statistically significant difference between two means (μ_{DT} - μ_{NN}) based on two independent samples, under the assumption that the variances of the two populations are equal. This will be achieved using a heteroscedastic t-test.

Conduct the hypothesis test:
Step 1: stating the null hypothesis(H_0) and alternative hypothesis(H_1):
H_0: μ_{DT} - $\mu_{NN} = 0$; H_1: μ_{DT} - $\mu_{NN} \neq 0$

Step 2: computing mean and standard deviation.

	Decision Tree (%)	Neural network (%)
Mean	71.1268	67.5567
Standard Deviation	3.4554	3.5299

Figure 7: calculate the mean and the standard deviation

Step 3: computing P-Value using a t-test

P-Value	0.0002	T.TEST(B2:B31,C2:C31,2,2)

Figure 8: calculate the P-Value

4.2 Conclusion

To sum up, this hypothesis test yielded a P-Value of 0.0002, indicating very strong evidence to reject H_0 in favor of H_1. Thus, we can conclude that there exists a statistically significant difference in the accuracy of breast-cancer detection between the Neural Network and Decision Tree algorithms. Furthermore, the Decision Tree outperformed the Neural Network with a mean accuracy of 71.1269%, compared to the Neural Network's mean accuracy of 67.5567%. It is worth noting that optimizing these models through various methods, such as increasing sample size, adjusting hyper-parameters, normalizing input data, etc., can result in a substantial improvement in accuracy. As a result, the use of machine learning is rapidly gaining momentum in the medical field.

References

[1] Asthana, M. Detecting breast cancer using machine learning. `https://medium.com/analytics-vidhya/detecting-breast-cancer-using-machine-learning-ab23e719f7fa`, 2020.

[2] Navlani, A. Decision tree clasification in python tutorial. `datacamp.com/tutorial/decision-tree-classification-python`, 2023.

[3] Sohrabi, S., Atahshi, A., Dadashi, A., and Marashi, S. M. A comparative study of multilayer neural network and c 4.5 decision tree models for predicting the risk of breast-cancer. 11–14.

[4] Soria, D., Garibaldi, J. M., Bigazoli, E., and Elliss, I. O. A comparison of three different methods for classification of breast-cancer data. IEEE - Seventh International Conference on ML and Applications (2008), 619–624.

[5] Synced. Three papers in the eye of the "ai breast cancer detection" storm. `medium.com/syncedreview/three-papers-in-the-eye-of-the-ai-breast-cancer-detection-storm-a63d2a2480ea`, 2020.

[6] Tarawneh, O., Otari, M., Husni, M., Abuadous, H., Tarawneh, M., and Almoumani, M. A. Breast cancer classification using decision tree algorithms. International Journal of Advanced Computer Science and Applications(IJACSA) 13, 4 (2022).

[7] Ting, F. F., and Sim, K. S. Self-regulated multilayer perceptron neural network for breast cancer classification. IEEE - 2017 International Conference on Robotics, Automation and Sciences (ICORAS) (2017).