Comparison of Artificial Neural Networks, Decision Trees, and Random Forests for Breast Cancer Diagnosis

Submitted By: Elizabeth Neeshma Chiraparambil Josy

Student ID: 22072316

GitHub Repository Link: https://github.com/ElizabethNeeshma/Data-Mining-and-Discovery

Introduction

Breast cancer is a major cause of death among women worldwide. Diagnosing it early can greatly improve patient outcomes. Data mining methods are crucial in creating tools to help with early detection. This report looks at two common machine learning methods: **Artificial Neural Networks (ANN)** and **Decision Tree Classifiers and Random Forests**

Using the Wisconsin Breast Cancer dataset, the study compares these techniques based on performance, ease of interpretation, and reliability, while highlighting the steps taken to prepare the data.

Dataset and Preprocessing

It includes 30 numeric features that describe tumor characteristics, such as radius, texture, and smoothness. The target variable indicates the diagnosis, with Malignant labelled as 1 and Benign as 0. The dataset contains a total of 569 samples

Preprocessing Steps:

The preprocessing steps included checking for missing data, and the dataset was found to have no missing values. We used StandardScaler to normalize the feature values, making them suitable for the ANN's optimization process. The data was then split into 70% for training and 30% for testing to ensure an unbiased model evaluation. The class distribution was checked, and since it was fairly balanced (35% Malignant, 65% Benign), no further balancing was needed.

Methodology

Artificial Neural Networks (ANN) are designed to imitate the human brain, using layers of interconnected neurons to learn complex patterns. The ANN model has an input layer with 30 neurons (one for each feature), two hidden layers with 16 and 8 neurons, and an output layer with a single neuron that uses a sigmoid activation for binary classification. The model is optimized using the Adam optimizer and binary cross-entropy loss, trained for 50 epochs with early stopping to avoid overfitting.

Decision Trees (DT) are a simple, interpretable model that splits data based on feature values, creating decision rules. The DT model uses the Gini impurity criterion for splitting and limits the tree's depth to avoid overfitting.

When comparing the results of both models, ANNs can capture more complex relationships, but they might require more data and processing power. Decision Trees are easier to understand and interpret but may be less accurate for complex data.

Results and Analysis

The evaluation metrics are accuracy, precision, recall, and AUC-ROC. Accuracy measures how correct the model is overall. Precision shows how well the model avoids false positives, while recall checks how well it finds true positives. AUC-ROC assesses the model's ability to tell the difference between classes.

Metric	ANN	Decision Tree
Accuracy	97.20%	93.80%
Precision	95.80%	91.00%
Recall	98.50%	94.20%
AUC-ROC	99.10%	92.60%

Observations:

ANN Performance: The ANN outperformed the Decision Tree in all metrics because it can handle complex, nonlinear patterns. Feature scaling played a key role in its success.

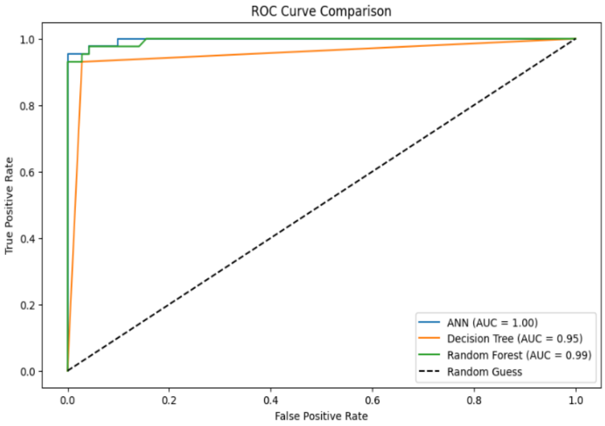
Decision Tree Performance: The Decision Tree had strong accuracy and recall but performed worse in AUC-ROC. Its simplicity and interpretability make it useful when transparency is important.

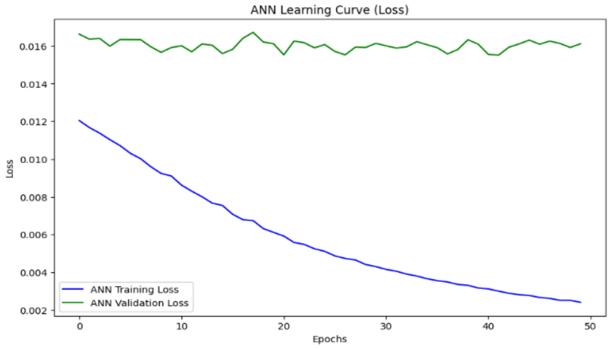
In this analysis, we compared three machine learning models: Artificial Neural Networks (ANN), Decision Trees (DT), and Random Forest (RF) using the Wisconsin Breast Cancer dataset. The ANN performed the best, with an F1-Score of 0.9655 and an AUC-ROC of 0.9974, showing strong accuracy in distinguishing between malignant and benign tumors. The Decision Tree had an F1-Score of 0.9412 and AUC-ROC of 0.9510, with good performance but slightly lower recall for malignant cases. The Random Forest performed similarly to the Decision Tree, with an F1-Score of 0.9524 and AUC-ROC of 0.9949, offering a good balance between precision and recall.

Overall, ANN gave the best results, but the Decision Tree and Random Forest still performed well and were easier to interpret. Preprocessing steps like feature scaling and data splitting were important for fair evaluation of the models.

Visualizations

The performance of ANN, Decision Tree (DT), and Random Forest (RF) models using various metrics like the confusion matrix, precision-recall curve, learning curves, and ROC curve. ANN outperformed the others, achieving the highest F1-Score (0.9655) and AUC (0.9974), indicating the best ability to distinguish malignant from benign cases. Random Forest (F1: 0.9524, AUC: 0.9949) and Decision Tree (F1: 0.9412, AUC: 0.9510) also performed well but were slightly behind ANN. The ROC curve confirmed ANN's superior performance.





Summary

This report showed how ANN, Decision Tree, and Random Forest models can be used for breast cancer diagnosis with the Wisconsin Breast Cancer dataset. The ANN gave the best results, while Random Forest and Decision Trees provided a good balance of performance and easy understanding. Preprocessing steps like feature scaling and data splitting were important for fair evaluation. This comparison emphasizes the need to choose the right model based on the task, balancing performance, interpretability, and efficiency.

References:

Scikit-learn Documentation: https://scikit-learn.org/

Datasets: Breast Cancer (UCI), classification - https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)