Machine Learning and Deep Learning Models for ‘Wisconsin Diagnostic Breast Cancer’ Dataset

Table of Contents

Machine Learning and Deep Learning Models for ‘Wisconsin Diagnostic Breast Cancer’ Dataset

Introduction:

In this report, we explore and compare the performance of various Machine Learning (ML) and Deep Learning (DL) models using the **Wisconsin Diagnostic Breast Cancer (WDBC)** dataset. The goal of this analysis is to classify whether a tumor is **malignant** or **benign** based on features extracted from digitized images of breast tissue samples.

We use a range of models including **Decision Tree, Random Forest, Naive Bayes,** and a **Multilayer Artificial Neural Network (ANN)** to evaluate their performance across several metrics—**Accuracy, Precision, Recall,** and **F1-Score**. The models are compared and the results are visualized through bar charts for an intuitive comparison.

Dataset Overview:

The **Wisconsin Diagnostic Breast Cancer (WDBC)** dataset was created by Dr. William H. Wolberg and colleagues. The dataset is composed of features extracted from fine needle aspirates of breast masses. The goal is to classify tumors into one of two classes:
- **Malignant** (denoted as "M")
- **Benign** (denoted as "B")

Attributes:

- **Number of Instances**: 569
- **Number of Features**: 32 (30 real-valued input features, 1 target variable, and 1 ID column)
- **Target Variable**: Diagnosis (M = Malignant, B = Benign)

Features:

Each feature is a characteristic of the cell nuclei present in the images. The features include:

1. **Radius**
2. **Texture**
3. **Perimeter**
4. **Area**
5. **Smoothness**
6. **Compactness**
7. **Concavity**
8. **Concave Points**
9. **Symmetry**
10. **Fractal Dimension**

For each of these features, the **mean**, **standard error (SE)**, and **worst value** (mean of the three largest values) were computed, resulting in 30 total features.

Machine Learning and Deep Learning Models for ‘Wisconsin Diagnostic Breast Cancer’ Dataset

Class Distribution

- **357** benign cases (62.7%)
- **212** malignant cases (37.3%)

Methodology:

Data Preprocessing:

- **Dropping ID Column**: The `id` column is removed as it does not contribute to the classification task.
- **Target Encoding**: The target variable `diagnosis` is converted to binary form: **M (Malignant)** is mapped to 1 and **B (Benign)** is mapped to 0.
- **Feature Scaling**: The features are standardized using `StandardScaler` to ensure that the neural network and some machine learning models (like Naive Bayes) perform optimally.

Models Used:

1. **Decision Tree Classifier**:
   - A tree-based model that splits data into branches based on feature values.

2. **Random Forest Classifier**:
   - An ensemble of decision trees, where the results of multiple decision trees are combined to improve the overall performance.

3. **Naive Bayes Classifier**:
   - A probabilistic classifier based on Bayes' Theorem, assuming independence between predictors.

4. **Artificial Neural Network (ANN)**:
   - A feedforward neural network with two hidden layers. The model is trained using the Adam optimizer and binary cross-entropy loss.

Evaluation Metrics

Each model's performance was evaluated using the following metrics:

- **Accuracy**: The proportion of correct predictions.
- **Precision**: The proportion of correctly predicted positive observations (malignant tumors) out of all predicted positives.
- **Recall**: The proportion of correctly predicted positives out of all actual positives.
- **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two.

Machine Learning and Deep Learning Models for ‘Wisconsin Diagnostic Breast Cancer’ Dataset

Model Performance Results:

Below are the results of the four models on the test dataset:

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.9474	0.9302	0.9302	0.9302
Random Forest	0.9649	0.9756	0.9302	0.9524
Naive Bayes	0.9649	0.9756	0.9302	0.9524
Artificial Neural Network (ANN)	0.9737	0.9762	0.9535	0.9647

Output Analysis:

1. Accuracy

- **ANN** achieved the highest accuracy (97.37%), indicating that it correctly classified 97.37% of the samples in the test set.
- **Random Forest** and **Naive Bayes** followed closely, both with an accuracy of 96.49%.
- **Decision Tree** had the lowest accuracy (94.74%), though still competitive for this task.

2. Precision

- **ANN** and **Random Forest** performed exceptionally well with precision scores above 0.975, which means they had a high proportion of correctly identified malignant cases among all predicted positives.
- **Naive Bayes** also achieved the same precision as Random Forest, while **Decision Tree** had a slightly lower precision (0.9302).

3. Recall

- **ANN** again outperformed the other models with a recall score of **0.9535**, meaning that it correctly identified 95.35% of the actual malignant cases.
- **Decision Tree**, **Random Forest**, and **Naive Bayes** had the same recall (93.02%), showing that they were still highly effective at detecting malignant tumors.

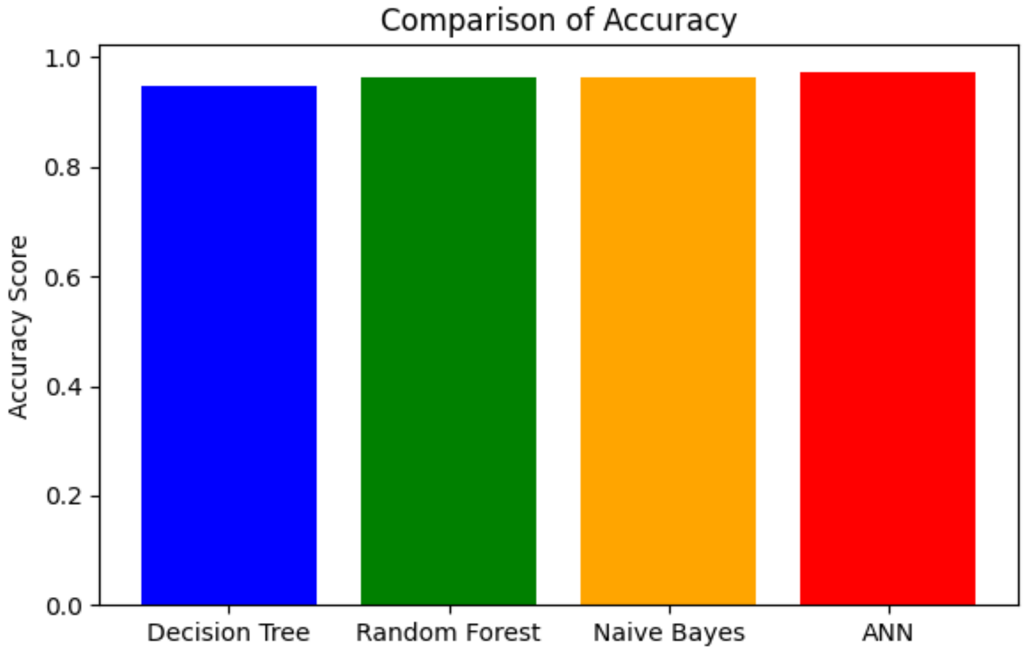
4. F1-Score

- The **ANN** model achieved the best F1-Score (0.9647), which indicates a good balance between precision and recall.
- **Random Forest** and **Naive Bayes** models followed with F1-scores of **0.9524**, while **Decision Tree** had the lowest F1-score (0.9302).

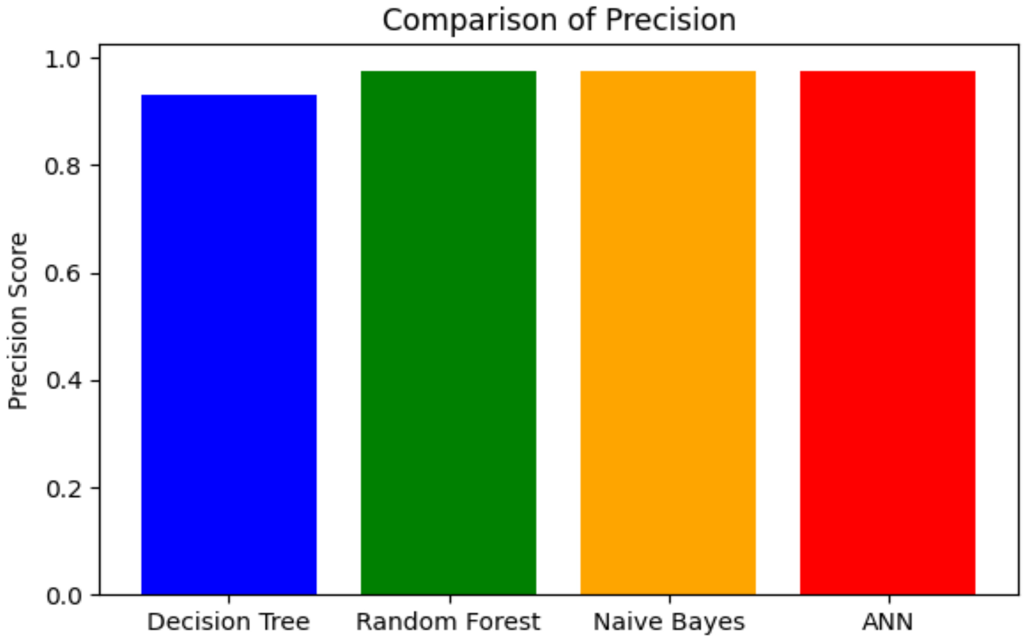
Visualization of Results

The performance of the models was visualized using bar charts, allowing for an easy comparison of each model's performance across the four metrics. The following bar charts were generated:

Comparison of Accuracy:

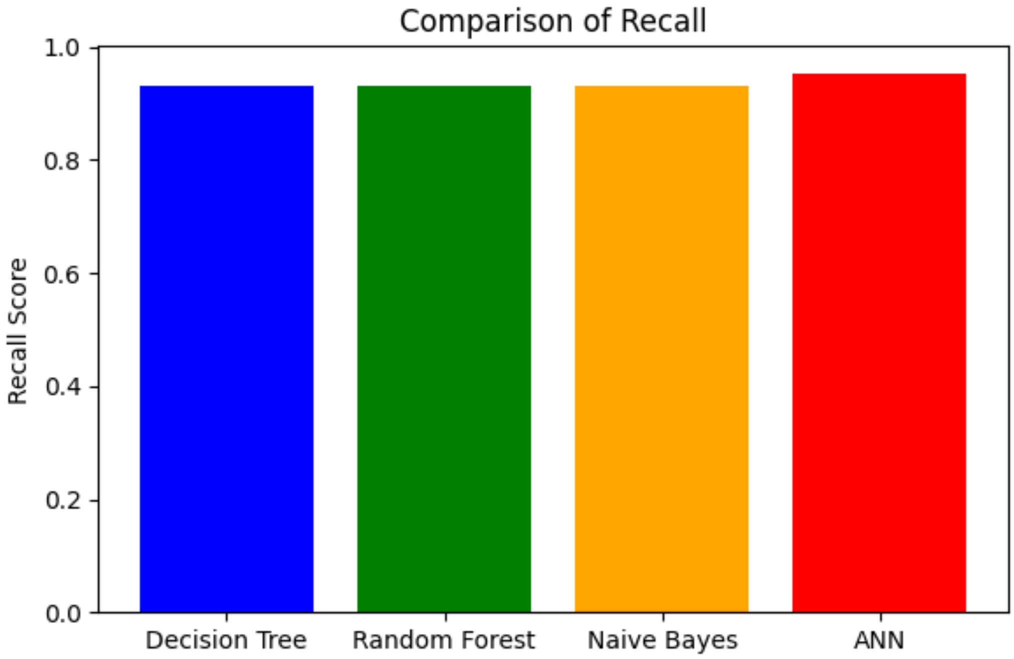


Comparison of Precision:

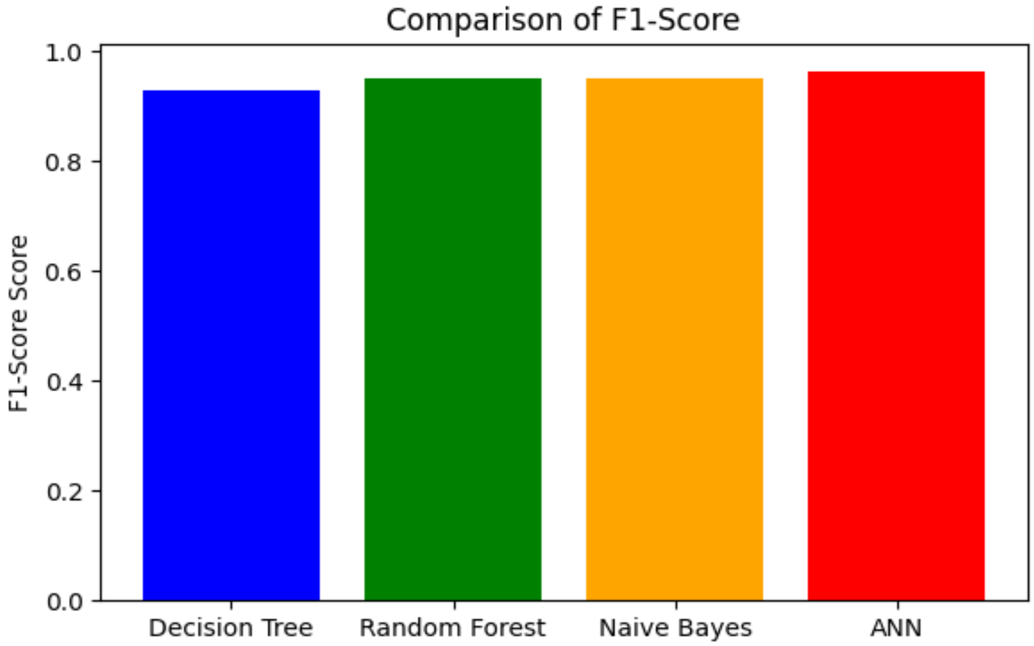


Machine Learning and Deep Learning Models for ‘Wisconsin Diagnostic Breast Cancer’ Dataset

Comparison of Recall:

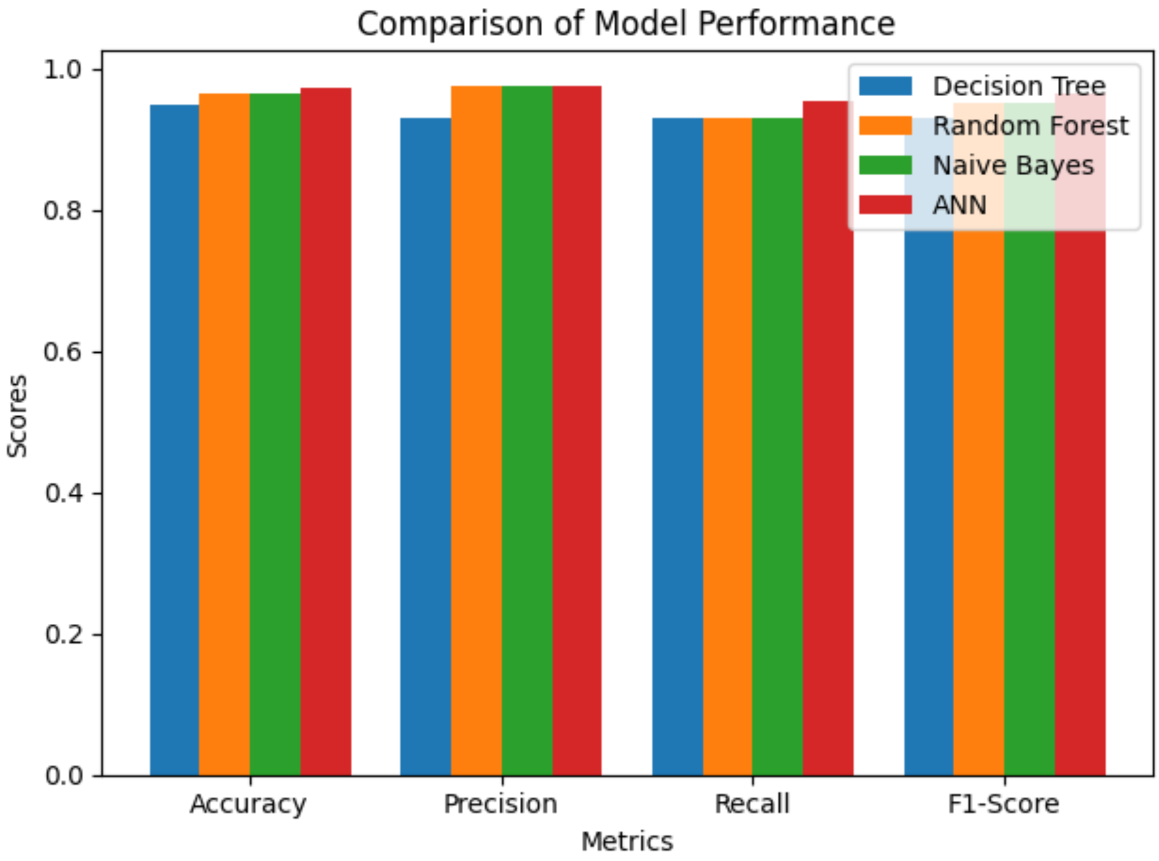


Comparison of F1-Score:



Machine Learning and Deep Learning Models for ‘Wisconsin Diagnostic Breast Cancer’ Dataset

Overall Model Performance Comparison:



Conclusion:

In this study, we analyzed the performance of four models—Decision Tree, Random Forest, Naive Bayes, and ANN—on the Wisconsin Diagnostic Breast Cancer dataset.

- **ANN** consistently outperformed the other models across all metrics, making it the most effective model for this classification task.
- **Random Forest** and **Naive Bayes** performed similarly well, both achieving high precision, recall, and F1-scores.
- **Decision Tree**, while still achieving reasonable accuracy and performance, was the least effective compared to the other models.

Based on these findings, it is recommended to use **ANN** for the task of classifying breast cancer from the WDBC dataset due to its superior performance, especially in terms of precision and recall. However, **Random Forest** and **Naive Bayes** also offer strong performance and could be considered for situations where computational efficiency is a concern, as these models generally require less training time than neural networks.

Machine Learning and Deep Learning Models for ‘Wisconsin Diagnostic Breast Cancer’ Dataset

References

- Wisconsin Diagnostic Breast Cancer (WDBC) Dataset: [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic)