Project Report: Comparison of SVM and Neural Network Models for Breast Cancer Classification

---

1. Introduction

The goal of this project was to compare the performance of two machine learning models, **Support Vector Machine (SVM)** and **Neural Network (NN)**, for classifying breast cancer data. The dataset consists of features related to cell characteristics, such as radius, texture, and smoothness, along with a binary target variable indicating whether a tumor is **benign** or **malignant**.

We explored the performance of these models using various activation functions in the neural network (ReLU, Sigmoid, and Tanh). We aimed to evaluate the models based on accuracy, precision, recall, F1-score, and other performance metrics. The final evaluation also included ROC curve analysis for each model to assess their classification effectiveness.

2. Dataset Overview

The dataset includes 30 features representing different tumor characteristics, and the target variable `diagnosis` has two classes:

- **Benign (0)**
- **Malignant (1)**

The dataset was split into **training** (80%) and **testing** (20%) sets. The features were normalized using **StandardScaler** to ensure efficient model training.

3. Model Implementation

We used two types of models for this project:

1. **Support Vector Machine (SVM)** with an RBF kernel, which is commonly used for classification tasks, especially in high-dimensional spaces.
2. **Neural Network (NN)** with different activation functions: **ReLU**, **Sigmoid**, and **Tanh**. The network architecture consists of:
    - One hidden layer with 16 neurons.
    - Output layer with a sigmoid activation function.

The models were trained using the **Adam optimizer** and **binary cross-entropy** loss function.

4. Results and Performance Evaluation

SVM with RBF Kernel

- **Accuracy**: 0.9825
- **F1 Score**: 0.98
- **Confusion Matrix**:

- `[[71  0]`
- `[ 2 41]]`
- The **SVM model** performed very well, achieving high accuracy and a balanced confusion matrix. The precision and recall for the malignant class (1) were both high, with only 2 false negatives.

Neural Network (Sigmoid Activation)

- **Accuracy**: 0.9912
- **F1 Score**: 0.9882
- **Confusion Matrix**:
- `[[71  0]`
- `[ 1 42]]`
- The **Neural Network** with the **Sigmoid activation** function outperformed the SVM model slightly in terms of accuracy and F1 score. It achieved near-perfect classification, with only 1 false negative, indicating excellent performance for detecting malignant tumors.

Neural Network (Tanh Activation)

- **Accuracy**: 0.9825
- **F1 Score**: 0.9767
- **Confusion Matrix**:
- `[[70  1]`
- `[ 1 42]]`
- The **Tanh activation** performed similarly to the ReLU model, with a high accuracy but slightly lower F1 score compared to the Sigmoid-based network. It had 1 false positive and 1 false negative, indicating a small compromise in precision and recall.

Neural Network (ReLU Activation)

- **Accuracy**: 0.9825
- **F1 Score**: 0.9762
- **Confusion Matrix**:
- `[[71  0]`
- `[ 2 41]]`
- The **ReLU activation** provided solid results, though slightly lower than the Sigmoid activation in terms of F1 score. It had 2 false negatives, which is still a low number but not as optimal as the Sigmoid model.

5. Model Comparison Summary

Model	Accuracy	F1 Score	Confusion Matrix
SVM (RBF Kernel)	0.9825	0.98	[[71, 0], [2, 41]]
Neural Network (Sigmoid)	0.9912	0.9882	[[71, 0], [1, 42]]
Neural Network (Tanh)	0.9825	0.9767	[[70, 1], [1, 42]]
Neural Network (ReLU)	0.9825	0.9762	[[71, 0], [2, 41]]

6. Visualizations

ROC Curve Comparison

The ROC curve analysis showed that the **Neural Network with Sigmoid** had the best performance, closely followed by the SVM. The **AUC (Area Under Curve)** for the Sigmoid network was the highest, which indicates that the Sigmoid model has the best ability to distinguish between benign and malignant cases.

---

7. Discussion

- **SVM** and **Neural Network (Sigmoid)** both performed exceptionally well in terms of classification accuracy, but the Sigmoid activation function in the neural network slightly outperformed SVM based on accuracy and F1 score.
- **ReLU and Tanh** performed well but showed minor compromises, such as higher false negatives or slightly lower F1 scores.
- The **Sigmoid network** is the most stable and accurate, making it the preferred model for this particular classification task.
- **Overfitting and Underfitting**:
    - There are no signs of overfitting in either model, as both showed high performance on the test set.
    - There is a possibility of **underfitting** in the ReLU and Tanh models, which can be addressed by tweaking the model architecture, using more layers or neurons, or experimenting with more advanced regularization techniques like **Dropout**..