Breast Cancer Classification with Support Vector Machines and Neural Networks

Breast cancer is one of the most common forms of cancer among women globally. Early detection is crucial for successful treatment. Machine learning algorithms can help physicians make more accurate diagnoses. This project focuses on comparing two powerful machine learning algorithms—Support Vector Machines and Neural Networks—for classifying breast tumors as benign or malignant.

Abstract

This report investigates breast cancer classification using state-of-the-art machine learning techniques namely Support Vector Machines and Neural Networks. The primary dataset employed is the Wisconsin Breast Cancer Dataset, a widely accepted benchmark in medical data science research. The study systematically evaluates the strengths of each model in identifying malignant and benign tumor types through rigorous preprocessing and model tuning.

Team Members: [Gehad Helmy Abdelrazik, 23011242] ,
[Alaa Adel Mohammed Ali, 23011045]

1. Introduction

Breast cancer remains one of the most prevalent and impactful health challenges worldwide. According to the American Cancer Society (ACS), approximately one in eight women will develop invasive breast cancer in their lifetime. In 2022 alone, there were an estimated 287,850 new cases of invasive breast cancer reported in the United States, highlighting the critical need for efficient diagnostic approaches.

Early detection and precise diagnosis are essential to improving patient survival rates and treatment outcomes. Traditional diagnostic methods can be time-consuming and prone to human error, thus motivating the integration of automated machine learning techniques in clinical workflows.

The objectives of this project are to compare the performance of Support Vector Machines and Neural Networks for breast cancer classification and to identify the most predictive features within the dataset. This comparative analysis aims to contribute valuable insights into the applicability of classical and modern machine learning algorithms in medical diagnostics.

2. Methodology

The dataset used for this study is the Wisconsin Breast Cancer Dataset, which contains 569 samples with 30 numeric features representing characteristics of cell nuclei extracted from breast mass images. The dataset is composed of two classes: benign and malignant tumors.

Data preprocessing is critical to ensure model efficacy. Missing values were addressed through appropriate data cleaning techniques, and feature scaling was performed using StandardScaler to normalize the input features, enhancing model convergence.

The study utilized two primary models:

- **Support Vector Machine (SVM):** Implemented with both linear and radial basis function (RBF) kernels to capture linear and non-linear boundaries respectively.
- **Neural Network:** A Multi-Layer Perceptron (MLP) architecture comprising [Number] hidden layers, each with [Number] neurons. The model employed ReLU activation functions in hidden layers for non-linearity and a sigmoid activation in the output layer to produce probabilistic class predictions.

Citations include relevant sources for dataset and model architectures to maintain scientific rigor.

3.. Data Preprocessing and Exploratory Data Analysis

Dataset Description

The Breast Cancer Wisconsin dataset contains features computed from digitized images of breast mass fine needle aspirates (FNA). The dataset includes measurements such as radius, texture, perimeter, area, smoothness, compactness, concavity, and symmetry of the cell nuclei.

Preprocessing Steps

1. **Data Loading and Cleaning**: The dataset was loaded and checked for missing values.
2. **Feature Selection**: The 'Unnamed: 32' and the 'ID' columns were removed as they don't contribute to the classification task.
3. **Target Encoding**: The diagnosis feature was encoded as binary values (B=0 for benign, M=1 for malignant).
4. **Train-Test Split**: The dataset was split into 70% training and 30% testing sets.
5. **Feature Standardization**: Features were standardized to have zero mean and unit variance.

Exploratory Data Analysis

Several visualizations were created to understand the data better:

- **Histograms**: To visualize the distribution of each feature
- **Pair Plots**: To observe interactions between features and their relationship with the target variable

3. Support Vector Machine Implementation

Model Configuration

A Support Vector Machine was implemented using scikit-learn's SVC class. GridSearchCV was used to tune hyperparameters with the following search space:

- **Kernel**: linear, polynomial, and radial basis function (RBF)
- **C parameter**: 0.1, 1, 10, 100
- **Gamma**: scale, auto, 0.1, 0.01

SVM Results

- **Best Parameters:** {'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}
- **Accuracy:** 0.9825
- **Precision:** 0.9839
- **Recall:** 0.9683
- **F1-Score:** 0.9760

The learning curve analysis indicated [whether the model was overfitting, underfitting, or well-balanced].

4. Neural Network Implementation

Model Architecture

A Multi-Layer Perceptron (MLP) neural network was implemented using scikit-learn's MLPClassifier. GridSearchCV was used for hyperparameter tuning with the following configuration options:

- **Hidden Layer Sizes**: Various architectures with different numbers of neurons
- **Activation Functions**: ReLU and Sigmoid
- **Regularization (alpha)**: 0.0001, 0.001

Neural Network Results

- **Best Parameters:** {'activation': 'tanh', 'hidden_layer_sizes': (50,)}
- **Accuracy:** 0.9708
- **Precision:** 0.9531
- **Recall:** 0.9683
- **F1-Score:** 0.9606

The loss curve showed [convergence behavior of the neural network].

5. Model Comparison and Analysis

Performance Metrics Comparison

Both models were compared using accuracy, precision, recall, and F1-score metrics:

Model	Accuracy	Precision	Recall	F1-Score
SVM	[0.9825]	[0.9839]	[0.9683]	[0.9760]
NN	[0.9708]	[0.9531]	[0.9683]	[0.9606]

Overfitting and Underfitting Analysis

Training and testing accuracies were compared to assess model fit:

- SVM: Training accuracy = [0.9824], Testing accuracy = [0.9825]
- Neural Network: Training accuracy = [0.9623], Testing accuracy = [0.9708]

5. Conclusion

This study compared the performance of SVM and Neural Network models on the Breast Cancer Wisconsin dataset. [Summarize which model performed better and key findings]. The results demonstrate that both models can effectively classify breast tumors with high accuracy, potentially assisting medical professionals in diagnosis. Further improvements could be achieved through feature engineering, ensemble methods, and more extensive hyperparameter tuning.

The clinical implications of these findings include the potential for early detection and improved diagnostic accuracy, ultimately contributing to better patient outcomes. Automated classification systems can support medical professionals by reducing diagnostic workload and minimizing errors.