Final Report

Introduction

Breast cancer accounts for one-fourth of all cancer cases worldwide and remains the most common cancer among women, making early and accurate detection critical for treatment. The main motivation of this project is to determine how machine learning methods can be used to support decision-making in medical applications. In this report, we aim to answer the following research question: Can we successfully predict whether a breast tumor is benign or malignant based on tumorous cell information? We compared and evaluated four different machine learning models - logistic regression, support vector machines, XGBoost, and a neural network - to determine which model could give us the best predictive accuracy.

Methods

The Wisconsin Breast Cancer Dataset consists of 30 numerical features extracted from digitized images of fine needle aspirate (FNA) biopsies of breast masses. Each feature represents computed characteristics of cell nuclei present in the images, including measurements of radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These 10 core measurements are provided as mean values, standard errors, and "worst" (largest) values. The target variable is binary: benign (0) and malignant (1). Data cleaning revealed no missing values, and all features were numerical, so encoding was not necessary. No additional data cleaning or outlier removal was performed.

We implemented 4 different machine learning models with tailored preprocessing. Logistic Regression used StandardScaler normalization in a Pipeline, tuning C, penalty, and solver via GridSearchCV. SVM employed identical scaling with an RBF kernel, optimizing C, gamma, and degree parameters. XGBoost required no scaling (tree-based), tuning n_estimators, max_depth, learning_rate, and regularization parameters across 972 combinations. Neural Network implemented PyTorch architectures with hidden layers, dropout, and early stopping across 243 configurations. All models used 5-fold StratifiedKFold cross-validation with StandardScaler fitted per fold to prevent data leakage and maintain class distribution. Recall (sensitivity) was the primary metric, prioritizing minimizing false negatives since missing cancer cases carries greater clinical cost than false positives. Hyperparameter optimization used GridSearchCV for Logistic Regression, SVM, and XGBoost, while Neural Network employed manual search with validation-based early stopping.

Results

We evaluated four models for this breast cancer classification task: Neural Network, XGBoost, Support Vector Machine, and Logistic Regression. Each model was trained on 455 samples and optimized using 5-fold stratified cross-validation. Recall as the primary scoring metric given that minimizing false negatives (missed malignant diagnoses) is critical in medical screening. This table below summarizes the cross-validated performance of each best model.

Model	Best CV Recall
Neural Network	0.9882

XGBoost	0.9588
SVM	0.9588
Logistic Regression	0.9529

The Neural Network achieved the highest cross-validation recall at 0.9882, approximately 3% higher than the other models. Based on this performance, the Neural Network was selected as the final model and evaluated on the held-out test set of 114 samples that was set aside at the beginning and never used during training or hyperparameter tuning. On the test set, the model achieved 97.37% accuracy, 95.35% precision, 97.62% recall, 96.47% F1-score, and 99.50% ROC-AUC. The confusion matrix revealed only 1 false negative (one malignant case missed out of 42) and 2 false positives, demonstrating strong generalization performance consistent with the cross-validation results. These findings confirm that the Neural Network effectively balances sensitivity and specificity for breast cancer classification.

Discussion

The results demonstrate that all four models achieved strong performance for breast cancer classification. The scores ranged from 95.29% – 98.82%, indicating great capability to minimize false negatives. The Neural Network had the best performance (98.92% recall), and this can be attributed to its ability to learn complex, non-linear feature interactions through multiple hidden layers, effectively capturing patterns in the 30-dimensional feature space that linear models may miss. XGBoost and SVM tied for second place with 95.88% recall. XGBoost's ensemble of 200 decision trees leveraged boosting to correct prediction errors iteratively, while SVM's RBF kernel mapped the data into a higher-dimensional space where classes became more separable. Logistic Regression, despite being the simplest model, achieved a performance of 95.29% recall. This may indicate that the relationship between features and diagnosis is not overly complex.

Despite our strong results, this project has a few limitations that would prevent immediate clinical deployment. We worked with a limited dataset, as our training dataset only had 455 samples and most likely did not capture the full diversity of breast cancer presentations. We also only had a single dataset source. The Wisconsin Breast Cancer Dataset is only one institution and limits generalization to other hospitals, populations, and geographic locations. The second limitation is with regard to our methodological and validation gaps. There was no external validation, as our models were only validated on splits from the same dataset and not tested on external/independent cohorts. There is also a risk of the neural network model overfitting, as the recall seems suspiciously high despite cross-validation.