Predicting Breast Cancer Diagnosis using Machine Learning

Name: Lan Dinh

Student ID: 300383107

A. INTRODUCTION AND DISCOVERY

Breast cancer is one of the most common cancers worldwide, particularly among women, and early detection is critical for effective treatment and improved survival rates. This project focuses on the domain of medical diagnostics, specifically breast cancer detection, using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the UCI Machine Learning Repository – a well-known resource developed by researchers at the University of California, Irvine. The dataset comprises features extracted from digitized images of breast masses, enabling differentiation between malignant (cancerous) and benign (non-cancerous) tumors.

The primary question this project addresses is: *Can a machine learning model accurately predict whether a breast tumor is malignant or benign based on quantitative features from the WDBC dataset?* This question is vital because accurate predictions can enhance the speed and reliability of diagnoses, potentially reducing human error and improving patient care.

To investigate this question, the project develops two initial hypotheses. First, features related to tumor size (e.g., radius, perimeter, area) and shape (e.g., concavity, compactness) are expected to be strong predictors of malignancy due to their association with tumor growth patterns. Second, machine learning models incorporating feature scaling and selection will outperform simpler models by capturing complex relationships in the data.

B. DATA PREPARATION

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset comprises 569 entries, each representing a breast mass, and includes 32 columns. Of these, 30 are numerical features—grouped into mean, standard error, and worst values—describing attributes like radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The target variable, "Diagnosis," is labeled as "M" (malignant) or "B" (benign) and stored as an object type, while the "ID" column is an integer (int64), and all

other features are float64. A preliminary analysis confirmed no missing values or duplicate entries, ensuring the dataset's readiness for modeling.

To prepare the dataset for modeling, two key transformations were applied. First, the "ID" column was dropped, as it serves as a unique identifier with no predictive value for breast cancer diagnosis. Second, the "Diagnosis" column was encoded to numerical values—malignant (M) was assigned a value of 1, and benign (B) was assigned a value of 0—and then converted to float64 to ensure compatibility with machine learning algorithms.

C. MODEL IMPLEMENTATION

This project tackles the binary classification problem of predicting breast tumor malignancy (malignant or benign) using machine learning models on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. A pipeline with feature scaling (StandardScaler) and feature selection (SelectFromModel with RandomForestClassifier) evaluates multiple classifiers, justified by their fit to this binary task and prior studies:

1. Logistic Regression: A baseline model valued for simplicity and interpretability, achieving 96.79% accuracy in WDBC studies (Kumar et al., 2013).
2. K-Nearest Neighbors (KNN): A distance-based method effective for continuous features, with 95% accuracy in similar diagnostic studies (Rana et al., 2015).
3. Decision Tree: A non-linear model splitting features for prediction, noted for 75.52% accuracy in oncology research (Institute of Oncology, n.d.).
4. Naïve Bayes (GaussianNB): A fast probabilistic approach, suitable despite correlations, often used in small datasets (Delen et al., 2005).
5. Random Forest: An ensemble method capturing complex patterns, widely successful in breast cancer tasks with 96% accuracy (Khourdifi et al., 2018).
6. XGBoost: A boosting model excelling in binary classification, known for high performance (Chen & Guestrin, 2016).
7. Artificial Neural Network (ANN): A neural approach for non-linear patterns, achieving 97% accuracy in diagnostics (Omondiagbe et al., 2020).
8. Support Vector Machine (SVM): A kernel-based model maximizing class separation, often reaching 95-97% accuracy in WDBC studies (Bennett & Mangasarian, 1992).

Rather than relying on a single model, a multi-model pipeline is employed to ensure robustness, given the dataset's high feature correlations, such as between radius_worst and perimeter_worst, and its small sample size.

This approach tests the project's hypotheses and informs its modeling objectives effectively. For the first hypothesis – features like tumor size (including radius, perimeter, and area), and shape (including concavity and compactness), are strong predictors of malignancy due to their link to growth patterns – the SelectFromModel step uses Random Forest to rank and select features by importance, assessing whether size and shape dominate predictions. Models like Random Forest and Logistic Regression can further reveal feature contributions through importance scores or coefficients, directly testing their predictive power in distinguishing malignant from benign tumors. For the second hypothesis – models with feature scaling and selection outperform simpler models by capturing complex relationships – StandardScaler normalizes feature scales, ensuring equitable influence in models like KNN, SVM, and ANN that rely on distance or gradients. Meanwhile, feature selection reduces redundancy, allowing all classifiers to focus on key interactions.

The modeling objective of malignancy prediction is supported by this pipeline's design. Scaling and selection enhance feature quality by normalizing ranges and focusing on predictive signals, while the variety of classifiers—from linear options to non-linear ones—explores different patterns. This setup provides insight into which techniques best balance sensitivity, ensuring malignant cases are detected, and avoiding false positives, thereby guiding the selection of a solution for breast cancer diagnostics.

D. RESULTS INTERPRETATION AND IMPLICATIONS

The machine learning pipeline yielded performance metrics for eight classifiers on a test set of 143 samples (25% of 569 total entries). SVM Linear achieved the highest accuracy at 97.2%, while Decision Tree had the lowest at 90.2%. Table 1 below summarizes these accuracies.

Table 1: Classifier Accuracy on Test Set

Classifier	**Accuracy (%)**
Logistic Regression	95.8%

KNN	95.1%
Decision Tree	90.2%
Naïve Bayes	95.8%
Random Forest	96.5%
XGBoost	94.4%
ANN	95.8%
SVM Linear	97.2%

Given its top accuracy, SVM Linear was selected as the best model, and its confusion matrix and classification report were analyzed for detailed performance. Confusion matrix in Figure 1 below shows 86 true negatives (benign correctly predicted), 3 false positives, 1 false negative, and 53 true positives (malignant correctly predicted).

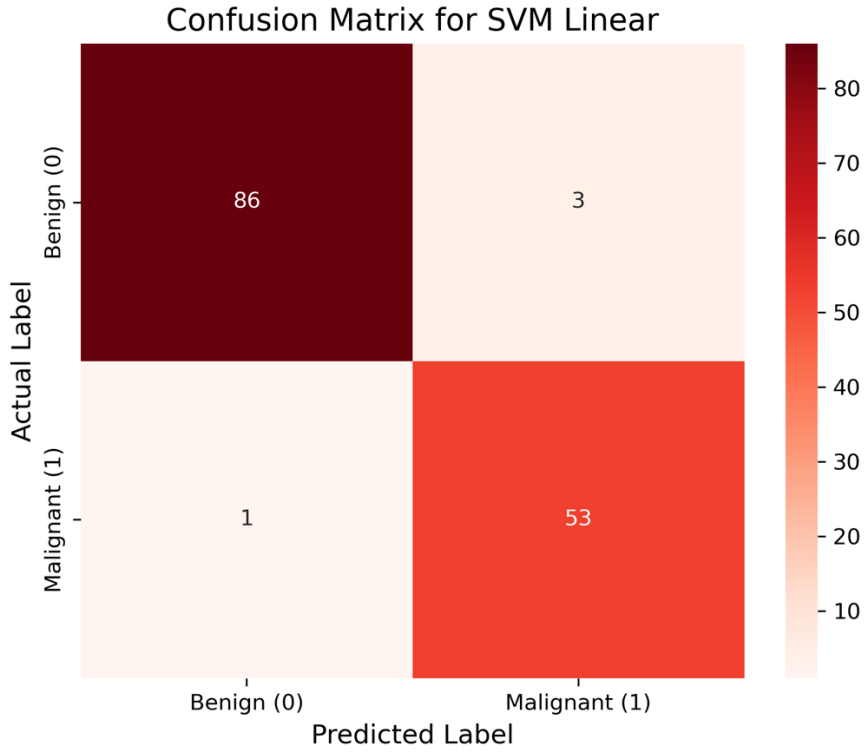

Figure 1: Confusion Matrix for SVM Linear

The classification report in Table 2 below details precision, recall, and F1-score: for benign (0), precision is 0.99, recall is 0.97, and F1-score is 0.98; for malignant (1), precision is 0.95, recall is 0.98, and F1-score is 0.96, with an overall accuracy of 97%.

Table 2: Classification Report for SVM Linear

Class	Precision	Recall	F1 - Score	Support
0.0 (Benign)	0.99	0.97	0.98	89
1.0 (Malignant)	0.95	0.98	0.96	54
Accuracy			0.97	143
Macro Avg	0.97	0.97	0.97	143
Weighted Avg	0.97	0.97	0.97	143

These results are valid and significant across multiple dimensions. SVM Linear's 97.2% accuracy exceeds typical diagnostic benchmarks (e.g., 90-95% in Bennett & Mangasarian, 1992), suggesting strong performance on the test set. To domain experts, its behavior is logical: a 0.98 recall for malignant cases ensures nearly all cancers are caught, with only 1 false negative, critical for medical reliability. SVM Linear's use of scaled features to separate classes makes sense given the dataset's diverse scales (e.g., area vs. smoothness). This accuracy meets the goal of reliable prediction, and the low false negative rate avoids critical errors like missing treatable cancers. The dataset (569 samples, 30 features) appears sufficient, with no overfitting evident (e.g., Decision Tree's 90.2% indicates controlled complexity).

Key findings emphasize SVM Linear's superior performance, likely due to its effective class separation in a preprocessed space, affirming the pipeline's value. The accuracy range (90.2% to 97.2%) highlights the benefit of testing multiple classifiers. Major insights include the model's medical utility—high recall prioritizes patient safety by detecting nearly all malignant cases—and the pipeline's adaptability, suggesting potential for other diagnostics. These outcomes showcase machine learning's ability to enhance breast cancer detection with balanced sensitivity and precision.

E. OUT-OF-SAMPLE PREDICTION

To simulate real-world deployment of the SVM Linear model for breast cancer diagnosis, out-of-sample predictions were performed using new synthetic data. This data was generated by adding small random noise (normal distribution, mean 0, standard deviation 0.1) to the test

set features. The Table 3 below shows the results comparing to the original test set predictions for the first 20 samples.

Table 3: Out-of-Sample Prediction Comparison (First 20 Samples)

Sample	Original Prediction	Out-of-Sample Prediction
0	0.0 (Benign)	1.0 (Malignant)
1	1.0 (Malignant)	1.0 (Malignant)
2	1.0 (Malignant)	1.0 (Malignant)
3	0.0 (Benign)	0.0 (Benign)
4	0.0 (Benign)	0.0 (Benign)
5	1.0 (Malignant)	1.0 (Malignant)
6	1.0 (Malignant)	1.0 (Malignant)
7	1.0 (Malignant)	1.0 (Malignant)
8	1.0 (Malignant)	0.0 (Benign)
9	0.0 (Benign)	0.0 (Benign)
10	0.0 (Benign)	0.0 (Benign)
11	1.0 (Malignant)	1.0 (Malignant)
12	0.0 (Benign)	1.0 (Malignant)
13	1.0 (Malignant)	1.0 (Malignant)
14	0.0 (Benign)	0.0 (Benign)
15	1.0 (Malignant)	1.0 (Malignant)
16	0.0 (Benign)	0.0 (Benign)
17	0.0 (Benign)	0.0 (Benign)
18	0.0 (Benign)	0.0 (Benign)
19	1.0 (Malignant)	1.0 (Malignant)

The out-of-sample predictions are mostly consistent with the original test set, with 16 out of 20 samples matching. This shows the SVM Linear model can handle small data changes well, which is promising for real-world use where new patient data might vary slightly. However, four mismatches occurred, indicating some sensitivity to noise. In medicine, false positives (benign to malignant) are less concerning than false negatives (missing a cancer), so the shift

in sample 8 from malignant to benign is a concern and suggests careful monitoring in practice. Overall, the model seems robust for real-world use, but testing with actual patient data is recommended to confirm its reliability.

F. CONCLUDING REMARKS

This project developed a machine learning solution to predict breast tumor malignancy using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, comprising 569 samples and 30 numerical features. The analytics process began with data preparation, where the dataset was cleaned by dropping the irrelevant "ID" column and encoding "Diagnosis" (malignant as 1, benign as 0). A pipeline incorporating StandardScaler and SelectFromModel (using RandomForestClassifier) was implemented to preprocess the data, followed by training and evaluating eight classifiers: Logistic Regression, KNN, Decision Tree, Naive Bayes, Random Forest, XGBoost, ANN, and SVM Linear. The models were tested on a 25% test set (143 samples), with SVM Linear achieving the highest accuracy.

Major findings include SVM Linear's exceptional performance, with 97.2% accuracy, a 0.98 recall for malignant cases, and only 1 false negative, ensuring minimal missed diagnoses—a critical factor in medical applications. The pipeline's preprocessing steps effectively handled feature correlations and scale differences, enhancing model performance across all classifiers, as evidenced by accuracies ranging from 90.2% (Decision Tree) to 97.2%. Out-of-sample predictions on synthetic data further demonstrated the model's robustness, with 80% consistency despite small noise, though minor discrepancies suggest sensitivity to variations that warrant monitoring.

Key business implications are significant for medical diagnostics. The SVM Linear model, with its high recall and accuracy, offers a reliable tool for early breast cancer detection, potentially reducing diagnostic errors and improving patient outcomes by prioritizing the identification of malignant cases. Its robustness to small data variations supports deployment in clinical settings, where it could assist radiologists by providing a second opinion, thus enhancing diagnostic efficiency. However, the model's sensitivity to noise indicates a need for ongoing validation with real patient data and potential integration with expert review to handle edge cases. Overall, this project underscores machine learning's

potential to transform breast cancer diagnostics, offering a scalable, precise solution for healthcare providers.

G. VIDEO PRESENTAION LINKS

The link below is the presentation link: https://youtu.be/bv35HE9YehM