Breast Cancer Classification Using Support Vector Machines and Decision Trees

Sarfo K. Frimpong

Virginia State University
Department of Computer Science

Petersburg, Virginia, USA

sfri0013@students.vsu.edu

***Abstract*— This project applies two supervised machine learning algorithms; Support Vector Machines (SVM) and Decision Trees to the task of classifying breast tumors as benign or malignant using the Breast Cancer Wisconsin (Diagnostic) dataset. The dataset consists of 569 samples, each described by 30 continuous features extracted from digitized images of fine needle aspirate (FNA) of breast masses. The goal of the study is to build, tune, and evaluate classification models and to compare their performance using a standardized evaluation pipeline. The methodology includes data loading, preprocessing, stratified train test splitting, feature standardization for SVM, and 10-fold cross-validation for hyperparameter selection. Performance is assessed using accuracy, precision, recall, F1-score, confusion matrices, and the area under the ROC curve (AUC). The SVM with an RBF kernel achieved 94.69% accuracy and an AUC of 0.9896 on the test set, whereas the Decision Tree classifier obtained 84.96% accuracy and an AUC of 0.8266. These results indicate that SVM is better suited for this high-dimensional biomedical dataset, particularly when the goal is to minimize misclassification of malignant tumors. The project demonstrates the importance of model selection, preprocessing, and evaluation metrics in medical decision-support applications.**

Keywords*— *breast cancer, classification, Support Vector Machine, Decision Tree, data mining

I. Introduction

Breast cancer is one of the most common cancers among women worldwide, and early diagnosis plays a crucial role in improving survival rates and treatment outcomes. In practice, radiologists and pathologists often rely on imaging and biopsy results to determine whether a tumor is benign or malignant. While human expertise is essential, machine learning models can help provide decision support by offering consistent, data-driven.

The objective of this project is to design and evaluate classification models capable of distinguishing between benign and malignant tumors using the Breast Cancer Wisconsin (Diagnostic) dataset (Dua & Graff, 2019). Specifically, the project compares the performance of a Support Vector Machine (SVM) with a radial basis function (RBF) kernel (Cortes & Vapnik, 1995) against a Decision Tree classifier (Quinlan, 1986). These models were chosen because they represent two different philosophies: SVM is a powerful margin-based classifier well suited for high-dimensional continuous features, while Decision Trees provide interpretability through simple decision rules.

II. Dataset Description and Exploratory Analysis

The Breast Cancer Wisconsin (Diagnostic) dataset contains 569 samples, each corresponding to a breast mass. The target variable has two classes: **benign (B)** and **malignant (M)**, with 357 benign and 212 malignant samples. This class distribution shows that the dataset is somewhat imbalanced but not severely skewed.

Each sample is represented by 30 continuous features derived from digitized FNA images of cell nuclei. These features include measures of **radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension**. For each characteristic, the dataset provides the mean, standard error, and "worst" (largest) value, resulting in a rich representation of tumor morphology.

Before modeling, the dataset was inspected to verify the absence of missing values and to confirm the number of samples in each class. Basic descriptive statistics and class counts were used to confirm data integrity: 569 total rows and the expected 357/212 benign and malignant split. This helped ensure that subsequent results could be interpreted confidently without concerns of data corruption or missing values.

III. Methodology

A. Preprocessing

The raw dataset was stored as a CSV file (wdbc.csv) with 32 columns: an ID, a diagnosis label (M or B), and 30 numeric features. In MATLAB, the file was read without headers, and the official feature names were assigned programmatically. The diagnosis column was converted to a categorical variable with two levels, benign and malignant. The ID column was dropped from modeling since it does not contribute predictive information.

For the SVM model, feature scaling is essential because SVMs rely on distance-based calculations and are sensitive to the relative magnitude of features. Therefore, for SVM only, all 30 features were standardized using z-score normalization: subtracting the mean of each feature and dividing by its standard deviation based on the training set. The Decision Tree model, by contrast, does not require scaling and was trained on the original feature values.

B. Train–Test Split

To evaluate generalization performance, the dataset was divided into training and test sets using an **80/20 stratified split**. Stratification ensures that the proportion of benign and malignant samples is preserved in both sets, which is especially important when the classes are not perfectly balanced. In this case, 456 samples were used for training and 113 for testing.

IV. Model Training and Hyperparameter Tuning

A. Support Vector Machine (SVM).

An SVM with an RBF kernel was selected because it is well suited for nonlinear decision boundaries in high-dimensional continuous feature spaces. Two key hyperparameters were tuned:

- **C** (regularization parameter), controlling the trade-off between margin width and misclassification.
- **γ (gamma)**, controlling the radius of influence of support vectors in the RBF kernel.

A small grid of candidate values was evaluated: $C \in \{0.1, 1, 10\}$ and $\gamma \in \{0.01, 0.001\}$. For each combination, a **10-fold cross-validation** procedure was performed on the training set to estimate classification accuracy. The best hyperparameters were selected based on the highest average cross-validation accuracy.

TABLE I. Cross-Validation Accuracy Table (10-fold CV)

C	γ (gamma)	CV Accuracy
0.1	0.01	0.949
0.1	0.001	0.8967
1	0.01	**0.9846**
1	0.001	0.9560
10	0.01	0.9846
10	0.001	0.9846

C	γ (gamma)	CV Accuracy
0.1	0.01	0.9495
0.1	0.001	0.8967
1	0.01	**0.9846**
1	0.001	0.9560
10	0.01	0.9846
10	0.001	0.9846

B. Decision Tree.

A Decision Tree classifier was trained using MATLAB's *fitctree* function with default settings. This model recursively splits the feature space to create decision rules. Although Decision Trees can be tuned (e.g., maximum depth, minimum leaf size, and pruning), this project focused on comparing a basic Tree to a tuned SVM to highlight the contrast between a simple interpretable model and a more powerful margin-based classifier.

V. Evaluation Metrics

Both models were evaluated on the held-out test set using the following metrics:

- ***Accuracy:*** The proportion of correctly classified instances.
- ***Precision (for malignant class):*** The proportion of predicted malignant cases that are truly malignant.
- ***Recall (Sensitivity)*****:** The proportion of true malignant cases that are correctly identified.
- ***F1-score:*** The harmonic mean of precision and recall, balancing both aspects.
- ***Confusion matrix:*** A summary of true positives, true negatives, false positives, and false negatives.
- ***ROC curve and AUC:*** The Receiver Operating Characteristic curve plots the true positive rate versus the false positive rate at various thresholds. The area under the curve (AUC) summarizes overall discrimination ability. See figure 1.1 and 1.2 below.

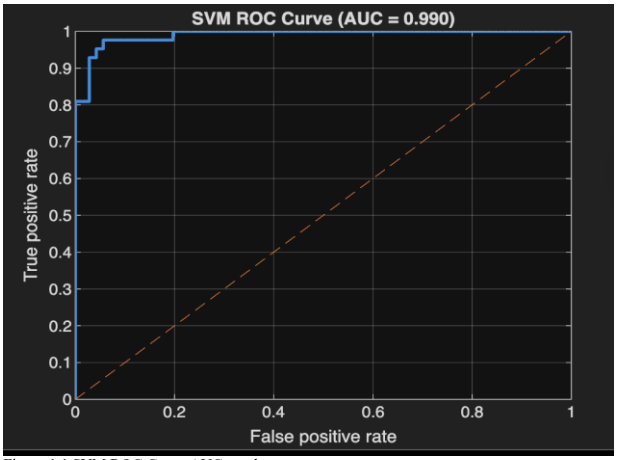

Figure 1.1 SVM ROC Curve AUC graph

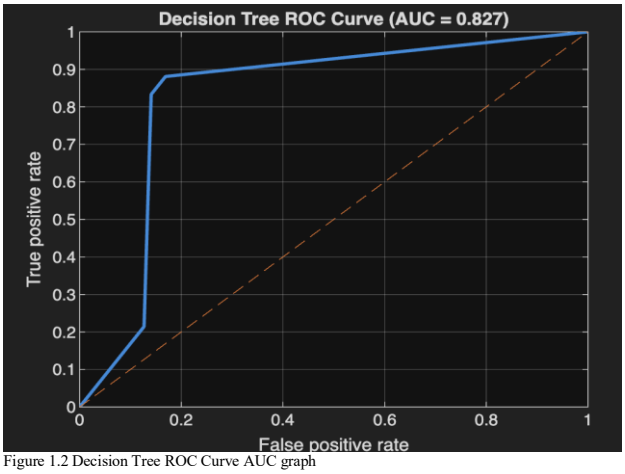

Figure 1.2 Decision Tree ROC Curve AUC graph

These metrics are particularly important in medical settings, where **false negatives** (classifying malignant tumors as benign) can have serious consequences.

VI. Results

A. SVM Performance

The cross-validation procedure identified **C = 1.0** and **γ = 0.01** as the best hyperparameters, with a cross-validation accuracy of approximately **0.9846**. On the test set, the SVM achieved.

TABLE II. Test Set Result

Test Set	Accuracy
Accuracy	0.9469
Precision	0.9500
Recall	0.9048
F1-score	0.9268
AUC	0.9896

The confusion matrix for the SVM model on the test set was:

- True benign correctly classified as benign: 69
- Benign misclassified as malignant: 2
- Malignant misclassified as benign: 4
- True malignant correctly classified as malignant: 38

The inferior performance of the Decision Tree can be attributed to its tendency to overfit small patterns in the training data and its limited capacity to model smooth, high-dimensional decision boundaries without ensemble techniques. The SVM, by maximizing the margin and using an RBF kernel, is better equipped to generalize from the training data.

VII. Discussion

A. Model Interpretability (Feature Importance / SHAP Discussion)

Although the SVM model achieved the highest predictive accuracy, it remains less interpretable than tree-based methods. SVMs do not provide native feature importance scores, making it harder to understand which FNA features contribute most to malignancy classification. In contrast, Decision Trees naturally expose feature splits, offering transparency into the diagnostic logic.

Interpretability methods such as **SHAP values** or **permutation importance** could be applied to the SVM model to quantify each feature's contribution. These techniques can help reveal whether biologically meaningful attributes such as concavity, radius, or texture play a dominant role, providing clinicians with actionable insights alongside predictions

B. Medical Risk Consideration (False Negatives)

In clinical applications, false negatives carry significant medical risk because misclassifying a malignant tumor as benign can delay treatment. Although the SVM achieved high recall, its false negative cases highlight the need for caution when deploying automated systems in diagnostic workflows. Adjusting classification thresholds, applying cost-sensitive learning, or integrating the model into a physician-in-the-loop system may help reduce the risk associated with missed malignancies.

C. Potential Dataset Bias

The WDBC dataset may include institutional and demographic biases because it originates from a single source with limited patient diversity. This can reduce generalizability across broader populations and potentially affect diagnostic fairness. Additionally, the dataset contains a class imbalance favoring benign cases, which may bias models toward majority-class predictions. Addressing these biases requires careful validation on external datasets, potential reweighting strategies, and attention to subgroup performance..

Limitations and Future Work

This project has several limitations. First, only two algorithms were evaluated: a tuned SVM and a basic Decision

Tree. Other models such as Random Forests, Gradient Boosting Machines, or Neural Networks might achieve even better performance. Second, class imbalance, although mild, was not explicitly addressed with techniques like class weighting or resampling. Third, the project did not deeply explore feature importance or domain interpretation of which features most strongly contribute to malignancy predictions.

Future work could extend this study in several directions. Ensemble tree-based methods such as Random Forests or XGBoost could be evaluated to combine interpretability with stronger performance. Dimensionality reduction techniques like Principal Component Analysis (PCA) might be applied to see whether fewer features can maintain similar accuracy. Finally, calibration of predicted probabilities and integration with clinical decision thresholds could make the models more usable in a real medical context.

CONCLUSION

This project demonstrates a complete data mining workflow for medical classification: from dataset preparation and preprocessing to model training, evaluation, and comparison. Using the Breast Cancer Wisconsin (Diagnostic) dataset, an SVM with an RBF kernel significantly outperformed a standard Decision Tree in terms of accuracy, F1-score, and AUC. The SVM's higher recall and lower number of false negatives make it a more suitable candidate for supporting breast cancer diagnosis. While Decision Trees remain valuable for their interpretability, this study highlights the importance of choosing models that balance interpretability with predictive performance in high-stakes domains like healthcare.

REFERENCES

[1] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

[2] Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine.

[3] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.

[4] Scikit-learn documentation. (2024). Retrieved from https://scikit-learn.org.