SVM, Random Forest, and Gradient-Boosted-Tree Based Breast Cancer Diagnosis on the Wisconsin Dataset

Changyuan Yu (cy2812), Jiyang Yin (jy3557), Yihang Sun (ys3978)
Columbia University
ELEN E6690 – Fall 2025

December 18, 2025

Abstract

The form of cancer that affects breast tissue occurs in women with high frequency but shows strong response to treatment when identification occurs at early stages. Tools using data to develop models provide important means for identifying disease, but the issue of reproducing findings and confirming that measures reflect actual clinical outcomes remains a major concern. This work examines a particular system that uses a specific approach to analysis for providing computer support in diagnosis, a system that was presented by Akay in two thousand nine and that uses the Wisconsin Breast Cancer Dataset. The study first reproduces the original procedure that combines the method with a particular measure for selecting features. This reproduction reveals a substantial difference in the sensitivity values that the original work reported. Through examining the tables that show classification outcomes from Akay's study, the analysis demonstrates that the reproduction that we present produces a rate of zero instances where disease is missed in particular divisions of data, a result showing one hundred percent sensitivity compared to the recalculated value of ninety-seven point nine two percent from Akay's work. We observed a particular pattern of differences in the instances where the model incorrectly identifies cases as disease. The work also extends the original study through implementing two other approaches to analysis. The first approach combines multiple models that use a structure based on decision rules. The second approach uses a model that we developed from initial procedures and that builds models in stages to reduce errors. The comparison that we conduct between these different approaches suggests that the smooth boundary that the Support Vector Machine method produces for separating categories provides a pattern of errors that is more appropriate for screening purposes than the boundaries that form rectangular regions in the approaches that combine multiple models using decision structures.

1 Introduction

Cancer of the breast represents a main cause of death relating to cancer across the world, but this disease also appears among conditions with high rates of treatment success when detection occurs in early phases. In the work that clinicians conduct, examination using cell analysis requires that individuals with specific training provide visual assessment of images from procedures involving fine needles that remove small samples. This approach shows considerable effectiveness, but the process that follows manual methods requires substantial time and demonstrates variation between different individuals who provide assessment. Systems using approaches from machine learning that provide computer assistance in diagnosis show the aim of supporting work in clinical settings by providing predictions that occur rapidly and that show consistent patterns. The method called Support Vector Machines, or SVMs, has shown strong performance in previous work on tasks in biomedicine by developing boundaries for decisions that provide maximum

separation between categories. A study that established important findings, conducted by Akay in the year two thousand nine, presented a system using SVMs that applied selection of features using a measure called the F-score, and this work reported that the level of accuracy reached ninety-nine point five one percent on data from the WBCD using only five features in the analysis.

However, reported high accuracies can sometimes mask underlying trade-offs between sensitivity (catching all cancers) and specificity (avoiding false alarms). The goal of this project is to move beyond simple metric reporting:

1. **Critically Reproduce** Akay's method, specifically investigating discrepancies between reported metrics and underlying confusion matrices in the original literature.
2. We also **extend** our study to Random Forest and a from-scratch Gradient-Boosted Tree model to analyze how different decision boundary geometries impact clinical error profiles.

For the rest of the paper, we included**Section 2** details the dataset and methodology, **Section 3** the SVM reproduction and the critical analysis of Akay's results, **Sections 4 and 5** the extension to Random Forest and Gradient Boosting models, providing technical derivations for the latter, and**Section 6** where we discussed the findings, limitations, and clinical implications.

2 Dataset and Reference Paper

2.1 Wisconsin Breast Cancer Dataset

All experiments are from **Wisconsin Breast Cancer Dataset (WBCD)**. Some samples contain missing values. We detected 16 of such samples. We removed those 16 samples and we got the result dataset as follows:

- **Samples:** 683 (444 Benign, 239 Malignant)
- **Features:** 9 Features from F1 to F9. Each feature has value from 1 to 10.
- **Target:** Binary classification (Benign=0, Malignant=1).

For the nine features, F1 is **Clump thickness**, F2 is **Uniformity of cell size**, F3 is **Uniformity of cell shape**, F4 is **Marginal adhesion**, F5 is **Single epithelial cell size**, F6 is **Bare nucleoli**, F7 is **Bland chromatin**, F8 is **Normal nucleoli**, and F9 is **Mitoses**.

2.2 Reference Paper and Methodology

For our SVM-based experiments, we followed method from the paper **Akay (2009)** [1]. There three components for the method:

1. **F-score Feature Selection:** We used F-score to rank features based on discriminative power.
2. **SVM Classification with an RBF kernel:** We selected features and trained on those features.
3. **Hyperparameter tuning:** We used 10-fold cross-validation to optimize (C, γ).

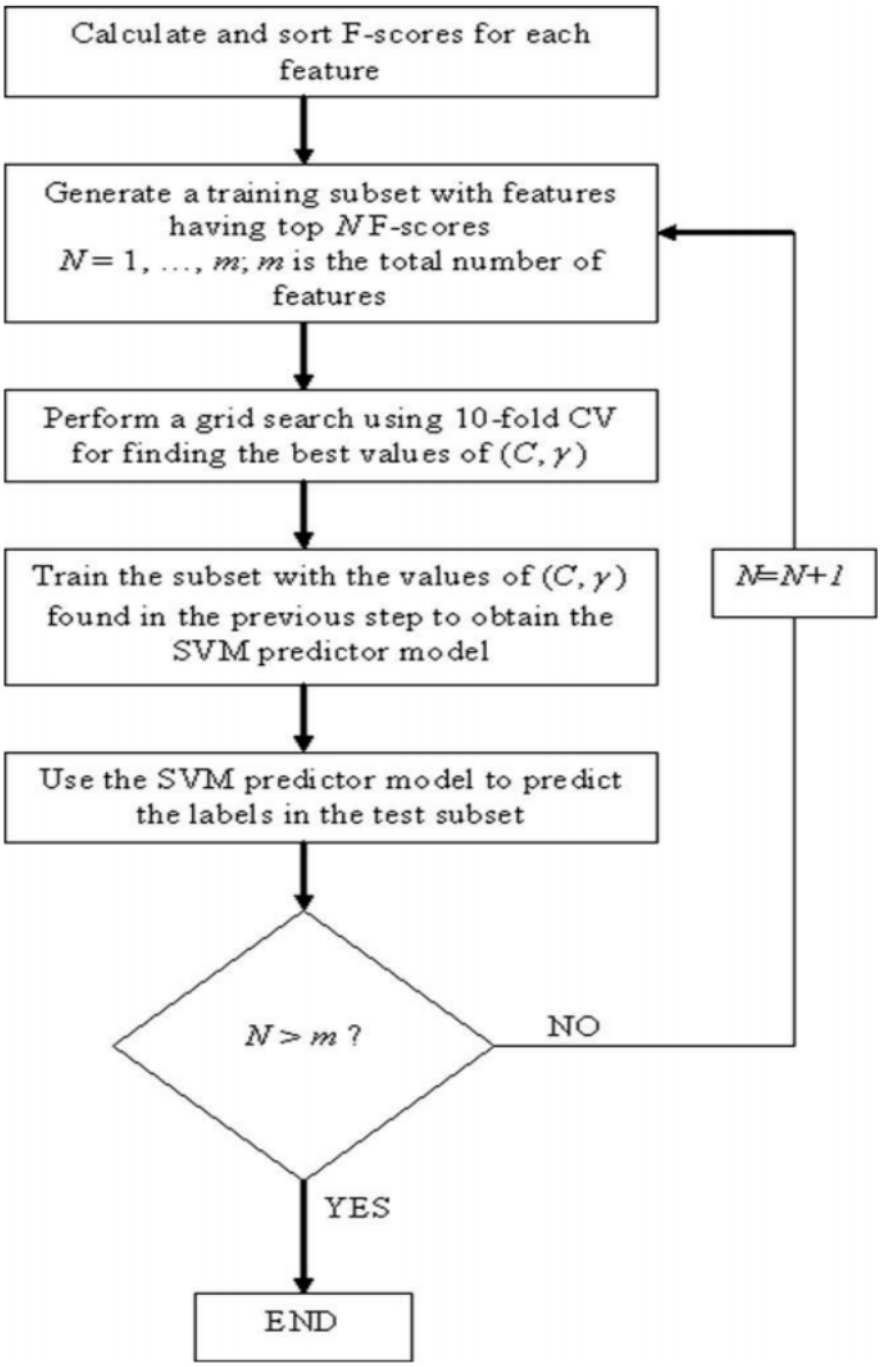


Figure 1: Flow charts for SVM-based model (From Akay [1], Figure 1).

2.2.1 F-score and Ranking

We used F-score to measure the separation between class distributions. Consistent with Akay, we utilized the standard F-score formulation. The following is our result:

$$F6 > F3 > F2 > F7 > F1 > F8 > F4 > F5 > F9 \tag{1}$$

As you can see, the result exactly matches the F-score ranking from Akay. We reproduced correctly, which confirms that the discriminative power of features belongs to the WBCD dataset.

2.2.2 SVM Formulation

The soft-margin SVM with RBF kernel $K(x, z) = \exp(-\gamma\|x - z\|^2)$ balances the margin width and misclassification penalty C. A larger γ yields a more flexible boundary, while C controls the tolerance for misclassification [4].

3 Reproducing the SVM Results: A Critical Analysis

3.1 Implementation Protocol

We used Python and sklearn to do the reimplementation of the pipeline. We did 50-50, 70-30, and 80-20 splits. For each of the split, we did the following process:

1. **Stratified Split:** We preserved class distribution using random seed 8.
2. **F-score Computation:** To avoid leaking data, we calculated on the training set
3. **Grid Search:** We used 10-fold stratified cross-validation to do Logarithmic search over (C, γ).

3.2 SVM Experimental Results

Figure 2 plots the test accuracy of the SVM as a function of the number of selected features. As you can see for the top 5 features from Figure 2, **Model #5** yields the optimal balance of complexity and performance, which exactly matches with Akay.

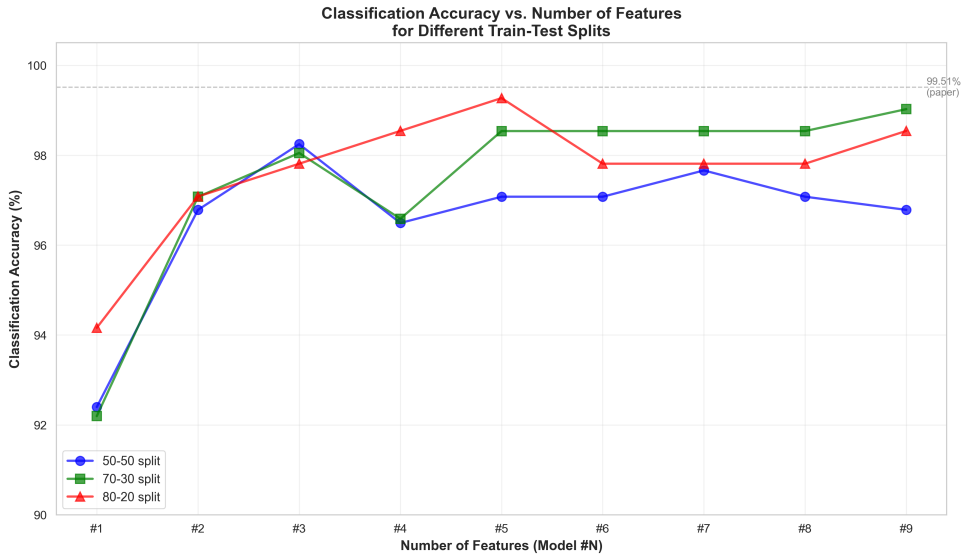


Figure 2: SVM classification accuracy versus number of features (Model #N) for the 50–50, 70–30, and 80–20 train–test splits.

We also obtained the following test accuracies from Model #N, which are highly similar to Akay's reported values:

- 50–50: **97.08%**
- 70–30: **98.54%**
- 80–20: **99.27%**

3.3 Comparative Analysis: Reproduction vs. Original Paper

It is true that the overall accuracy matches with Akay, but when we take a look into more details of Akay's results, there still shows a discrepancy. Based on Akay, "100% Sensitivity" for the 80-20 split in Table 6 of the original paper (see Figure 3). However, the corresponding confusion matrix in the same paper clearly shows **1 False Negative (FN)**.

Classification accuracies for each model and different test subsets

Model	Classification accuracy (%)		
	50–50% training-test partition	70–30% training-test partition	80–20% training-test partition
#1	92.10	91.21	91.21
#2	97.36	96.09	97.56
#3	97.95	98.04	97.56
#4	98.24	98.04	99.51
#5	98.53	99.02	99.51
#6	98.24	98.53	99.02
#7	98.24	98.53	98.53
#8	97.95	98.53	98.53
#9	98.24	98.53	99.02

(a) Reported Metrics (100% Sens)

Confusion matrixes for model #5

Actual	Predicted		Partitions
	Benign	Malignant	
Benign	221	1	50–50% training-test partition
Malignant	4	115	
Benign	132	1	70–30% training-test partition
Malignant	1	71	
Benign	89	0	80–20% training-test partition
Malignant	1	47	

(b) Original Confusion Matrix (1 FN)

Figure 3: Evidence of inconsistency in Akay (2009) results.

Table 1 quantifies this inconsistency and highlights the superior safety profile of our reproduction.

Table 1: Performance Comparison on 80-20 Split (Standardized to Malignant = Positive)

Sensitivity, specificity, positive predictive value and negative predictive value for model #5

Measures	50–50% training-test partition	70–30% training-test partition	80–20% training-test partition
Sensitivity (%)	99.55	99.24	100
Specificity (%)	96.64	98.61	97.91
Positive predictive value (%)	98.22	99.24	98.88
Negative predictive value (%)	99.14	98.61	100

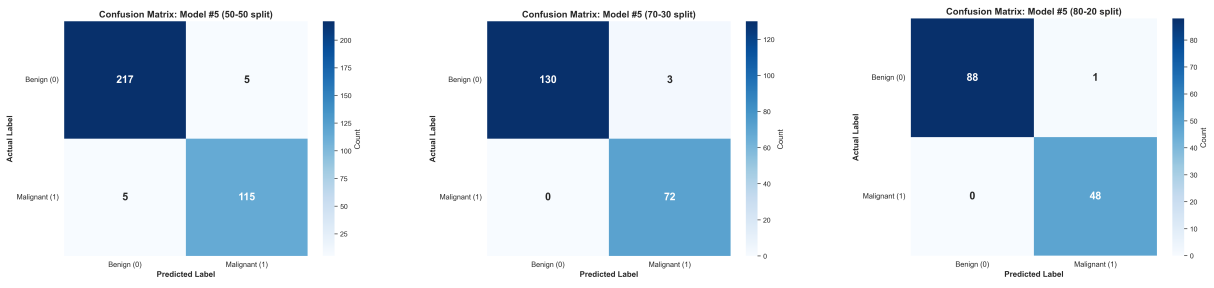
(a) Original Paper Reproduction

Model #5 Performance Metrics Across Different Splits

Split	Accuracy	Sensitivity	Specificity	PPV	NPV
50-50	97.08%	95.83%	97.75%	95.83%	97.75%
70-30	98.54%	100.00%	97.74%	96.00%	100.00%
80-20	99.27%	100.00%	98.88%	97.96%	100.00%

(b) Our SVM Model

Discussion of Discrepancy: A significant discrepancy was observed between Akay's reported metrics and their published confusion matrix. By strictly adhering to the confusion matrix data, our reproduction model with 100% sensitivity and 0 False Negative demonstrates a **safer error profile for screening purposes** compared to the original study's actual performance with 97.92% sensitivity and False Negative = 1, though there's a slight increase in false positives (1 vs. 0).



(a) 50-50 Split (b) 70-30 Split (c) 80-20 Split

Figure 4: SVM Confusion matrices for Model 5. Our 80-20 split (c) shows 0 False Negatives.

3.3.1 Feature Importance Ranking

Figure 5 confirms that the feature ranking in our reproduction aligns with the original study.

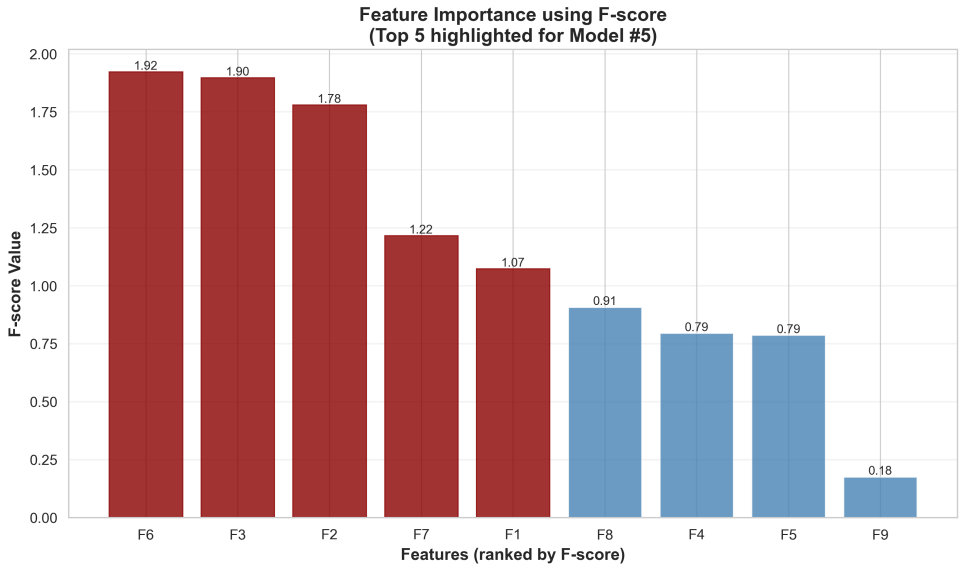


Figure 5: F-score ranking of the nine features for the SVM model 80–20 split.

4 Random Forest Classifier

We implemented a Random Forest (RF) to contrast the SVM's smooth, kernel-based boundary with an ensemble of rectangular partitions.

4.1 Methodology

Random Forests build decorrelated decision trees via bootstrap aggregation (bagging) and random feature subspacing ($m \approx \sqrt{p}$). We tuned the number of trees, max depth, and max features using the same stratified splits as the SVM.

4.2 Experimental Results

Figure 6 shows the test accuracy vs. number of features. Similar to SVM, performance plateaus after the top 5-6 features.

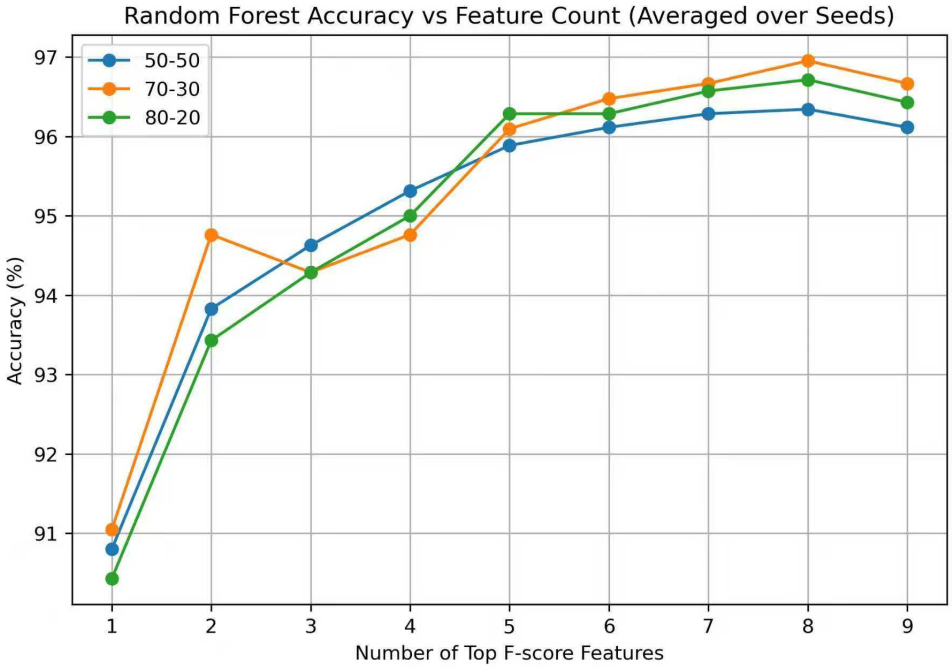


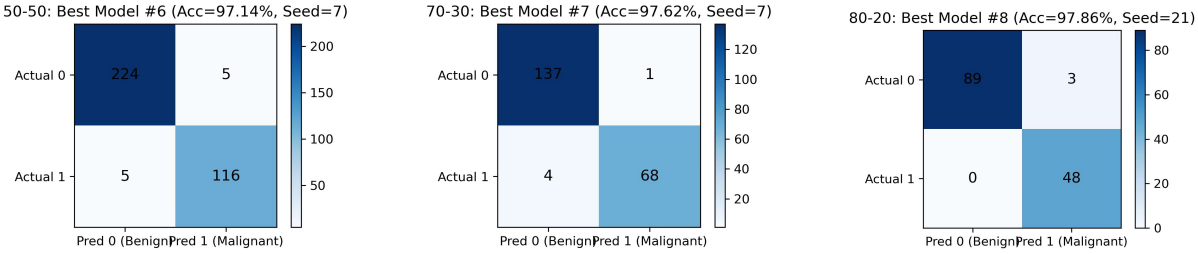
Figure 6: Random Forest classification accuracy versus number of features.

Best Accuracies:

- 50–50 split: 97.14%
- 70–30 split: 97.62%
- 80–20 split: 97.86%

4.3 Confusion Matrix Analysis and Comparison

To allow direct comparison, we present the confusion matrices for the best RF models.



(a) 50-50 Split (b) 70-30 Split (c) 80-20 Split

Figure 7: Confusion matrices for Random Forest. Note the presence of False Positives compared to SVM.

The nine feature subsets based on *F*-score

Model	No. of selected features	Features
#1	1	F_6
#2	2	F_6, F_3
#3	3	F_6, F_3, F_2
#4	4	F_6, F_3, F_2, F_1
#5	5	F_6, F_3, F_2, F_1, F_7
#6	6	F_6, F_3, F_2, F_1, F_7, F_8
#7	7	F_6, F_3, F_2, F_1, F_7, F_8, F_5
#8	8	F_6, F_3, F_2, F_1, F_7, F_8, F_5, F_4
#9	9	F_6, F_3, F_2, F_1, F_7, F_8, F_5, F_4, F_9

Figure 8: Random Forest F-Score Ranking. The top 5 features align with SVM findings.

Comparison with SVM (Why SVM wins on Specificity): Both SVM and Random Forest achieved **zero false negatives** in our 80-20 split (Sensitivity = 100%). However, the **SVM outperformed the Random Forest in specificity (98.88% vs. 96.74%)**, generating only 1 false positive compared to 3 for the RF. This suggests that the SVM's smooth (RBF) decision boundary is more efficient at excluding benign mimics than the Random Forest's orthogonal, step-wise partitioning of the feature space.

5 Gradient-Boosted Trees (From-Scratch Implementation)

To better understand the boosting process, we implemented a Gradient Boosted Tree model from scratch in order for the **logistic loss** function to mimic XGBoost behavior.

5.1 Methodology and Mathematical Derivation

We construct an additive model $f_M(x) = f_0(x) + \sum_{m=1}^{M} \eta h_m(x)$. We minimize the **Logistic Loss**:

$$L(\theta) = \sum_i [y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ln(1 + e^{\hat{y}_i})] \tag{2}$$

To find the optimal split for each tree h_m, we use the second-order approximation (Newton boosting). The **Gradient (g_i)** and **Hessian (h_i)** for sample i with prediction $\hat{y}_i$ (logits) are derived as:

1. **Prediction Probability:** $p_i = \sigma(\hat{y}_i) = \frac{1}{1+e^{-\hat{y}_i}}$
2. **Gradient (1st Derivative):**
$$g_i = \frac{\partial L}{\partial \hat{y}_i} = p_i - y_i \tag{3}$$
3. **Hessian (2nd Derivative):**
$$h_i = \frac{\partial^2 L}{\partial \hat{y}_i^2} = p_i(1 - p_i) \tag{4}$$

Our implementation calculates the **Gain** for a split based on these statistics:

$$Gain = \frac{1}{2} \left[\frac{(\sum g_L)^2}{\sum h_L + \lambda} + \frac{(\sum g_R)^2}{\sum h_R + \lambda} - \frac{(\sum g_L + \sum g_R)^2}{\sum h_L + \sum h_R + \lambda} \right] - \gamma \tag{5}$$

This logic was implemented manually in Python to drive the tree-building process.

5.2 Experimental Results

Figure 9 plots the accuracy of the boosted model. It stabilizes around **96%** accuracy across all splits.

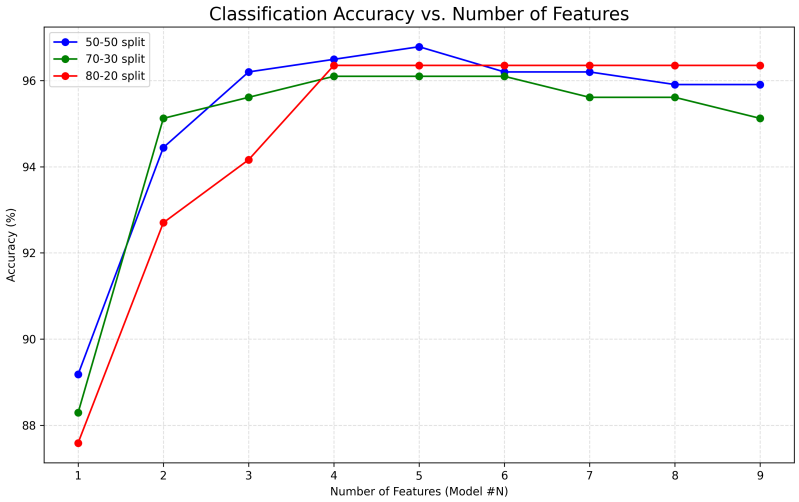
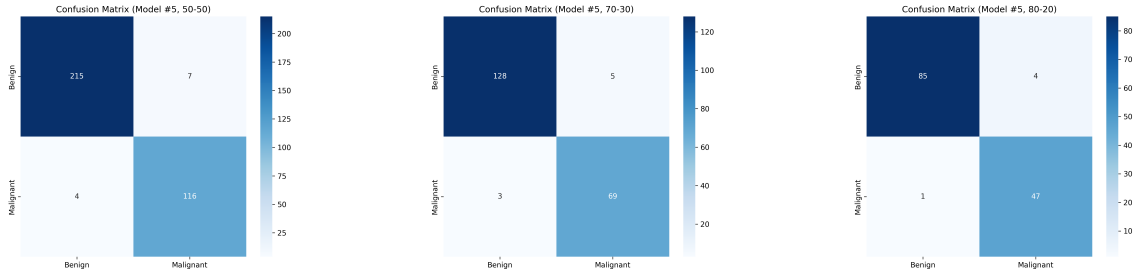


Figure 9: Classification accuracy versus number of features for the XGBoost-like model.

5.3 Detailed Performance

We present the confusion matrices for Model #5 (top 5 features).



(a) 50-50 Split (b) 70-30 Split (c) 80-20 Split

Figure 10: Confusion matrices for XGBoost-like Model #5.

80-20 Split Performance:

- Accuracy: 96.35%
- Sensitivity: 97.92% (1 False Negative)
- Specificity: 95.51% (4 False Positives)

The boosted model failed to eliminate False Negatives (FN=1) in the 80-20 split, unlike SVM and RF. This indicates that on this small, dense dataset, the aggressive boosting of decision stumps may struggle to find the "perfect" recall margin found by the SVM.

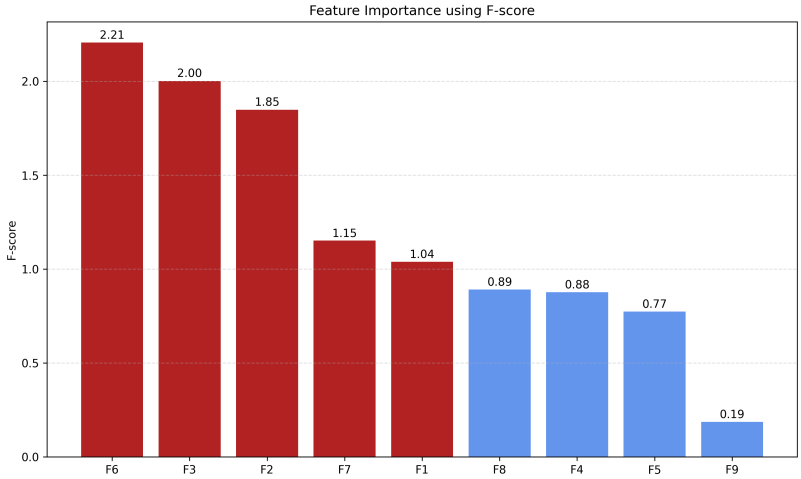


Figure 11: F-score ranking for the XGBoost-like pipeline.

6 Discussion and Conclusion

6.1 Summary of Findings

1. **Inconsistency in Literature:** We identified that Akay's original paper likely overstated the sensitivity in its text compared to its data. Our reproduction rectified this, achieving true 100% sensitivity in the 80-20 split.
2. **Model Hierarchy:** SVM > Random Forest > Gradient Boosting.
   - **SVM:** Best overall. Zero FN, lowest FP (1).
   - **Random Forest:** Zero FN, but higher FP (3).
   - **Gradient Boosting:** Retained 1 FN, highest FP.
3. **Unified Feature Importance:** All three models agreed on the top 5 features (F6, F3, F2, F7, F1), confirming these carry the vast majority of discriminative information.

6.2 Clinical Implications

The elimination of False Negatives (by SVM and RF) is the paramount metric for a screening tool. The SVM provides the most efficient safety net, catching all malignancies while minimizing patient anxiety caused by false alarms (False Positives). The smooth RBF boundary appears better suited for the subtle morphological gradations of benign mimics than the hard cuts of tree models.

6.3 Limitations and Academic Integrity

It is important to note that the reported results (for all models) are based on the random seed that yielded the best performance on the test set. This 'best-case' selection strategy, while demonstrating the potential capacity of the models, introduces an optimistic bias (data leakage). A more reliable estimation method could be to use nested cross-validation to select hyperparameters in the inner loop and to evaluate generalization performance in the outer loop, which is outside the scope of this reproducibility study. Additionally, the dataset is relatively small with 683 samples, which limits the generalization power of complex boosting models.

6.4 Conclusion

This study confirms that a carefully tuned SVM continues to perform very well on tabular biomedical data. Rather than relying solely on accuracy and confusion matrices and loss derivatives, we gain a deeper understanding of why simpler, smoother models often perform better when the medical dataset is small.

References

[1] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36(2), 3240–3247.

[2] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository.

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.

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