Breast Cancer Classification Using Support Vector Machines (SVM)

Team YOLO - s19 (205, 355, 408, 420)

1. Introduction

Support Vector Machines (SVM) are supervised learning models used for classification and regression tasks. They work by finding an optimal hyperplane that maximizes the margin between different classes in the feature space. This report explains the implementation of SVM for a classification problem using features selection method called SHAP (SHapley Additive exPlanations). Following a structured approach that includes dataset selection, model training, and performance evaluation.

Key Definitions

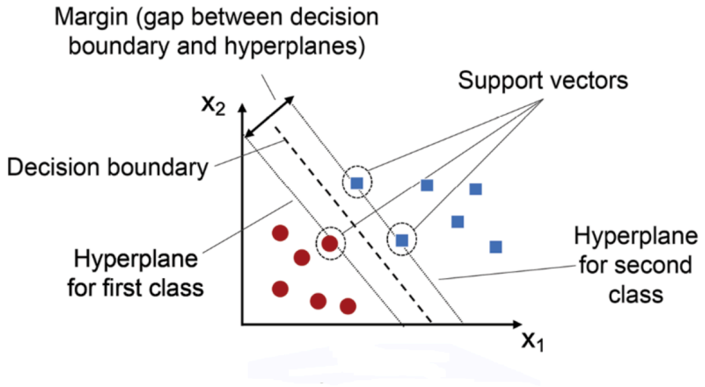

Margin (gap between decision boundary and hyperplanes)
Support vectors
x_2
Decision boundary
Hyperplane for first class
Hyperplane for second class
x_1

Figure 1 : SVM for binary classification

- **Hyperplane** : A decision boundary that separates different classes in an SVM model. The optimal hyperplane is chosen to maximize the margin between classes.
- **Support Vectors** : Support vectors are the data points closest to the hyperplane. They are important for defining the margin in SVM classification.
- **SHAP** (SHapley Additive exPlanations) : SHAP is an explainability technique for machine learning models, based on Shapley values. It identifies the most important features by computing the mean absolute SHAP values.
- **PCA Visualization** : Principal Component Analysis (PCA) is a dimensionality reduction technique. That transforms high-dimensional data into 2D or 3D representations. (making clusters or patterns easier to identify)

2. Project Workflow

I. Problem Identification & Project Planning
II. Data Collection
III. Algorithm Development
IV. Testing with Breast Cancer dataset
V. Report Writing

3. Dataset Details

The dataset used in this project comes from the **Breast Cancer dataset** in `sklearn.datasets`. It contains numerical features that describe tumor characteristics, and the target variable represents **benign (0)** and **malignant (1)** cases. (contains 569 samples with 30 features)

4. Methodology

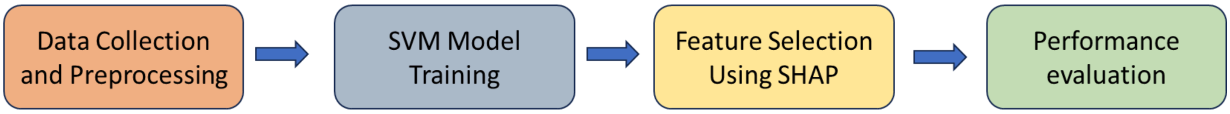

Data Collection and Preprocessing
SVM Model Training
Feature Selection Using SHAP
Performance evaluation

Figure 2: Followed methodology

4.1. Data Collection

We used a publicly available breast cancer dataset for our study. This dataset consists of various features that describe tumor characteristics, which help in classifying whether a tumor is malignant or benign.

4.2. Data Preprocessing

To ensure data quality and improve model performance, we removed null values from the dataset to prevent inconsistencies

4.3. Model Development

For classification, we used a **Support Vector Machine (SVM)** model. The steps involved were:

- Training the SVM model using the dataset.
- Optimizing the model based on feature selection to enhance performance.
- Fine-tuning hyperparameters if necessary to improve classification results.

4.4. Feature Selection

Feature selection was performed to identify the most important features contributing to the model’s prediction. We followed these steps,

- Initial Model without Feature Selection: We first trained the model using all available features.

- Feature Selection using SHAP (SHapley Additive Explanations): SHAP values were used to determine feature importance by analyzing their impact on model predictions.

- Reducing Features Step by Step: Based on SHAP results, we reduced the number of features to 15, 10, and 5, progressively refining the model with fewer features while maintaining performance.

4.5. Performance Evaluation

To assess the effectiveness of the model, we evaluated its performance using several metrics such as Precision, Recall, F1-Score, and Accuracy.

5. Results and Discussion

5.1. SVM decision boundary for all features

When all features were used to train the model, it achieved the highest accuracy compared to other methods (using 15, 10, and 5 features).

SVM model Accuracy is 98% and also, it has one FN and the best precision/recall balance.

5.2. SVM decision boundary for selected features

First, let's see how SHAP values vary from feature to feature. According to that we have selected the top features for training the model.

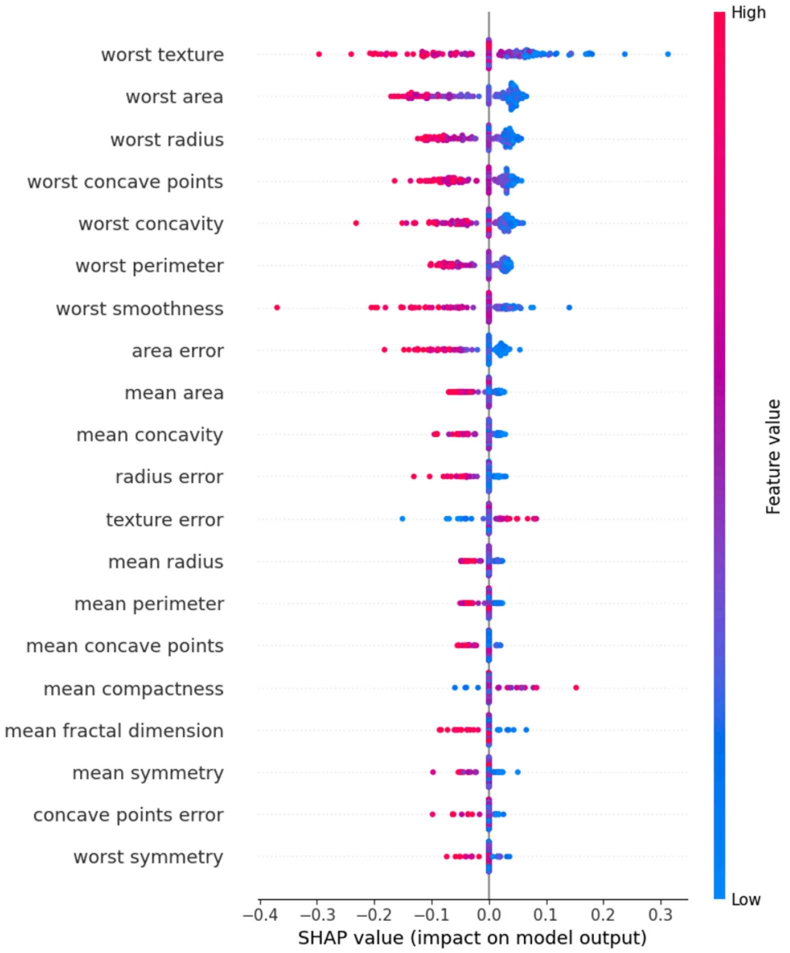
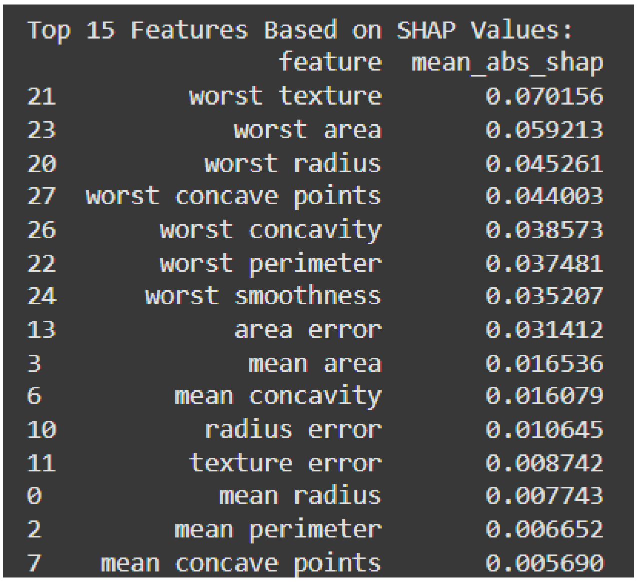


Figure 3 : SHAP for feature importance measure

- **SVM decision boundary for top 15 features**

These are the top 15 features that are chosen by the SHAP.



```
Top 15 Features Based on SHAP Values:
                 feature  mean_abs_shap
21         worst texture       0.070156
23            worst area       0.059213
20          worst radius       0.045261
27  worst concave points       0.044003
26       worst concavity       0.038573
22       worst perimeter       0.037481
24      worst smoothness       0.035207
13            area error       0.031412
3              mean area       0.016536
6         mean concavity       0.016079
10          radius error       0.010645
11         texture error       0.008742
0            mean radius       0.007743
2         mean perimeter       0.006652
7    mean concave points       0.005690
```

Figure 4 : Top 15 features selected by SHAP

Then using the above features we trained the SVM for classification. Accuracy is 96%. It has a lower recall for benign (94%). But the recall of full features is 97%.

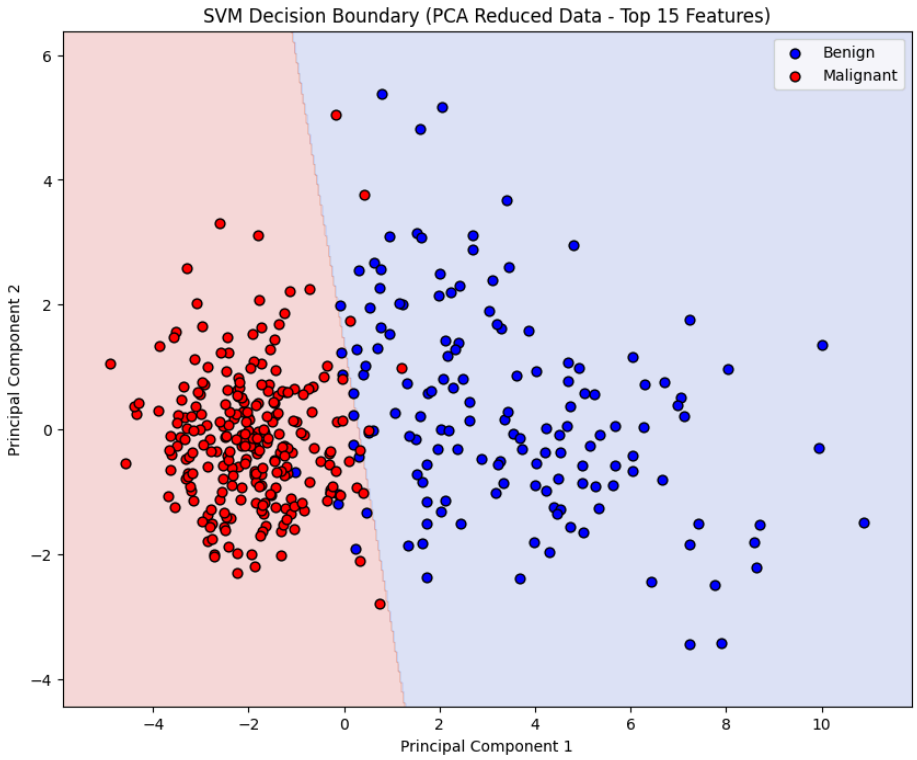
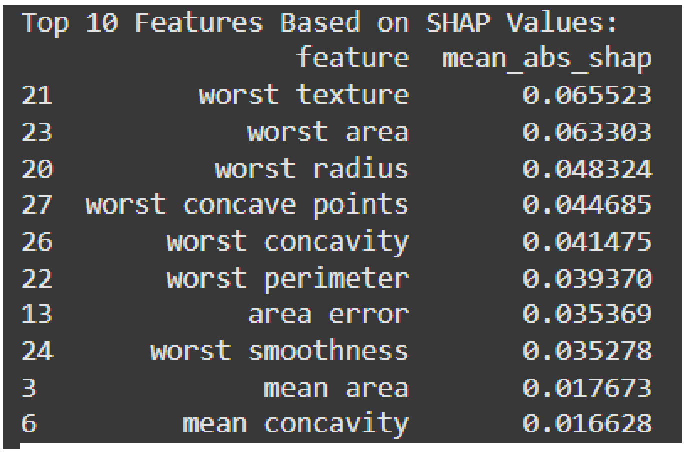


Figure 5 : Decision boundary using top 15 features

- **SVM decision boundary for top 10 features**

These are the selected 10 features according to SHAP values.



```
Top 10 Features Based on SHAP Values:
                 feature  mean_abs_shap
21         worst texture       0.065523
23            worst area       0.063303
20          worst radius       0.048324
27  worst concave points       0.044685
26       worst concavity       0.041475
22       worst perimeter       0.039370
13            area error       0.035369
24      worst smoothness       0.035278
3              mean area       0.017673
6         mean concavity       0.016628
```

Figure 6 : Top 10 features selected by SHAP

Model accuracy is 96%. It matches with the top 15 features' accuracy. Spatially this gives the better precision for malignant (0.97) than Top 15 (0.96). Better precision for malignant (0.97) than Top 15 (0.96).

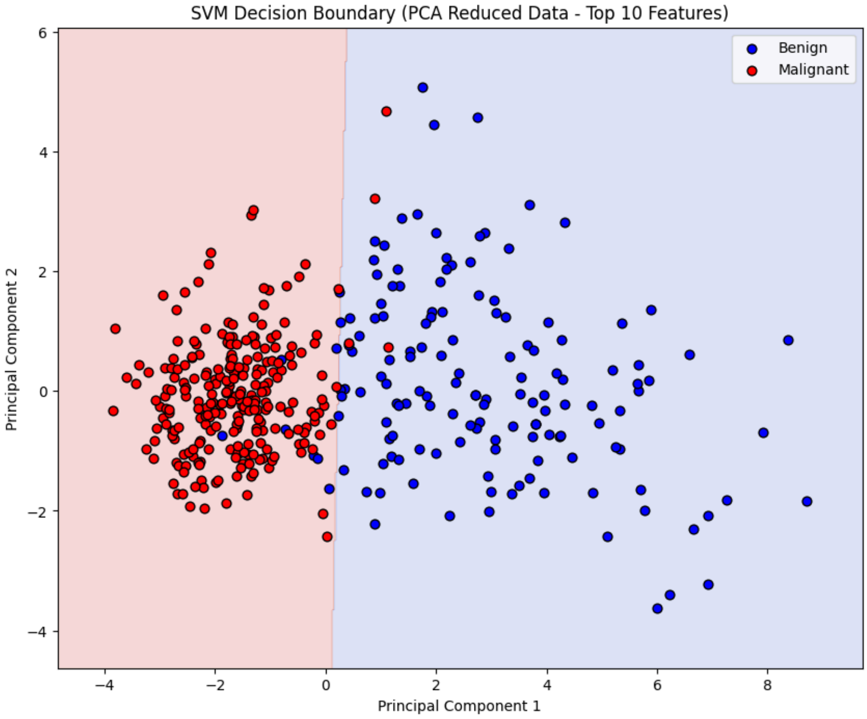
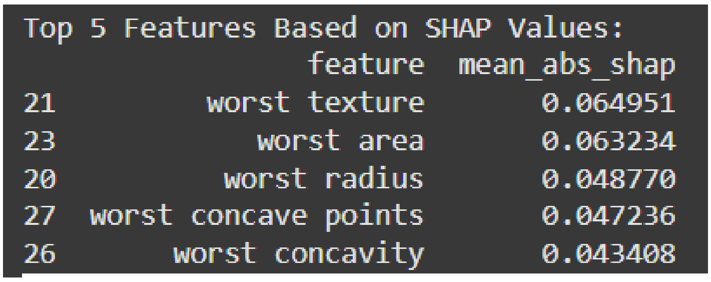


Figure 7 : Decision boundary using top 10 features

- **SVM decision boundary for top 5 features**

These are the top 5 features selected by calculating SHAP values.



```
Top 5 Features Based on SHAP Values:
                 feature  mean_abs_shap
21         worst texture       0.064951
23            worst area       0.063234
20          worst radius       0.048770
27  worst concave points       0.047236
26       worst concavity       0.043408
```

Figure 8 : Top 5 features selected by SHAP

Model accuracy is 96%. It has the highest malignant recall (98%) among reduced-feature models. But it has the lowest benign recall (92%).

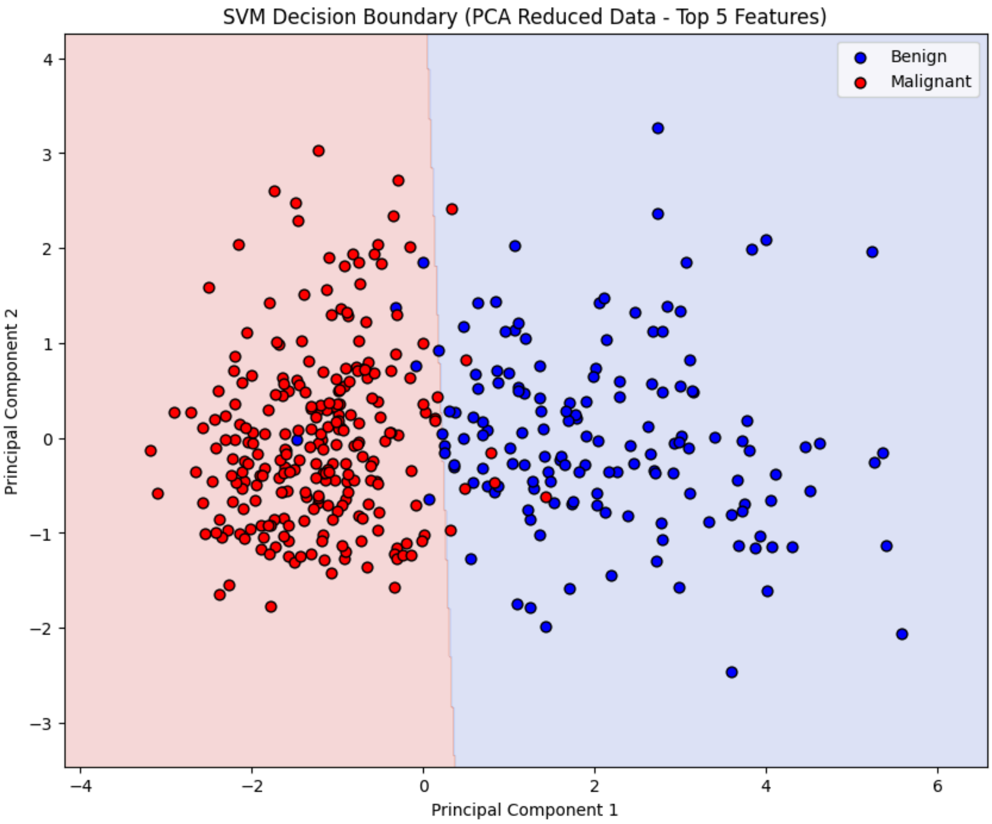
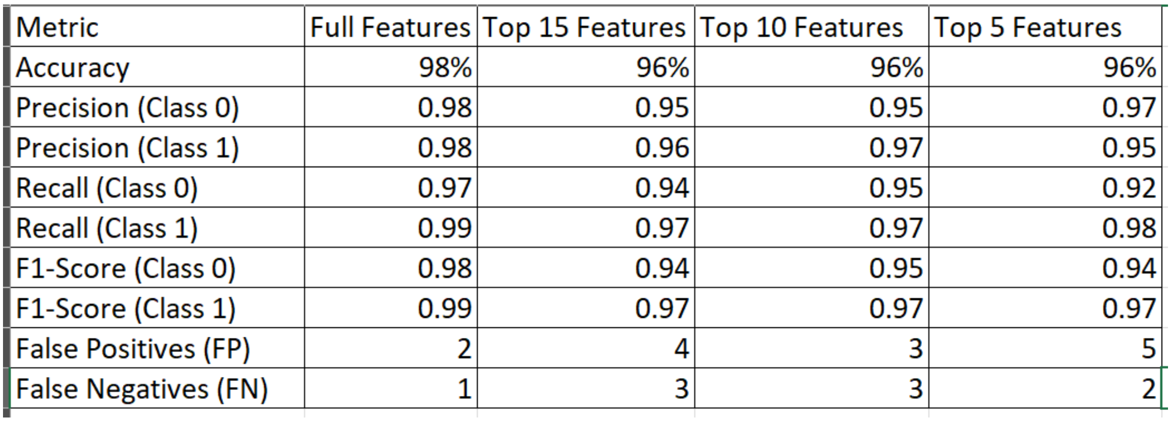


Figure 9 : Decision boundary using top 5 features

Result Summary



Metric	Full Features	Top 15 Features	Top 10 Features	Top 5 Features
Accuracy	98%	96%	96%	96%
Precision (Class 0)	0.98	0.95	0.95	0.97
Precision (Class 1)	0.98	0.96	0.97	0.95
Recall (Class 0)	0.97	0.94	0.95	0.92
Recall (Class 1)	0.99	0.97	0.97	0.98
F1-Score (Class 0)	0.98	0.94	0.95	0.94
F1-Score (Class 1)	0.99	0.97	0.97	0.97
False Positives (FP)	2	4	3	5
False Negatives (FN)	1	3	3	2

Table 1 : Feature accuracy

Class 0 - benign
Class 1 - malignant

All reduced-feature models (15, 10, 5) have approximately the same accuracy, 96%. Among the top 10 features, the model provides the best balance for most use cases.

Finally, we can identify that when reducing the number of features, we could not notice any significant drop in accuracy.

6. More details about SVM

SVM applies for 3 class problems

SVM is fundamentally binary classification but it can be extended to handle 3 class problems. We can use mainly 2 strategies to achieve this,

1. **One vs one -** Modeles are trained for every pair of classes and the final prediction is determined by majority voting.
2. **One vs rest** - Models are trained for a class against all others and resulting in three hyperplanes for a three-class problem.

Advantages and disadvantages of SVM

i) Advantages:

- High Accuracy in Small Datasets
  Example: Classifying tumor types (benign vs. malignant) with 100 patient records.

- Handles complex patterns effectively
  Example: In spam email detection

ii) Disadvantages

- Struggles with large datasets.
  Example: Classifying millions of social media posts for sentiment analysis.

- Overlaps in data degrade performance
  Example: In customer churn prediction with messy data

Computational Efficiency in Big Data

SVM does not perform well with big data because it takes a lot of time and computing power to process tons of samples. We can use decision trees or neural networks to handle big data better and faster.

SHAP Instead of Other Feature Selection Methods

- SHAP looks at all possible combinations of features, not just one at a time. Since it is more fair to use.

- We can perform more calculations in less time since SHAP calculates contributions more efficiently.
- SHAP catches tricky relationships in data but other methods don't have that capacity.

QP Problem in SVM

This math concept tries to maximize the margin between classes as big as possible. QP solves a math problem with a curvy (quadratic) formula that measures the margin size. It follows straight-line rules (linear constraints) to classify things correctly.

When we have big QP problems we can split them into smaller chunks using tricks like Sequential Minimal Optimization (SMO). This speeds things up instead of solving it all at once.

Example: For 1,000 pictures, SMO tackles a few at a time to find the line faster.

7. Task Breakdown

The project was divided into three main tasks,

Task	Team Member
Project Reporting & Documentation	S19408, S19420
Code Development, Continuation & Evaluation	S19205, S19355
Presentation Creation & Presenting	S19205, S19355

8. Conclusion

We successfully implemented an SVM model for classification using the Wisconsin dataset. By selecting different features, we trained various models for classification. For feature selection, we used the SHAP method. We trained an SVM classifier on the Wisconsin Breast Cancer dataset and achieved 98% accuracy. Then, we simplified the models using SHAP-selected features (15, 10, and 5), which resulted in 96% accuracy. Finally, we can say that feature reduction guided by SHAP preserves performance while improving efficiency.

9. References

- Vapnik, V. N. et al., *A Training Algorithm for Optimal Margin Classifier*
- MIT OpenCourseWare on SVM
- OpenCV Documentation: https://docs.opencv.org/4.x/d1/d73/tutorial_introduction_to_svm.html
- Kaggle Notebook on SVM: https://www.kaggle.com/code/pierra/credit-card-dataset-svm-classification