Experiment: Support Vector Machine (SVM) for Classification with Hyperparameter Tuning

Dataset: Breast Cancer Wisconsin (Diagnostic)

1. Dataset Source

The dataset used in this experiment is the **Breast Cancer Wisconsin (Diagnostic) Dataset**, available on Kaggle.

Dataset Link:
https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

This dataset was originally collected by researchers from the University of Wisconsin–Madison to classify tumors as malignant or benign.

2. Dataset Description

Overview

The Breast Cancer Wisconsin dataset is used for **binary classification**, where the goal is to predict whether a tumor is:

- Malignant (Cancerous) → M
- Benign (Non-cancerous) → B

Dataset Characteristics

Property	Value
Number of instances	569
Number of features	30
Target variable	diagnosis (M or B)
Problem type	Binary classification
Missing values	None

Feature Description

The features are computed from digitized images of breast mass cell nuclei.

Examples of features include:

- radius_mean – Mean distance from center to points on the perimeter
- texture_mean – Standard deviation of gray-scale values
- perimeter_mean – Perimeter of tumor
- area_mean – Area of tumor
- smoothness_mean – Local variation in radius lengths
- compactness_mean
- concavity_mean
- symmetry_mean
- fractal_dimension_mean

There are 30 numerical features in total.

Target Variable

Value	Meaning
M	Malignant
B	Benign

For machine learning, this is converted to:

- Malignant = 1
- Benign = 0

3. Mathematical Formulation of SVM

Support Vector Machine is a supervised learning algorithm used for classification. It finds the optimal hyperplane that separates classes with maximum margin.

Hyperplane Equation

A hyperplane is defined as:

$$w \cdot x + b = 0$$

Where:

- w = weight vector
- x = input feature vector
- b = bias

Classification Rule

$$y = \begin{cases} +1 & \text{if } w \cdot x + b \geq 0 \\ -1 & \text{if } w \cdot x + b < 0 \end{cases}$$

Margin Maximization

SVM maximizes margin:

$$Margin = \frac{2}{||w||}$$

Optimization objective:

$$\min \frac{1}{2} ||w||^2$$

Subject to:

$$y_i (w \cdot x_i + b) \geq 1$$

Soft Margin SVM (with regularization parameter C

$$\min \frac{1}{2} ||w||^2 + C \sum \xi_i$$

Where:

- C = Regularization parameter
- ξ = Error penalty

Kernel Trick

For non-linear data:

$$K(x_i, x_j)$$

Common kernels:

- Linear
- Polynomial
- Radial Basis Function (RBF)

RBF kernel:

$K(x_i,x_j)=e^{-\gamma||x_i-x_j||^2}$K(xi,xj)=e−γ||xi−xj||2

4. Algorithm Limitations

SVM has several limitations:

1. Sensitive to feature scaling

Requires normalization or standardization.

2. Slow for large datasets

Training complexity:

$O(n^2)$ to $O(n^3)$O(n2) to O(n3)

3. Difficult to interpret

Unlike decision trees, SVM is not easily explainable.

4. Sensitive to hyperparameters

Incorrect values of C or gamma reduce performance.

5. Not suitable for very large datasets

Memory consumption is high.

5. Methodology / Workflow

Step 1: Import Dataset

Load dataset using pandas.

Step 2: Data Preprocessing

- Remove unnecessary columns (id)
- Convert diagnosis to numeric
- Split features and target
- Train-test split (80% training, 20% testing)
- Feature scaling using StandardScaler

Step 3: Model Training

Train SVM using:

SVC()SVC()

Step 4: Hyperparameter Tuning

Use GridSearchCV to tune:

- C
- gamma
- kernel

Step 5: Model Testing

Predict test data:

*ypred*ypred

Step 6: Performance Evaluation

Calculate:

- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix

6. Performance Analysis

Performance is evaluated using classification metrics.

Accuracy

Accuracy=TP+TNTP+TN+FP+FNAccuracy = \frac{TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN

Measures overall correctness.

Precision

Precision=TPTP+FPPrecision = \frac{TP}{TP + FP}Precision=TP+FPTP

Measures correctness of positive predictions.

Recall

Recall=TPTP+FNRecall = \frac{TP}{TP + FN}Recall=TP+FNTP

Measures ability to detect positive cases.

F1 Score

F1=2×Precision×RecallPrecision+RecallF1 = \frac{2 \times Precision \times Recall}{Precision + Recall}F1=Precision+Recall2×Precision×Recall

Balanced measure.

Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	TP	FN
Actual Negative	FP	TN

Interpretation

High accuracy (>95%) indicates SVM performs well for breast cancer classification.

Low false negatives are especially important because missing cancer detection is critical.

7. Hyperparameter Tuning

Hyperparameter tuning improves model performance by selecting optimal parameters.

Tuned Parameters

Parameter	**Description**
C	Regularization parameter

gamma	Kernel coefficient
kernel	Type of kernel

GridSearchCV Method

Example parameter grid:

C=[0.1,1,10,100]C = [0.1, 1, 10, 100]C=[0.1,1,10,100] gamma=[1,0.1,0.01,0.001]gamma = [1, 0.1, 0.01, 0.001]gamma=[1,0.1,0.01,0.001] kernel=[′linear′,′rbf′]kernel = ['linear', 'rbf']kernel=[′linear′,′rbf′]

Process

GridSearchCV:

1. Tries all parameter combinations
2. Uses cross-validation
3. Selects best parameters

Impact of Hyperparameter Tuning

Before tuning:

Accuracy ≈ 94%

After tuning:

Accuracy ≈ 97–99%