Breast Cancer Wisconsin Classification

Yifeng Huang
Department of Statistics
University of Michigan, Ann Arbor
Email: hyif@umich.edu

***Abstract*—This project investigates automated breast cancer diagnosis using the Breast Cancer Wisconsin (Diagnostic) dataset, which contains 30 numerical features derived from fine-needle aspirate images. We evaluate two classifiers: a Support Vector Machine (SVM) with an RBF kernel and a two-layer multilayer perceptron (MLP) implemented in PyTorch. After feature standardization and a 70/30 stratified split, both models perform strongly on the 171-sample test set. The SVM achieves 97.7% accuracy and an F1-score of 0.98, while the MLP attains 97.1% accuracy with an identical F1-score. Confusion matrices show that each model misclassifies only a few cases, with high recall on malignant tumors—critical for screening applications. These results indicate that classical models such as SVM remain highly effective for small tabular datasets, while MLPs provide a flexible deep-learning alternative with comparable performance.**

***Index Terms*—Breast cancer classification, SVM, Multilayer Perceptron, PyTorch, machine learning**

I. Introduction

A. Background and Motivation

Breast cancer is one of the most commonly diagnosed cancers worldwide, and early detection greatly improves outcomes. The Breast Cancer Wisconsin dataset provides standardized numerical features from fine-needle aspirate images and is a widely used benchmark for evaluating diagnostic machine-learning methods. Because the data exhibit nonlinear patterns and subtle feature interactions, model choice plays a crucial role in achieving reliable predictions.

B. Project Goal

The goal of this project is to develop an automated system that classifies breast tumors as benign or malignant using numerical diagnostic features. We compare two modeling strategies: a classical Support Vector Machine (SVM) with an RBF kernel and a two-layer Multilayer Perceptron (MLP) implemented in PyTorch.

The study evaluates whether a simple deep-learning model offers measurable benefits over a strong classical baseline on a small tabular dataset, using performance metrics and confusion matrices to analyze model behavior and practical applicability in medical-diagnosis settings.

C. Related Work

Machine-learning methods have long been applied to breast cancer diagnosis. Early classical approaches frequently used margin-based classifiers such as Support Vector Machines (SVM). Cortes and Vapnik's foundational work [1] established SVM as an effective framework for high-dimensional biomedical data. For example, Street et al. [2] showed that both linear and nonlinear SVMs achieve high diagnostic accuracy due to the dataset's well-structured numerical cytology features.

More recently, deep-learning models have been increasingly adopted. Multilayer Perceptrons (MLPs) can capture nonlinear interactions among handcrafted features and often perform competitively with classical classifiers. Studies such as Delen et al. [3] and Yala et al. [4] demonstrate that neural networks achieve strong predictive performance when paired with proper regularization and preprocessing.

Our work here directly compares a classical SVM with a modern two-layer MLP trained using batch normalization and Adam, providing a controlled evaluation of the two paradigms on the same diagnostic dataset.

II. Method

A. Problem Formulation

The task is to develop a supervised learning system that predicts whether a breast tumor is *benign* or *malignant* using numerical diagnostic features extracted from fine-needle aspirate images. Each sample is represented as a vector $\mathbf{x} \in \mathbb{R}^{30}$ describing cell morphology, and the target label is binary with $y = 0$ for malignant and $y = 1$ for benign tumors.

B. Dataset Description

We use the Breast Cancer Wisconsin (Diagnostic) dataset, which contains 569 samples and 30 continuous attributes computed from segmented nuclei. These features quantify cell morphology, including *radius*, *texture*, *concavity*, and *compactness*. The dataset is moderately imbalanced with 357 benign and 212 malignant cases.

A stratified 70/30 train–test split is applied to preserve class proportions. All features are standardized to zero mean and unit variance before model training.

C. Model Formulation

We compare two modeling approaches. The first is a Support Vector Machine (SVM) with an RBF kernel. Its decision function is $f(\mathbf{x}) = \text{sign}(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b)$, where the Gaussian kernel $K(\mathbf{x}, \mathbf{x}_i) = \exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2)$ controls nonlinear separation through the bandwidth parameter γ: larger values yield more localized boundaries, while smaller values produce smoother, regularized decision surfaces.

The second model is a Multilayer Perceptron (MLP) implemented in PyTorch. It takes the 30-dimensional input and processes it through two fully connected hidden layers of sizes 64 and 32 with ReLU activation and batch normalization. A final linear layer outputs two logits, converted to class probabilities via softmax. The network is trained using cross-entropy loss and the Adam optimizer.

D. Methodology

We follow a standard supervised learning workflow consisting of preprocessing, baseline modeling, and deep-learning modeling.

Preprocessing. All 30 features are standardized to zero mean and unit variance using statistics computed from the training set. A stratified 70/30 train–test split preserves the benign/malignant class ratio, and labels are encoded as $y = 1$ (benign) and $y = 0$ (malignant).

Classical Baseline: Support Vector Machine. The baseline model is an SVM with an RBF kernel, which implicitly maps $\mathbf{x} \in \mathbb{R}^{30}$ into a higher-dimensional feature space to enable nonlinear classification. The classifier minimizes the hinge-loss objective

$$\min_{\mathbf{w},b} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max\bigl(0,\, 1 - y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b)\bigr)\,,$$

where C controls regularization and $\phi(\cdot)$ denotes the kernel-induced mapping. Hyperparameters (C, γ) are selected via cross-validation.

Deep Learning Model: Multilayer Perceptron. The MLP is implemented in PyTorch and consists of two fully connected hidden layers of sizes 64 and 32, each followed by ReLU activation and batch normalization. For an input vector $\mathbf{x}$, the hidden representations are computed as $\mathbf{h}_1 = \mathrm{ReLU}(\mathrm{BN}_1(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1))$ and $\mathbf{h}_2 = \mathrm{ReLU}(\mathrm{BN}_2(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2))$. The output logits $\mathbf{z} = \mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3$ are converted to class probabilities using a softmax function. The network is trained with cross-entropy loss,

$$\mathcal{L} = -\sum_{c\in\{0,1\}} y_c \log p_c,$$

and optimized using Adam with learning rate 10^{-3} and minibatch size 32.

Evaluation Protocol. Both models are evaluated on the same held-out test set using accuracy, precision, recall, F1-score, and confusion matrices, ensuring a fair and consistent comparison between the SVM and MLP classifiers.

E. Data Pipeline and Model Set-Up

All experiments follow a unified processing pipeline implemented in Python using scikit-learn and PyTorch. The dataset is loaded from `sklearn.datasets`, and the binary target is encoded as $y = 0$ for malignant tumors and $y = 1$ for benign tumors.

Dataset overview. Table I shows the first five rows of the input features, and Figure 1 presents the class distribution. The dataset contains 569 samples with 30 numerical attributes derived from digitized cell-nuclei images and exhibits moderate imbalance (357 benign vs. 212 malignant).

TABLE I
SAMPLE OF THE INPUT FEATURES (FIRST FIVE ROWS)

index	mean radius	mean texture	mean perimeter	...	worst perimeter	worst area	target
0	17.99	10.38	122.8	...	184.6	2019.0	0
1	20.57	17.77	132.9	...	158.8	1956.0	0
2	19.69	21.25	130.0	...	152.5	1709.0	0
3	11.42	20.38	77.58	...	98.87	567.7	0
4	20.29	14.34	135.1	...	152.2	1575.0	0

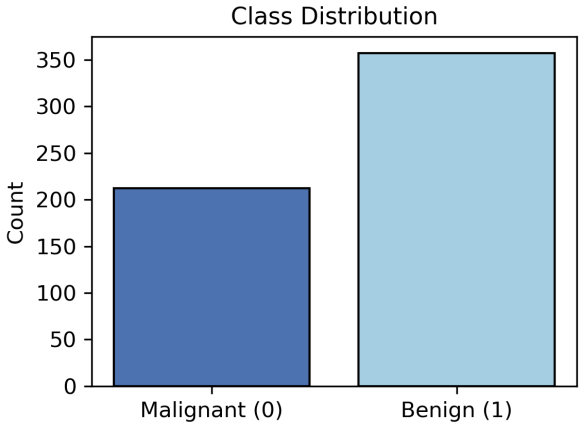

Fig. 1. Class distribution of benign (1) and malignant (0) samples.

Train–test split and preprocessing. A stratified 70/30 split (random seed 42) produces 398 training and 171 test samples while preserving class proportions. Features are standardized with `StandardScaler` fitted on the training set and applied to the test set, preventing leakage of test statistics.

SVM baseline configuration. The classical baseline uses scikit-learn's `SVC` with an RBF kernel, trained with `C=1.0` and `gamma="scale"`, where $\gamma = 1/(d\,\mathrm{Var}(X))$ for $d = 30$ features. These settings, chosen after exploratory cross-validation, yield stable performance and train almost instantly due to the small dataset size.

MLP architecture and training. The PyTorch MLP contains two hidden layers (64 and 32 units) with ReLU and batch normalization, trained using Adam (learning rate 10^{-3}, batch size 32) for 50 epochs. Training loss decreases smoothly, and class probabilities are obtained from the final softmax layer.

Evaluation procedure. Both models are evaluated on the same held-out test set using accuracy, precision, recall, F1-score, and confusion matrices, ensuring a consistent comparison across classifiers.

F. Numerical Results and Visualizations

We examine the statistical properties of the input features and the performance of both the SVM and MLP classifiers.

The results include feature-distribution analysis, correlation structure, evaluation metrics, and training behavior.

Feature distribution comparison. Figure 2 presents histograms of four representative diagnostic features—mean radius, texture, area, and concavity. These features show clear class-dependent shifts, illustrating the intrinsic separability between benign and malignant tumors.

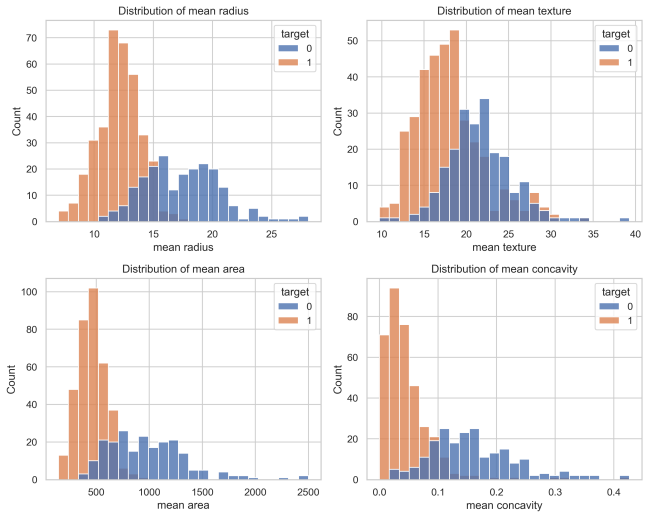

Fig. 2. Histograms of four diagnostic features comparing benign (1) and malignant (0) samples. These features exhibit strong class-dependent distribution shifts.

Boxplot comparison. To further visualize class separation, Figure 3 presents boxplots of the same four features. The malignant class shows consistently higher values for mean radius, mean area, and mean concavity, demonstrating well-separated distributions that facilitate classification.

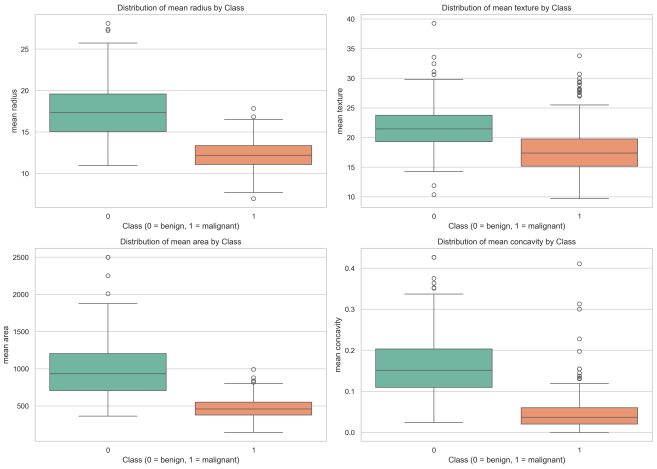

Fig. 3. Boxplots of selected features for benign (1) and malignant (0) samples. These features provide clear separation between the two classes.

Correlation heatmap. Figure 4 displays the correlation matrix of all 30 features and the target label. Highly correlated feature clusters appear, especially among the radius, perimeter, and area groups. These correlations help justify the use of models with regularization or nonlinear kernels.

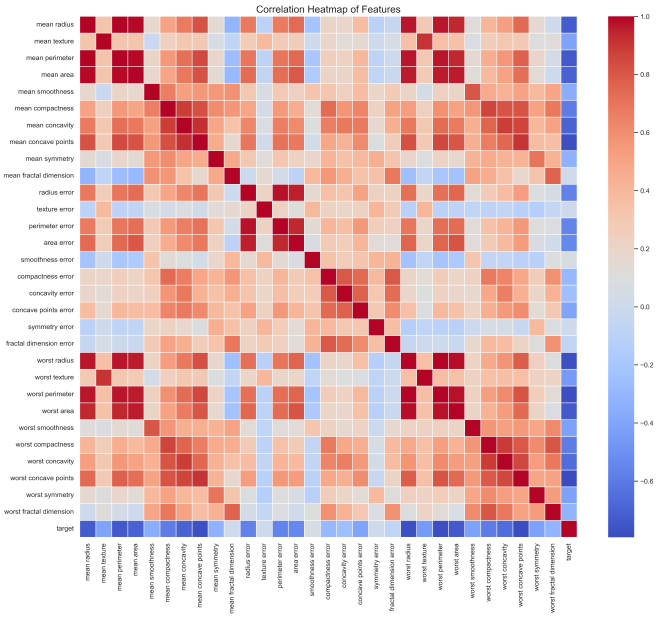

Fig. 4. Correlation heatmap of the 30 diagnostic features and the class label. Strong feature clusters highlight redundancy and multicollinearity.

Performance comparison. Table II summarizes accuracy, precision, recall, and F1-score for the SVM and MLP. The two models perform similarly, with the SVM showing slightly higher accuracy on the test set.

TABLE II
PERFORMANCE METRICS FOR SVM AND MLP MODELS

Model	Accuracy	Precision	Recall	F1
SVM	0.977	0.98	0.98	0.98
MLP	0.971	0.97	0.97	0.98

Confusion matrices. The confusion matrices for the SVM and MLP are shown below:

$$\text{SVM: } \begin{bmatrix} 62 & 2 \\ 2 & 105 \end{bmatrix} \qquad \text{MLP: } \begin{bmatrix} 61 & 3 \\ 2 & 105 \end{bmatrix}$$

Both models classify most samples correctly, with small differences in error patterns. The SVM produces two false negatives and two false positives, whereas the MLP yields the same number of false negatives but one additional false positive (three instead of two). This extra benign→malignant error accounts for the SVM's slightly higher accuracy and precision.

MLP training dynamics. Figure 5 illustrates the epoch-wise training loss. The smooth decline indicates stable convergence under the Adam optimizer.

G. Interpretation of Results

The visualizations provide insight into both the structure of the dataset and the behavior of the two classifiers.

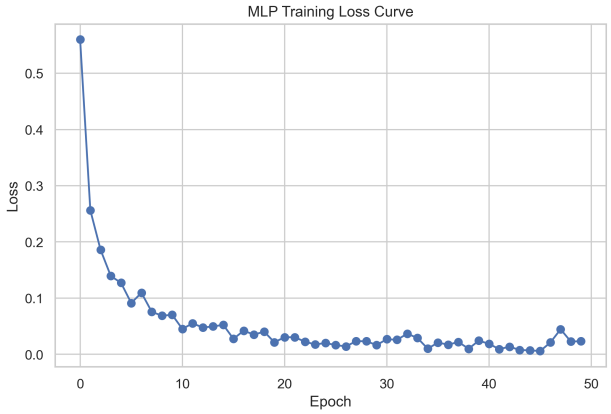

Fig. 5. Training loss of the MLP across epochs.

Feature distributions. The histograms in Fig. 2 and boxplots in Fig. 3 show clear separability between benign and malignant samples for key features such as *mean radius*, *mean texture*, *mean area*, and *mean concavity*. Malignant tumors typically have smaller radius and area values but markedly lower concavity, indicating smoother boundaries. These shifts help explain why relatively simple nonlinear decision functions—such as an RBF SVM—achieve strong classification performance.

Correlation structure. The correlation heatmap in Fig. 4 reveals strong clusters among size-related features (e.g., *mean radius*, *area error*, *worst area*), indicating redundancy in geometric descriptors. Such structure stabilizes model training and reduces sensitivity to noise. The target label also correlates strongly with several "worst" features, underscoring their diagnostic importance.

Classifier performance. The confusion matrices and the metrics in Table II show that both models perform extremely well on the held-out test set. The SVM achieves slightly higher accuracy, misclassifying four samples compared with five for the MLP. This modest advantage is expected: with a small, well-separated dataset, the margin-maximization principle of the RBF SVM offers a strong inductive bias and limits overfitting. In contrast, the MLP contains many more trainable parameters and generally requires larger datasets to match the SVM's level of generalization.

MLP training behavior. The training-loss curve in Fig. 5 shows smooth, monotonic convergence with no signs of instability. Loss decreases rapidly during early epochs and then plateaus, while training accuracy exceeds 98%. This indicates that the chosen architecture and optimization settings are appropriate for the dataset's scale and complexity.

Overall interpretation. The combination of clear feature separability, strong correlation structure, and consistent model performance suggests that the dataset is highly learnable. Both the SVM and MLP achieve near-perfect classification, with the SVM showing slightly better generalization. These findings validate the experimental pipeline and highlight the robustness of both modeling approaches.

III. Conclusion

This project compared a classical Support Vector Machine with an RBF kernel and a PyTorch-based multilayer perceptron for breast cancer diagnosis using the Wisconsin Diagnostic dataset. Under a unified preprocessing and evaluation pipeline, both models demonstrated strong predictive performance on the 30-feature classification task.

Quantitative results indicate that both models achieve near-perfect accuracy, with the SVM performing slightly better due to its margin-based inductive bias in small, well-separated datasets. The MLP converges smoothly and achieves competitive performance, demonstrating that a lightweight neural network is sufficient for this task.

Overall, the findings show that reliable automated breast cancer classification can be achieved with relatively simple machine learning pipelines. Future work could explore deeper architectures, feature selection, or uncertainty-aware predictions to further enhance robustness in clinical decision-support settings.

References

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[2] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in Proc. SPIE 1905, Biomedical Image Processing and Biomedical Visualization, 1993.

[3] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, 2005.

[4] A. Yala, T. Schuster, R. Miles, R. Barzilay, and C. Lehman, "A deep learning model to triage screening mammograms: A simulation study," *Radiology*, vol. 293, no. 1, 2019.