Pattern Recognition Project Report

Breast Cancer Detection using KNN and SVM

A Comparative Study on the Wisconsin Diagnostic Dataset

Yasir Alalawi & Ajab Aldosari
447540166 & 447540089

Abstract

This report compares two supervised pattern recognition algorithms — K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) — applied to breast cancer diagnosis on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset (569 samples, 30 features). After feature standardization and a stratified 80/20 train-test split, both classifiers were tuned with 5-fold cross-validation. KNN reached 97.37% test accuracy with k = 8, while SVM with a linear kernel and C = 0.1 reached 98.25% test accuracy. SVM outperformed KNN on accuracy, malignant recall, malignant F1-score, and malignant AUC. Most importantly, SVM missed only 1 malignant case, while KNN missed 2, making SVM slightly preferable for this medical classification task. Overall, the results indicate that classical, well-tuned algorithms can reach strong benchmark performance on the WDBC dataset.

1. Introduction

Breast cancer is one of the most common cancers worldwide, and early detection is strongly associated with better survival outcomes. Manual cytology/pathology assessment, however, is time-consuming and subject to inter-observer variability. Computer-aided diagnosis (CAD) systems built on pattern recognition techniques offer a consistent, scalable, and objective second opinion that can support clinicians in their decision-making process.

This project investigates two classical pattern recognition algorithms taught in the course — K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) — and compares their ability to discriminate between malignant and benign tumors based on quantitative features extracted from breast biopsy images. The objectives are: (1) to design a reproducible classification pipeline including preprocessing and hyperparameter tuning; (2) to evaluate both algorithms using accuracy, malignant precision, malignant recall, malignant F1-score, specificity, and malignant AUC; and (3) to discuss the trade-offs between the two methods in the context of a medical screening problem where false negatives are particularly costly.

2. Dataset Description

Data Source: This study uses the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, a publicly available, peer-reviewed clinical dataset created by Dr. William H. Wolberg and colleagues at the University of Wisconsin Hospitals, Madison (1995). The dataset is hosted on the UCI Machine Learning Repository (archive.ics.uci.edu/dataset/17) and is also distributed with the scikit-learn library, where it was loaded for this project. All 569 samples are real anonymized patient records — no synthetic or generated data was used at any stage of this work.

Each sample corresponds to a digitized fine-needle aspirate (FNA) image of a breast mass. From every cell nucleus visible in the image, ten real-valued features are computed (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal

dimension), and three summary statistics — mean, standard error, and worst (largest) value — are reported, yielding 30 features per sample.

The dataset contains 569 samples described by 30 continuous numerical attributes and split into two classes: 357 benign tumors (62.7%) and 212 malignant tumors (37.3%), with no missing values.

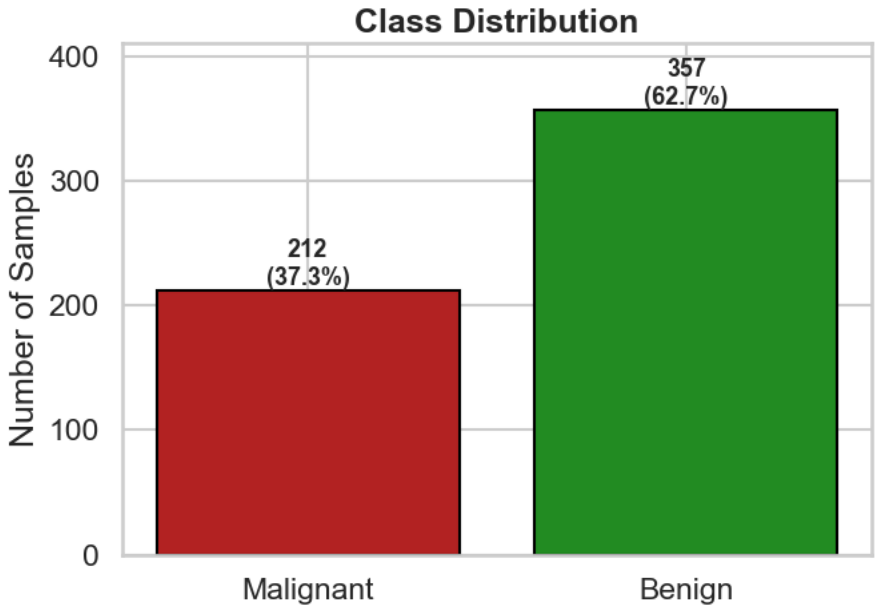


Figure 1. Class distribution of the WDBC dataset.

The class distribution is moderately imbalanced but not severely so, which means standard stratified evaluation protocols remain appropriate without explicit resampling. Several features exhibit strong pairwise correlations — for example, mean radius, mean perimeter and mean area form a near-collinear cluster, since they all describe the size of the nuclei. This redundancy motivates the use of feature scaling to ensure that distance-based and margin-based methods do not over-weight large-magnitude attributes.

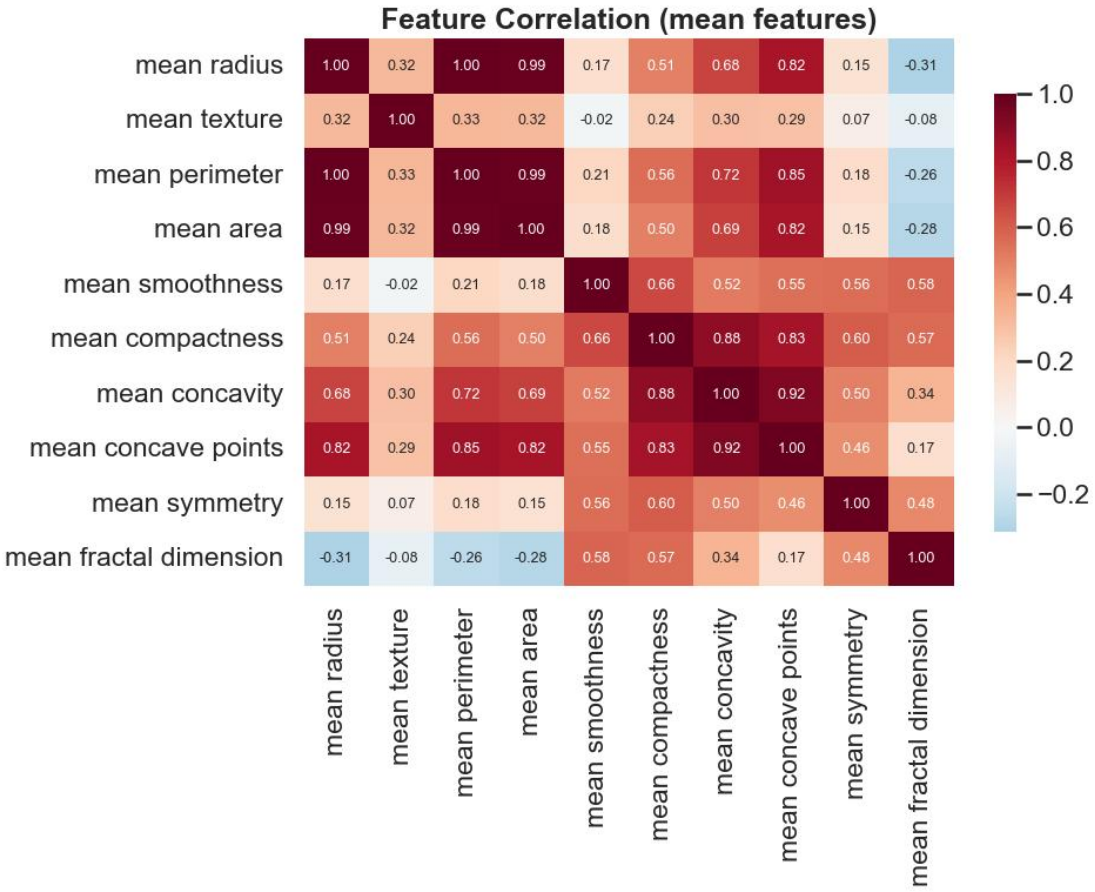


Figure 2. Pearson correlation among the ten mean-value features.

3. Methodology

3.1 Pipeline

The full classification pipeline consists of: (i) loading the WDBC dataset, (ii) splitting into training and test sets using a stratified 80/20 split, (iii) standardizing each feature to zero mean and unit variance using a StandardScaler fitted only on the training set, (iv) tuning model hyperparameters with 5-fold stratified cross-validation on the training set, (v) refitting the best configuration on the full training set, and (vi) evaluating on the held-out test set. Standardization is essential because both KNN (Euclidean distance) and SVM with an RBF kernel are sensitive to differences in feature scale.

3.2 K-Nearest Neighbors (KNN)

KNN is a non-parametric, instance-based (lazy) learning algorithm. To classify a new sample, the algorithm finds the k closest training points in feature space — typically using Euclidean distance — and assigns the majority class among those neighbors. Because there is no explicit training phase, KNN stores the entire training set and defers all computation to inference time. The main hyperparameter, k, controls the bias–variance trade-off: small values of k produce flexible, low-bias decision boundaries that may overfit; large values yield smoother but potentially under-fit boundaries. We searched $k \in \{1, 2, \ldots, 30\}$ and selected the value that maximized 5-fold cross-validation accuracy on the training set.

3.3 Support Vector Machine (SVM)

SVM is a discriminative, margin-based classifier that finds the hyperplane separating the two classes with the largest possible margin. For non-linearly separable data, SVM uses the kernel trick to

implicitly map inputs into a higher-dimensional space where a linear separator may exist. The regularization parameter C controls the penalty for misclassified training points (larger C → harder margin → higher variance), while the kernel and its associated parameters (e.g., γ for RBF) control the shape of the decision boundary. We performed an exhaustive grid search over C ∈ {0.1, 1, 10, 100}, γ ∈ {scale, 0.001, 0.01, 0.1}, and kernel ∈ {linear, RBF}, scoring on 5-fold cross-validated accuracy.

3.4 Evaluation Metrics

We report six metrics. Accuracy is the overall proportion of correct predictions. Because this is a cancer-detection task, malignant tumors are treated as the clinically important positive class. Malignant precision measures how reliable malignant predictions are, while malignant recall/sensitivity measures the model's ability to detect actual malignant cases. Specificity measures how well the model correctly identifies benign cases. F1-score is the harmonic mean of malignant precision and malignant recall. Malignant AUC summarizes the ROC curve for malignant detection and reflects the classifier's ability to separate malignant from benign samples across decision thresholds. In this clinical context, malignant recall is especially important because missing a malignant tumor is more harmful than flagging a benign case for further review.

4. Results

4.1 KNN Tuning

Cross-validation accuracy peaked at k = 8, with a CV score of 97.14%. As shown in Figure 3, performance is fairly stable in the range k = 5 to k = 12; very small values of k introduce noise sensitivity, while very large values of k start to lose discriminative power.

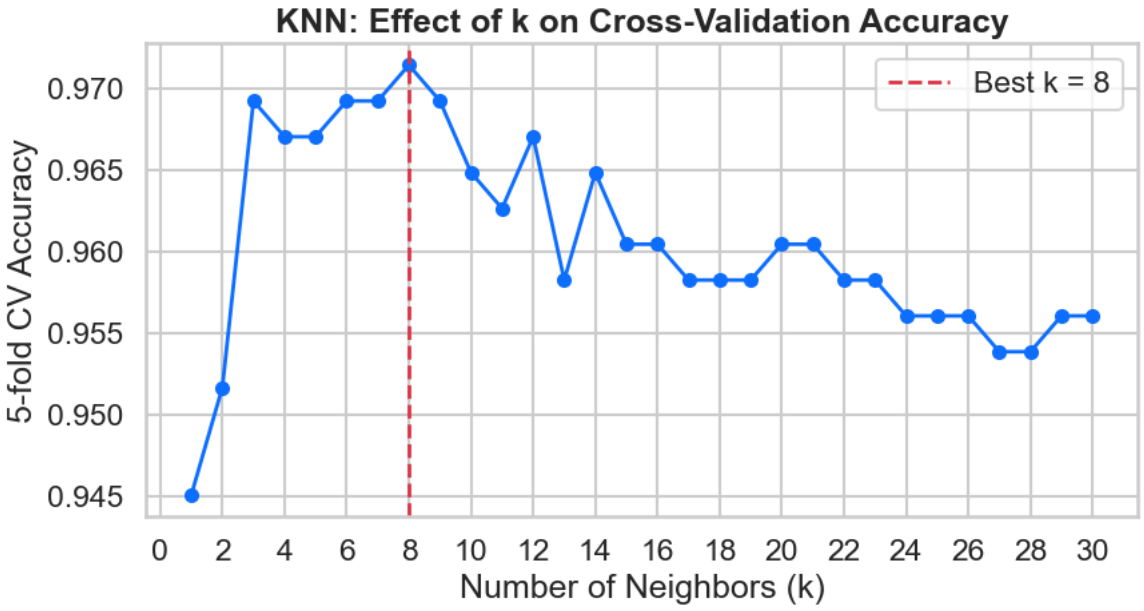


Figure 3. KNN cross-validation accuracy as a function of k.

4.2 SVM Tuning

Grid search selected a linear kernel with C = 0.1, achieving 98.02% cross-validation accuracy. The fact that a linear SVM with low C (i.e., a soft margin) is optimal indicates that, after standardization, the two classes are largely linearly separable in the original 30-dimensional space, and that allowing some training-set misclassification produces a more robust decision boundary.

4.3 Test-Set Performance

Metric	KNN (k = 8)	SVM (linear, C = 0.1)	Winner
Accuracy	0.9737	0.9825	**SVM**
Malignant Precision	0.9756	0.9762	**SVM**
Malignant Recall	0.9524	0.9762	**SVM**
Malignant F1-Score	0.9639	0.9762	**SVM**
Specificity	0.9861	0.9861	**Tie**
Malignant AUC	0.9909	0.9937	**SVM**

Table 1. Test-set performance comparison.

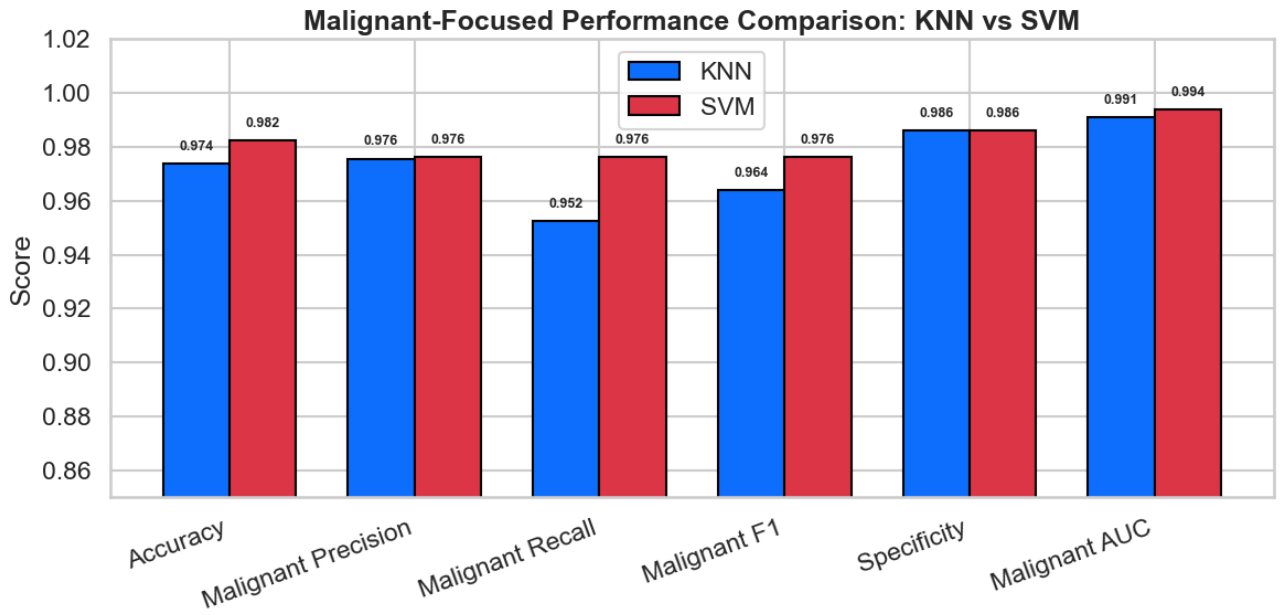


Figure 4. Side-by-side performance comparison of KNN and SVM.

4.4 Confusion Matrices and ROC Curves

On the 114-sample test set, KNN correctly detected 40 malignant cases and 71 benign cases. It missed 2 malignant cases by predicting them as benign, and it incorrectly flagged 1 benign case as malignant. SVM correctly detected 41 malignant cases and 71 benign cases. It missed 1 malignant case and incorrectly flagged 1 benign case as malignant. Therefore, SVM is preferable in this clinical context because it reduces missed malignant diagnoses while maintaining the same number of false alarms.

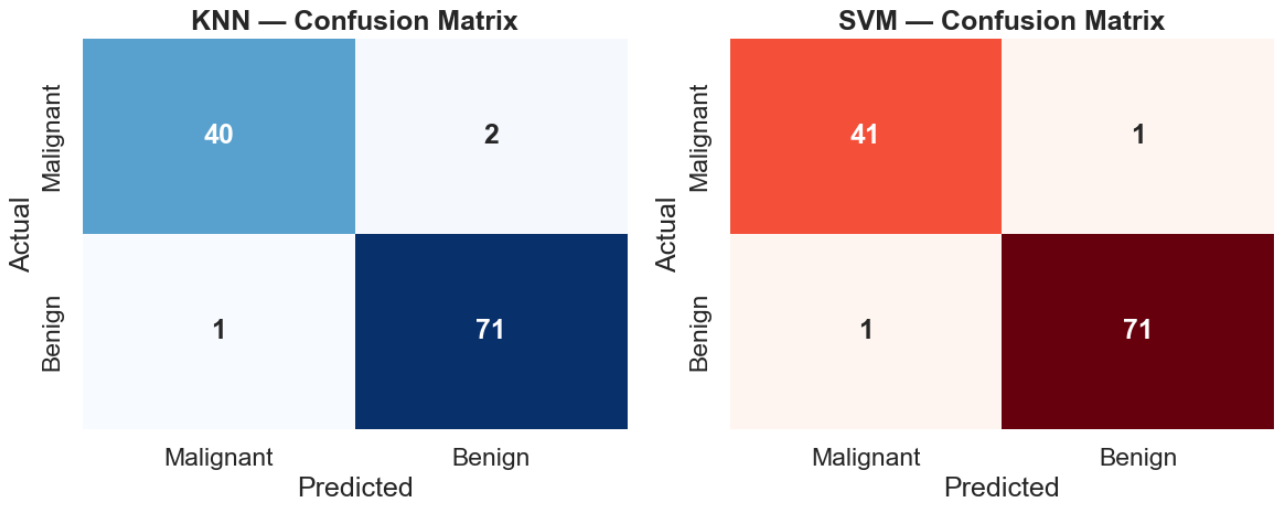


Figure 5. Confusion matrices for KNN (left) and SVM (right).

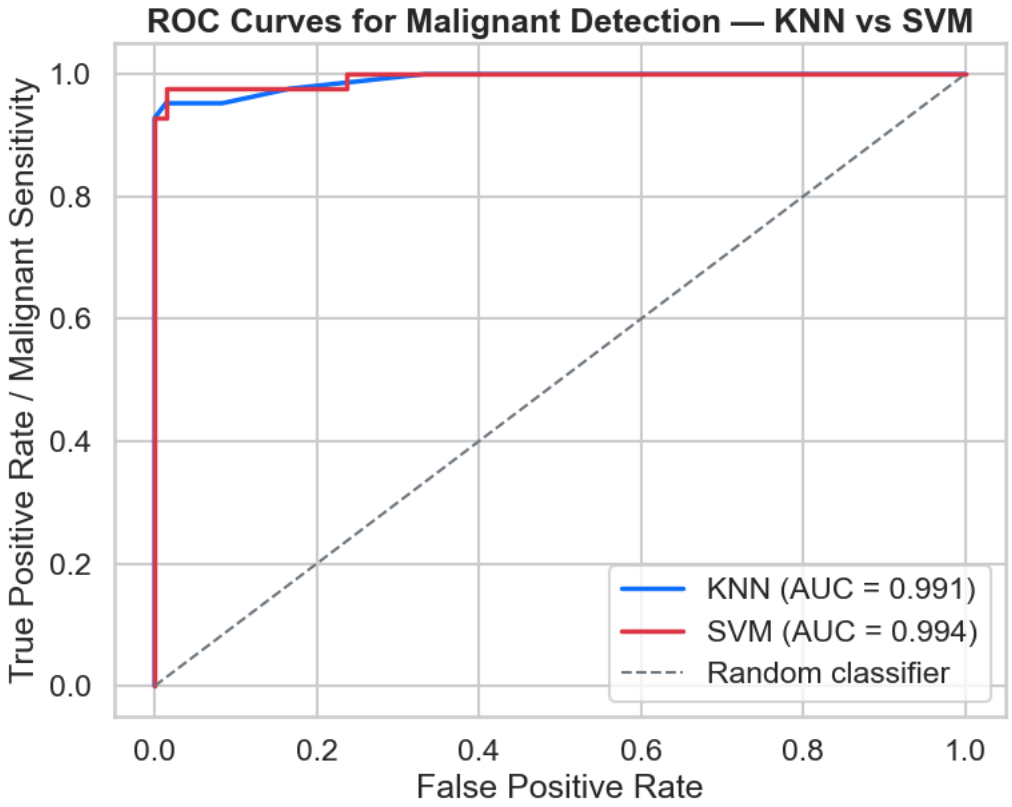


Figure 6. ROC curves for KNN (AUC = 0.991) and SVM (AUC = 0.994).

5. Discussion

Both classifiers achieved strong performance on the WDBC dataset, with most metrics above 97% and malignant AUC above 0.99 for both models. This confirms that the WDBC features are highly discriminative once standardized. SVM held a small but consistent advantage. Three factors plausibly explain this: (1) SVM optimizes a global, margin-based criterion, while KNN makes local decisions that can be perturbed by individual noisy training points; (2) the linear kernel matches the underlying geometry of the data, as suggested by the strong linear correlations between many of the size-related features; and (3) SVM's regularization parameter C provides a principled way to balance margin width against training error, whereas KNN's only flexibility is the choice of k. The two methods

also differ in computational cost: KNN has zero training cost but its inference cost grows linearly with the training set, since every prediction requires distances to all stored examples; SVM is more expensive to train but produces compact models that depend only on the support vectors, leading to fast and predictable inference — an attractive property for a real-time clinical CAD system.

6. Conclusion

We applied two classical pattern recognition algorithms — KNN and SVM — to breast cancer diagnosis on the Wisconsin Diagnostic Breast Cancer dataset. Both methods produced strong results, with KNN reaching 97.37% test accuracy at k = 8 and SVM reaching 98.25% test accuracy with a linear kernel. SVM outperformed KNN on accuracy, malignant recall, malignant F1-score, and malignant AUC. Clinically, the most important difference is that SVM missed 1 malignant case, while KNN missed 2. The combination of robust margin-based optimization and feature standardization made SVM the better choice for this binary medical classification task. Future work could extend the comparison to ensemble methods such as Random Forests and gradient-boosted trees, explore feature-selection or dimensionality-reduction stages to reduce the 30-feature input space, and validate the pipeline on independent external cohorts to assess its generalization beyond the WDBC benchmark.

References

[1] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Wisconsin Diagnostic Breast Cancer (WDBC). UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[2] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.

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

[4] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[5] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern Classification (2nd ed.). Wiley-Interscience.