Exploring Morphological Signatures of Malignancy: A Nested Cross-Validation Approach to Breast Cancer Classification

Kotsovilis Yiannos

Abstract

Breast cancer remains a global health burden and the most frequently diagnosed cancer among women, emphasizing the importance of accurate and accessible diagnostic tools. This study explores the application of machine learning to classify breast tumors as benign or malignant using morphological features extracted from fine needle aspirate (FNA) images. A repeated nested cross-validation (rnCV) strategy was employed to compare six classification algorithms and ensure reliable model selection with unbiased performance estimates. Models were evaluated using metrics aligned with binary classification tasks relevance, Matthews Correlation Coefficient (MCC). Among the evaluated models, Logistic Regression (LR) consistently delivered top-tier performance, achieving high MCC scores while demonstrating competitive consistency over multiple folds and repetitions. Notably, LR outperformed more complex models such as LightGBM and Random Forest, reinforcing the value of well-tuned simpler models in structured biomedical settings. Limitations include the relatively small sample size (512 instances) and mild class imbalance. These findings demonstrate that with proper tuning and validation, simpler models can achieve great performance in structured biomedical classification tasks, offering a promising direction for the integration of machine learning in clinical diagnostic workflows.

Introduction

Breast cancer is the most common cancer among women worldwide and remains a leading cause of cancer-related mortality [1]. According to the World Health Organization (WHO), in 2020 alone, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally. Early detection and accurate diagnosis are essential to improving breast cancer outcomes. Current diagnostic pathways typically involve clinical evaluation, imaging, and tissue sampling procedures, which may be limited by access, cost, or variability in clinical interpretation. In this context, there is a growing interest in leveraging computational methods, particularly machine learning, to support and enhance diagnostic accuracy using minimally invasive data sources [2].

The dataset analyzed in this study consists of morphological features extracted from digitized images of fine needle aspirates (FNA) of breast masses. Each tumor sample is represented by thirty numerical features describing various structural properties of cell nuclei, such as radius, texture, perimeter, area, and fractal dimension. These measurements provide a rich, quantitative representation of tissue abnormalities that can potentially be exploited for automatic classification into benign and malignant categories.

This report presents a machine learning pipeline designed to systematically evaluate and compare multiple classification algorithms for breast cancer diagnosis. The pipeline employs a repeated nested cross-validation (rnCV) framework, allowing for unbiased estimation of

generalization performance and systematic hyperparameter tuning through Optuna optimization. Special emphasis is placed on selecting the best-performing model based on several metrics, including the Matthews Correlation Coefficient (MCC). The goal is to identify a predictive model that balances accuracy, interpretability, and stability, contributing to the broader effort of integrating machine learning tools into biomedical diagnostic workflows.

Materials and Methods

Repeated nCV Pipeline Design

Figure 1 presents a conceptual overview of the repeated nested cross-validation (rnCV) framework. The design separates model evaluation into two nested loops: an outer loop used for estimating generalization performance and an inner loop responsible for hyperparameter optimization. The input consists of a 512×30 feature matrix and a list of candidate untrained classifiers.

Each outer loop iteration begins by splitting the dataset into training and testing subsets. Within the training subset, the inner loop performs cross-validation by splitting again into training and validation folds. In each inner fold, the pipeline applies preprocessing (median imputation, standard scaling, and PCA), followed by model training and evaluation. Optuna-based hyperparameter tuning ensures optimal model configuration is selected per repetition.

Once the inner loop concludes, the best hyperparameters are used to train a new pipeline on the full outer training data. This model is then evaluated on the outer test fold to obtain unbiased performance estimates. Metrics are stored after each outer fold, and the process is repeated for R rounds to ensure statistical reliability. The final model is trained using optimal parameters and performance is averaged across all outer loops.

This structure ensures strict separation between training and evaluation stages, effectively mitigating the risk of data leakage and promoting a reliable estimate of model generalization.

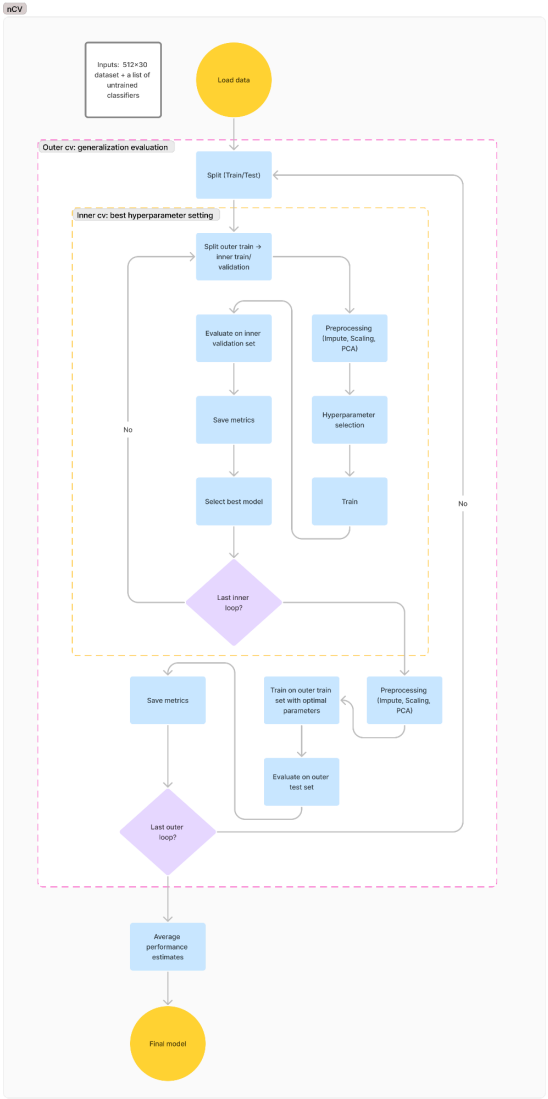

Figure 1: Abstraction of one complete iteration of the repeated nested cross-validation (rnCV) pipeline, illustrating the separation of outer and inner cross-validation loops.

Dataset Description

The dataset comprises 512 samples in total. Each sample is labeled with a binary classification: benign (B) or malignant (M). The target variable diagnosis is supported by 30 numerical features, grouped across ten primary morphological properties of the cell nuclei: radius, texture, perimeter, area, smoothness,

compactness, concavity, concave points, symmetry, and fractal dimension. Each of these properties is further described in three forms: mean, standard error (SE), and worst. An additional id column is used solely as a unique identifier.

Feature	Type
diagnosis	object
radius_mean	float
texture_mean	float
perimeter_mean	float
area_mean	float
…	…

Table 1: Summary of selected feature names and their corresponding data types.

For brevity, only a subset of features is shown in Table 1. Each numerical predictor is represented as floating-point values. These features were omitted from the table to maintain clarity and conserve space, but all were retained and analyzed in the exploratory phase.

Dataset Exploration

Initial inspection revealed no duplicate rows in the dataset. A small number of missing values were detected in several numerical features, for example concavity_mean, radius_mean and concave_points_mean.

Additionally, a check for zero values was conducted to identify potential placeholder values that could interfere with model learning. Notably, some features — such as concavity_mean, concave_points_mean, and their SE and worst counterparts — contain a dozen or more zeros.

The diagnosis distribution reveals a mild class imbalance, with 63% of samples labeled as benign and 37% as malignant.

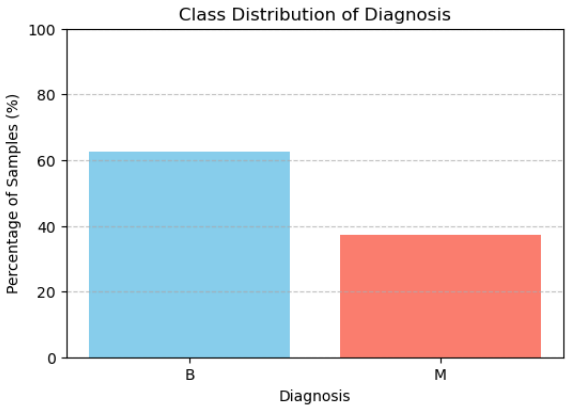

Figure 2: Class distribution of the target variable "diagnosis".

The majority of features exhibit right-skewed distributions (Figure 4, page 8). This skewness suggests the presence of outliers or non-Gaussian distributions, which may adversely affect some algorithms if not properly normalized. A few features such as texture_mean, symmetry_mean, and fractal_dimension_mean approximate more symmetric, bell-shaped curves.

Outliers were systematically assessed using boxplots, which reveal a number of extreme values in standardized space (Figure 5, page 8). Features such as perimeter_se, area_se, and concavity_se display particularly heavy tails, with multiple points falling well beyond the ±3 standard deviation range.

Correlation Analysis

To understand the relationships between the 30 numerical features, a correlation heatmap was created (Figure 6, page 9). In this heatmap, dark red areas represent strong positive correlations (i.e., when one feature increases, the other tends to increase too), while dark blue represents strong negative correlations.

One of the most noticeable observations is the high correlation between radius_mean, perimeter_mean, and area_mean. This makes sense from a biological perspective, as larger tumors will naturally have a bigger radius, a longer perimeter, and occupy more area [3]. These features may essentially capture the same structural aspect of the tumor — size —

suggesting some redundancy in the dataset. Recognizing this helps us later on when we want to reduce the dimensionality of the data or choose a subset of features for modeling.

To examine these patterns more closely, scatter plots were also used. These plots give us a more intuitive feel for how two features behave in relation to each other (Figure 7, page 9). For example:

- radius_mean and perimeter_mean showed a strong linear relationship, confirming what we saw in the heatmap.
- radius_mean and area_mean had a nonlinear but clear increasing trend, suggesting a more complex mathematical relationship.
- radius_mean and radius_se showed a weaker, scattered relationship, suggesting they capture different aspects of tumor structure — the former representing the typical cell size, and the latter reflecting how much the sizes vary within the sample. This variability could be linked to structural irregularities often associated with malignancy.

Principal Component Analysis (PCA)

Once we understood which features were strongly related, we applied Principal Component Analysis (PCA) to explore whether we could reduce the dataset's dimensionality without losing too much information.

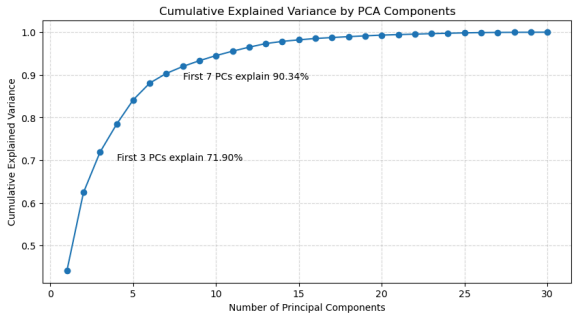

Figure 3: Cumulative explained variance from PCA.

The cumulative explained variance plot shows how much of the dataset's total variance is captured by the first few components. Here, we observed that:

- The first 3 principal components (PCs) explain about 71.9% of the total variance.
- The first 7 components capture over 90% of the variance.

This tells us that although we started with 30 features, most of the information can be preserved by using just a handful of components. This will prove useful in order to reduce noise, improve model speed, or visualize the data in fewer dimensions.

To visualize this idea in practice, we created a 3D scatter plot using the first three principal components (Figure 8, page 10). Even though PCA is unsupervised (it doesn't use the diagnosis labels), the plot showed that samples from benign and malignant classes tend to separate quite well in this reduced space. This is a strong indication that the features we have do in fact carry meaningful information for classification — and that dimensionality reduction could help in modeling without sacrificing performance.

Pipeline Structure

To ensure robust model evaluation and prevent information leakage, a Repeated Nested Cross-Validation (rnCV) strategy was implemented. This approach combines an outer cross-validation loop for unbiased generalization assessment and an inner cross-validation loop for hyperparameter tuning. The pipeline was designed to systematically tune multiple classifiers using Optuna optimization, and to evaluate model performance across several critical metrics.

Each iteration of the pipeline follows a structured sequence of transformations and evaluations. First, missing values are handled using median imputation, which was selected for its robustness to outliers and skewed distributions commonly observed in biological data. Next, all features are standardized to zero mean and unit variance using StandardScaler, a

crucial step to ensure comparability across features measured on different scales. Following scaling, Principal Component Analysis (PCA) is applied, retaining enough components to explain 95% of the variance. PCA serves to reduce feature redundancy and mitigate multicollinearity identified during exploratory analysis.

During the outer loop, the dataset is split into N=5 stratified folds, ensuring that the class imbalance observed in the original dataset is maintained within each fold. For each outer training split, an inner cross-validation (K=3 folds) is performed, where an Optuna study explores the hyperparameter space of each candidate estimator. The objective function used for tuning is the mean F1 score across the inner folds, balancing precision and recall under the mild class imbalance. Once the best hyperparameters are identified, the model is retrained on the outer training fold and evaluated on the corresponding outer test fold. This process is repeated over R=10 rounds to account for variability due to random partitioning.

Performance is tracked using a set of metrics: Matthews Correlation Coefficient (MCC), Area Under the ROC Curve (AUC), Balanced Accuracy (BA), F1 score, F2 score (favoring recall), Recall, and Precision. These metrics were chosen to provide a comprehensive view of model behavior, especially considering the critical importance of minimizing false negatives in a medical diagnostic context.

LLM Usage

Throughout the course of this project, a large language model (ChatGPT) was used as a supportive tool to enhance development efficiency and deepen understanding of machine learning concepts. It was particularly useful in exploring and clarifying documentation related to hyperparameter optimization, especially in understanding the structure and functionality of Optuna's API. In addition to assisting with technical concepts, the model was consulted for guidance on software engineering best practices. This included object-oriented programming (OOP) principles such as encapsulation, abstraction, inheritance, and polymorphism, as well as design philosophies like favoring composition over inheritance. These insights informed the architectural design of the code and contributed to a cleaner, modular implementation in order to achieve maintainability and reusability goals.

Results and Discussions

EDA

The id column was excluded from modeling because it carries no biological meaning or predictive value.

The class distribution analysis revealed a mild imbalance favoring benign samples (63% benign vs 37% malignant). Although not extreme, this imbalance could cause classifiers to bias predictions toward the majority class. Consequently, it became important to incorporate evaluation strategies sensitive to imbalance, such as balanced accuracy, precision-recall curves, and class-weighted metrics, rather than relying solely on overall accuracy which might be misleading.

Outlier analysis showed that a substantial portion of malignant samples exhibited extreme feature values, often beyond ±3 standard deviations. Initially, a filtering step was tested to remove these outliers; however, this approach resulted in a significant loss of malignant cases (~44.5%), compared to benign ones (~28%). Given that malignant tumors are expected to present greater morphological variability, removing these cases risks discarding biologically meaningful and diagnostically important patterns. For this reason, it was decided to retain all samples, accepting the inherent variability as part of the data's natural structure rather than treating it as noise.

Finally, examination of zero values revealed isolated concentrations of zeros in features related to tumor concavity and concave points. These zeros likely represent true biological

measurements — instances where the tumor surface was perfectly smooth without noticeable concavity. Their presence was not treated as missing data. For this reason, careful consideration during preprocessing was essential to ensure that true biological zeros were respected and not inadvertently imputed or altered.

Algorithms Comparison

The classification performance of six different algorithms — Gaussian Naive Bayes (GNB), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Random Forest (RF), LightGBM (LGBM), and Support Vector Classifier (SVC) — was systematically evaluated using repeated nested cross-validation.

When selecting the most suitable model, emphasis was placed on metrics that remain robust in the presence of mild class imbalance. Although Area Under the ROC Curve (AUC) is widely used due to its ability to summarize separability across all decision thresholds, recent literature has highlighted limitations of AUC, especially in imbalanced settings, where it can provide an overly optimistic view of model performance [4]. Therefore, while AUC was still reported for completeness, it was interpreted with caution.

Instead, the Matthews Correlation Coefficient (MCC) was prioritized as the most meaningful metric. MCC integrates all values from the confusion matrix and remains reliable regardless of class distribution. As emphasized in Chicco & Jurman (2020) [4], MCC is uniquely suited for imbalanced binary classification problems because it balances the performance on both majority and minority classes and is invariant to label switching. It thus provides a unified and intuitive measure of classification quality. In addition, metrics like Balanced Accuracy (BA), F1, F2, and Precision were also examined to capture complementary aspects of performance.

Among the evaluated models, Logistic Regression (LR) and Support Vector Classifier (SVC) consistently ranked highest across the most important metrics. Both achieved median AUC values above 0.993 and MCC values exceeding 0.93 (Figure 9, page 10). LightGBM also performed competitively, particularly in terms of Precision, but showed slightly broader confidence intervals and lower MCC scores than LR and SVC. In contrast, Gaussian Naive Bayes (GNB) had the weakest overall performance, exhibiting lower median values and wider variability across all metrics.

To substantiate the selection of the final model, 95% confidence intervals for each metric were analyzed (Table 2). LR and SVC demonstrated not only high median performance but also narrower confidence intervals, indicating more stable predictions across repetitions. Ultimately, Logistic Regression (LR) was selected as the best-performing model. This decision was based on LR having the highest lower bounds in BA and MCC, which offers stronger guarantees in worst-case scenarios. This finding reinforces conclusions from prior studies, such as Bhagat et al. (2023) [5], which demonstrated the strong predictive power of regularized LR in breast cancer classification tasks.

	LR	**SVC**
AUC	0.9932-0.9965	0.9935-0.9968
BA	0.9592-0.9734	0.9663-0.9811
F1	0.9580-0.9726	0.9484-0.9621
F2	0.9462-0.9591	0.9438-0.9669
MCC	0.9350-0.9570	0.9192-0.9403
Precision	0.9591-0.9737	0.9487-0.9754
Recall	0.9460-0.9474	0.9211-0.9474

Table 2: 95% Confidence Intervals for LR and SVC across key evaluation metrics.

Conclusions

This study demonstrated the successful application of repeated nested cross-validation to evaluate and select the best machine learning

algorithm for breast cancer diagnosis based on morphological features extracted from fine needle aspirates. Logistic Regression (LR) emerged as the best-performing model, achieving strong generalization performance across key evaluation metrics such as MCC. The study emphasized that, when properly tuned and validated, even relatively simple models like regularized Logistic Regression can match or outperform more complex ensemble methods in structured biomedical datasets.

However, several limitations were identified. The dataset used, although well-curated, was relatively small, containing only 512 samples. Furthermore, a mild class imbalance favoring benign tumors was present, which, although not extreme, may influence model training and evaluation. Larger, more diverse datasets with more balanced class distributions would likely improve model robustness and generalization in real-world settings.

Future research directions include exploring methods to make the repeated nested cross-validation framework more computationally efficient, such as by reducing the number of hyperparameter trials dynamically or by leveraging optimization techniques. Additionally, expanding the hyperparameter search space using conditional logic — where different parameters are tuned depending on prior choices — could provide better configurations and further improve model performance. Another promising avenue is the inclusion of additional algorithms that have demonstrated strong results in similar biomedical classification studies, such as those discussed in [6], to assess whether they offer competitive or superior performance in this context.

References

[1] World Health Organization. “Breast cancer”. In: *WHO.int*. url: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (visited on 27/04/2025).

[2] Cleveland Clinic. “Fine-Needle Aspiration (FNA)”. In: *my.clevelandclinic.org*. url: https://my.clevelandclinic.org/health/diagnostics/17872-fine-needle-aspiration-fna (visited on 04/05/2025).

[3] T. Islam et al, “Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI”, Scientific Reports, vol. 14, no. 1, Art. no. 8487, Apr. 2024, doi: 10.1038/s41598-024-57740-5.

[4] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation” BMC Genomics, vol. 21, no. 6, 2020, doi: 10.1186/s12864-019-6413-7.

[5] A. Bhagat, S. Shekhar, R. Ac, and V. W., “Breast Cancer Classification Using Logistic Regression,” High Technology Letters, vol. 29, no. 8, pp. 204–209, Aug. 2023. [Online]. Available: https://www.researchgate.net/publication/372956293

[6] I. Ozcan, H. Aydin, and A. Cetinkaya, “Comparison of Classification Success Rates of Different Machine Learning Algorithms in the Diagnosis of Breast Cancer”. In: Asian Pacific Journal of Cancer Prevention 23, vol. 23, no. 10, pp. 3287–3297, Oct. 2022, doi: 10.31557/APJCP.2022.23.10.3287.

Figures

The figures referenced across the report are shown in pages 8-10.

Code Access

The code to reproduce the results of this report and possibly add new features can be found in this Github repository: https://github.com/ikotsov/MLCB-Assignment-2

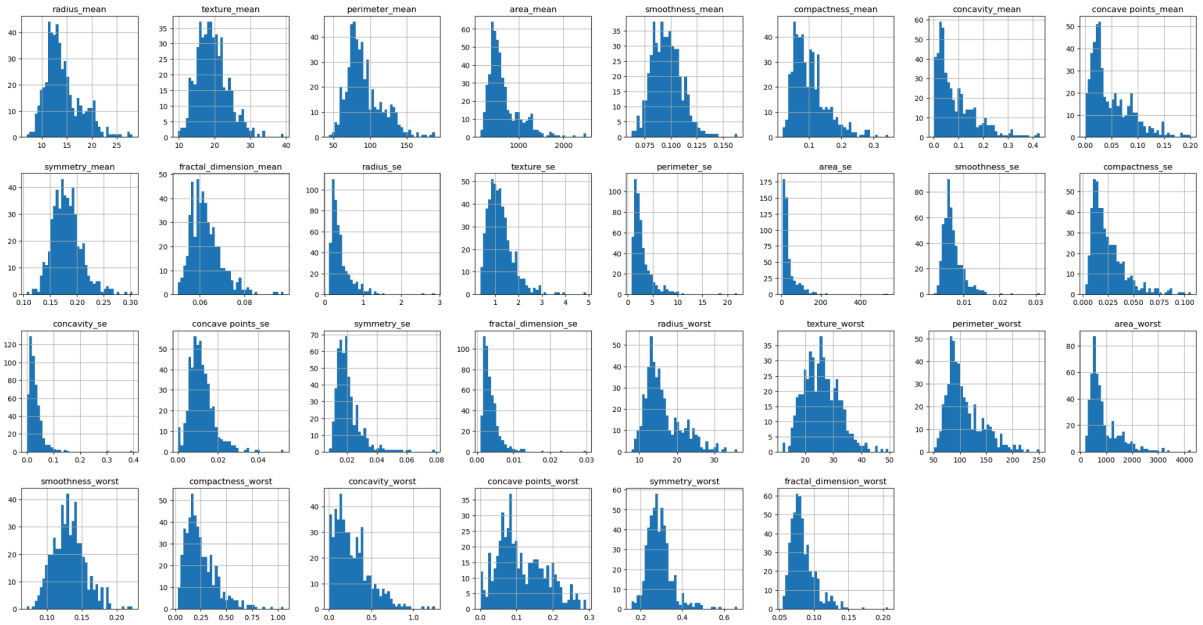

Figure 4. Distribution histograms of all numerical features before standardization, revealing right-skewed patterns and potential outliers.

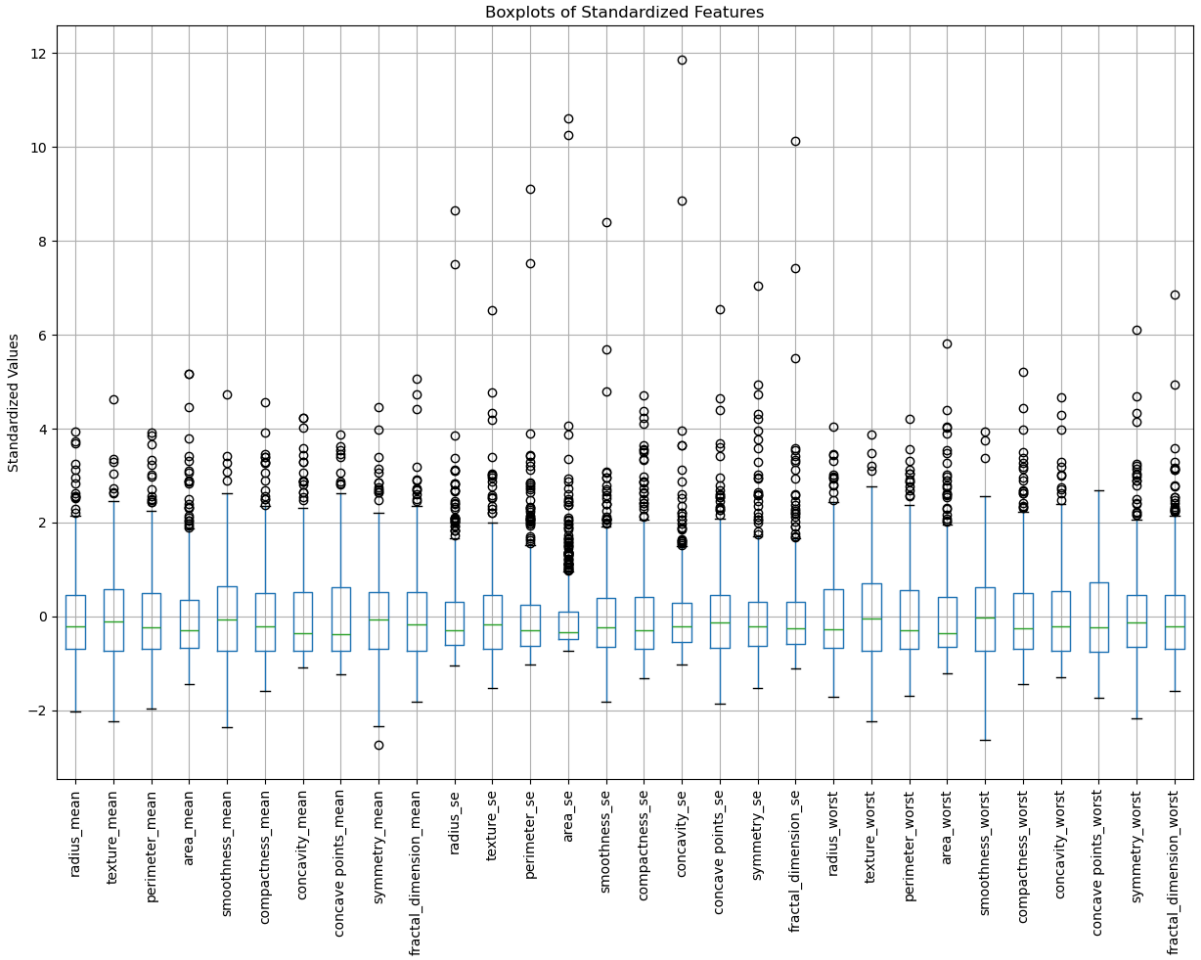

Figure 5. Boxplots of standardized feature values, highlighting the presence of extreme values (outliers) across multiple features.

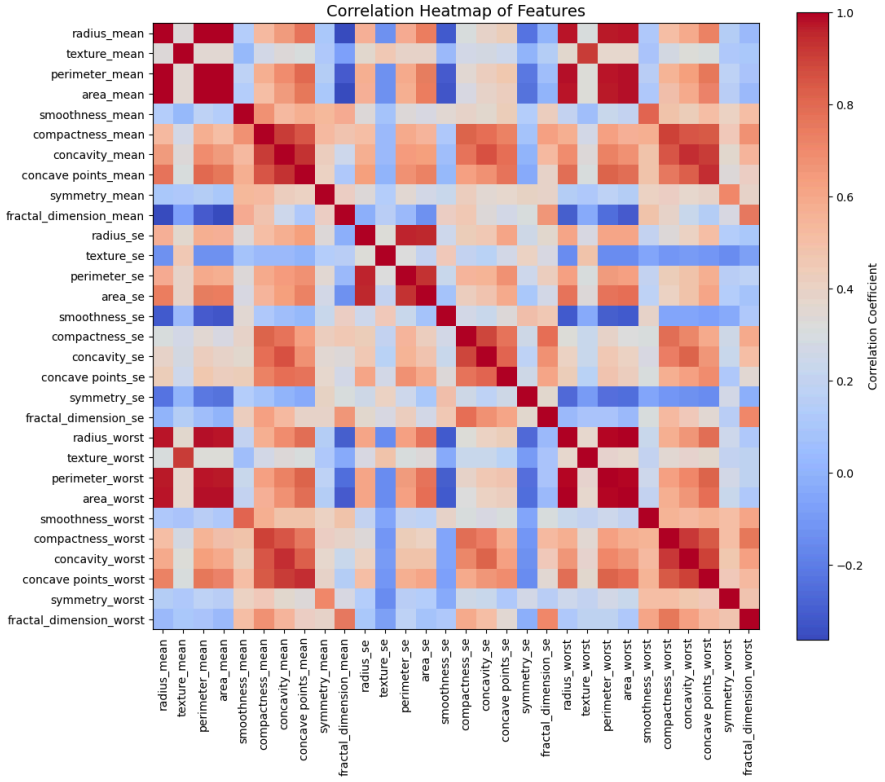

Figure 6. Correlation heatmap of numerical features. Strong correlations suggest potential redundancy among size-related attributes.

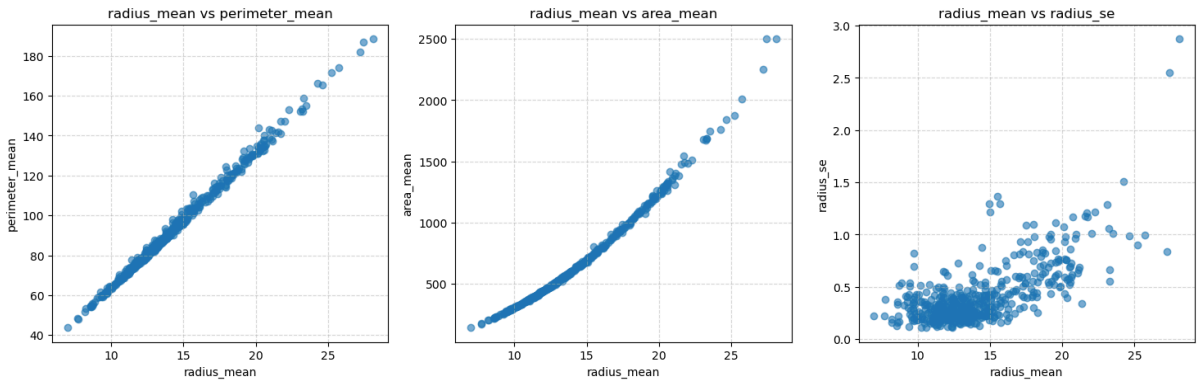

Figure 7. Scatter plots of selected feature pairs. Linear and nonlinear trends are observed, revealing the nature of inter-feature dependencies.

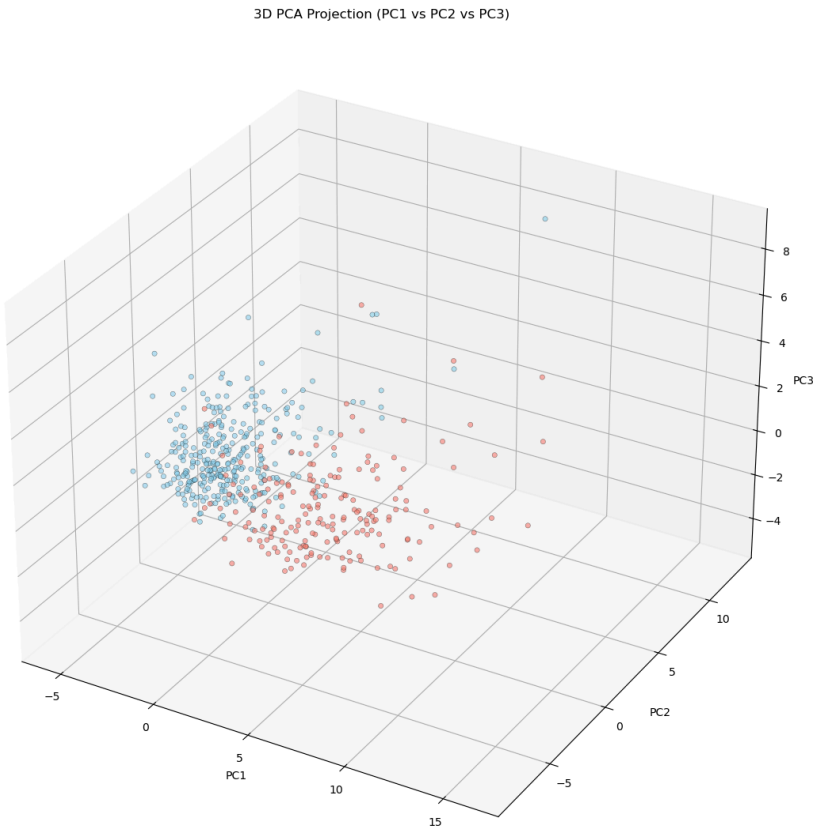

Figure 8. 3D PCA projection using the first three components. Shows partial class separation without using class labels during decomposition.

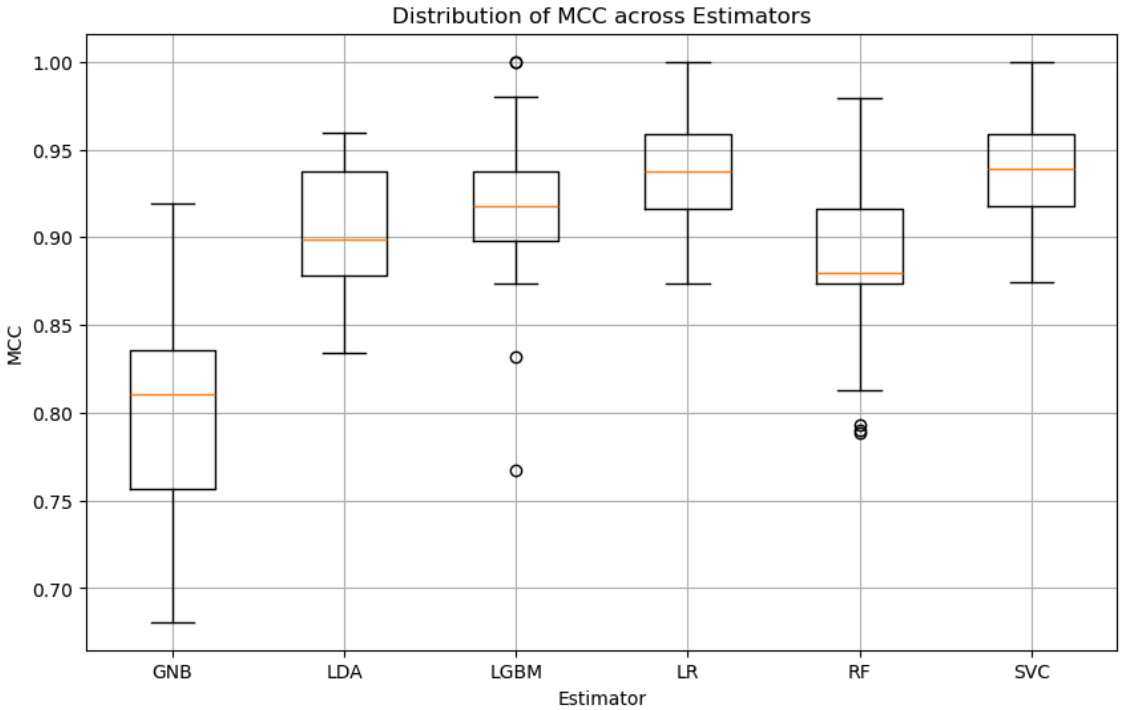

Figure 9: Distribution of Matthews Correlation Coefficient (MCC) across classifiers.