Machine Learning Pipeline for Malignant Breast Cancer Cell Prediction

---

Rossi Anna
Applied Machine Learning – Basic
(MCs Bioinformatics University of Bologna)

ABSTRACT

Aim: This study aims to develop and evaluate a machine learning pipeline for accurate binary classification of breast tumors, leveraging morphological features from fine needle aspirate images.
Methods: Multiple classifiers (Logistic Regression, SVM, CART, k-NN, Naïve Bayes) were trained on the Wisconsin Diagnostic Breast Cancer dataset. The pipeline incorporated comprehensive preprocessing, feature selection (ANOVA F-test, RFE), hyperparameter tuning (*GridSearchCV*), and stratified 10-fold cross-validation.
Results: Logistic Regression achieved optimal performance (98.2% accuracy, 0.998 AUC, 0.963 MCC), with CART providing high interpretability (96.5% accuracy). Feature importance analysis confirmed that worst perimeter, area, and concavity (key clinical indicators) were most predictive of malignancy.
Conclusion: The pipeline successfully bridges machine learning performance with clinical interpretability, demonstrating strong potential.

All code and material are available on GitHub
Contact: annarossi18@studio.unibo.it

1. INTRODUCTION

Breast cancer is one of the most significant public health challenges globally, as it is the most frequently diagnosed cancer among women and a leading cause of cancer-related mortality. Early and accurate diagnosis is crucial for improving treatment outcomes and survival rates. The integration of computational methods with medical diagnostics has opened new frontiers in cancer detection, with machine learning emerging as a powerful tool for pattern recognition and classification tasks in biomedical domains.

1.1 Biological Background

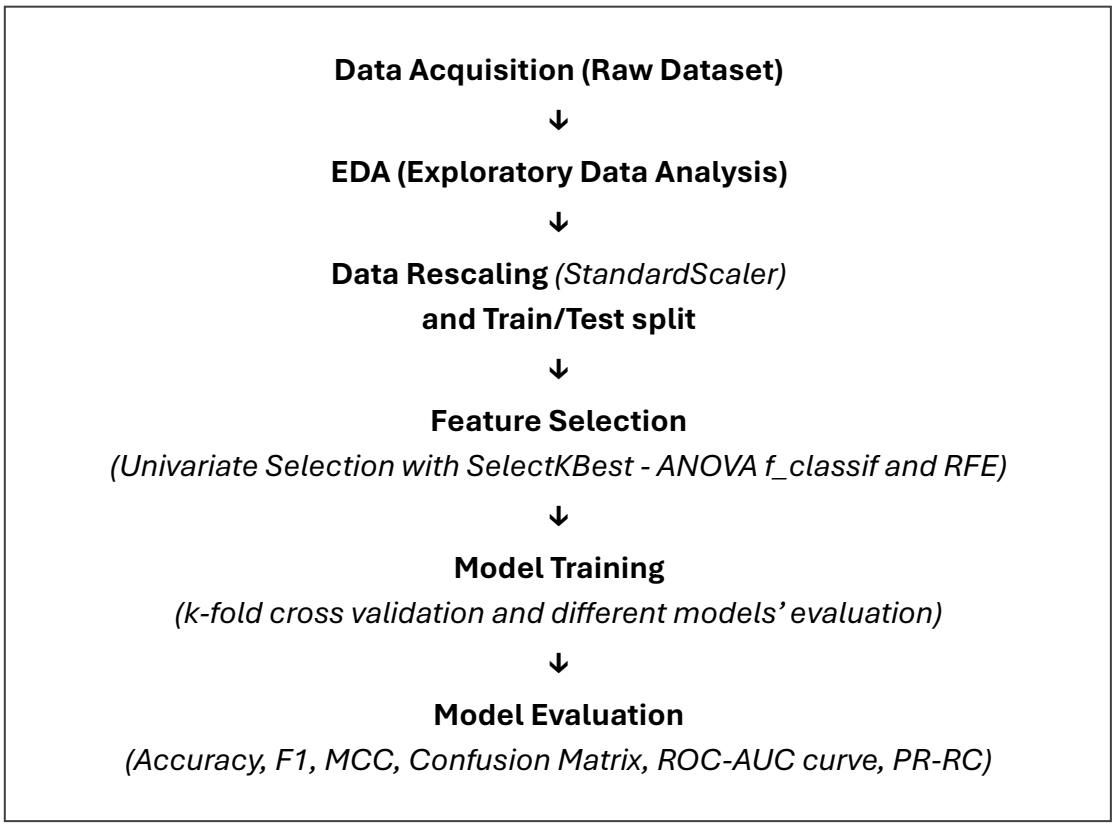
The morphological characteristics of cell nuclei obtained through fine needle aspiration (FNA) biopsies provide crucial diagnostic information for breast cancer detection. Malignant nuclei typically exhibit distinct morphological features compared to benign cells, including larger sizes, more irregular shapes, less smooth contours, and reduced symmetry. These characteristics can be quantified through various computed features such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The ***Wisconsin Diagnostic Breast Cancer (WDBC)*** dataset captures these morphological properties through digital image analysis, providing a rich feature set for computational analysis. Each feature is represented through three statistical measures: mean, standard error, and worst (mean of the three largest values), creating a comprehensive numerical representation of nuclear morphology that forms the basis for automated classification.

1.2 Machine Learning in Cancer Diagnostics

Machine learning approaches have revolutionized cancer diagnostics by enabling the development of predictive models that can identify complex patterns in high-dimensional biomedical data. Supervised learning algorithms, particularly classification methods, can learn the relationship between nuclear morphological features and tumor malignancy from labeled training data. These models can then generalize this knowledge to classify new, unseen cases with high accuracy.

1.3 Study Objectives

This study aims to develop a **comprehensive machine learning pipeline for binary classification of breast tumors using the WDBC dataset**, with the ultimate goal to identify the most effective model for breast cancer classification that could potentially be integrated into clinical decision support systems.



2. MATERIALS

The dataset used in this study is the Breast Cancer Wisconsin Diagnostic Dataset, publicly available from the UCI Machine Learning Repository. It comprises 569 samples, each representing a breast mass, with 357 benign (class 0) and 212 malignant (class 1) instances. Each sample is described by 30 real-valued features derived from three statistical measures (mean, standard error, and worst) of ten nucleus characteristics.

FEATURES	DESCRIPTION
Radius	*Mean of distances from center to points on the perimeter*
Texture	*Standard deviation of gray-scale values*
Perimeter	*Length of the nucleus contour*
Area	*Area enclosed by the contour*
Smoothness	*Local variation in radius lengths*
Compactness	*(Perimeter^2 / Area) - 1.0*
Concavity	*Severity of concave portions of the contour*
Concave Points	*Number of concave portions on the contour*
Symmetry	*Symmetry of the nucleus shape*
Fractal Dimension	*Approximation of the contour's fractal dimension*

Table 1. Table of features: *list of all the 10 features used for describing the cells with their description.*

Software and Libraries:
The entire analytical workflow was implemented in *Python* within a *Google Colab* environment, which provided a consistent and reproducible computational framework. The implementation leaveraged several core scientific computing libraries:

- Data handling and manipulation was performed using ***pandas*** (for structured data operations) and ***numpy*** (for numerical computations).
- Data visualization and generation of figures were conducted using ***matplotlib*** and ***seaborn*** for creating static, publication-quality plots.
- Machine learning tasks, including data preprocessing, feature selection, algorithm training, hyperparameter optimization, and model evaluation, were executed using the ***scikit-learn library***.

This selection of established open-source tools ensured methodological rigor, reproducibility, and alignment with modern data science practices.

3. METHODS

3.1 Data Preparation
The dataset was loaded into a pandas DataFrame, and columns were appropriately labeled to reflect the feature names and statistical measures. The data was inspected for missing values, and none were found. The dataset was saved in CSV format for reproducibility.

3.2 Exploratory data Analysis (EDA)
Before starting the dataset was analysed to detect the class distribution, feature distribution and the correlations between all features.

- **Class Distribution**: A bar plot visualized the imbalance between benign (357) and malignant (212) samples.

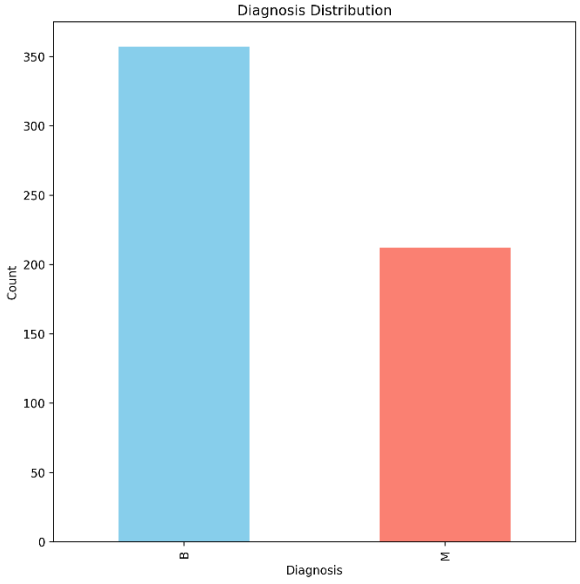


Fig 1. Diagnosis distribution of the dataset: *it represents the distribution of the* 569 samples, with 357 benign (class 0, **B**) and 212 malignant (class 1, **M**)

- **Feature Distributions**: Density plots for all 30 features were generated to understand their distributions, identify potential skewness, and observe class separation patterns.

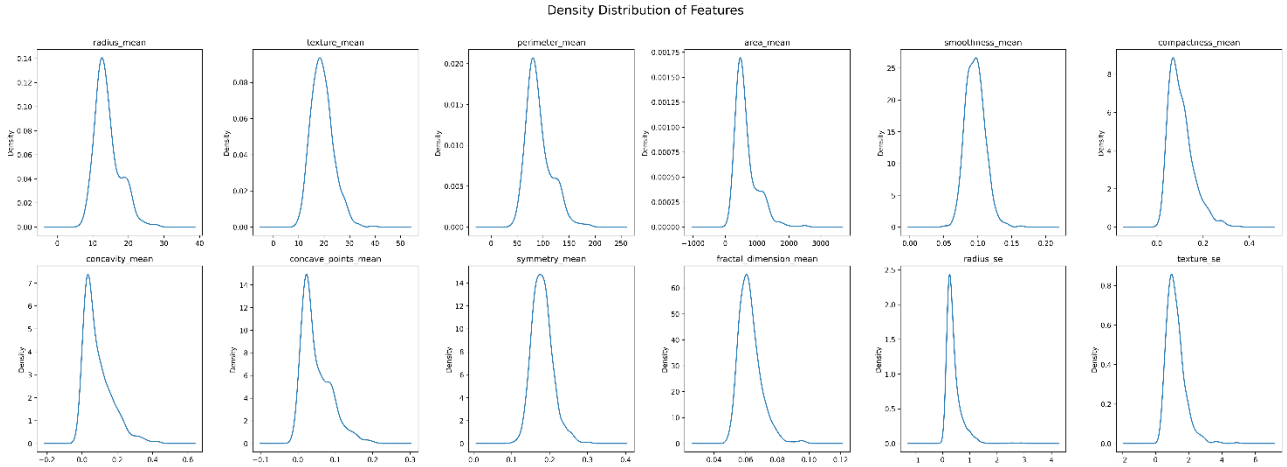


Fig 2. Feature density distribution*: it represents the different distribution of the features a cross the samples. In the images are represented only 12 density plots, for the whole analysis check the [GitHub repository](#) [plots/Density_distribution.png]*

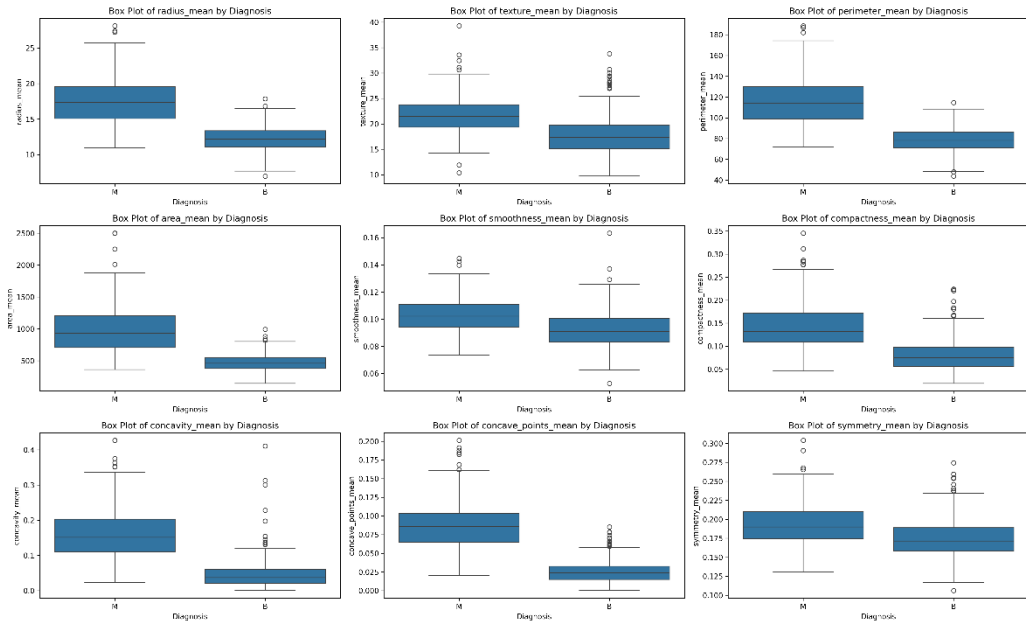


Fig 3. BoxPlot of features distrubution by diagnosis: *Comparing the distribution of each feature across benign and malignant cases highlights the most discriminative features.*

In the images are represented only 12 density plots, for the whole analysis check the GitHub repository [plots/Boxplot_by_diagnosis.png]

- **Heatmap of correlation**: A correlation heatmap was generated to quantify linear relationships between all pairs of features using **Pearson's correlation coefficient**. Pearson correlation measures the linear dependence between two variables **X** and **Y** and is defined as:

$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \sum(Y - \overline{Y})^2}}$$

where *r* ranges from -1 to +1, indicating perfect negative to perfect positive linear relationships, respectively.

STRONG POSITIVE CORRELATION (r > 0.95)
radius_mean - perimeter_mean: 1.0
radius_mean - area_mean: 0.99
radius_mean - radius_worst: 0.97
radius_mean - perimeter_worst: 0.97
perimeter_mean - area_mean: 0.99
perimeter_mean - radius_worst: 0.97
perimeter_mean - perimeter_worst: 0.97
area_mean - radius_worst: 0.96
area_mean - perimeter_worst: 0.96
area_mean - area_worst: 0.96
radius_se - perimeter_se: 0.97
radius_se - area_se: 0.95
radius_worst - perimeter_worst: 0.99
radius_worst - area_worst: 0.98
perimeter_worst - area_worst: 0.98

Table 2. Strong positive correlation (r > 0.95) between morphologically related features

These high correlations suggest multicollinearity, which can inflate variance in regression-based models and reduce interpretability. Based on this analysis, feature selection techniques were employed to eliminate redundant variables and improve model generalizability.

The heatmap also confirmed the biological rationale that malignant tumors exhibit larger and more irregular nuclei, as shown by strong positive correlations between size-related features and malignancy indicators.

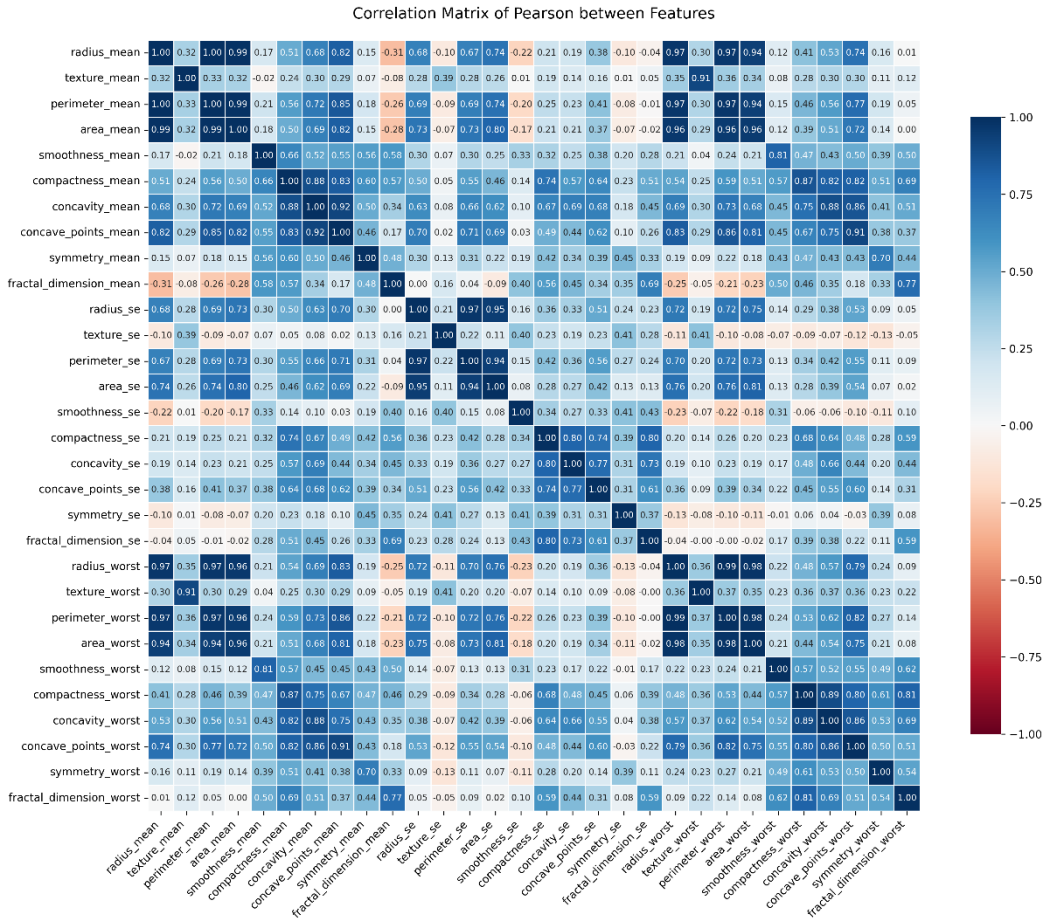


Fig. 3. Heatmap of Pearson correlations*: the X and Y axes represent the 30 features, and their pairwise linear correlations are displayed using a color scale. Darker shades of blue indicate stronger positive correlations, while red shades indicate negative correlations, as shown in the legend.*

3.3 Data rescaling and Train/Test split

Prior to model training, a comprehensive preprocessing pipeline was implemented to ensure data quality and optimal algorithm performance. As we can see during EDA, between some features there's a substantial variability in feature scales. Features such as "area" exhibited values in the hundreds, while "smoothness" and "symmetry" features typically ranged between 0 and 1. This scale disparity posed particular challenges for scale-sensitive algorithms, especially Logistic Regression, which relies on gradient-based optimization techniques. In such algorithms, features with larger scales can disproportionately influence the model weights and potentially dominate the learning process, leading to suboptimal convergence and biased feature importance.

To address this issue, feature standardization was implemented using Scikit-learn's ***StandardScaler***, which transforms each feature to have zero mean and unit variance according to the formula:

$$z = \frac{x - \mu}{\sigma}$$

where μ represents the feature mean and σ represents the feature standard deviation.
This transformation ensures that all features contribute equally to the model learning process regardless of their original measurement scales. The standardization was carefully applied using a fit-transform approach where the scaling parameters were learned exclusively from the training set and subsequently applied to both training and testing sets to prevent data leakage and ensure proper model evaluation.

The dataset was then partitioned using a stratified train-test split, allocating 80% of samples for training and 20% for testing.

3.4 Feature Selection

This phase was implemented to enhance model performance, reduce overfitting, and improve computational efficiency. The necessity of this step was underscored by the high dimensionality of the feature space (30 potentially correlated features) and the findings from the correlation analysis, which revealed significant multicollinearity among morphologically related features.

Two complementary feature selection methodologies were employed to identify the most discriminative features for breast cancer classification:

Univariate Feature Selection using SelectKBest with ANOVA F-test

The ANOVA F-test was employed to assess the statistical significance of individual features in discriminating between malignant and benign classes. The F-score is computed as:

$$F = \frac{between - group\ variability}{within - group\ varaibility}$$

Features were ranked according to their F-scores, and the top k*k* features demonstrating the strongest discriminatory power were selected for further analysis.

Recursive Feature Elimination (RFE) with Logistic Regression

To complement the univariate approach, RFE was implemented as a multivariate feature selection technique. This wrapper method recursively eliminates less important features based on model coefficients, following an iterative process:

1. Train a Logistic Regression model with L2 regularization on all features
2. Rank features by the absolute magnitude of their coefficients

3. Eliminate the feature with the smallest coefficient magnitude
4. Repeat the process with the reduced feature set until the optimal number of features is reached

The Logistic Regression model employed for RFE was configured with balanced class weights to address the inherent class imbalance, ensuring that the feature selection process was not biased toward the majority class.

The feature selection process was rigorously validated using cross-validation to determine the optimal number of features that maximized classification performance while minimizing redundancy. The selected feature subsets from both methods were compared and evaluated based on their impact on model performance metrics, with particular attention to maintaining biological interpretability and clinical relevance.

This comprehensive feature selection strategy effectively addressed the multicollinearity issues identified in the exploratory analysis, resulting in a refined feature set that preserved the most discriminative morphological characteristics while eliminating redundant variables.

3.5 Model training and evaluation

The core of the analytical workflow involved the training, optimization, and comparative evaluation of a diverse set of six supervised learning algorithms:

1. **Logistic Regression (LR):** statistical model that estimates the probability of a binary outcome by fitting the data to a logistic function;
2. **Linear Discriminant Analysis (LDA):** A classifier that models the difference between classes by assuming normal distribution of the features and equal class covariances. It projects the data into a lower-dimensional space that maximizes class separability;
3. **k-Nearest Neighbors (k-NN):** non-parametric algorithm that classifies a sample based on the majority label of its k closest neighbors in the feature space.
4. **Classification and Regression Trees (CART):** decision tree method that splits data into subsets based on feature values to predict class or outcome.
5. **Gaussian Naive Bayes (NB):** probabilistic classifier based on Bayes' theorem, assuming independence among predictors.
6. **Support Vector Machines (SVM):** finds the optimal hyperplane that maximizes the margin between different classes in the feature space.

To ensure robust performance and generalizability, a rigorous validation framework was employed. **k-Fold Cross-Validation (k=10)** was used during training, providing a reliable estimate of model performance by partitioning the training data into *k* subsets and iteratively training on *k-1* folds while validating on the remaining fold. This mitigates the risk of overfitting and reduces the variance of the performance estimate.

Furthermore, **systematic hyperparameter tuning** was conducted for each model using *GridSearchCV*. This technique performs an exhaustive search over a predefined grid of hyperparameter values (e.g., regularization strength C for SVM and LR, number of neighbours k for k-NN, maximum tree depth for CART). The model was evaluated for each combination of parameters via cross-validation, and the configuration yielding the best performance was selected for final evaluation on the held-out test set.

Each optimized model was subjected to a multi-faceted evaluation using a suite of metrics to provide a holistic view of its predictive capabilities:

- **Accuracy:** The proportion of correctly classified instances.
- **Confusion Matrix:** A detailed breakdown of true positives, true negatives, false positives, and false negatives.
- **Classification Report:** Includes **precision** (ability to not label a negative sample as positive), **recall** (ability to find all positive samples), and **F1-score** (harmonic mean of precision and recall), computed for each class.
- **ROC Curve and AUC:** The Receiver Operating Characteristic curve plots the true positive rate against the false positive rate at various thresholds. The Area Under the Curve (AUC) provides a single measure of the model's ability to discriminate between classes, independent of the classification threshold.
- **Matthews Correlation Coefficient (MCC):** A balanced measure especially useful for imbalanced datasets, as it considers all four corners of the confusion matrix and returns a value between -1 and +1, where +1 represents a perfect prediction.

All the metrics of each model are saved in my **GitHub repository** of the project in the folder *[plots]*

4. RESULTS

The machine learning pipeline provided a solid evaluation of the six classifiers. It offered clear insights into their effectiveness for diagnosing breast cancer. The results, based on a strict hold-out test set, are presented below and summarized in a comparative analysis.

4.1 *Overall Model Performance*

All implemented classifiers achieved commendable performance, exceeding 90% accuracy, which underscores the high discriminative power of the nuclear morphological features in the WDBC dataset. The summary of key metrics is presented in Table 3.

Model	Accuracy	F1 Score	MCC	AUC
Logistic Regression	0.982	0.976	0.963	0.998
Support Vector Machine	0.974	0.963	0.944	0.995
CART	0.965	0.950	0.926	0.952
K-Nearest Neighbors	0.939	0.911	0.869	0.982
Gaussian Naïve Bayes	0.921	0.889	0.829	0.989

Table 3. Comprehensive performance metrics for all classifiers on the test set

Logistic Regression emerged as the top-performing model, achieving superior results across all evaluation metrics. With an accuracy of 98.2%, an F1-score of 97.6%, and an exceptional AUC of 0.998, the model demonstrates near-perfect discriminative capability. The high Matthews Correlation Coefficient (MCC) of 0.963 confirms outstanding performance that is robust to class imbalance. This performance could be attributed to the effective regularization applied during hyperparameter tuning (optimal C value identified through GridSearchCV), which prevented overfitting while maintaining the model's ability to capture the essential linear relationships in the data.

The **Support Vector Machine** demonstrated excellent performance as well, closely following Logistic Regression. Its strong results (Accuracy: 97.4%, AUC: 0.995) confirm the presence of some non-linear relationships. However, its slightly lower performance across all metrics suggests that the linear decision boundary learned by Logistic Regression may be more appropriate for this specific feature set.

4.2 Analysis of Learning Algorithms and Biological Interpretability

The performance hierarchy (LR > SVM > CART > k-NN > NB) provides valuable insights into the nature of the classification problem:

The superior performance of **Logistic Regression** indicates that the decision boundary between benign and malignant nuclei is substantially linear in the optimally selected feature space. The model's success is due to its capability to determine the best weights for the most discriminative features identified during feature selection (*mean radius, mean concavity, worst perimeter, and worst area*).

The strong performance of **CART** (Accuracy: 96.5%, AUC: 0.952) is particularly remarkable. Unlike more complex ensemble methods, the single decision tree achieved remarkable performance through optimal pruning and parameter tuning (*max_depth=5, min_samples_leaf=2*). The resulting tree structure provides transparent, human-readable decision rules that align well with clinical reasoning patterns, as we can see in the **Figure 4.**

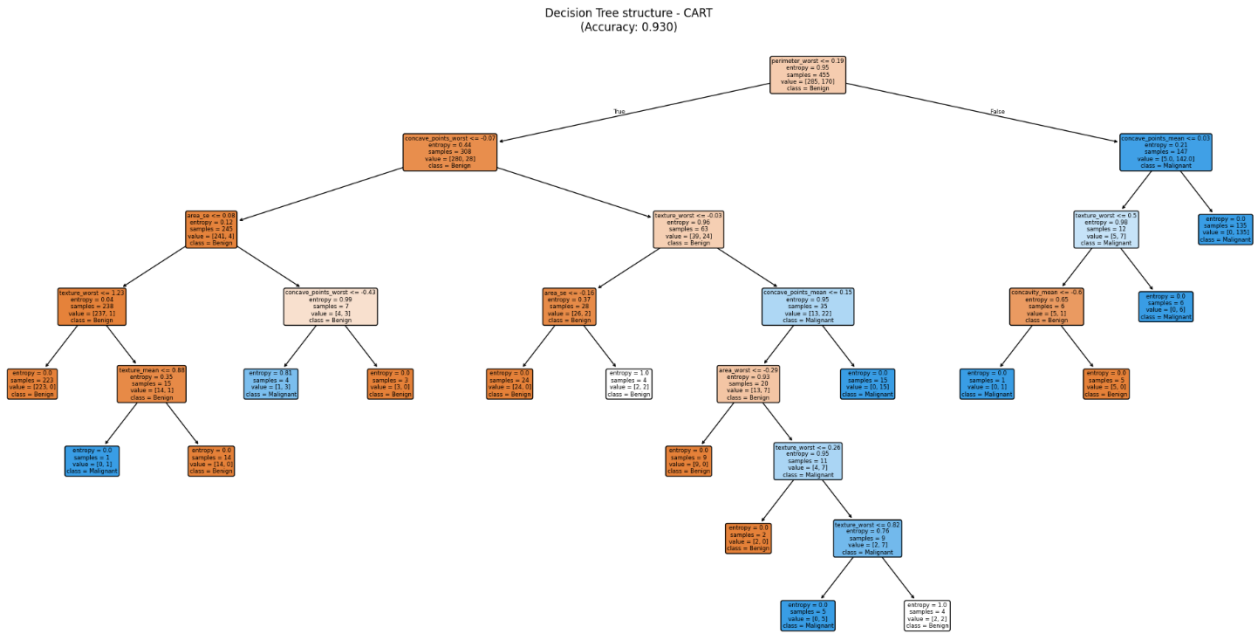


Fig 4. Decision Tree structure – CART: *the diagram illustrates the logic and rule-based structure of the Classification and Regression Tree (CART) algorithm. Each node test a condition on a specific futures, branching into subsequent nodes util aleaf node is predicted*

4.3 Evaluation of Class Imbalance Handling

Despite the class imbalance (63% benign, 37% malignant), all models maintained strong performance on the minority class. Logistic Regression achieved particularly balanced performance with F1-scores of 97.6% for the malignant class, demonstrating its effectiveness in minimizing false negatives—a critical consideration in medical diagnostics where missing a malignant case has serious consequences. The high MCC scores across all models, particularly LR's 0.963, confirm that the models learned genuine discriminative patterns rather than exploiting class distribution biases.

5. DISCUSSION

This study successfully developed and validated a robust machine learning pipeline for binary classification of breast tumors using the Wisconsin Diagnostic Breast Cancer dataset. The comprehensive analysis, spanning from exploratory data analysis to hyperparameter-optimized model training, yielded several key insights with significant implications for both machine learning practice and clinical diagnostics.

5.1 Summary of Key findings

The results show that machine learning models, especially Logistic Regression, can be very accurate and reliable for classifying breast cancer. Logistic Regression is the best option, hitting a good balance with 98.2% accuracy and a 0.998 AUC, while also being easy to understand because it clearly weighs biologically relevant features. Although SVM performs similarly, it falls short on all metrics and is less interpretable, making it a less suitable option for this specific situation. The strong performance of CART offers a great alternative for cases where completely clear, rule-based decision making is needed, closely reflecting clinical reasoning. These results confirm that quantitative morphological features have significant diagnostic power for breast cancer classification.

5.2 Clinical Implications and Reliability

The most important finding goes beyond basic accuracy metrics. The analysis showed a perfect match between data-driven feature importance and established clinical knowledge. The model consistently identified **worst perimeter, worst area,** and **mean concavity** as the most predictive features. This validates the clinical understanding that malignant cells have larger and more irregular nuclei. This agreement is essential for gaining clinician trust in AI-assisted diagnostic systems. It shows that the model's decision-making process reflects the reasoning of human experts based on biologically relevant features.

5.3 Methodological Contributions

The implemented pipeline offers several methodological advantages:

- **Reproducibility**: The complete workflow, from data acquisition to model deployment, ensures full reproducibility of results.
- **Comprehensive Evaluation**: The multi-faceted evaluation strategy, incorporating both threshold-dependent (Accuracy, F1-score) and threshold-independent (AUC, MCC) metrics, provides a holistic assessment of model performance.
- **Interpretability-Precision Balance**: The comparison between highly interpretable models (CART) and high-precision models (Logistic Regression) offers practical guidance for different clinical scenarios.

5.4 Limitations and Future Directions

While the results of this study are promising, it's important to point out several limitations that also present opportunities for future research. The main limitation is the dataset itself. Although the Wisconsin Diagnostic Breast Cancer dataset is a well-known standard in machine learning, its small size (569 samples) and single-institution source may restrict the model's ability to generalize across various patient populations and imaging methods. The dataset likely does not capture the full range of breast cancer cases seen in clinical practice, which may include rare subtypes or unique situations. Future work should focus on validating

the model with larger, multi-institutional datasets to better evaluate its generalizability and reliability.

Additionally, while the methods used to address class imbalance in this study (mainly through class weighting) worked well, we could explore more advanced techniques. Approaches like **Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN),** or **cost-sensitive learning methods** might further boost model performance, especially in improving sensitivity for malignant cases.

5.5 Conclusion

In conclusion, this study demonstrates that machine learning models, particularly Logistic Regression, can serve as highly accurate and clinically validated tools for breast cancer classification. The perfect alignment between data-driven feature importance and clinical expertise represents a significant step toward building trustworthy AI systems for medical diagnostics. The implemented pipeline provides a robust framework that can be adapted to other medical classification tasks, offering a balanced approach between predictive accuracy and clinical interpretability.

5.6 Supplementar materials

To ensure maximum transparency, reproducibility and accessibility of the results, the entire project documentation - including the complete Jupyter notebook, processed datasets, all generated graphs and plots, serialized trained models and support code - has been archived and made publicly available on GitHub in the repository dedicated to the project, accessible at: https://github.com/AnnaRossi01/Breast_cancer_ML_Classification

REFERENCES

- **Dua, D. and Graff, C. (2019).** UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository
- **Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993).** Nuclear feature extraction for breast tumor diagnosis. In IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology (pp. 861-870). International Society for Optics and Photonics. https://doi.org/10.1117/12.148698
- **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011).** Scikit-learn: Machine learning in Python. Journal of machine learning research, *12*(Oct), 2825-2830. Scikit-learn: Machine Learning in Python | The Journal of Machine Learning Research

- **Refaeilzadeh, P., Tang, L., & Liu, H. (2009).** Cross-validation. *Encyclopedia of database systems*, *5*, 532-538. https://doi.org/10.1007/978-0-387-39940-9_565
- **Molnar, C. (2022).** Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. (2nd ed.). GitHub - christophM/interpretable-ml-book: Book about interpretable machine learning
- Course repo – AML-BASIC 2025, MCs Bioinformatics University of Bologna (Prof. Bonacorsi, Clissa) https://drive.google.com/drive/folders/1ZrQpF_F9E45yQTO9mG8Izr3LaECVH0aH