Breast Cancer Diagnosis using Machine Learning

By Aylmer Liang (agl2162), Caterina Almazan (cga2133),
Jason Manuel (jkm2191), Sejal Mittal (sm5756)

COMS 4995 Applied Machine Learning
Fall 2024

I. Background & Dataset Context

Breast cancer affects millions of people worldwide. In fact, according to the Breast Cancer Research Foundation, every 14 seconds a woman is diagnosed with breast cancer. Countless researchers have worked to either discover a new method or improve upon previous ones for diagnosing and detecting breast cancer. In this project, given features derived from images of breast masses, we **built and trained machine learning models** to accurately diagnose cancers as **malignant** or **benign**.

We used a Kaggle dataset on breast cancer detection, originally sourced from the University of Wisconsin Madison and now being maintained by the University of California Irvine's Machine Learning Repository. The dataset includes **569 instances** of breast cancer cases, with each instance having **10 main features** derived from a digitized image of a fine needle aspirate (FNA) of a breast mass. These features describe characteristics of the cell nuclei, such as:

- Radius: *Mean of distances from the center to points on the perimeter*
- Texture: *Standard deviation of gray-scale values*
- Perimeter
- Area
- Smoothness: *Local variation in radius lengths*
- Compactness: *Perimeter² / area - 1.0*
- Concavity: *Severity of concave portions of the contour*
- Concave Points: *Number of concave portions of the contour*
- Symmetry
- Fractal Dimension: *Approximation of coastline dimension ("coastline" of the cell nuclei)*

Finally, the target variable for our project is **diagnosis**, which is a binary classification of either **Malignant (M)** or **Benign (B)**.

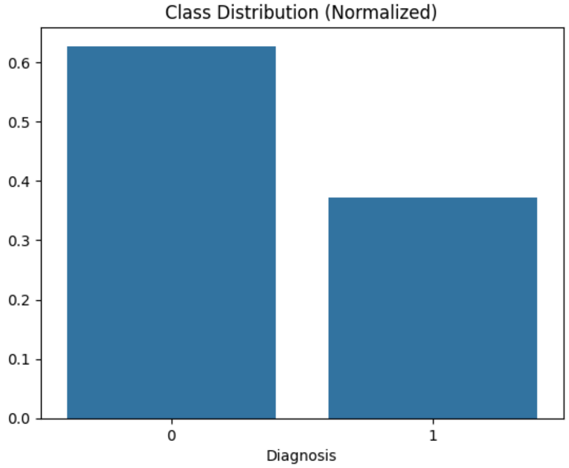

Fig. 1: The class distribution of the 569 cases in our dataset.

II. Preprocessing

For our preprocessing, we checked if our dataset included any missing values as well as any outliers. We did not find any missing values, and there were no features that had such extreme outliers that it would help to remove them from the training set. We also examined the correlation of our features to determine if there were features that could be dropped, but we decided to keep all of them, especially since multicollinearity would not affect our tree-based models; we performed feature importance analysis for our models to later check for potential features to be dropped.

We performed **stratified splitting** on our dataset because, as we saw from Fig. 1, the dataset is imbalanced. Then, we proceeded to scale our train, validation, and test sets by using a **standard scaler** fitted on the train set to ensure that no one feature dominates the others due to magnitude.

III. *Models*

Using the prior modeling results from University of California Irvine's Machine Learning Repository as our baseline, we implemented a variety of modeling techniques that differed both in complexity and explainability. We started with a very simple and explainable logistic regression model, then moved on to a more complex but still explainable set of tree-based approaches, before finally trying a neural network that trades explainability for high model complexity. Since we are dealing with an imbalanced dataset, and our main objective is to **detect malignant tumors**, we prioritized **F1-score and recall** during evaluation. Across all models, we achieved high F1-scores, which we were able to further improve with hyperparameter tuning and, in the case of logistic regression, with regularization as well.

IV. *Model Results*

Model	F1-score before tuning	F1-score after tuning
Logistic Regression	0.963	**0.975**
Decision Tree	0.9024	0.925
Random Forest	0.938	0.963
XGBoost	0.928	0.962
Neural Network	N/A	0.951

The above results demonstrate that the simplest and most explainable model, **logistic regression**, performs the best, achieving the highest F1-score of **0.975** on the test set. We moved on to trying tree-based methods, starting with a decision tree, but even after tuning it lagged behind logistic regression. Random forest and XGBoost were closer in performance to logistic regression, achieving a large performance boost via hyperparameter tuning on parameters such as *n_estimators* and *max_depth*. In the Random Forest model, we used grid search with out-of-bag error for evaluating and selecting the optimal hyperparameters, while in XGBoost, we used grid search with F1-score; both models achieved ~0.03 increases in F1-score due to increasing model complexity. Finally, the neural network not only did not outperform logistic regression but also did not improve upon the tree-based methods.

The low F1-score of the neural network is not surprising since we did not have access to nearly enough data for a neural net to generalize well to the test set. Furthermore, given the already great performance using a simple model, the high complexity of both the neural network and tree-based methods probably resulted in overfitting and capturing more noise than signal in the training data.

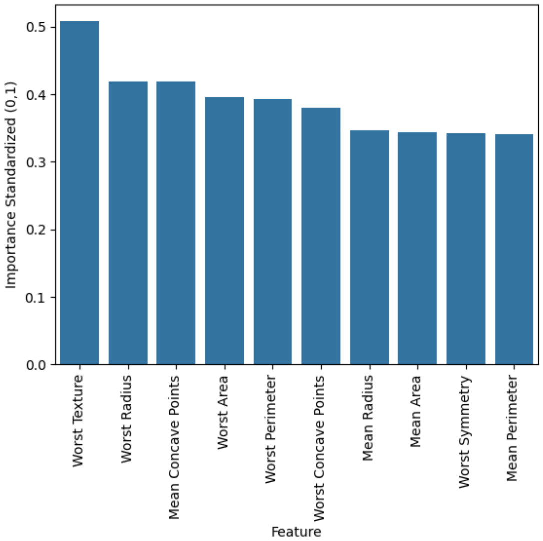

Fig. 2: The graph of standardized feature importance for the tuned logistic regression model.

The above feature importance plot shows the most important features in descending order of importance, which is computed based on the magnitude of their coefficients in the logistic regression model. The top ten variables here mainly consist of the *mean* and *worst* variables, where the "worst" is defined as the mean of the three largest values for a given feature. The importance of these *worst* variables, two of which (*worst texture* and *worst radius*) are the most important, highlights the integral role of the outliers for each feature and why we decided to not mitigate their impact during preprocessing.

V. Conclusions

In this project, we explored five different machine learning models to classify breast cancer cases as malignant or benign. Our primary goal was to maximize F1-score and recall to address the class imbalance and the critical need for accurate malignant case detection. Our experiments revealed that simpler models like logistic regression performed remarkably well, achieving the highest F1-score of 0.975 after hyperparameter tuning. This strong performance is likely due to the relatively **small size** and **linear separability** of this specific dataset, which favored less complex models. The more complex models saw significant gains after hyperparameter tuning, but logistic regression already had a high F1-score before any tuning or regularization.
We also saw when analyzing feature importance in other models, such as XGBoost, that some pairs of variables with high correlation with one another that were important in the tree-based models, such as *worst texture* and *mean texture*, were not emphasized by logistic regression, suggesting that multicollinearity could be affecting logistic regression. In next steps, we would explore dropping at least one of the variables from these pairs. Having the opportunity to explore and test the performance of our models on a larger, more comprehensive dataset may mitigate these issues and improve model generalization and robustness further.

If we were to continue working on this study, we would love to further observe the impacts that machine learning can have on breast cancer prevention through combining these machine learning models with computer vision techniques. Using an additional dataset with images, such as from screening mammograms, provides researchers the opportunity to employ deep learning models and allows doctors to have the opportunity to be in a position where they can provide more accurate diagnoses, detect earlier stages of breast cancer, and create more strategic health plans for their patients.