Predicting Breast Tumor Malignancy Using Generalized Additive Models

Ava, Laura, Abby, Ella, Grady

Project Overview

Objective: Build an interpretable ML model that can assess whether a breast mass is benign or malignant using measurable characteristics such as radius, area, smoothness, and concavity extracted from digitized medical images.

Prediction: Train a classifier to estimate the probability that a new mass is malignant.

Inference: Determine and interpret features which have the greatest influence on determining the assigned probabilities.

Target Audience: clinicians, researchers who are responsible for diagnosing breast masses..

Other Notes: The project also addresses several relevant challenges, including potential noise in imaging, derived features, limitations associated with an older dataset, and the need for models that doctors can trust when making time-sensitive decisions. Trust and interpretability are central concerns in medical applications, and our modeling pipeline is intentionally designed to balance performance with transparency, ensuring that the results are clinically meaningful and accessible.

Importance and Relevance

Breast cancer is one of the most pressing health challenges worldwide…

- 30% of all new female cancer cases/year
- 2nd leading cancer related deaths among women
- Incidence rates rise approximately 1%/year

Our model is vital as early detection of breast cancer has been shown to boost survival rates by 20%. ML models can meaningfully strengthen diagnostic processes more than traditional techniques.

Data Overview

- Data originated from the Diagnostic Wisconsin Breast Cancer Database (downloaded from Kaggle)
- Source dataset donated on October 31, 1995
  - Our model serves as a **proof-of-concept framework**
- No missing values, predictors standardized to four significant digits.
- 30 continuous predictors from 10 cell nucleus features
  - Each of the ten features recorded in 3 forms (see table)
- Response variable is diagnosis: malignant or benign
  - 569 observations - 357 benign (62.7%), 212 malignant (37.2%)

Features
Radius
Texture
Perimeter
Area
Smoothness
Compactness
Concavity
Concave points
Symmetry
Fractal dimension

Forms	Description
Mean	Average value across sample
Standard Error	Variability of that feature
Worst (maximal)	Mean of highest 3 values

Table 1: Key Feature Variables & Forms

EDA & Preliminary Findings (1/2)

- No variables raised concerns enough to justify transformations, especially since GAMs allow for nonlinearity.
- Outliers were examined, but retained as they reflect real biological variation rather than measurement error.

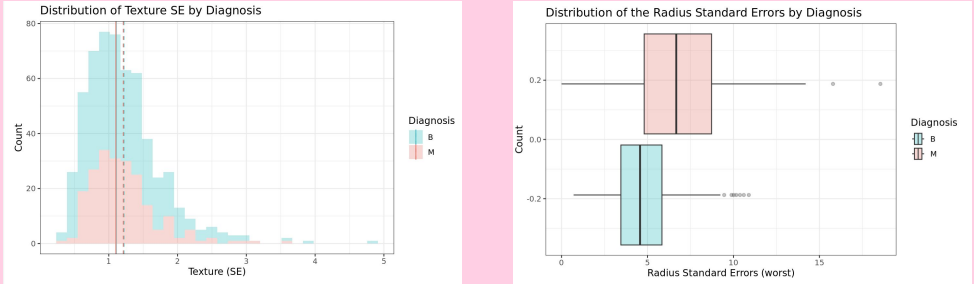

Figure 1: Preliminary EDA

- We fit several classification models including KNN, Random Forest, Lasso, Logistic Regression, and Decision Trees.
- Based on domain knowledge, interpretability, and predictive performance, we performed variable selection with LASSO logistic models, including interaction effects, and then fit a GAM informed by these selected features.

EDA & Preliminary Findings (2/2)

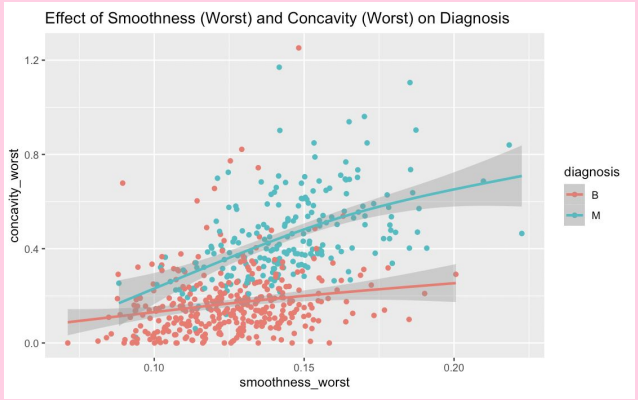

Figure 2: Smoothness & Concavity

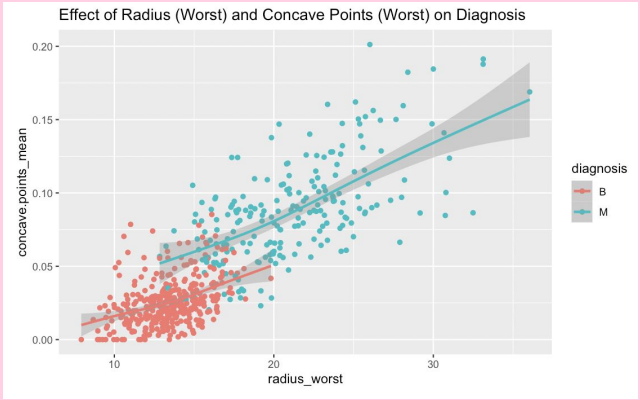

Figure 3: Radius & Concave Points

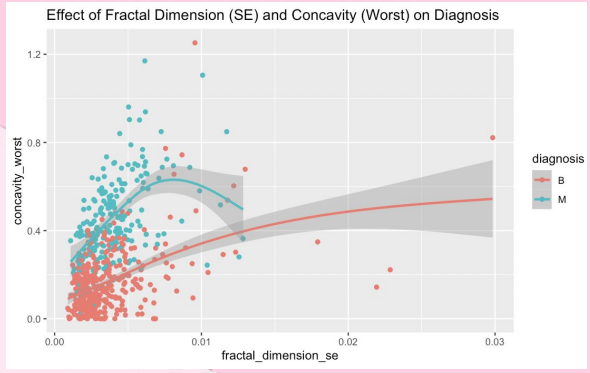

Figure 4: Fractal Dimensions & Worst

We also explored bivariate EDA to examine potential interaction effects between predictors. In these plots, there appears to be interaction effects between each of the listed pairs of predictors since the relationships between each pair differs with the benign and malignant tumors.

Cost Analysis

Goal: determine an optimal threshold that prioritizes maximizing sensitivity, while maintaining high specificity

- Prioritize maximizing sensitivity to minimize false negatives (missing a malignant tumor)
  - Trade-off: lower specificity (unnecessary further testing)
- False-negative results lead to delayed detection and treatment —> greater long-term costs, morbidity & mortality rates
  - $26 billion per year in national cost-savings from early diagnosis of cancers
- Standard mammography screening: 76%-86% in sensitivity, 87%-99% in specificity
  - Gap in sensitivity rate

LASSO Logistic Model

To identify key variables for our final model, we used a LASSO logistic model to perform variable selection.

LASSO Model:

- Use a 80/20 train-test split to assess model performance
- Fit the model using a 5-fold CV on the train dataset

Accuracy	0.992
Sensitivity	1
Specificity	0.983

		Truth	
		Benign	Malignant
Predict	Benign	60	0
	Malignant	1	53

Table 2: LASSO Model Performance and Confusion Matrix

Key Features:

Mean:
- Compactness
- Concave points

Standard Error:
- Radius
- Texture
- Smoothness
- Fractal dimension
- Compactness

Worst:
- Texture
- Radius
- Smoothness
- Concave points
- Symmetry
- Concavity

Setting Optimal Threshold

For a clinical diagnosis problem, we want to minimize the false negative rate (sensitivity) since the cost of incorrectly predicting someone doesn't have cancer when they do is greater than the cost of a false positive.

To do this, we used Youden's J statistic:

- Method for choosing the "best" classification threshold by maximizing the combined performance of sensitivity and specificity
- We want the threshold to be less than 0.5 since we want to maximize sensitivity and reduce the rate of false negatives
- Despite maximizing sensitivity, we still observed high specificity giving us confidence in our final threshold.

Optimal Threshold = 0.351

GAM (1/2)

Model: Generalized Additive Model with a binomial logit link to flexibly model the log-odds of malignancy using nonlinear predictors.

Methodology:

1. Compare different splines (cubic spline, thin-plate spline, thin-plate shrinkage) using AIC, REML ~ selected cubic spline due to lower AIC and REML values indicating greater stability
2. Select nonlinear features
   a. Visualize linearity of different features using the binned logit plots for the log-odds of predicting malignancy
   b. Fit a GLM and GAM with a cubic spline for each of the features → select the features with a statistically significant change in AIC (delta AIC > 4) when comparing the GLM and GAM model
   c. Nonlinear features: texture_se, compactness_se, fractal_dimension_se, texture_worst, concavity_worst

GAM (2/2)

Methodology (continued):

3. Build the simple GAM with the cubic splines for the previously identified nonlinear terms and the rest of the features as main effects

4. Assess Model Performance

Accuracy	Sensitivity	Specificity
0.98	0.981	0.983

	Truth		
		Benign	Malignant
Predict	Benign	60	1
	Malignant	1	52

Table 3: GAM Model Performance and Confusion Matrix

Next Step: Add interaction terms to the model to capture interactions between different variables based on the EDA.

Adding Interaction Terms

We included tensor-product interaction terms between clinically meaningful feature pairs.

- **smoothness_worst x concavity_worst**: Low smoothness (coarse boundary) and high concavity is a malignancy marker. There are different risks associated with smooth but concave masses compared to rough but concave ones.
- **radius_worst x concave.points_mean**: Large lesions with many concave points are signals of malignancy.
- **fractal_dimension_se x concavity_worst:** Fractal dimension relates to edge complexity and high fractal dimension with strong concavity could indicate irregular growth.

Final GAM Model

1. Adding the interactions demonstrated a statistically significant improvement in fit over a purely additive GAM when examining AIC and p-value from Chi-squared (0.002).

	df <dbl>	AIC <dbl>
mod_all	15.93296	68.20717
mod_best	15.93357	68.19962

Table 4: GAM AIC Comparison

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	
1	437.13	36.341				
2	437.13	36.332	0.0008306	0.0087786	0.002015	**

Table 5: GAM P-Value Significance

2. The Confusion matrix on the test set is identical for GAM without interactions vs with interactions, likely due to the small sample size.
3. We chose the GAM with selected nonlinear terms and interactions.
   a. Strong predictive performance
   b. Captured clinically plausible nonlinear and interaction effects
   c. Remained interpretable
4. Final sensitivity of 0.981, so 98.1% of malignancy cases were correctly identified in the model.

Key Takeaways

Model: Our GAM model with interactions has very similar predictive performance to the GAM model without, so both can be used by clinicians depending on whether they prioritize complexity or interpretability.

Clinical:

- The features from the LASSO model provide clinicians with an idea of which features are most important when initially diagnosing a mass before receiving biopsy results.
- Using these features, the clinician can make a more informed decision regarding the malignancy probability if the values of these specific features are outside of expected ranges.
- We have strong confidence that the false positive rate is minimized with this model since the optimal threshold was set to maximize sensitivity. Although there is a small chance of false positives, it is unlikely for the model to falsely predict a mass isn’t malignant when it is.

Downstream Uses of the Model

Purpose: Our project serves as a proof of concept illustrating how interpretable ML methods can support clinical reasoning.

- Inform research
- Guide clinicians' early decision making
- Support development of future diagnostic tools
- Guide resource allocation

By ensuring that our model is interpretable and clearly justified, it can be used ethically and effectively in supporting clinical workflows, improving early detection, and helping reduce false negatives, a challenge in mammographic screening.

Limitations

- **Outdated Data:** The dataset was acquired in 1995. Consequently, we are using it as a proof of concept model which we believe will generalize to more recent data.
- **Small Sample Size:** We only have 569 observations so our model is developed using a small sample of observations.
- **Generalization:** Our features are based on summary statistics which may fail to capture specific nuances in tumor characteristics.
- **Patient Stage:** We are only working with masses which are in Stage 1 or Stage 2 if they are determined to be malignant.
- **Modeling assumption:** Nonlinearities or interaction effects may be overlooked due to the nature of selected models.

Conclusion & Future Work

Conclusion: Our model serves as a framework for clinicians to use to predict malignancy probabilities with future datasets.

- We provide insight on which variables to focus on, and guide direction of healthcare professionals next steps

Future Work: Our recommendation is to include additional variables related to other health diagnostics.

- We recommend assessing the interaction between Stage of Cancer
- We also suggest including variables for Race, Ethnicity, Socioeconomic Class, and Age

Thank you!