Predicting Breast Tumor Malignancy Using Generalized Additive Models

Laura Cai, Ava Exelbirt, Abby Li, Grady Purcell, Ella Tillinghast

2025-12-12

Introduction

Project Goals

The primary goal of our project is to build an interpretable machine learning 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. Our analysis has two complementary aims: prediction, where we train a classifier to estimate the probability that a new mass is malignant, and inference, where we identify and interpret the features that have the greatest influence on these probability assignments. Because our target audience includes clinicians, researchers, and individuals without a technical background, we emphasize transparency and interpretability, focusing not only on whether the model predicts accurately, but also on why it arrives at its predictions.

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 and remains the second most common cancer among women, accounting for roughly 30% of all new female cancer cases each year, and the second leading cause of cancer related deaths among women (American Cancer Society, 2024). As incidence rates continue to rise by approximately 1% annually, improving tools for early differentiation between benign and malignant tumors is critical for timely intervention. Early detection of breast cancer has been shown to boost survival rates by up to 20%, further underscoring the clinical relevance of high quality diagnostic support tools (ClinicalKey, 2025).

Traditional diagnostic techniques like mammography, ultrasound, MRI, and biopsy are essential but can also be expensive, time consuming, and prone to human error, particularly false positives and false negatives. Recent research demonstrates that machine learning (ML) can meaningfully strengthen diagnostic processes. For instance, ML models have shown higher accuracy than clinicians in predicting several cancers, including breast, brain, and lung cancer (PMC10312208). A 2020 DeepMind based system outperformed human specialists in breast cancer detection (Nature, 2020), and other ML systems have reached 97% accuracy in identifying common types of lung cancer (PMC10312208). These findings highlight the potential for ML driven tools to enhance clinical workflows and reduce diagnostic uncertainty.

Even within the specific Wisconsin Breast Cancer dataset used here, prior research has achieved strong performance: Support Vector Machines have reached 87–89% accuracy in distinguishing malignant from benign masses, and deep learning models such as CNNs have successfully extracted subtle features (e.g., microcalcifications or architectural distortions) that radiologists may overlook (ClinicalKey, 2025).

Although the dataset we analyze is more than two decades old, our project serves as a proof of concept illustrating how interpretable ML methods can support clinical reasoning. Because modern imaging systems continue to quantify many of the same structural features used in this dataset, our modeling framework remains relevant, scalable, and adaptable. By identifying the features most strongly associated with malignancy, we provide insights that can inform research, guide clinicians' early decision making, and support the development of future diagnostic tools designed to complement, not replace, medical expertise.

Overview of Proposed Methodology

Since we are building our model to assist doctors with diagnosing cancer patients, our model needs to be interpretable so doctors are more confident in their diagnosis and can explain the factors which led to the final diagnosis. To do this, we will first use a LASSO logistic regression model to shrink the features so that only relevant features are included in the final model. Shrinkage is important in this setting because it is likely that there is multicorrelation between many of the features so we don't want our final model to include all of the features. Once the LASSO logistic model provides the relevant features, we will build a Generalized Additive Model (GAM) since our EDA has indicated nonlinearity between the predictors and state of the tumor, so we want our final model to have a balance between flexibility and interpretability. After selecting which features should be included as non-linear splines in the model, we will then include interactions based on exploratory analysis, domain knowledge, and model performance.

Data

The dataset used in this analysis comes from the Diagnostic Wisconsin Breast Cancer Database (WBCD), originally created from digitized images of fine needle aspiration (FNA) biopsies taken from breast masses. The version used here was downloaded from Kaggle, but the source dataset was donated on October 31, 1995. Although imaging modalities have improved significantly since 1995, modern systems continue to quantify similar structural and morphological features, meaning that this dataset remains an informative foundation for building a proof-of-concept modeling framework that can later be extended to more advanced imaging technologies.

Each observation corresponds to a single breast mass and contains 30 continuous predictors, derived from 10 underlying cell nucleus features: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension. Each of these ten features was recorded in three forms:

1. Mean – the average value across the tumor sample
2. Standard Error (SE) – the variability of that feature
3. Worst (maximal) – the mean of the highest three values

The response variable is the diagnosis: malignant (M) or benign (B).

There are 569 observations, including 357 benign (62.7%) and 212 malignant (37.2%) cases. Although benign tumors are more common in the dataset, the dataset is only moderately imbalanced, so rebalancing procedures such as SMOTE or class weighting are not necessary for this analysis.

The data contains no missing values, and all predictors were standardized with four significant digits in the version we downloaded. Because the features are continuous, structured, and well behaved, this dataset is well suited for multiple modeling approaches including logistic regression, LASSO, KNN, random forests, and decision trees.

Finally, this dataset provides a solid platform for exploratory data analysis (EDA) and model interpretation because each predictor has a clear biological meaning and can be reasoned about in clinical terms (e.g., radius relates to tumor size, concavity relates to irregularity).

The 13 predictors selected from the LASSO logistic model can be grouped into three categories: those which measure mean, standard error, and worst. The mean variables include compactness and concave points. The standard error features are radius, texture, smoothness, compactness and fractal dimension. Lastly, the worst variables are radius, texure, smoothness, concavity, concave points and symmetry.

Modeling Methodology

Our methodology consists of five main stages: exploratory data analysis, threshold optimization and evaluation, preliminary model comparison, regularized model selection, and nonlinear modeling with generalized additive models (GAMs).

Exploratory Data Analysis (EDA)

We began by conducting a comprehensive exploratory data analysis to understand the distribution of each of the 30 predictors and to identify any necessary transformations or outlier adjustments. Visualizations included histograms, boxplots, and scatterplots across both classes (benign vs. malignant).

Although certain variables showed skewed distributions, which is expected in biological measurements, none raised concerns severe enough to justify transformation. Outliers were examined but ultimately retained, as they reflect real biological variation rather than measurement error. No missing data were present, so no imputation or deletion procedures were required. The EDA thus served primarily to guide expectations about predictor behavior and to inform our choice of appropriate models.
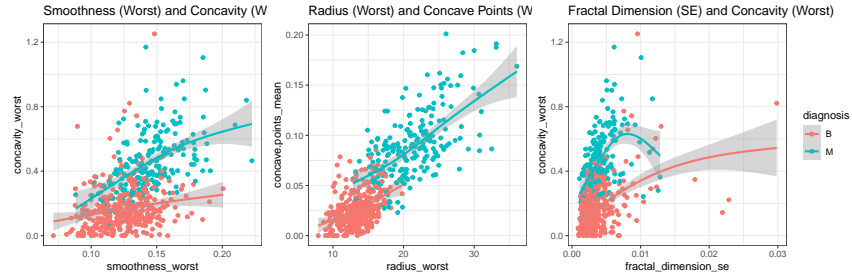

Figure 1: EDA for Interaction Terms

We also explored bivariate EDA to examine potential interaction effects between predictors. In Figure 1, 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.

Building a Model of a Mass

For one observation of the data, we have attempted to translate the variables into the model into a 3D representation of a mass. The purpose of this exercise was to gain intuition into how different variables translate into features of the mass.

** ABBY INSERT WORK HERE

Cost Analysis and Threshold Selection

We researched the cost of false positives and negatives to compute a cost based loss function to determine the optimal classification threshold. This threshold will reflect the asymmetric cost of false negatives (missing a malignant tumor) versus false positives (unnecessary further testing), and will guide our final decision rule applied to our model.

As a result, our threshold selection prioritizes maximizing sensitivity to minimize false negatives. False-negative results can lead to delayed detection and treatment, resulting in greater long-term costs, morbidity, and mortality. One study estimated the national cost-savings in the U.S. from early diagnosis of cancers to be $26 billion per year due to reduced treatment intensity and improved outcomes (Brill, 2020).

Prior research also shows that the accuracy of mammography screening increases with patient age, with sensitivity ranging from 76% to 86% and specificity ranging from 87% to 99% across age groups. (Newton, 2025). This gap in threshold shows why maximizing sensitivity matters: since detecting malignant tumors is the harder task, optimizing sensitivity is essential to improving screening effectiveness. Thus, our decision to choose a lower threshold is consistent with addressing current screening practices and reducing future costs associated with delayed diagnoses.

Preliminary Model Fitting

We next fit several classification models to evaluate baseline predictive performance and to assess the trade-offs between accuracy, interpretability, computational complexity, and clinical trustworthiness. The models included: K-Nearest Neighbors (KNN), Random Forest, LASSO, Standard Logistic Regression, and Decision Trees. Each model was then evaluated using standard metrics including accuracy, sensitivity, specificity, and confusion matrices. Cross validation was used where appropriate to reduce sampling variability.

Although the random forest achieved strong predictive accuracy, it lacked interpretability which should be prioritized since this model will be used by doctors to diagnose whether a breast cancer mass is benign or malignant at an early stage. KNN performed reasonably but does not provide coefficients, feature weights, or insight into mechanism. Standard logistic regression was interpretable but tended to overfit without regularization due to the high number of correlated predictors.

Model Selection and Rationale

Based on accuracy, interpretability, stability, and feature selection capability, we selected the LASSO logistic regression model as our primary predictive and inferential tool. LASSO automatically performs variable selection by shrinking less important coefficients to zero, providing a concise, clinically meaningful set of predictors and potential interactions.

After selecting these top predictors, we further fit a GAM informed by the LASSO selected features. We selected a GAM model for interpretability and flexibility in including nonlinear terms. GAM models are more interpretable than many other regression techniques since they enable visualization between features and the predicted malignancy probability with interpretability being essential for clinical decisions. Additionally, the nonlinearlity of GAMs is advantageous as tumor characteristics rarely relate to malignancy in linear ways, so GAMs can model predictors using a smooth function like splines to learn curves without assuming linearity which improves predictive performance.

Regularized Logistic Regression and Feature Selection

We first fit a LASSO logistic model to perform variable selection and establish a strong, interpretable baseline classifier. The diagnosis variable was encoded as 1 for malignant and 0 for benign. Identifier columns were removed, and the remaining predictors were standardized. The data were randomly split into training (80%) and testing (20%) sets. A LASSO logistic regression model was fit using 5-fold cross-validation to select the optimal penalty parameter with `cv.glmnet`. This approach addresses multicollinearity among predictors and prevents overfitting by shrinking less informative coefficients to zero. The final LASSO logistic model achieved approximately 99% accuracy on the held-out test set, with perfect sensitivity and strong specificity.

Setting the Optimal Threshold

From the LASSO logistic regression, we determined the optimal threshold to predict whether a mass is benign or malignant based on the assigned probability. From the cost analysis and domain knowledge, we knew that the optimal threshold would maximize sensitivity. Using the `proc` library, we found that the optimal threshold to achieve perfect sensitivity is 0.351 using Youden's J statistic. Youden's method is a way to choose the "best" classification threshold by maximizing the combined performnace of sensitiivty and specificity. The optimal threshold is one that maximizes J = Sensitivity + Specificity - 1. For this diagnostic classification problem, Youden's J statistic chooses the threshold that provides a balance weight to correctly identifying malignant and benign cases. Since cancer diagnosis requires higher sensiviity to avoid missing malignant cases, we want the threshold to be less than 0.5 so the false negative rate is decreased, so 0.351 seems a reasonable value. Other benefits for using Youden's statistics is that it is not distorted by how common cancer is in the dataset and it is transparet for clinicians.

Although perfect sensitivity often comes at the cost of poor specificity and many false positives, we also observe a high specificity of 0.98 with perfect sensitivity. Hence, we are less concerned with the model classifying every observation as benign at this threshold given that the specificity is also very high. For the rest of the paper, 0.351 will be the threshold used for classifying a tumor.

Table 1: Confusion Matrix

	Outcome	Benign	Malignant
0	Predicted Benign	60	0
1	Predicted Malignant	1	53

Model Diagnostics

The Confusion Matrix above uses the optimal threshold and achieves perfect sensitivity, a specificity of 0.98, and a 99% accuracy. This indicates strong overall model performance and although there are possible concerns about model overfitting, we are confident in our results given that we used 5-fold CV to train the data and then tested the data on the hold-out set.

Model performance was further evaluated using an ROC curve and a measure of AUC, providing a threshold-independent measure of discrimination. The AUC for the ROC curve is 0.9994 which is fairly high, giving us confidence in our model.

The last step after building the LASSO logistic model was extracting the features which had non-zero coefficients for predicting the probability of a tumor being benign or malignant. Below is a list of the 13 features (out of 30) which the LASSO model indicated were important for predicting the probabilities ordered in descending order of magnitude.

Table 2: Selected Features

Mean	SE	Worst
concave.points_mean	smoothness_se	concave.points_worst
compactness_mean	fractal_dimension_se	smoothness_worst
	compactness_se	symmetry_worst
	radius_se	concavity_worst
	texture_se	radius_worst
		texture_worst

In summary, the LASSO logistic model yielded a parsimonious subset of clinically interpretable predictors, which served two purposes:

1. Establishing a transparent, high-performing baseline classifier.
2. Guiding subsequent nonlinear modeling by identifying variables most strongly associated with malignancy.

Nonlinearity Assessment and Motivation for GAMs

While LASSO logistic regression provides interpretability through coefficients, it assumes linear relationships between predictors and the log-odds of malignancy. To assess whether this assumption was appropriate, we conducted a systematic investigation of potential nonlinearity.

For each predictor selected by the LASSO model, we binned the predictor into quantiles, computed observed malignancy proportions within each bin, and examined both probability-scale and log-odds-scale plots. Several predictors, most notably texture_se, compactness_se, fractal_dimension_se, texture_worst and concavity_worst, exhibited clear nonlinear patterns, suggesting that a purely linear model may be misspecified.

Generalized Additive Models (GAMs)

To flexibly model these nonlinear effects while preserving interpretability, we employed Generalized Additive Models (GAMs) with a binomial logit link. GAMs model the log-odds of malignancy as a sum of smooth functions of predictors, allowing nonlinear relationships to be captured without requiring explicit interaction terms.

We compared different spline bases (thin-plate regression splines, cubic regression splines, and thin-plate shrinkage splines) for individual predictors using AIC, REML scores, and visual inspection. Restricted Maximum Likelihood (REML) was used to compare the splines instead of Maximum Likelihood because REML provides less biased estimates for variance components. Although no spline performed definitively better across all metrics, cubic splines generally had lower AIC and REML across variables. Additionally, cubic splines had a lower range of plausible predicted probabilities and less wiggly than the other two splines. Thus, we chose to use cubic regression splines since consistently provided better stability and more realistic probability behavior, and were therefore selected.

After determining the appropriate spline to be cubic splines, we selected which features show strong violation of nonlinearity so we can include them in splines. All other variables which don't show strong evidence of nonlinearity will be included as main effects or interaction terms. To do this, we first binned each of the predictors and calculated the appropriate log-odds of malignancy within the training data. We then visualized the logit plot for each predictor to determine where we could see strong violations of nonlinearity.

In Figure 3, we observe that many of the variables to violate nonlinearity. Thus, we also want to quantitatively determine which features should be included as splines by comparing a GLM model to predict malignancy probability with a GAM using only a cubic spline with the selected feature. Then, we compared the change in AIC between the GLM and GAM as well as the p-value for the Chi-Squared test where the cutoff was set at $\alpha = 0.01$. Features which had a drop in AIC with a magnitude of at least 4 and a significant p-value were identified as features with strong evidence of non-linearity. The delta AIC value of 4 was selected using the convention that a drop in AIC greater than 4 is evidence that the GAM meaningfully improves the model fit.

variable	delta_AIC
fractal_dimension_se	15.179108
compactness_se	32.201843

variable	delta_AIC
concavity_worst	40.642836
texture_se	4.392266
texture_worst	5.291678

From the results above, the following features were included in the GAM as cubic splines: texture_se, compactness_se, fractal_dimension_se, texture_worst, and concavity_worst.

From here, the initial GAM model was built using cubic splines for 5 nonlinear features and including the other 8 features as main effects.

Table 4: Confusion Matrix

	Outcome	Benign	Malignant
0	Predicted Benign	60	1
1	Predicted Malignant	1	52

The model performance for this base GAM was assessed using the test dataset. The model has a 98% accuracy as well as a sensitivity of 0.981 and a specificity of 0.983. The model also has an area under the curve extremely close to 1 indicating the model has a strong ability to distinguish between classes. For cancer diagnoses we want to maximize sensitivity, so we next consider adding interactions to improve our model to further increase sensitivity.

Adding Interaction Effects

Although GAMs capture nonlinear marginal effects, additive models may miss important interactions. Guided by domain knowledge and prior findings on the Wisconsin Breast Cancer Dataset, we extended the additive GAM to include tensor-product interaction terms between clinically meaningful feature pairs like tumor size and boundary irregularity. Candidate interactions included: smoothness × concavity, radius × concave points, and fractal dimension × concavity.

Justification for these interactions are as follows: 1. 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. 2. radius_worst x concave.points_mean: Concave points measure sharp inward curves on the tumor boundary whereas radius measures size. Large lesions with many concave points are signals of malignancy. 3. fractal_dimension_se x concavity_worst: Fractal dimension relates to edge complexity and high fractal dimension with strong concavity could indicate irregular growth.

These interactions were implemented using tensor-product smooths to accommodate differing scales and nonlinear dependence. Model comparison using AIC and likelihood ratio tests demonstrated a statistically significant improvement in fit over the purely additive GAM. We fit several different models with different interaction terms, but the final model was selected for having the largest drop in AIC as well as a significant p-value from the Chi-Squared test.

The final p-value from the Chi-squared test for the model with interactions compared to the model without was 0.002 which is significant past the $\alpha = 0.01$ level.

We then assessed the model performance of this GAM with interactions on the test dataset.

Table 5: Confusion Matrix

	Outcome	Benign	Malignant
0	Predicted Benign	60	1
1	Predicted Malignant	1	52

We observe that the Confusion matrix is identical to the GAM without interactions, but this is likely due to the small sample size. Additionally, the accuracy, sensitivity, and specificity remain the same but this can likely also be attributed to the small sample size and the strong performance of the original model without interactions.
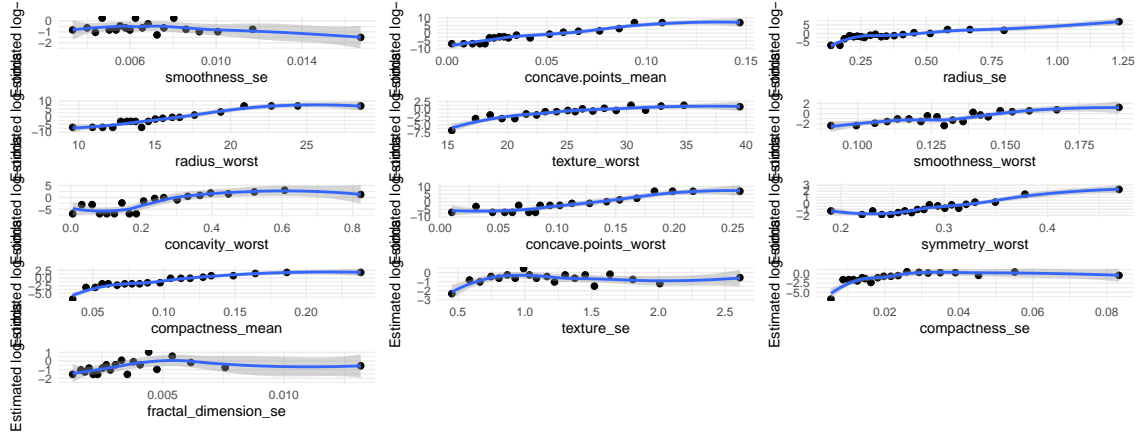
Figure 2: Log-Odds Plots

Final Model Selection

The GAM with selected nonlinear terms and domain-motivated interactions was chosen as the final model, as it achieved strong predictive performance (high AUC and balanced sensitivity/specificity), captured clinically plausible nonlinear and interaction effects, remained interpretable through smooth and interaction visualizations, and aligned with the project's emphasis on transparency and trust in medical decision making. The final sensitivity was 0.981 which is very high since 98.1% of malignancy cases were correctly identified in the model, so only 1.9% of cases are false negatives which we want to minimize.

Downstream Uses of the Model

Our final predictive model can serve several downstream purposes. Clinicians may use a model derived malignancy probability as a preliminary triage tool to help determine urgency of follow up imaging or need for biopsy. Researchers may use our inferred feature importance patterns to better understand structural characteristics of malignant tissue, signaling which measurements warrant closer attention in future imaging protocols. In healthcare settings, such tools can also guide resource allocation by highlighting high risk cases sooner, especially when imaging workloads are high (with over 42 million mammograms performed annually across the US and UK, Nature 2020).

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 positives which have been long recognized as a challenge in mammographic screening.

Results

Key Findings

Since both our final GAM models have essentially identical predictive results, we have two potential models for clinicians to choose between. If they want a simpler model with fewer features, the GAM without interactions would be a great predictive model to use. On the other hand if they want to prioritize more complex models to explain ways that certain features interact, they should use the GAM with interactions.

Although the LASSO logistic model wasn't selected as our final model for predictions, the features it identified as key features are very helpful for understanding which features of a mass are most indicative of the mass' probability of being malignant. Identifying these features allows clinicians to grasp which aspects of the mass they should focus on when they receive the data by seeing if any of these selected 13 features have values outside of a normal range.

Limitations

One limitation is that the data is from 1995. While our work provides proof of concept modeling, future work can expand upon these findings by utilizing modern imaging technology (3D mammography, MRI, MBI) that can can detect additional

features. Additionally, we are working with a fairly small sample size of observations (569 total). We would expect to see better predictive accuracy and improvement in sensitivity when adding interactions if we had a larger test dataset.

Another limitation is that the predictors are based on computed summary statistics for each tumor (mean, SE, "worst"). Although having these values provide general trends, they may fail to capture specific nuances to tumor characteristics compared to using raw imaging data. An additional limitation to the models selected, lasso logistic regression and GAMs, is that important nonlinear or interaction effects may be overlooked, since regression relies on linear assumption and GAMs may fail to capture complex, higher-order interactions between tumor features.

Conclusion

Future Analysis

References

Github: https://github.com/gradypurcell/STA-325-Final-Project.git