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.

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.

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, the dataset is only moderately imbalanced, so rebalancing procedures such as SMOTE or class weighting are not necessary for this analysis.

The dataset 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).

Methodology

Our methodology consists of three 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.

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 clear interpretability. 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. Decision trees offer high interpretability, producing decision rules that patients and clinicians can easily understand. This structure directly supports the project's interpretability goals and meets the rubric requirement for communicating results to individuals outside the field.

Regularized Logistic Regression and Feature Selection

We first fit a LASSO 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 model achieved approximately 98% accuracy on the held-out test set, with strong sensitivity and specificity. Model performance was further evaluated using the ROC curve and AUC, providing a threshold-independent measure of discrimination.

Importantly, the LASSO 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 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 smoothness_se, texture_worst, concavity_worst, and radius_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 and cubic regression splines) for individual predictors using AIC, REML scores, and visual inspection. Cubic regression splines consistently provided better stability and more realistic probability behavior, and were therefore selected.

Formal model comparisons between generalized linear models and GAMs using AIC differences and likelihood ratio tests, confirmed strong evidence of nonlinearity for several predictors. Based on these results, smooth terms were retained for smoothness_se, texture_worst, and concavity_worst. Other predictors entered the model linearly.

Final Additive GAM Specification

The resulting additive GAM incorporated nonlinear smooth terms for predictors with strong evidence of nonlinearity, along with linear effects for remaining LASSO-selected features. Model diagnostics, including residual plots, k-index checks, and concurvity analysis, indicated adequate fit and no severe violations of modeling assumptions. Model performance was evaluated using confusion matrices, ROC curves, and AUC. The GAM demonstrated strong discriminatory ability and improved calibration relative to the purely linear model, while remaining interpretable through smooth effect plots.

Incorporating 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: concave points × radius, concavity × radius, smoothness × concavity, and texture × concavity.

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.

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.

Cost Analysis and Threshold Selection (add here once done)

We then 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.