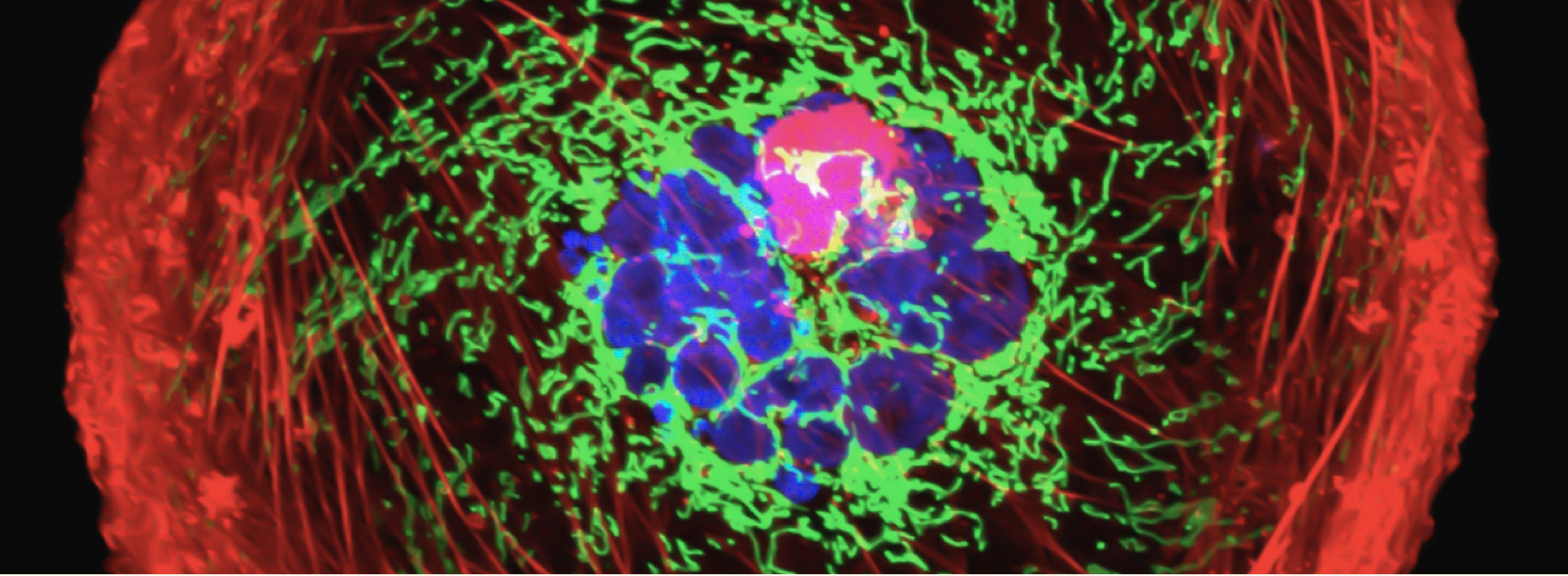


Breast Cancer Prediction Using Bayesian Logistic Regression and Generalized Additive Models

ALI RAJABIMEHR & CRAIG MANNING

STAT802 ADVANCED TOPICS IN ANALYTICS

Introduction

Breast cancer prevalence

Breast cancer is the world's most prevalent cancer, with 2.3 million women diagnosed and 685,000 deaths in 2020.

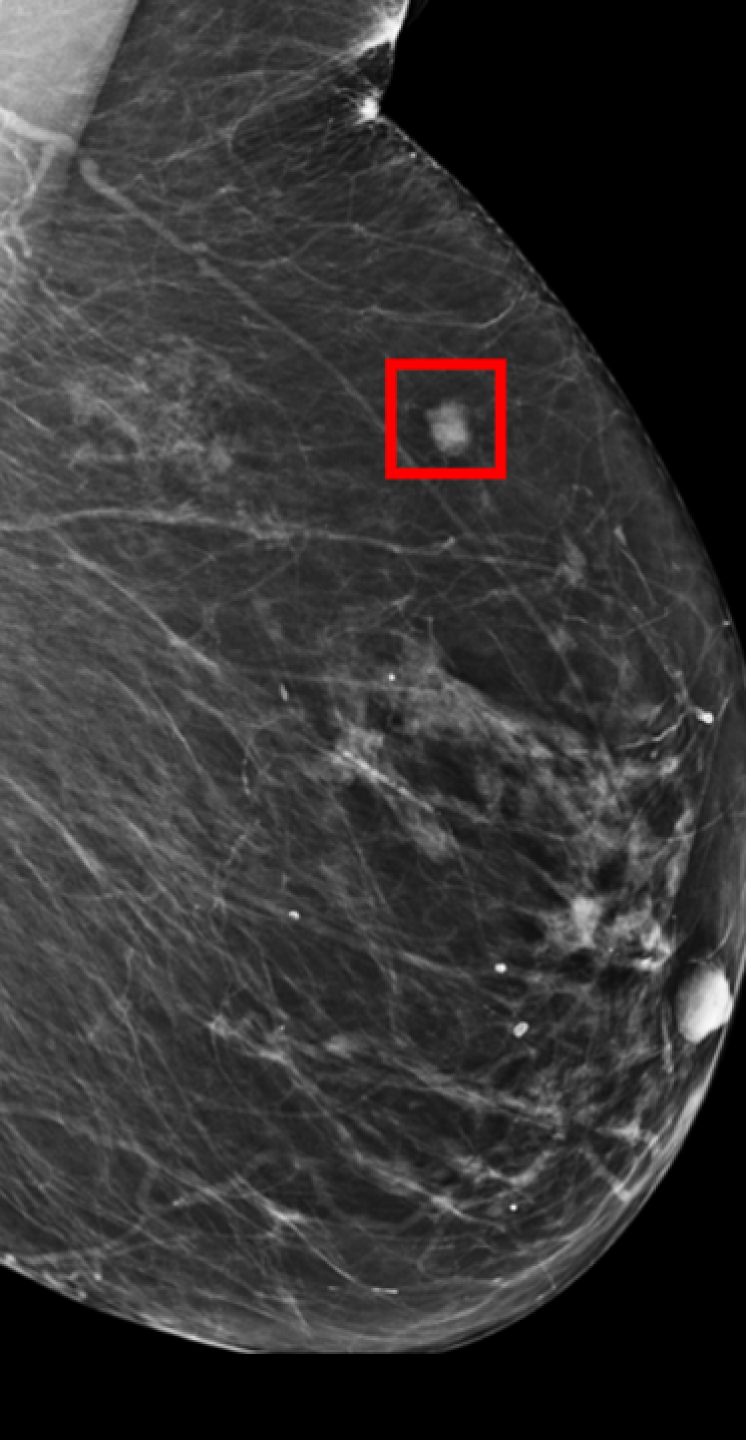
Importance of early detection

Early detection is crucial for successful treatment and improved survival rates.

Research study

This study aims to improve breast cancer diagnosis by predicting breast cancer using features derived from Fine Needle Aspiration (FNA).

Fine Needle Aspiration (FNA) is a diagnostic procedure where a thin needle is inserted to extract cells from a suspicious breast lump for microscopic examination.



Study overview

Identify the predictors of breast cancer

Identify the key features derived from FNA imagery, that could significantly predict breast cancer

Estimate the probability of cancer

Estimate the probability of cancer by varying the predictor values

Bayesian Logistic Regression model

Assess the predictive accuracy of the Bayesian Logistic Regression model

Generalised Additive Models (GAMs)

Compare the results of the Bayesian Logistic Regression Model with GAMs and discuss the advantages and disadvantages of each model

FNA features

- **Radius**

  The average distance from the center to the points on the cell nucleus perimeter.

- **Texture**

  The variance of the gray-scale values within the cell nucleus, indicating coarseness of the cell's interior.

- **Perimeter**

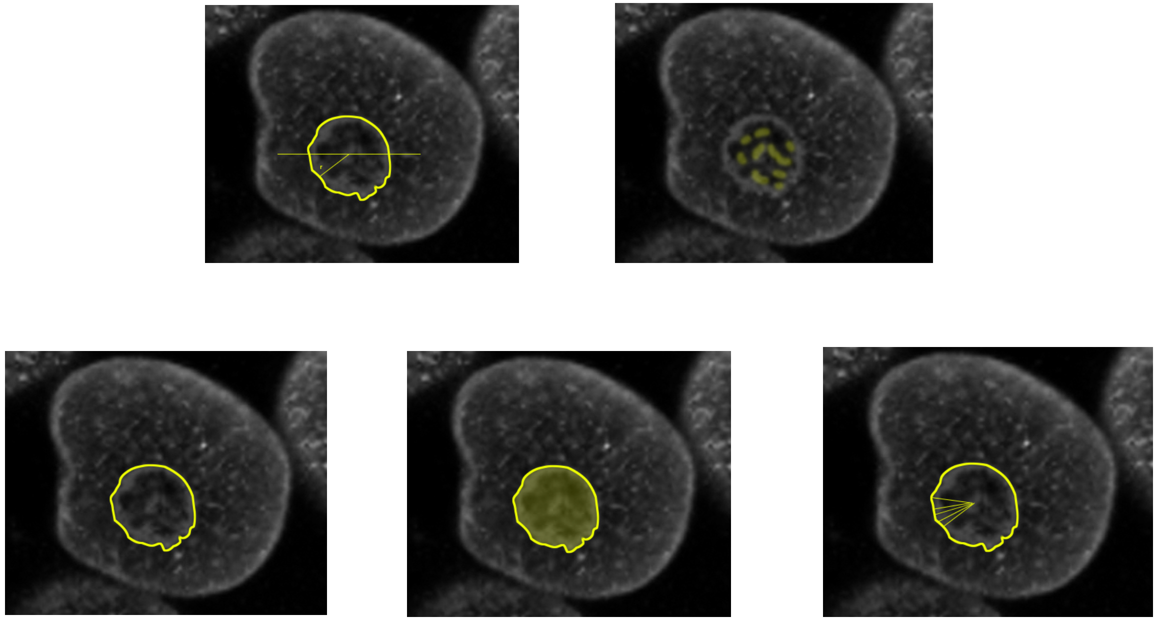
  The total distance between points on the cell nucleus boundary, which provides information about the irregularity of the cell shape.

- **Area**

  The number of pixels within the cell nucleus boundary, representing the overall size of the cell nucleus.

- **Smoothness**

  The variation in the radius lengths, reflecting the smoothness or irregularity of the cell nucleus boundary.



FNA features

- **Compactness**

  The ratio of the perimeter squared to the area, providing a measure of the cell's shape and deviation from a perfect circle.

- **Concavity**

  The severity of the concave portions of the cell nucleus contour, indicating abnormal cell shape.

- **Concave Points**

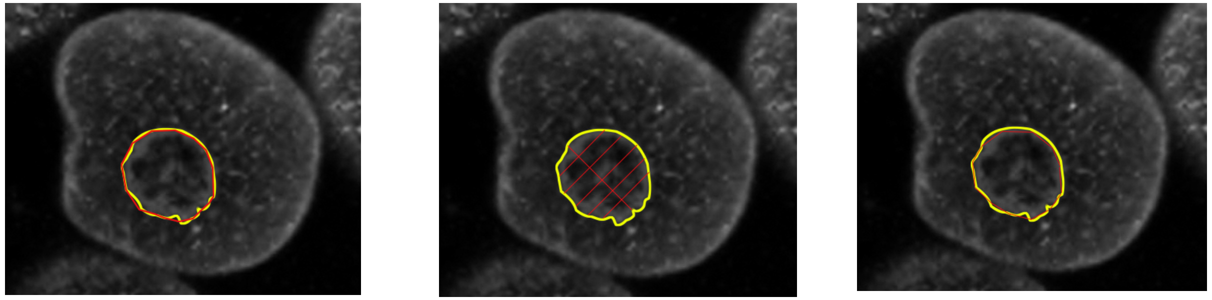
  The number of concave portions of the cell nucleus contour, further characterizing the abnormality of the cell shape.

- **Symmetry**

  The similarity in shape and size between opposite sides of the cell nucleus, reflecting the overall symmetry of the cell.

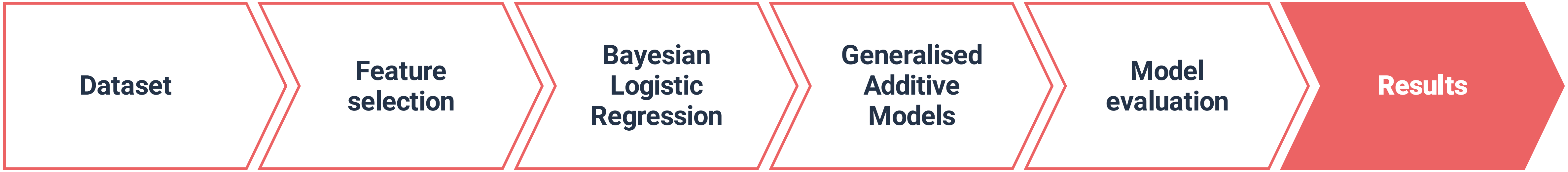
- **Fractal Dimension**

  A measure of the complexity of the cell nucleus boundary, providing information about the irregularity and roughness of the cell outline.



The **mean**, **standard error** (Variability across different cells) and **worst** values (The mean of the three largest values) were recorded for each feature.

Research design



Dataset

A dataset of fine needle aspiration (FNA) images of breast tissue, obtained from the University of Wisconsin Hospitals, Madison. 569 cases (357 benign, 212 malignant) with 30 features per case

Feature selection

Feature selection using correlation, ANOVA, and VIF identified the most relevant predictors for breast cancer diagnosis.

Standardisation ensured all features had a similar influence on the model by removing scaling difference

Bayesian Logistic Regression

Bayesian logistic regression model fitted using JAGS in R

Five sets of priors: non-informative to informative with varying certainty

MCMC settings: 4 chains, 5,000 iterations, burn-in of 1,000

Generalised Additive Models

AMs fitted to the selected features to capture non-linear relationships

Smoothing functions used to model the effect of each predictor on the diagnosis

Model performance compared to the Bayesian logistic regression

Model evaluation

Bayesian Information Criterion (BIC) used for model comparison

95% credible intervals identified significant predictors

Posterior predictive checks demonstrated the effect of varying predictor values

Convergence diagnostics (trace plots, effective sample sizes, Gelman diagnostics) assessed MCMC convergence

Results

Significant predictors identified.

Predictive probabilities calculated for varying predictor mean values

Discussion on the potential of these models for aiding in early breast cancer detection and diagnosis

Key findings

Bayesian Logistic Regression

- Best model: Informative prior (3) with the lowest BIC
- Accuracy: 93.15%, Sensitivity: 91.79%, Specificity: 93.92%
- Significant predictors: concave points worst, radius mean, compactness mean, perimeter standard error

Model performance

- Both Bayesian logistic regression and GAMs demonstrated high accuracy, sensitivity, and specificity in predicting breast cancer diagnosis
- GAMs showed slightly better performance, suggesting the importance of capturing non-linear relationships

Generalised Additive Models

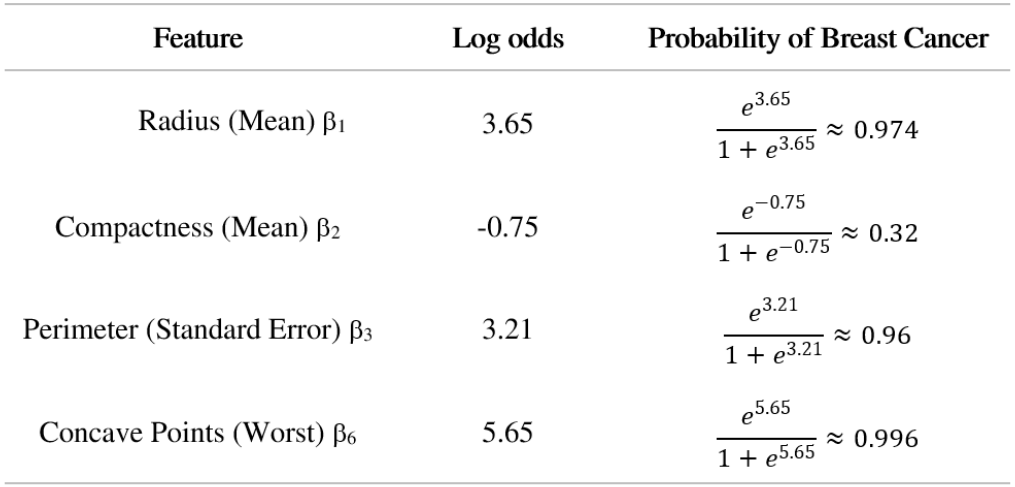
- Slightly outperformed the Bayesian model
- Accuracy: 94.90%, Sensitivity: 94.63%, Specificity: 95.05%
- Non-linear relationships captured using smoothing functions

Significant predictors

- Concave points (Worst)
- Radius (Mean)
- Compactness (Mean)
- Perimeter (Standard Error)

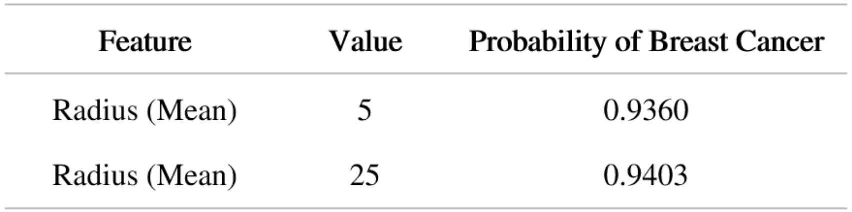
Key results

Bayesian model - Informative prior (3)



Feature	Log odds	Probability of Breast Cancer
Radius (Mean) β_1	3.65	$\frac{e^{3.65}}{1+e^{3.65}} \approx 0.974$
Compactness (Mean) β_2	-0.75	$\frac{e^{-0.75}}{1+e^{-0.75}} \approx 0.32$
Perimeter (Standard Error) β_3	3.21	$\frac{e^{3.21}}{1+e^{3.21}} \approx 0.96$
Concave Points (Worst) β_6	5.65	$\frac{e^{5.65}}{1+e^{5.65}} \approx 0.996$

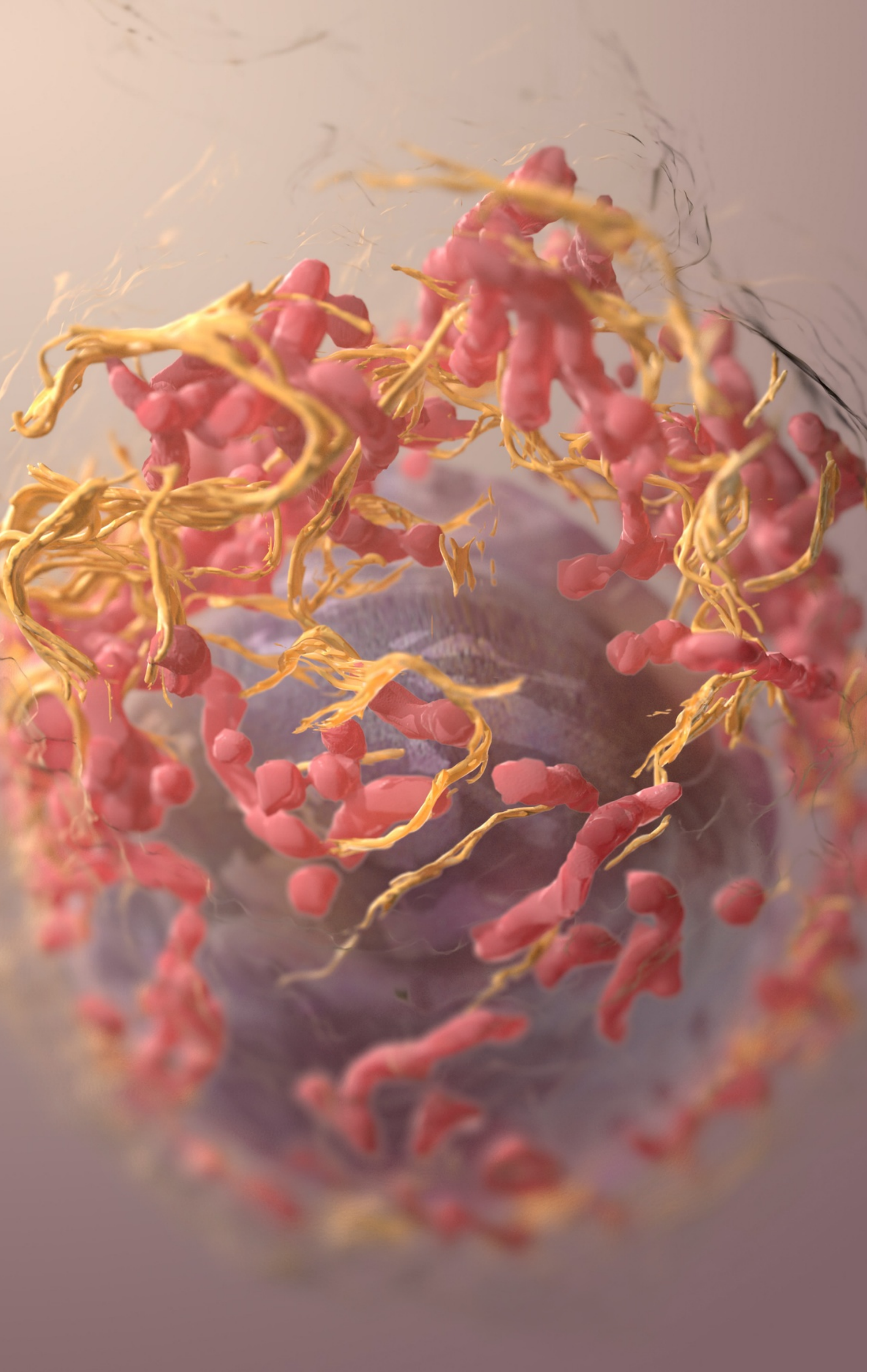
Predictive probabilities of breast cancer



Feature	Value	Probability of Breast Cancer
Radius (Mean)	5	0.9360
Radius (Mean)	25	0.9403

Most Significant Feature: Concave Points (Worst)

- Highest absolute coefficient value (β6) with credible interval [2.83, 5.65]
- Indicates that a higher number of concave points significantly increases the likelihood of malignancy
- Aligns with the irregular shape often seen in malignant tumors
- A one-unit increase in Concave Points (Worst) results in a 99.60% probability of breast cancer, holding all other features constan



Discussion

Key findings

- Bayesian logistic regression and GAMs effectively predicted breast cancer diagnosis using FNA image features
- The Bayesian model with informative prior (3) proved to be the best fit, highlighting the importance of incorporating prior knowledge
- GAMs slightly outperformed the Bayesian model, suggesting the value of capturing non-linear relationships

Interpretation

- Concave points worst, radius mean, compactness mean, and perimeter standard error were significant predictors
- Higher values of concave points worst and radius mean increased the probability of malignancy
- These findings align with the irregular shape and size of malignant tumors observed in clinical practice

Limitations

Single source

The dataset was produced in 1993, which may not reflect most recent advancements in breast cancer diagnosis and imaging techniques

Further validation on more up-to-date datasets is needed to confirm results

Small dataset

The study used a relatively small dataset, which may not represent the general population.

Conclusion

Implications

- The findings highlight the potential of using FNA image features and advanced statistical models to support early breast cancer detection and diagnosis
- The identification of significant predictors can guide clinical decision-making and help prioritise patients for treatment
- The age of the dataset should be considered when interpreting the results and their applicability to current clinical practic

Recommendations

- Validate the results on larger, more diverse, and up-to-date datasets to assess the generalizability of the findings
- Collaborate with healthcare professionals to develop and evaluate user-friendly tools that integrate predictive models into clinical practice
- Conduct further research using more recent datasets, whilst considering the existing work in the field

Both Bayesian Logistic Regression and Generalised Additive Models show promise in identifying key features that can aid in the early detection and diagnosis of breast cancer