Predicting Breast Cancer: A Ridge Regression Approach

Kyle Paolo Bautro[1], Joshua David Oraiz[1], Symond Ridge Fernandez[1], Yochanan Bangoy[1], Carl Dane Penano[1]

[1] BS Data Science, USTP – Cagayan de Oro Campus

Abstract. *An accurate tumor prediction is very important for the treatment of breast cancer. The present study builds a Logistic Ridge Regression model based on the six predictors from the Wisconsin Breast Cancer Dataset. The method Logistic Ridge Regression makes use of L2 regularization and results in stable coefficients, thus, performance is not hindered. With respect to logistic regression, stability and interpretability are enhanced. VIF values from 2.53 to 11.43 support the appropriateness of regularizing the correlated tumor features. The model shows that regularization has a significant impact on stabilization of the correlated features for prediction*

Keywords: *Breast cancer prediction, Ridge regression, Multicollinearity, Tumor characteristics, L2 regularization*

1 Introduction

Breast cancer remains a leading cause of mortality in women worldwide, making early detection crucial for treatment success. While imaging and pathology have improved, predicting tumor behavior still requires time-consuming clinical testing. Predictive models using tumor features can support earlier clinical decision-making.

This study develops a tumor prediction model using six diagnostic features: radius_worst, concavity_mean, concave.points_worst, compactness_mean, fractal_dimension_mean, and radius_mean. These features represent tumor size, shape irregularity, and structural complexity are all linked to malignancy. We selected highly correlated but non-redundant variables to balance clinical interpretability with statistical validity.

Ridge Regression serves as our primary method because it stabilizes coefficient estimates when predictors are multicollinear, unlike ordinary least squares (OLS). While OLS produces unstable coefficients with correlated features, Ridge applies L2 regularization to reduce overfitting and *improve generalization. This approach handles interrelated tumor features while maintaining predictive performance. Using the Wisconsin Breast Cancer Dataset from Kaggle, we implement Ridge Regression in R and compare its coefficients against OLS to demonstrate regularization benefits.*

This study is significant because it shows how using tumor-derived features and regularized regression techniques can enhance predictive modeling in breast cancer. By concentrating on quantitative measurements of breast tissue tumors, the method facilitates clinical decision-making with minimal and easily accessible data, vital for early risk assessment.

2 Literature Review

2.1 Predictors of Breast Cancer

Breast cancer exhibits various biological and morphological characteristics that influence prognosis and treatment. This section reviews key quantitative predictors used in diagnostic models.

Radius Measurements

Radius_mean and radius_worst consistently demonstrate strong predictive power across classification models, with both metrics appearing clearly in machine learning approaches [1,2,3].

Concavity Features

Tumor boundary irregularity is captured through concavity_mean and concave.points_worst. Research shows these features correlate with tumor aggressiveness and significantly impact classification accuracy [2,4,5].

Compactness

Compactness_mean measures how closely tumor shape resembles a defined geometric form. This feature correlates with tumor behavior and has been successfully applied in machine learning models for risk assessment [3,6].

Fractal Dimension

This metric represents the complexity of the boundary, which is closely related to the potential of the tumor to grow. The application of fractal *geometry to diagnostic procedures reveals the irregularity of the tumor margins, hence enhancing the diagnostic process [7].*

The morphological characteristics that are interlinked this way provide the basis for precise detection and prediction of breast cancer [2,3,7].

2.2 Predictive Modeling of Breast Cancer

Predictive modeling for breast cancer has advanced through both traditional statistical methods and modern machine learning techniques.

Traditional Regression Models

Logistic regression remains popular in breast cancer prognosis due to its interpretability in binary classification while effectively identifying key predictors [8].

Machine Learning Approaches

Random Forests, SVMs, and Neural Networks have shown superior performance by capturing non-linear relationships and complex interactions in high-dimensional tumor data [9,10].

Regularization Techniques

Ridge regression stabilizes correlated predictors by reducing variance without substantially increasing bias [11]. Lasso performs variable selection by eliminating irrelevant features [11]. Elastic Net combines both approaches, particularly useful for high-dimensional datasets [11]. These techniques manage bias-variance tradeoffs, which is critical when prediction accuracy directly affects patient outcomes [11].

Handling Highly Correlated Predictors

Ridge and Elastic Net specifically address multicollinearity through coefficient limitation or grouped variable selection. Careful feature selection remains essential for building accurate and interpretable models [11,12].

Stabilizing Coefficient Estimates

Regularization techniques like Ridge and Lasso stabilize coefficient estimates by penalizing regression loss functions, managing bias-variance tradeoffs to produce reliable, interpretable models. This is critical in breast *cancer detection where prediction accuracy greatly affects patient management and outcomes [11].*

2.3 Research Gaps

Despite advances, several areas need investigation. Ridge regression effectively handles multicollinear data but remains underutilized in clinical settings compared to other methods [13]. A persistent challenge is balancing prediction accuracy with interpretability—complex algorithms provide precision but complicate clinical adoption [9]. New opportunities exist in integrating minimally invasive predictors like biosignatures with imaging data, potentially enabling earlier detection and personalized treatment approaches [14].

3 Methodology

3.1 Research Approach

This study uses a comparative approach to evaluate Logistic Ridge Regression for breast cancer prediction. Ridge addresses multicollinearity among tumor predictors, stabilizing coefficients, reducing overfitting, and retaining all relevant features. The study is guided by three research questions:

1. Can Ridge Regression stabilize coefficient estimates for correlated tumor features?
2. How much does Ridge reduce multicollinearity based on VIF analysis?
3. How do Ridge and ordinary logistic regression predictions compare?

3.2 Problem Breakdown

Diagnosing breast cancer is significantly reliant on the shape of the tumor, but determining through various factors the predictability of it being cancerous or not is not easy. In this study, six quantifiable factors are used, with the malignancy label represented by either 0 for a benign condition or 1 for malignant growths. This approach is data-driven for accurate diagnostic purposes.

3.3 Model Selection: Ridge Regression

Linear regression was initially considered, but the categorical outcome made OLS unsuitable. Ridge-regularized logistic regression was chosen instead, predicting probabilities while the L2 penalty stabilizes coefficients and improves generalization.

$$-\sum_{i=1}^{n}\Big[y_i\log(p_i)+(1-y_i)\log(1-p_i)\Big]+\lambda\sum_{j=1}^{p}\beta_j^2$$

$$\textit{where:}\quad p_i=\frac{1}{1+e^{-(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}}$$

Notation & Explanation

y_i → observed outcome (0 = benign, 1 = malignant)

x_i□ → value of the j^{th} predictor for the i^{th} observation

β□ → coefficient/weight for predictor j

β_0 → intercept (baseline log-odds when all predictors are zero)

p_i → predicted probability of malignancy for observation i

Objective function

First term:
- Σ [y_i log(p_i) + (1 - y_i) log(1 - p_i)] → negative log-likelihood, measures model fit

Second term:
λ Σ β□^{2} → Ridge (L2) penalty, shrinks coefficients to prevent overfitting

λ → regularization strength; higher λ → more shrinkage

Why Logistic Ridge Regression is appropriate:

Because of the strong multicollinearity within the tumor predictors, ordinary logistic regression produced unstable coefficients. Ridge Logistic Regression shrinks the coefficients by imposing an L2 penalty, reduces overfitting, and keeps all the predictors. Cross-validation decided the best regularization that improved the model's generalizability and reliability for malignancy prediction.

3.4 Workflow

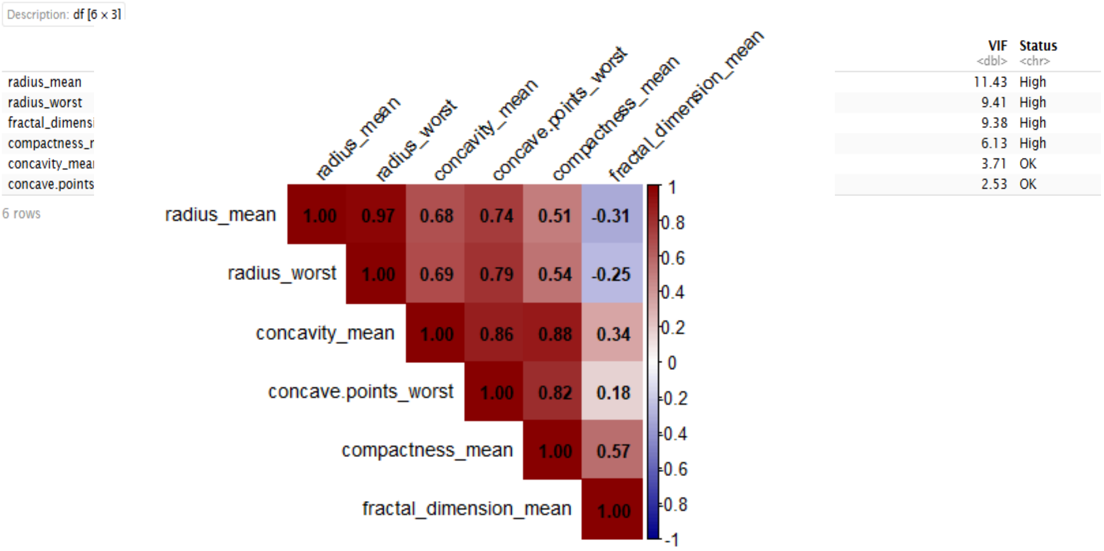
Initial Variable Screening

All numeric variables in the dataset were first identified and evaluated using an initial Logistic Regression model. This step provided an overview of variable behavior and allowed early detection of unstable coefficients or potential multicollinearity issues.

Multicollinearity Assessment Using VIF and Correlation

VIF values were computed to assess multicollinearity, and variables with high VIF were flagged for review. Pairwise correlations were also analyzed, with a correlation matrix visualizing predictor relationships to guide variable selection.

***Table 1**. Variance Inflation Factors (VIF) for breast cancer predictors*



Description: df [6 × 3]

	VIF <dbl>	Status <chr>
radius_mean	11.43	High
radius_worst	9.41	High
fractal_dimensi	9.38	High
compactness_r	6.13	High
concavity_meai	3.71	OK
concave.points	2.53	OK

6 rows

	radius_mean	radius_worst	concavity_mean	concave.points_worst	compactness_mean	fractal_dimension_mean
radius_mean	1.00	0.97	0.68	0.74	0.51	-0.31
radius_worst		1.00	0.69	0.79	0.54	-0.25
concavity_mean			1.00	0.86	0.88	0.34
concave.points_worst				1.00	0.82	0.18
compactness_mean					1.00	0.57
fractal_dimension_mean						1.00

Fig 1. *Pearson correlation coefficients among selected predictors.*

Data Pre-processing for Ordinary and Ridge Logistic Regression

The dataset's numeric variables were recognized and assessed initially with the help of a basic Logistic Regression model. This phase not only gave a glimpse of the variables' behavior but also made it possible to spot early on the unstable coefficients or the existence of multicollinearity problems.

Lambda Values for Logistic Ridge Regression

In the case of Ridge Logistic Regression, lambda.min produces bigger coefficients, which help to detect even the slightest patterns resulting in a higher predictive accuracy. On the other hand, lambda.1se makes the coefficients small leading to a simple and robust model. As our goal was to obtain the best prediction, lambda.min was selected for the final model.

Model Fitting and Summary Interpretation

Ordinary and Ridge Logistic Regression models were fitted using the selected predictors. Comparing coefficients shows how Ridge reduces inflation from multicollinearity while retaining each feature's influence on malignancy risk.

Evaluation Framework

No external validation set was utilized, yet model stability and multicollinearity were evaluated via comparisons of coefficients between ordinary and Ridge Logistic Regression. Furthermore, the likelihood of overfitting was tested by contrasting training error with cross-validated deviance.

4 Results

This section presents the model outputs obtained from fitting Logistic Ridge Regression compared to Ordinary Logistic Regression. The analysis focuses on multicollinearity assessment, coefficient comparison, and model stability.

4.1 Coefficient Comparison (OLR vs Ridge)

Table 2. *Differences in Coefficients: OLR vs Ridge Logistic Regression*

Predictor	OLS	Ridge	Difference
Intercept	**-13.0692**	**-7.5009**	**-7.5009**
radius_mean	**-1.5444**	**0.1811**	**0.1811**
radius_worst	**2.1799**	**0.1785**	**0.1785**
concavity_mean	**21.4992**	**6.6560**	**6.6560**
concave.points_worst	**34.2848**	**13.3451**	**13.3451**
compactness_mean	**-4.7525**	**6.5510**	**6.5510**
fractal_dimension_mean	**-103.9200**	**-23.5245**	**-23.5245**

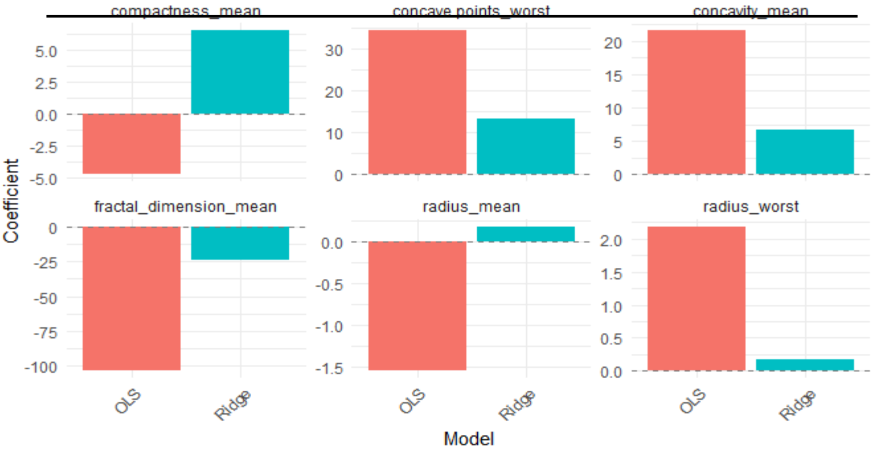


Fig 2. Bar plot of OLR and Ridge coefficients for selected predictors.

Interpretation

The intercept has moved from (-13.069 to -7.501) thus its pull towards zero has resulted in the reduction of extreme baselines. The variations in the mean radius and worst radius (0.181 & 0.178) are also mitigated, thus having a stabilizing effect. The large positive coefficients associated with concavity_mean and concave.points_worst (21.499 → 6.656; 34.285 → 13.345) are lessened which in turn prevents overfitting. On the other hand, compactness_mean does a complete turnaround from negative to positive (-4.753 → 6.551), while fractal_dimension_mean experiences a dramatic drop (-103.92 → -23.525), all these contributing to the stabilization of the model. In the end, Ridge reduces extreme OLS coefficients, variances and overfitting, and preserves the directions of predictions, which makes it easier to see the features that are vulnerable to collinearity or outliers.

4.2 Model Stability

Ridge Logistic Regression was responsible for coefficient variance reduction and extreme value moderation leading to more stable and interpretable estimates than the ordinary logistic regression. All predictors were kept and the negative effect of multicollinearity was reduced by the process of overfitting minimization.

4.3 Cross-Validation Outcome

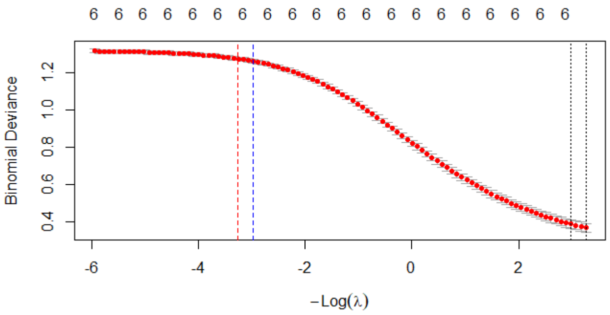


Fig 3. *Cross-validation results for Ridge Logistic Regression with selected λ = 0.0384.*

Through cross-validation, the optimal regularization strength was determined to be λ=0.0384 (lambda.min), which resulted in the least binomial deviance. This number was then utilized to train the final Ridge Logistic Regression model, resulting in stabilized coefficients and greater generalization.

4.4 Model Generalization and Overfitting Assessment

Table 3. *Overfitting and Regularization Assessment Summary*

Metric	*Value*
Training Log Loss (lambda.min)	***0.1813***
Cross-Validated Log Loss (lambda.min)	***0.3665***
Cross-Validated Log Loss (lambda.1se)	***0.3665***
Difference (lambda.min vs lambda.1se)	***0.0196***

Training vs Cross-Validation Performance

The training log loss is lower than the cross-validated log loss, which is expected and indicates that the model does not overfit the training data. The moderate gap suggests effective regularization and good generalization.

Lambda.min vs Lambda.1se Comparison

The cross-validated losses for lambda.min and lambda.1se are very close, indicating strong model stability. Lambda.min was selected as it achieves the lowest validation error while maintaining robustness.

5 Conclusion

The researchers developed a Logistic Ridge Regression model aiming to predict breast cancer malignancy based on six attributes of tumor structure obtained from medical imaging. The predictors chosen showed high multicollinearity which caused ordinary logistic regression to give unstable and inflated estimates of coefficients. The use of Ridge regularization, however, eliminated this problem by reducing the coefficient of the extreme values, hence lowering the variance of the model, and still retaining all the biologically relevant predictors. The parameters that were estimated by the Ridge model were more stable and interpretable when compared to those produced by ordinary logistic regression while the overall directional influence of every feature was maintained.

Cross-validation confirmed strong generalization performance, supporting lambda.min for maximum predictive accuracy. Importantly, significant tumor features in terms of size, concavity, compactness, and surface roughness retained their diagnostic power after regularization. The downside is the limitation of lacking an external validation, though internal diagnostics indicate that the model gives good generalization. In the future, it would be great that possibly there will be works validating the findings of this model on breast cancer dataset from different medical centers, to compare the *performance with Lasso and Elastic Net models using the same features would clarify when each method works best for better predictive results.*

References

1. Author, F.: Article title. Journal 2(5), 99–110 (2016).
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).
4. Author, F.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010).LNCS Homepage, http://www.springer.com/lncs, last accessed 2016/11/21.

Ref list

[1] Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. SN Applied Sciences, 3(2). https://doi.org/10.1007/s42452-021-04148-9

[2]Sahu, D., & Pandey, R. (2025). Research based exploration of breast cancer data ease on machine learning outlook. Journal of Neonatal Surgery, 14(7S), 396–404. https://doi.org/10.52783/jns.v14.2422

[3]Abdolrazzagh-Nezhad, M., & Izadpanah, S. (2023). A new fuzzy bio-inspired based classification to cancer detection. https://doi.org/10.21203/rs.3.rs-3376596/v1

[4]Gastounioti, A., Conant, E., & Kontos, D. (2016). Beyond breast density: A review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. Breast Cancer Research, 18(1). https://doi.org/10.1186/s13058-016-0755-8

[5]Nguyen, B., Le, N., Trinh, N., Le, T., & Pham, T. (2022). Breast cancer diagnosis based on detecting lymph node metastases using deep learning. Science and Technology Development Journal. https://doi.org/10.32508/stdj.v25i2.3894

[6]Mazo, C., Aura, C., Rahman, A., Gallagher, W., & Mooney, C. (2022). Application of artificial intelligence techniques to predict risk of recurrence of breast cancer: A systematic review. Journal of Personalized Medicine, 12(9), 1496. https://doi.org/10.3390/jpm12091496

[7]Fuhr, M., Meyer, M., Fehr, E., Ponzio, G., Werner, S., & Herrmann, H. (2015). A modeling approach to study the effect of cell polarization on keratinocyte migration. PLOS ONE, 10(2), e0117676. https://doi.org/10.1371/journal.pone.0117676

[8]Li, Y., et al. (2023). Logistic regression and stepwise approaches for breast cancer prediction using clinical and imaging features. Computers in Biology and Medicine, 165, 107298. https://doi.org/10.1016/j.compbiomed.2023.107298

[9]Patra, A., Behera, S., Sethy, P., Barpanda, N., & Mahapatra, I. (2023). Breast tumor detection using efficient machine learning and deep learning techniques. Machine Learning and Applications: An International Journal, 10(2/3), 17–33. https://doi.org/10.5121/mlaij.2023.10302

[10]JB, A., T, D., & L, N. (2024). A comprehensive review of breast cancer detection using machine learning and deep learning classifiers. https://doi.org/10.59544/exzw6527/icgmes24p2

[11]Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[12]Elguoshy, A., Zedan, H., & Saito, S. (2025). Machine learning-driven insights in cancer metabolomics: From subtyping to biomarker discovery and prognostic modeling. Metabolites, 15(8), 514. https://doi.org/10.3390/metabo15080514

[13]Wang, J., et al. (2022). Two-stage penalized regression screening to detect biomarker-treatment interactions in randomized clinical trials. Biometrics, 78(1), 141–153. https://doi.org/10.1111/biom.13424

[14]Hassan, A., Naeem, S., Eldosoky, M., & Mabrouk, M. (2024). Multi-omics-based machine learning for the subtype classification of breast cancer. Arabian Journal for Science and Engineering, 50(2), 1339–1352. https://doi.org/10.1007/s13369-024-09341-7