Identification of a Predictive Model for Breast Tumor Classification

Sritan Vemuru, Sreekar Kompella, Srikar Sannidhi, Aaron Patel

December 4, 2024

Abstract

This study examined the application of machine-learning techniques to classify breast tumors into malignant and benign tumors. The Breast Cancer Wisconsin (Diagnostic) dataset from the UC Irvine Machine Learning Repository was used to conduct this research, which aimed to develop a model to optimize the classification of breast tumors via imaging. Accurate tumor classification is crucial for developing personalized patient treatments and predicting patient outcomes. After data scaling was applied to each attribute, feature selection was applied using LASSO. Then, models including Logistic Regression, SVM, Random Forest, KNN, and Naive Bayes were used to predict tumor classification. The results showed that the Random Forest model exhibited superior performance relative to the other models. This study demonstrates that the selected attributes of this particular dataset are strong predictors of the accurate classification of tumors, but results are limited by the dataset having fewer malignant tumors and overall fewer samples. Future work will utilize larger, balanced breast tumor datasets to better examine the differences in these models' performances.

1 Introduction

Breast cancer is the second leading cause of cancer death for women in the United States[1], and it presents itself in patients with varying levels of progression. Benign breast tumors are non-cancerous and typically pose a lesser risk, often requiring less aggressive treatment. On the other hand, malignant breast tumors are cancerous, often growing rapidly, invading surrounding tissue, and potentially spreading to other parts of the body, meaning they require more intensive treatment. Accurate differentiation between benign and malignant breast tumors is a crucial component of developing personalized treatment plans and foreseeing patient prognoses, allowing for optimized care and better survival rates.

2 Literature Review

Machine learning has advanced cancer classification by utilizing pathological diagnosis (recognizing patterns from tissue samples) and imaging diagnosis. There are two primary types of cancer: malignant and benign. Benign tumors tend to be harmless, while malignant tumors are closely monitored. (Patel). Studies have shown that the sigmoid kernel-based Support Vector Machine (SVM) achieves incredible accuracy rates, but this does not account for future optimizations. Employing these optimization techniques or incorporating neural network models could significantly enhance the classification abilities of SVM for breast cancer diagnosis (Jia et al.). In the future, deep learning is expected to play a pivotal role in clinical diagnostics. It uses three primary categories: Supervised learning, which uses labeled data to predict outcomes

based on prior information; Unsupervised learning, which uncovers hidden patterns in data without relying on labeled responses; and Reinforcement learning, which employs a system of rewards or penalties to guide actions. Using these approaches will enable the generation of diverse datasets and improve the performance of predictive models. By recognizing complex patterns, deep learning techniques have the potential to detect cancers that were previously overlooked by traditional methods (Iqbal et al.).

3 Dataset Preparation and Preprocessing

The data selected for this exploration, the Breast Cancer Wisconsin (Diagnostic) datasets, was sourced from the UC Irvine Machine Learning Repository, funded by the National Science Foundation. The dataset includes 569 instances and 30 attributes, all of which are numeric, and the outcome variable of interest—benign_0__mal_1. These numerical attributes describe the characteristics of cell nuclei present in each digitized image of a fine needle aspirate (FNA) of a breast mass. The outcome variable, benign_0__mal_1, is the variable that the model is designed to predict, and its possible values are 1 or 0 (1 = "Malignant" and 0 = "Benign").

Data Scaling: Min-max normalization was applied to transform the data to better suit a wider range of models due to its increased numerical stability, feature compatibility, and avoidance of dominance. The values for each attribute then ranged between 0 and 1.

Correlation Matrix: A correlation matrix (Figure 1) was produced to evaluate the correlations between predictors. As observed in Figure 1, the correlations varied greatly, not exhibiting any particular pattern except for a trend of the attributes having stronger correlations with the outcome variable, benign_0__mal_1.

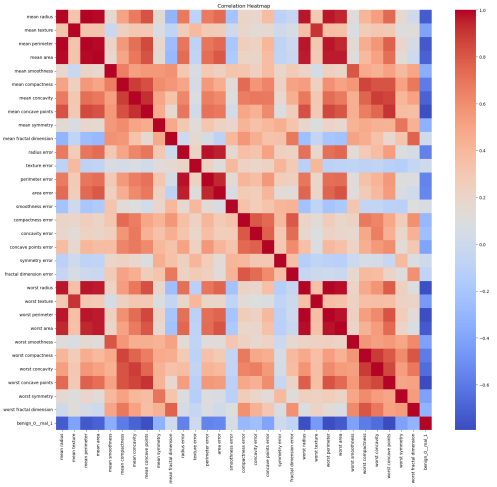

Figure 1: Correlation matrix for predictors and classification variable of Breast Cancer Wisconsin (Diagnostic) dataset.

Feature Selection: A heatmap of LASSO and Ridge Regression coefficients (Figure 2) was created to identify the best feature selection method and reduce the number of predictors. The LASSO technique uses a slightly different penalty than the Ridge penalty which is used in Tikhonov regularization. Instead of reducing the magnitudes but keeping all features, it zeroes out (removes) some features and keeps a select subset of them. On the other hand, Ridge regression simply follows the steps of Tikhonov regularization. As observed in Figure 2, LASSO has zeroed out significantly more

Table 1: Comparison of Model Performance

Method	Accuracy	ROC-AUC	False Negative Rate
Logistic Regression	0.956	0.998	0.0282
SVM	0.974	0.996	0.0141
Random Forest	0.974	0.998	0.0141
KNN	0.965	0.983	0.0282
Naive Bayes	0.965	0.997	0.0141

features than Ridge Regression, so it is the better feature selection method to use. By mapping the zeroed-out values to coefficients close to 0 under Ridge Regression, we can identify features that would have minimal effects on outcome variables. We can say that the mean smoothness, mean symmetry, and texture error features can be removed without extensively affecting model behavior. This reduces our number of predicting features to 27.

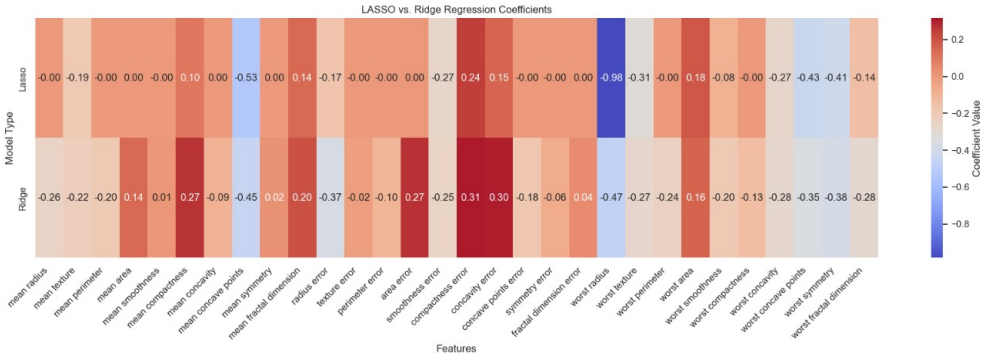

Figure 2: LASSO vs. Ridge Regression Coefficients

4 Model Selection and Development

Logistic Regression, Support Vector Machine (SVM), Random Forest, K-nearest neighbors (KNN), and Naive Bayes models were used for this data. To select the most appropriate model, the performance of the models with a split of the dataset into 80% training and 20% testing were determined and shown in Table 1.

5 Results

Logistic Regression Model: The Logistic Regression model presented a balanced performance with the lowest accuracy relative to the other models and the highest ROC-AUC score. It showed a good trade-off between precision and recall, with effective classification performance in both classes. It had a relatively higher false negative rate, suggesting that Logistic Regression could be a less valuable option for clinical applications.

Support Vector Machine (SVM) Model: The SVM had a high accuracy and a high ROC-AUC using the RBF kernel, but its strong performance can be computationally expensive. Therefore, it may be less ideal when real-time prediction is necessary or when interpretability is crucial. However, it was effective in classification, with a very low false negative rate, meaning it correctly classified most cases.

Random Forest Model: The Random Forest Model worked well with precision and

the same ROC-AUC score as the Logistic Regression model. This method coped well with the complex patterns in the data. The low false negative rate indicates that it performed well in classifying cases, so the model had the overall highest statistics. However, feature importance suggests that it might have over-relied on a few important features, which would hamper the interpretability of the decisions made by the model.

K-Nearest Neighbors Classification (KNN) Model: KNN had a strong accuracy rate and a strong ROC-AUC score, but both values were weaker than the other models. It demonstrated subpar performance in classification, with a relatively higher false negative rate, indicating that the model misclassified cases more.

Naive Bayes: Naive Bayes resulted in a relatively lower accuracy with a high ROC-AUC score. This simple model produced competitive performance by effectively utilizing its probabilistic assumptions. The low false negative rate suggests it performed competitively with other models at proper classification. On the other hand, this may be prone to over-simplification, since the model relies heavily on the independence hypothesis across where feature dependency exists.

6 Discussion

Overall, the model with the strongest overall performance was the Random Forest model, with the overall highest accuracy and ROC-AUC, and with the overall lowest false negative rate of any of the other models. All models exhibited high accuracy, high ROC-AUC, and low false negative rates, which was likely a result of the dataset including a smaller amount of instances (which would reduce differences in the models' values) and the overall strength of the data, as the variables had a high correlation to the outcome variable that was being predicted, as seen in Figure 1. Additionally, it is important to note that since misdiagnosing a benign tumor as malignant is less dangerous than misdiagnosing a malignant tumor as benign, the false negative value has high importance in selecting the most appropriate model (Patel). This supports the choice of the Random Forest model.

Data Limitations: The dataset contained an imbalanced number of malignant and benign samples, with more benign (N=357) than malignant (N=212). This imbalance likely reduced the capacity of tumor classification of these models, as more data exists for benign samples. Additionally, the sample contained only 569 instances, which is a relatively lower sample considering the high prevalence of breast cancer. Therefore, these results can be improved with datasets that have more balanced numbers of the two tumor types and an overall increased number of samples.

7 Appendix

The following link is the Github repository containing the code used to examine this dataset and develop these models: https://github.com/svemuru15/562project

References

1. "Breast Cancer." Cleveland Clinic, 9 Sept. 2024, my.clevelandclinic.org/health/diseases/3986-breast-cancer.

2. Patel, Aisha. Benign vs Malignant Tumors — Oncology — JAMA Oncology — Jama Network, 30 July 2020, jamanetwork.com/journals/jamaoncology/fullarticle/2768634.

3. Jia, Xiao, et al. "Breast Cancer Identification Using Machine Learning." Wiley Online Library, 3 Oct. 2022, onlinelibrary.wiley.com/doi/full/10.1155/2022/8122895.

4. Iqbal, Muhammad Javed, et al. "Clinical Applications of Artificial Intelligence and Machine Learning in Cancer Diagnosis: Looking into the Future - Cancer Cell International." SpringerLink, BioMed Central, 21 May 2021, link.springer.com/article/10.1186/s12935-021-01981-1.