Breast Cancer Prediction

Bokai Hu
University of California San Diego
La Jolla, California
boh001@ucsd.edu

Yanchen Jing
University of California San Diego
La Jolla, California
yajing@ucsd.edu

ABSTRACT

In the contemporary landscape of healthcare, breast cancer has emerged as a critical concern affecting millions globally, so the proportion of different factors that contribute to the breast cancer could be a potential help for both the patient and doctors. We propose to leverage analysis to gain insight into breast cancer based on clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses to get a classification of breast cancer. Our code is available here: Github repository.

1 INTRODUCTION

Breast cancer is one of the most common malignant tumors among women and affects the health of many people worldwide. According to the World Health Organization (WHO), breast cancer is one of the tumors with the highest cancer incidence and mortality rates worldwide, and its harm to individuals and society cannot be ignored. Therefore, the prediction and diagnosis of early breast cancer becomes crucial to take appropriate treatment measures early and improve the cure and survival rates.

In the field of medical image analysis based on machine learning and deep learning, many studies utilize multimodal inputs, combining images with other indicators for disease diagnosis. In our project, considering that training image models requires significant computational resources, we primarily used numerical indicators (including clump thickness and uniformity of cell shape, etc.) for breast cancer classification.

We started by conducting exploratory data analysis and visualized the dataset by reducing its dimensionality using principal component analysis. Then, we fitted the dataset using both Lasso regression and ensemble models, showcasing the performance of different models. Additionally, we analyzed the importance of features in the dataset.

2 DATASET

Our dataset is from Breast cancer Wisconsin (original) dataset, which contains real data of 683 observations with independent variables that allows us to classify dependent variable into malignant or benign. The dataset contains Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, which could give an numerical value towards the analysis. The number of data in the class benign is 444, and the number of data in the class maglinant is 239. The data here exhibits a certain degree of class imbalance, which might be a common and natural occurrence in medical data, where the number of cases with varying degrees of disease is unbalanced. An illustration of the dataset is shown in Figure 1.

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
count	683.000000	683.000000	683.000000	683.000000	683.000000	683.000000	683.000000	683.000000	683.000000	683.000000
mean	4.442167	3.150805	3.215227	2.830161	3.234261	3.544656	3.445095	2.869693	1.603221	2.699854
std	2.820761	3.065145	2.988581	2.864562	2.223085	3.643857	2.449697	3.052666	1.732674	0.954592
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	2.000000
25%	2.000000	1.000000	1.000000	1.000000	2.000000	1.000000	2.000000	1.000000	1.000000	2.000000
50%	4.000000	1.000000	1.000000	1.000000	2.000000	1.000000	3.000000	1.000000	1.000000	2.000000
75%	6.000000	5.000000	5.000000	4.000000	4.000000	6.000000	5.000000	4.000000	1.000000	4.000000
max	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	4.000000

Figure 1: An illustration of the breast cancer dataset

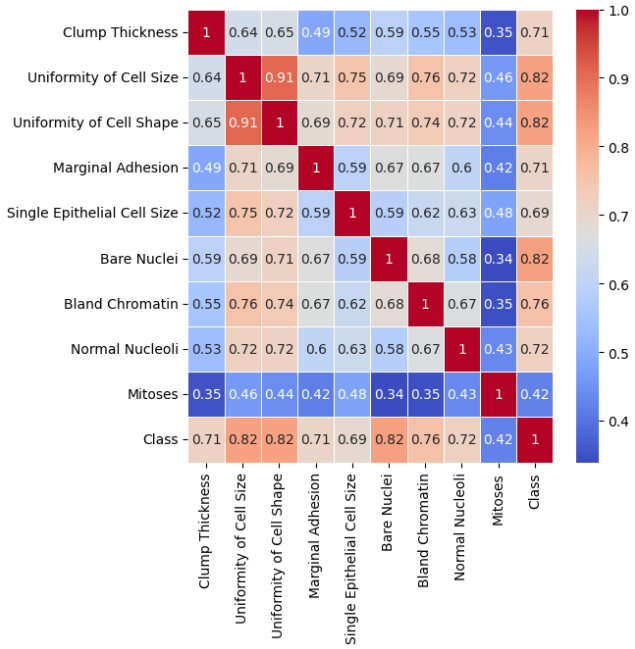

Figure 2: Heatmap of the breast cancer dataset

3 DATA ANALYSIS AND VISUALIZATION

The very first thing we start is to find out the relation between each variables, so we calculate the pearson's correlation coefficient and the results are demonstrated in Figure 2. Except for the Mitoses feature, all other features show a relatively strong correlation with the class label. Therefore, we will discuss the model's performance both with and without the use of the Mitoses feature. Additionally, the features Uniformity of Cell Size and Uniformity of Cell Shape is highly correlated, leading us to believe they might exhibit slight co-linearity and could contain redundant information. Consequently, we have decided to remove the Uniformity of Cell Shape feature

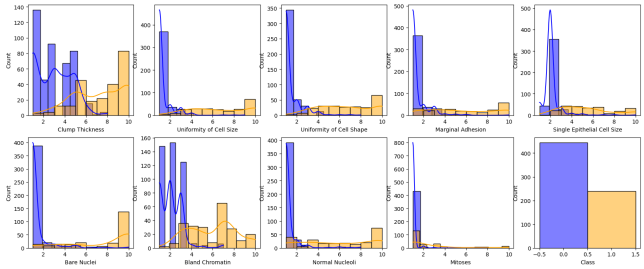

Figure 3: Histogram of each feature

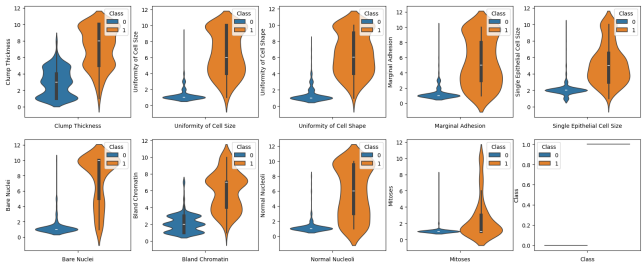

Figure 4: Violin plot of each feature

and only use the Uniformity of Cell Size feature.

The next step is to analyze the skewness and kurtosis of our data. Skewness measures the asymmetry of the distribution of a random variable, which is the third central moment of the data.

$$Skew(X) = E[(\frac{(X-\mu)}{\sigma})^3]$$

It indicates whether the data is skewed to the left (negatively skewed), symmetric, or skewed to the right (positively skewed), and Kurtosis measures the tailedness or sharpness of the distribution's peak, which is the fourth central moment.

$$Kurt(X) = E[(\frac{(X-\mu)}{\sigma})^4]$$

It helps identify whether the data has heavy tails or is more concentrated around the mean[2]. As is shown in Figure 3, we could find that several features, such as "Single Epithelial Cell Size" and "Mitoses," have high positive skewness and positive kurtosis, suggesting distributions with longer tails and more extreme values. From the Figure 4, it can be observed that the majority of features for data with label 0 are highly concentrated and close to 0. Conversely, for data with label 1, the features either tend to be near the opposite extreme or are distributed relatively evenly.

3.1 PRINCPLE COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a common dimensionality reduction technique used for data processing. It first calculates the covariance matrix between different features in the data and then computes the eigenvalues λ and eigenvectors v of the covariance matrix.

$$\Sigma v = \lambda v$$

The eigenvalues represent the values of variances, and the eigenvectors indicate the directions of variances. In PCA, we select the top k eigenvectors with the highest variances as k principal components. Typically, k is less than n, which achieves the dimensionality reduction effect.

We form a projection matrix P using these k principal components,

$$P = \begin{bmatrix} | & | & | & | \\ v_1 & v_2 & \dots & v_n \\ | & | & | & | \end{bmatrix}$$

and use this matrix to transform the original data X into a lower-dimensional space Y.

$$Y = X \cdot P$$

By retaining the top k directions with the highest variances from the original data, we can achieve dimensionality reduction while preserving most of the information in the data.

In this project, we employed PCA to reduce the dimensionality of breast cancer data features. The primary reason for this was our inability to visualize multi-dimensional feature data exceeding three dimensions effectively. For visualization purposes, we chose to utilize only the first two principal components. If $\{\sigma_1^2, ..., sigma_n^2\}$ are in descending order, and we calculated the information captured by the principle components using the following formula:

$$Ratio = \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{n} \sigma_i^2}$$

According to the calculation, the first two principle component can capture more than 75% percent of the variance information. We separated these two principal components based on categories, and the results are depicted in Figure 5.

From the Figure 5, we can observe that the majority of data points with label 0 cluster together in one group and can be effectively distinguished from points with label 1. This indicates that the two principal components can efficiently capture information related to the categories, and the features present in the dataset are sufficient for classifying these data points into their respective categories.

4 METHODS

In this section, we will discuss the classification models we employed, including logistic regression and random forest models. We will evaluate the classification performance of these models on the dataset and present our conclusions.

4.1 Logistic Regression with Lasso (L1) Regularization

We first use logistic regression with Lasso (least absolute shrinkage and selection operator) regularization to fit the data. Lasso regularization is a regularization method used in linear regression. Mathematically, Lasso achieves regularization by adding an L1-norm term to the classic linear regression loss function. Specifically, the objective function for Lasso regression is as follows:

$$L_{\text{LASSO}}(\beta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\beta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\beta_j|$$

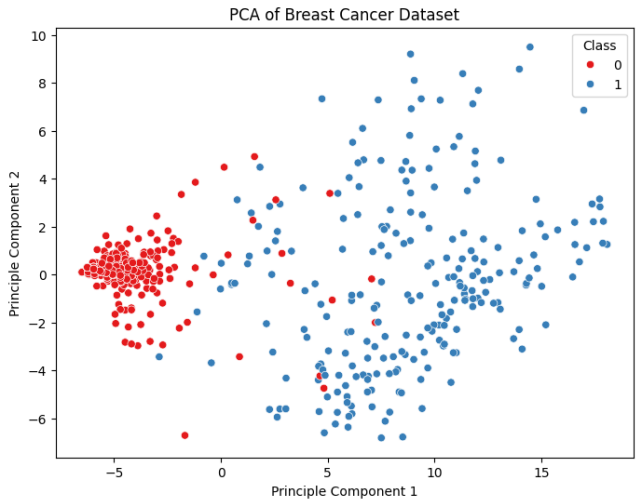

Figure 5: Visualization of the first two principle components of the breast cancer dataset

where m for the number of training examples, $h_\beta(x)$ for the logistic hypothesis function, β for the vector of coefficients, λ for the regularization parameter, n for the number of features, $|\beta_j|$ for the absolute value of the j-th coefficient.

Unlike Ridge (L2) regularization, Lasso regularization does not yield an analytical solution and tends to result in a sparse weight vector θ, effectively performing automatic feature selection [3]. Therefore, Lasso regularization is suitable for handling datasets with multicollinearity or a much larger number of features than samples. As the regularization coefficient λ increases, more coefficients are pushed towards zero, leading to a reduction in the model's complexity.

Table 1: Comparison of Lasso regression with and without the Mitoses

	With Mitoses	Without Mitoses
Accuracy	0.964	0.964
Precision	0.934	0.934
Recall	0.964	0.964
F1 Score	0.948	0.948

Next, we compared the performance of the Lasso classifier trained with and without Mitoses as a feature and displayed the weights of each feature when Mitoses is used as a feature. From Table 1, it can be seen that the use of Mitoses does not significantly impact the model's performance. In our experiments, the differences in various metrics between these two models are all less than 0.001.

Figure 6 and 7 display the confusion matrices obtained from the last trail of 100 repeated experiments for the two models. It can be observed that their performance is identical and both exhibit good performance.

Table 2: Weights of the features in Lasso regression

Features	Weights
Clump Thickness	3.734
Uniformity of Cell Size	2.717
Marginal Adhesion	1.655
Single Epithelial Cell Size	0.328
Bare Nuclei	3.437
Bland Chromatin	2.237
Normal Nucleoli	1.769
Mitoses	0.430

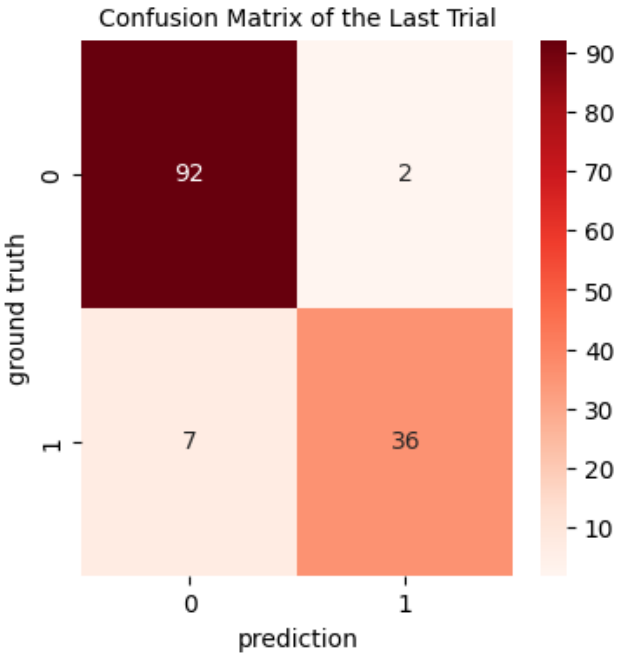

Figure 6: Confusion matrix using mitoses

Table 2 presents the average weight values of different features over 100 repeated experiments. It is evident that the weights of Mitoses and Single Epithelial Cell Size are much smaller than those of other features, and in over one-third of the total experiments, the weights of these two features are zero. We believe this situation may arise because the correlation between Mitoses and the class labels is not strong, and the smaller weight of Single Epithelial Cell Size could be due to multicollinearity with other weights.

4.2 Ensemble Method (Random Forest Classifier)

We begin with a brief introduction to random forests. Random forest is a combination of tree predictors where each tree depends

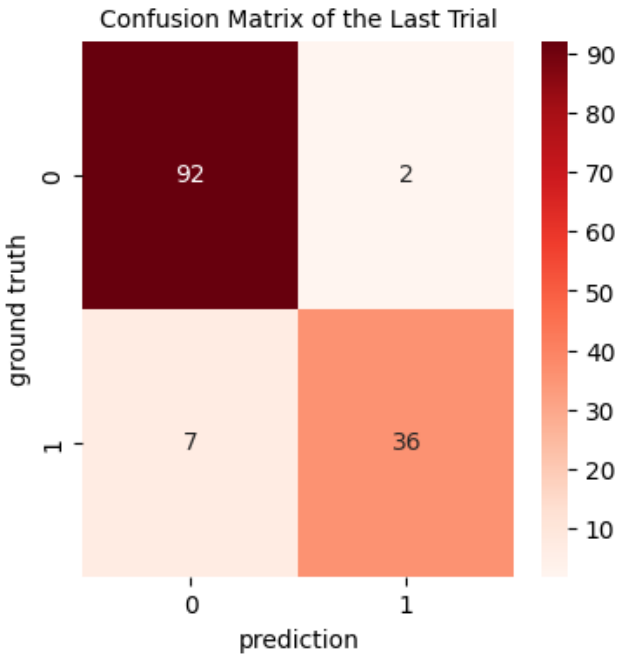

Figure 7: Confusion matrix without using mitoses

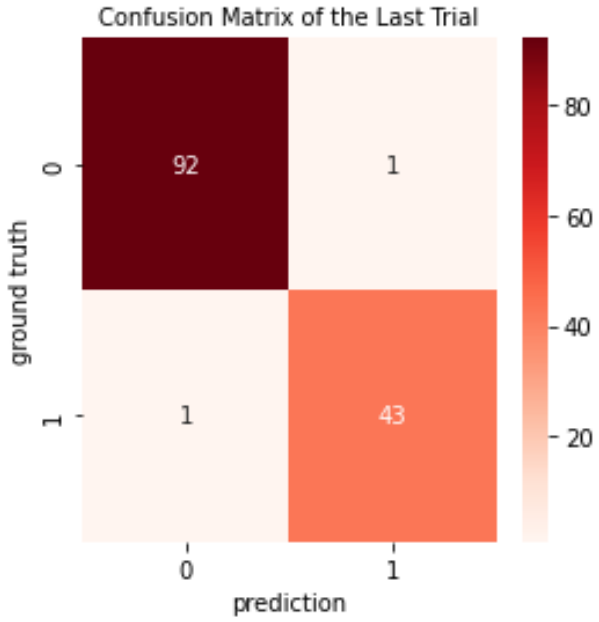

Figure 8: Confusion matrix of random forest classifier

on the features that is sampled independently and identically distributed for all trees in the forest. The generalization error of a forest converges to a limit as the number of trees in the forest increases. The generalization error of a tree classifier forest depends on the strength of the individual trees in the forest and the correlation between them.[1] In this section, we apply ensemble method using decision trees to perform classification. Since decision tree and random forest may easily fall into the pit of overfitting, we first adjust the hyperparameters of the model. We find that the cross-validation process revealed the optimal number of subtrees for maximizing performance is 101. In our experiment, the maximum depth of individual trees is 11 and the max features used in a single tree is 2, which these hyperparameters resulted in a maximum cross-validated accuracy score of approximately 0.9751. To validate its robustness, we evaluate the model with over 100 random seeds and find out the average accuracy is 0.971, the average precision is 0.9654, the average recall is 0.9565, and the average F1 Score 0.9603. From the results, it is evident that the Random Forest classifier shows better performance than logistic regression. This is reflected in higher accuracy, precision, F1 score, and the confusion matrix shown in Figure 8. However, the recall of the Random Forest is lower, so there may need to be some trade-offs when choosing a model, or a comprehensive consideration of the results of both models might be necessary. We also analyze the feature importance using the random forest and the results are show in Table 3, and it is obvious that the importance of Mitoses is smaller than those of the other features, which is consistent with our results shown using logistic regression.

Table 3: Feature importance calculated using random forest

Features	Weights
Clump Thickness	0.084
Uniformity of Cell Size	0.231
Marginal Adhesion	0.056
Single Epithelial Cell Size	0.153
Bare Nuclei	0.176
Bland Chromatin	0.197
Normal Nucleoli	0.093
Mitoses	0.009

5 CONCLUSION

In this project, we started by analyzing the features in the dataset, selecting the relevant ones, and then trained logistic regression with Lasso regularization and a random forest model to validate our hypothesis. Through the analysis of experimental results, we confirmed our hypothesis that there might be features in this dataset with weak correlations to the labels and a presence of multicollinearity. Furthermore, the multiple metrics we calculated also demonstrated that using these features allowed for a good classification between benign and malignant breast tumors.

6 FUTURE WORK

For future work, we can consider incorporating more data and features to mitigate overfitting issues. Additionally, exploring the integration of medical imaging data for multimodal fusion could enhance the model's informativeness.

REFERENCES

[1] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (October 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[2] S. Jhajharia, S. Verma, and R. Kumar. 2016. A cross-platform evaluation of various decision tree algorithms for prognostic analysis of breast cancer data. IEEE, Coimbatore. https://doi.org/10.1109/inventive.2016.7830107

[3] Dinesh Kumar. 2023. *Understanding Lasso Regression.* https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/