Classification of Breast Cancer Histopathology Images

Department of Applied Mathematics and Statistics

Anusha Rao, Gracelyn Shi, Krutal Patel
Fadil Santosa Ph.D., Jeremias Sulam Ph.D.

Objective

- Investigate image data from the Breast Cancer Histopathological Database and develop a supervised classification algorithm to accurately predict image labels as malignant (1) or benign (0).
- Database contains 9,109 images of breast tumor tissue, collected from 82 patients in 4 magnifications (40, 100, 200, 400).
- For both training and analysis, **40X magnification images** are used, containing 585 Benign and 1370 malignant images.

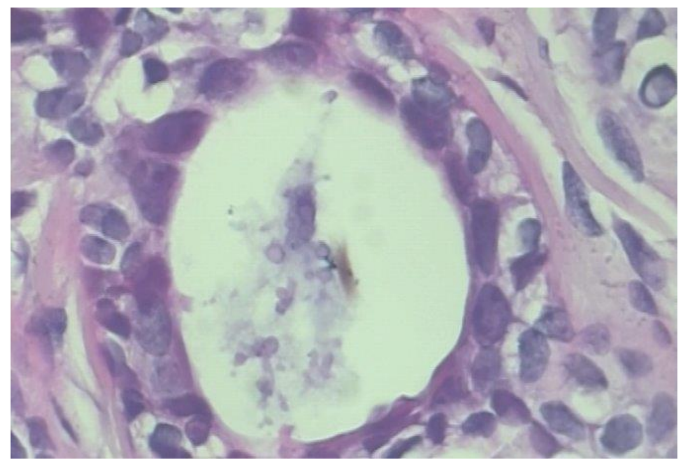

Figure 1. Image of a benign tumor (40x)

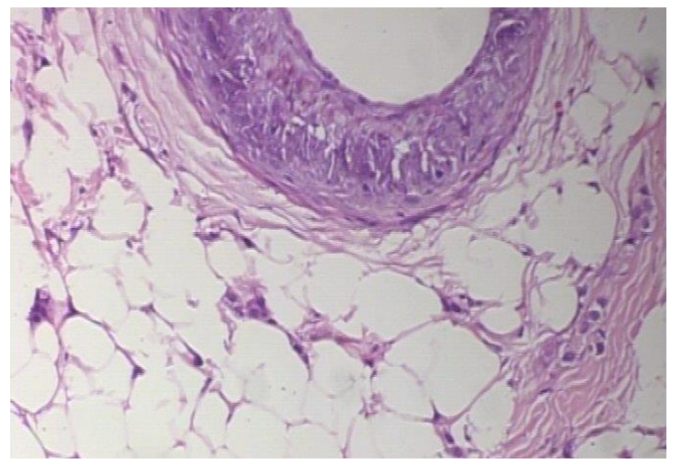

Figure 2. Image of a malignant tumor (40x)

Source: Department of Informatics and Graduate Program in Computer Science of the Federal University of Parana.

Feature Extraction using ResNet-18

- Extracted features (weights) using a pretrained convolutional neural network, **ResNet-18** (pretrained on ImageNet).
- Generated 512 total features (weights) for each image from the last layer of network, before the fully connected layer (see Fig. 3).

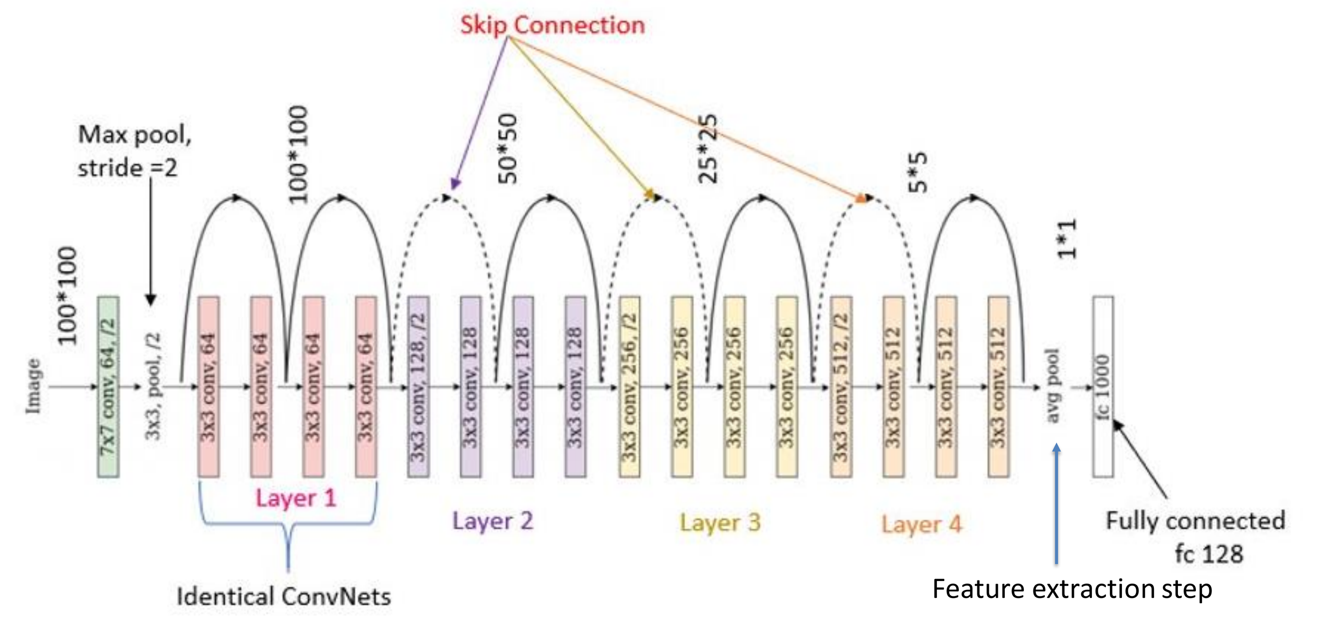

Figure 3. Network architecture of a ResNet-18 (signifying 18 layers), a convolutional neural network utilizing skip connections to fix the vanishing gradients problem

Scan to view our presentation with additional information on the data and methods used

Supervised Learning: Logistic Regression

- Logistic regression model is ideal for binary classification problems.
- Trained a logistic regression model (using 'liblinear' optimization solver) on the 512 features.
- **Obtained accuracy of 94.9%.**
- Multivariate linear model performed with 62.8% accuracy.

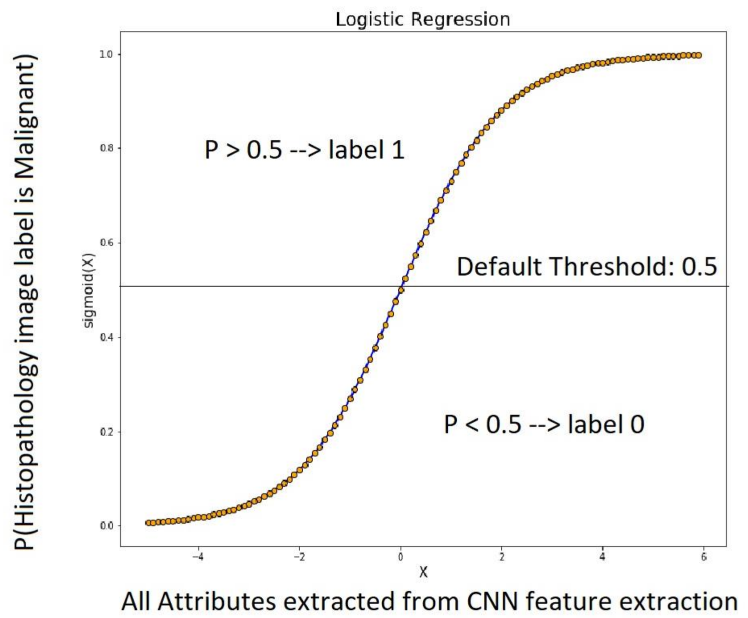

Figure 4. Visualization of threshold value for logistic regression model

Comparison of LogReg and Multivariate Linear Models

Model	True Positives	True Negatives	False Positives	False Negatives	Accuracy	AUC
Logistic Regression (train/test 40x)	89%	96%	3%	2%	0.949	0.987
Linear Regression	44.6%	45.3%	5.4%	4.7%	0.628	0.964
Logistic Regression (train 40x/test all)	91.9%	95.9%	9.6%	3.4%	0.804	0.985
Logistic Regression (train/test all)	90.1%	88.9%	11.2%	9.77%	0.906	0.968

Figure 4. Table containing performance metrics for each model

- Tested the (40x) model on images of all resolutions (80.4% accuracy).
- Model trained on 40x images generalizes well to other resolutions.
- Finally trained model on images of all resolutions, with a test accuracy of 90.6%.

Principal Component Analysis

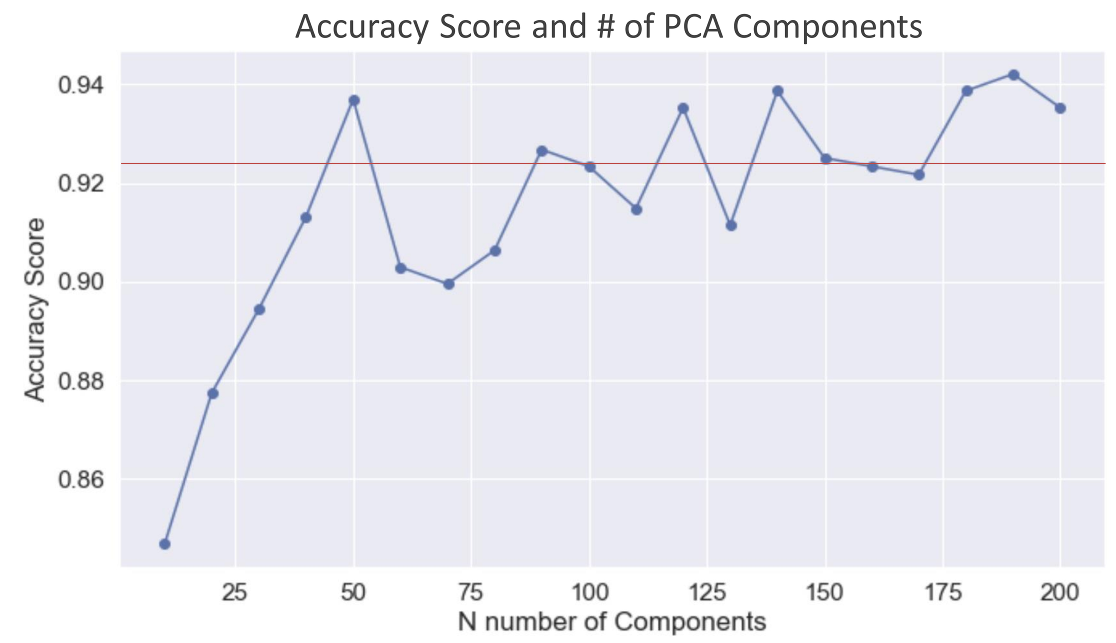

Figure 5. Accuracy score as a function of number of PCA components

- Reduced the number of attributes needed for classification by applying principal component analysis and using maximum explained variance ratio.
- **512 features → 120 principal components** with ~91% explained variance.
- Model Evaluation: 93.9% accuracy with comparable sensitivity ratios.

Optimal Threshold Analysis

- Goal: reduce the false negative rate of model.
- Optimize threshold using ROC analysis to **reduce the number of false negatives** while maintaining as high of an accuracy as possible.
- Achieve this by maximizing Youden's J-statistic (see Fig. 6).

Youden's J-Statistic for Optimal Threshold Determination

$$J = sensitivity + specificity - 1 \implies J = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1$$

$$Define\ FPR = false\ positive\ rate = \frac{FP}{TN + FP}$$

$$\frac{FP}{TN + FP} + \frac{TN}{TN + FP} = 1 \rightarrow J = \frac{TP}{TP + FN} + 1 - \frac{FP}{TN + FP} - 1$$

$$J = \frac{TP}{TP + FN} + \frac{FP}{TN + FP} \implies J = TPR - FPR\ (maximize\ J)$$

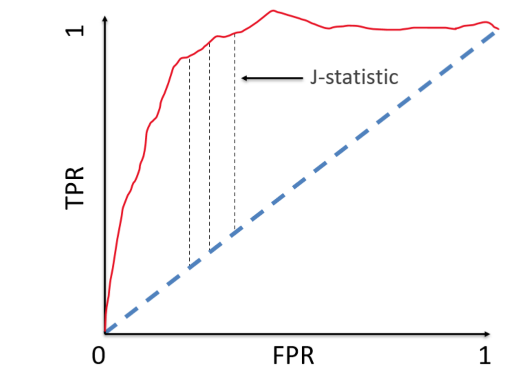

Figure 6. Visualization of Youden's J-statistic on a ROC curve

- The model with the optimal threshold produced an accuracy of 93.4% and reduced number of FN to 6.

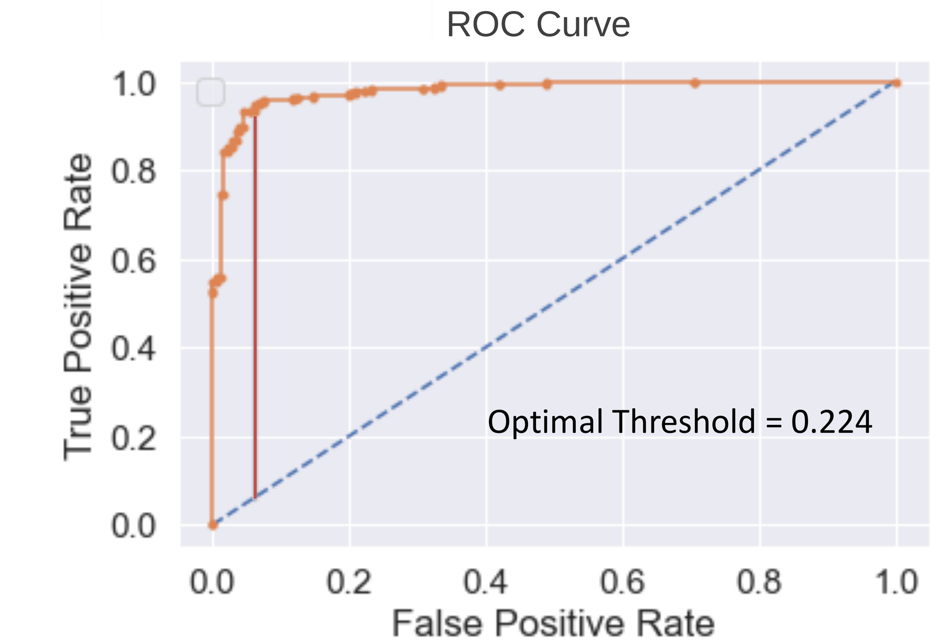

Figure 7. ROC curve of the 40x resolution logistic regression model

Conclusion

- Logistic regression model performs strongly both in terms of accuracy and sensitivity, can be a powerful tool for breast cancer diagnosis to both improve accuracy and decrease time.
- 40x model generalizes well to higher resolutions, could be used to reduce labor and costs due to decreasing need for high-labor high-resolution imaging.
- **Remaining Challenges:** black box problem/interpretability, liability

Acknowledgements

- We would like to thank Professor Fadil Santosa, Professor Jeremias Sulam, and Jacopo Teneggi for their guidance, advice, and mentorship on this project.