Machine Learning Approaches to Breast Cancer Diagnosis and Treatment Response Prediction

Katie Planey, Stanford Biomedical Informatics

1. Introduction

Given the disease severity and well-publicized philanthropic efforts concerning breast cancer, it is not a new arena for medical researchers, computational or otherwise. The goal in most breast cancer classification problems is to determine whether a patient's lesion is malignant or benign.

Machine learning has been successfully applied to this problem in recent years; for example, a group in Turkey reported higher than 99% accuracy for SVM classification on the widely used Wisconsin University breast cancer dataset. However, the holy grail of machine learning techniques that fuse high or even reasonable accuracy with readily accessible features from the average clinic has proved elusive. The previously mentioned group used carefully measured pathological features such as single cell epithelial size and mitoses that are not always included in a standard clinical report.[1] Thus, it is still highly clinically relevant to search for breast cancer machine learning features that are highly predictive of disease state.

This project lays the foundation for continued research on two machine learning applications to breast cancer: predicting malignant vs. benign tumors to aide in biopsy decisions, and predicting whether a patient's cancer will successfully respond to specific treatment regimens.

2. Methods

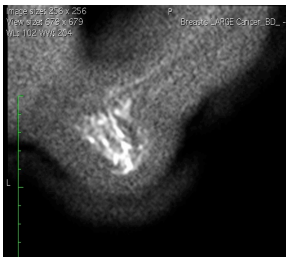
2.1 Diagnosis Dataset

This dataset from Stanford Radiology includes patients who had suspicious breast lesions and underwent MR scans. After radiologist inspection of the MRI, results were still inconclusive for 77 patients, and thus these patients had biopsies and pathology reports done to determine if the tumor was indeed malignant. Thus, these patients did not have "obvious" malignant or benign tumors; the tumors themselves were often difficult to even identify on a scan.

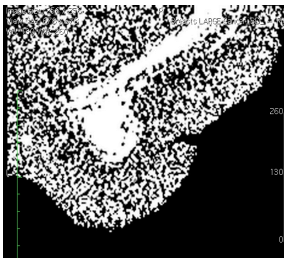
The dataset includes 78 lesions total (one patient had two lesions), and 64.9% were path-proven malignant tumors. This means that there was an unnecessary biopsy for 35.1% of tumors.

2.1.1 Features

The currently available data beyond the biopsy-confirmed response variable of malignancy status incudes the three basic measurements of age, longest tumor length, and ADC, or Apparent Diffusion Coefficient. ADC maps are created from Diffusion Weighted Images (DWI). ADC measurements provide a quantitative measure of the underlying vasculature in tumor structures by measuring water flow patterns through the tumors.



DWI

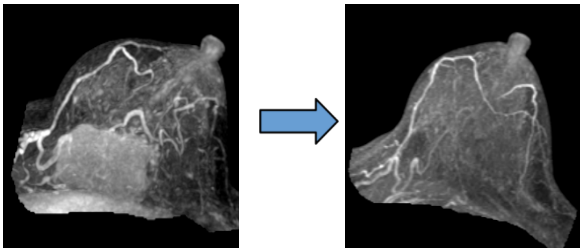


ADC

DWI scans, and therefore ADC values, are not yet widely used in clinical practice, mainly due to DWI being a relatively new MR technique. However, ADC values have proven correlative to tumor status[2] and further supporting evidence such as its utility as a machine learning algorithm feature could impact clinical imaging protocols.

2.2 Treatment Dataset

Stanford is the main treatment center for a Phase II neoadjuvant breast cancer study of gemcitabine, carboplatin, and poly (ADP-Ribose) polymerase (PARP) inhibitor BSI-201. Neoadjuvant therapy implies that chemotherapy or other drugs are given to a patient before surgical removal of the tumor, in an attempt to shrink the tumor and prevent re-growth, and/or avoid a complete mastectomy. In certain patient groups, neoadjuvant therapy has proven efficacious in eradicating cancer growth.[3]



Pre-treatment **Post-treatment**

While gemcitabine and carboplatin are standard chemotherapy treatments, the PARP inhibitor is a novel therapeutic agent recently shown to increase progression-free and overall survival in triple-negative breast cancer patients.[4] This class of breast tumor is known to be extremely aggressive, and, until the PARP inhibitor drug, it was often a grim diagnosis for this patient subset.

Some patients have had dramatic results on the drug, such as the patient above, whose tumor essentially vanished after PARP inhibitor treatment.

2.2.1 Features

The trial is ongoing; at this point in time, the final dataset used included 43 patients and 15 multinomial features, outlined in Table I. All input features were measured before treatment or only after 1 cycle of treatment (patients had a total of 4-6 treatment cycles). The response variable was a binary complete or incomplete pathologic response.

Unfortunately, most of the quantitative MRI measurements are not yet complete (due to HIPAA/anonymization pre-processing requirements), and thus could not be included in this initial experiment. However, the current dataset is still quite rare, given that it includes pathology/tissue results of both the lymph nodes and main tumor, and gene tests.

Table I: Key PARP Features

Term	Definition
Demographics	Race/ethnicity of patient
Age	Age in Years
Clinical Stage	0-IV, IV= most aggressive
Node Sample	Biopsy result: cancerous or benign?
Node Imaging	Nodes present on MRI?
Node Baseline	Nodes palpable to clinican?
Node C2D1	Palpable after 1 drug cycle?
Initial Measurement	Initial tumor size
C2D1 Measurement	Size after 1 drug cycle
Total Drug Cycles	4 or 6 cycles
Pathology Grade	1-3, 3 = fastest cell growth
BRCA 1 Gene Test	Positive for known harmful BRCA 1 gene mutations?
BRCA 2 Gene Test	Positive for BRCA 2 gene mutations?
MRI Tumor Length	Length (cm) on MRI
Pathologic Response	Response variable: patient fully responded to treatment or had an incomplete response (i.e. cancer is still present after all cycles completed)

3. Pre-Processing

3.1 Diagnosis Dataset

Following HIPAA protocol, I downloaded all patient scans from Stanford PACS (Picture Archiving Communication System) and anonymized them. I was given a brief tutorial by Stanford radiologist Dr. Bruce Daniel on how to identify breast lesions so that I could compute length and

ADC measurements using Osirix, an open-source DICOM viewing software.

3.2 Treatment Dataset

For the treatment dataset, I had access to Stanford's REDCap Database, a secure online site that stores IRB-approved study information. There were approximately 40 features total, and I consulted Stanford radiologist Dr. Jafi Lipson to understand the clinical significance of each feature. I then chose features based on both clinical significance, and how many patients already had a specific feature measured, in an attempt to maintain a reasonably sized dataset.

3.3 Discretization

In order to test classification rates using models such as Naïve Bayes, Logistic Regression, and SVM classification, all features had to be discretized. In the diagnostic dataset, all three input features, age, tumor length, and ADC value, were discretized, and in the treatment dataset, age and any tumor measurements were discretized.

In an attempt to create biologically intuitive models, I did not simply assign a blanket number of discrete states to all features, as this would mean that for any categories in which no observed feature occurred, I would have to assign this zero state a small nonzero probability to avoid computational issues. Yet consider the tumor length feature in the treatment dataset: none of the patients had tumors less than 2 cm in diameter, as this was a patient entrance criterion for the trial.

Assigning a separate state to say tumors below 1cm would not create computational errors, but it would create a model that is fitting parameters based on a nonsensical medical setting for this subset of patients. Thus, I took care to look at every single feature in both datasets, continuous or discrete, to decide the best discrete categories. Attempting to retain medical intuition in breast cancer datasets is no new phenomenon; gene signatures from machine learning techniques have already come under scrutiny.[5]

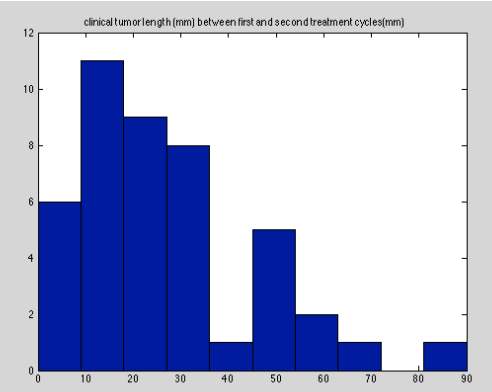


Figure 1: An example of the difficulty of biologically intuitive discretization. Here, one can see that 9 separate categories for 0-90 does not reflect the true patient distribution and will result in lower statistical power for categories with extremely sparse data points.

Even in the already discrete category of stage in the treatment dataset, no patients had a stage of 1. Therefore, I discretized the lowest state to <= stage 2. Because this data is only for extremely high-risk patients, it should be fit to optimize this patient cohort, even if this limits model usage beyond triple-negative patients.

3.4 Matrix Preparation for LibSVM

Once discretized, each category within a multinomial feature was represented as one binomial feature. While not a requirement for all the classification software packages used, the LibSVM development group at National Taiwan University recommend on their website to prep the data matrices in this manner. Sparse matrices were then created using the SparseM package in R.

3.5 Classification Packages Used

The diagnostic dataset was initially run on a Naïve Bayes multinomial model

with Laplace smoothing coded in MATLAB and tested on a randomly selected 20% holdout sample.

Further classification attempts on both datasets were done in R, using the LibSVM package, GLM, and knn (K-Nearest Neighbor). For cross-validation on GLM and knn, the cv.GLM and knn.cv functions were used. Leave-one-out cross-validation (LOOCV) was the only accuracy measure employed on all models due to the small size of these datasets.

4. Classification & Results

4.1 Varying SVM and GLM Inputs

For SVM and GLM, a wide array of function inputs was tested. In SVM, both classification Type I and Type II (labeled C and nu in R) and all available kernels (radial, sigmoid, linear, polynomial) were tested. A range of function inputs such as gamma and Cost were also tested. In GLM, distribution models of Binomial and Gaussian were tested.

Testing various scenarios of these function parameters did not lead to significant increases in accuracy. This is most likely due to the small size of the datasets – no specific combination of classification function inputs led to a serendipitous jump in accuracy.

Because logistic regression and GDA assume continuous outputs, I tested both the discretized and original datasets. The discrete diagnostic dataset had been split into greater than five categories for each of the three variables, leading to little difference between discrete and continuous states. Most features in the treatment dataset were already binary and could not be construed as true continuous random variables.

4.2 Diagnosis Predictions

The key results from classification models on the diagnostic dataset are listed in table 3. Logistic regression performed the best with 75% LOOCV accuracy. It is intuitive that a regression model performed best, given that the original features were indeed continuous.

It is reassuring that the Naïve Bayes model obtained 70% accuracy both with a random 20% hold-out sample in my own MATLAB code, and with LOOCV in R.

Table II: Diagnostic Classification Results

Model	**LOOCV Accuracy**
Naïve Bayes & Laplace Smoothing	70%
K-Nearest Neighbor (k=1)	70%
K-Nearest Neighbor (k=3)	61%
Logistic Regression	**75%**
SVM Type 1, Sigmoid Kernel	66%

4.3 Treatment Predictions

LOOCV accuracy rates deviated on a wider scale for the treatment dataset than the diagnostic dataset. For example, the LibSVM SVM I function with a linear input kernel resulted in 47% accuracy, while the polynomial kernel with degree 4, the best performer, resulted in 69% accuracy.

Table III: Treatment Classification Results

Model	**LOOCV Accuracy**
K-Nearest Neighbor (k=1)	65%
K-Nearest Neighbor (k=3)	48%
Logistic Regression	67%
SVM Type 1, Linear Kernel	47%
SVM Type 1, Radial Kernel	67%
SVM Type 1, Polynomial, d=4	**69%**

The high variability is most likely due to the small sample size of 43 patients with 16 multinomial features (resulting in about 30 total binomial features) and the dataset being a mix of continuous and discrete variables. Attempting to employ KNN beyond k=1 resulted in a significant drop in accuracy, indicating that the dataset

did not contain many repeat patients of similar feature inputs and response results. Because knn.cv saves the prediction results, I was, however, able to note that there was not a significant difference in the number of false positives and false negatives predicted, on either dataset.

Expanding the features in a higher-dimensional space with a polynomial kernel of 4 proved to provide the best results, with 41 support vectors being chosen by LibSVM for the model. However, both logistic regression and SVM with a radial kernel provided similar results.

5. Discussion & Future Work

5.1 Diagnostic Data

Even though a predictive accuracy of 75% is not high enough for widespread clinical use, it is surprising that with my own untrained eye both identifying and outlining tumors, my logistic regression model performed better than the clinician (65%) in deciding which tumors should be biopsied.

This result emphasizes the potential utility of ADC values in diagnostics. Additionally, a logistic regression using the ADC feature alone resulted in 70% LOOCV accuracy, highlighting its predictive strength as a stand-alone feature.

I plan to present these initial results to Stanford breast cancer radiologists with two aims: 1) Re-identify and outline tumors with the aide of a trained radiologist to reduce noise/ decrease error 2) Recruit radiologists to participate in studies that compare their performance alone to combined performance with machine learning aids that incorporate ADC values.

Additionally, the diagnostic dataset has a host of quantitative features in DCE-MRI images that can be data mined. I hope to employ unsupervised clustering techniques to find quantitative, unbiased feature inputs I can add to the model.

5.2 Treatment Data

I am particularly excited about the groundwork laid for the PARP treatment dataset; while the accuracy is only 69% at the moment, imaging features have yet to be added. I will also be working on another treatment dataset from UCSF next quarter that has hundreds of patients, is extremely well curated, includes a wider range of breast cancer types, and is longitudinal.

Given that this small initial dataset still achieved 69% predictive accuracy, work on the more comprehensive UCSF dataset could result in models that are truly implemented in clinical practice.

6. Acknowledgements

I would like to thank Dr. Bruce Daniel and Dr. Jafi Lipson for their continued research collaboration. I am also very grateful to Dr. Daniel Rubin, my research advisor, for guiding me through both the trials of MRI pre-processing and the design of novel feature experiments.

References

[1] Akay M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *36*, 3240-3247.

[2] Woodhams R., et al (2005). Diffusion-weighted imaging of malignant breast tumors: the usefulness of Apparent Diffusion Coefficient (ADC) value and ADC map for the detection of malignant breast tumors and evaluation of cancer extension. *Journal of Computer Assisted Tomography , 29* (5), 644-649.

[3] Liedtke C., et al (2008). Response to neoadjuvant therapy and long-term survival in patients With triple-negative breast cancer. *Journal of Clinical Oncology , 26* (19), 3286.

[4] O'Shaughnessy J., et al (2009). Efficacy of BSI-201, a poly (ADP-ribose) polymerase-1 (PARP1) inhibitor, in combination with gemcitabine/carboplatin (G/C) in patients with metastatic triple-negative breast cancer (TNBC): Results of a randomized phase II trial. *Journal of Clinical Oncology Meeting Abstracts , 27* (18S), 3.

[5] Drier Y, Domany E. (2011). Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? *PLoS ONE , 6* (3), e17795.