Diagnosing Malignancy in Breast Cancer Using Machine Learning

Sophie Shi, Patrick Garr, Yu Miao

We adhered to the Duke honor code during the completion of this assignment.

Abstract

Breast cancer is the most frequent malignancy in women worldwide, with an estimated 2.1 million women who were newly diagnosed with breast cancer and 626,679 women with breast cancer who died in 2018 [1][2]. While there are many different ways of predicting malignancy statuses, there are concerns with mixed success in diagnosis with some methods [3]. Due to the accuracy and speed of computer aided diagnosis, machine learning is gaining more attention for assisting disease prediction in this area. In this study, we used multiple machine learning algorithms to analyze malignancies in breast cancer. These techniques were applied on a diagnostic dataset of features from breast tissue cells obtained from Fine Needle Aspiration (FNA) to determine malignancy status. By creating confusion matrices and evaluating across several metrics, we found that the k-nearest neighbor (KNN) model had the highest performance among all tested models and scored the highest accuracy (97.66%). We were also able to create a simplified generalized additive model (GAM) with comparable accuracy (95.91%) that only relied on 3 features. All experiments were carried out in R Studio.

Introduction

Breast cancer is the most common cancer in women, accounting for about 30% of all new female cancers each year [4]. According to an estimate by the American Cancer Society, there will be around 43,250 deaths caused by breast cancer in the US in 2022 [5]. Additionally, the cancer burden in women is increasing globally, due to population growth and aging populations [6]. Therefore, the accurate diagnosis of malignancies after symptomatic indications is of great importance.

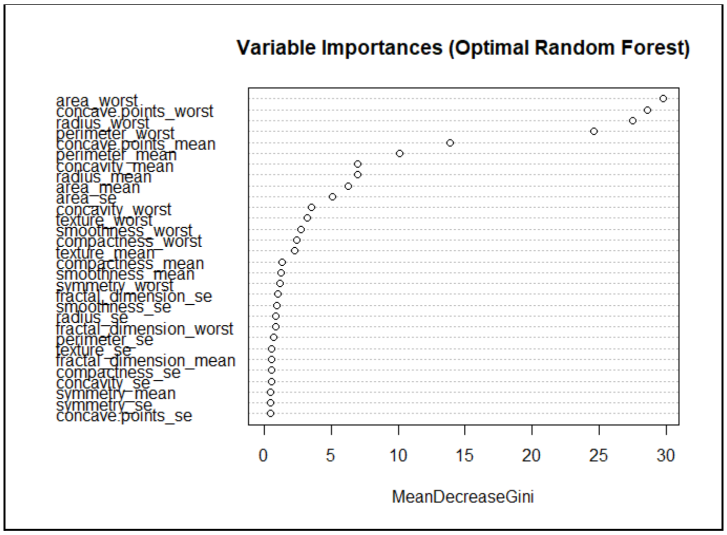
Among all the clinical techniques, FNA biopsy is one of the most common methods in diagnosing cancer, lymph, nodules, and other tumors [7]. It is less painful, more consistent, and less expensive than other methods. FNA biopsy is an effective cytological method, and it achieves adequate diagnostic precision clinically, and is often implemented in prostate and lung cancer cases [8]. However, using it for conventional breast cancer diagnosis relies highly on the experience of the experts, and it is estimated that diagnostic errors for medical patients in general range between 5% - 28% [9]. Due to these factors, we believe that it is critical that better methods are developed to speed up this process, make it more accessible to non-experts, and provide more accurate results, and hypothesize that we can use machine learning algorithms to achieve these goals. Feasibility of model construction has also been proved by earlier studies [10][11]. The model we constructed was created by splitting the data for training and testing sets using a 70/30 ratio, then building the model with the training set and using the testing set for evaluation. Our metrics included accuracy, sensitivity, specificity, Cohen kappa value and F-1 score; the number of features is also considered as a metric of model complexity. Because a malignant diagnosis would result in further treatments including surgeries and chemotherapy, we considered an emphasis on specificity to reduce those risks.

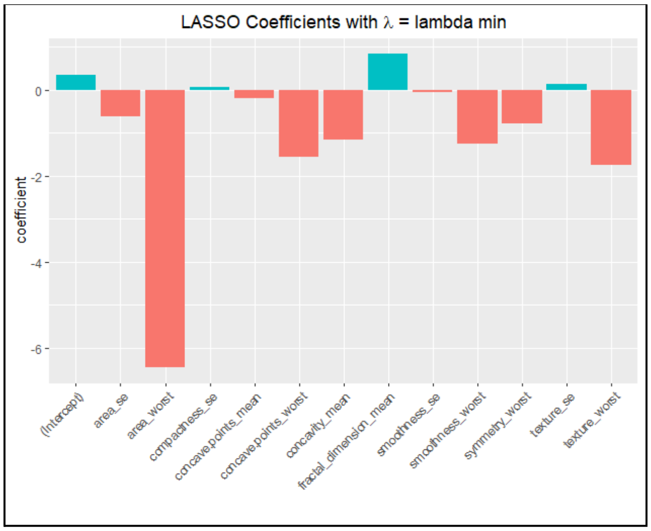
Data Description and Preprocessing

The dataset we used is a diagnostic data set for breast cancer with features obtained from processed cell images obtained at the University of Wisconsin Hospitals, Madison. We found this dataset from the UCI machine learning repository [12]. This dataset includes 569 patients and 32 attributes. There were 357 benign cases, and 212 malignant ones in this dataset. All features were computed from digitized images of a FNA of a breast mass and these features depict the characteristics of cell nuclei in the image [13]. 10 real-valued features were collected from the images for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension. Each of these 10 anatomical features has 3 categorical values: mean, standard error and “worst” (the mean of the three largest values present in the image). More information on how each of these were computed can be found from WN Street et. al [13].

One benefit of this dataset was that there was no missing data and thus no need for imputation. Thus, in our initial data exploration, The only step we took initially was to scale and standardize the data for some of the models because distance based algorithms and gradient descent based algorithms are especially sensitive towards scaling. After some of our initial model building, we also looked at feature reduction of our dataset to simplify our models. In order to do this, we implemented a few different methods of feature extraction. We wanted a holistic view, so we constructed decision tree (DT), random forest (RF), backwards elimination feature selection, and LASSO regression models to determine what coefficients were most important and see if the results were similar (Figure 2 & Figure 3). We used this information for the model simplification discussed in the next section. While there was general agreement in which variables were most important, there were some differences, which informed our simplified model building later on.

Figures 2 and 3: Feature Extraction from Random Forest and LASSO Models





Core Methods

Following our initial data analysis and pre-processing, we were not sure what models would perform best, and so began with ones that incorporated every feature. These included k-means cluster analysis (KMCA), hierarchical cluster analysis (HCA), KNN, linear discrimanatory analysis (LDA), logistic regression (Log. R), and ridge regression. Once we had analyzed and optimized these models using our evaluation criteria, we looked into using fewer features for prediction. Thus, we also created models using DT, LASSO regression, support vector machine (SVM), and GAM methodologies. The top level results are shown in Table 1 and will be discussed more in the next section, but due to the number of models analyzed, we will only discuss the best performing ones from both groups - KNN and GAM.

	KMCA	HCA	KNN	Ridge	Log. R	LDA	LASSO	SVM	DT	GAM
Accuracy	0.9051	0.8893	0.9766	0.9649	0.9415	0.9591	0.9708	0.9649	0.9181	0.9591
Sensitivity	0.8302	0.9670	0.9365	0.9524	0.8889	0.8889	0.9524	0.9206	0.8889	0.9365
Specificity	0.9496	0.8431	1.0000	0.9722	0.9722	1.0000	0.9815	0.9907	0.9352	0.9722
Kappa	0.7934	0.7738	0.9491	0.9246	0.8727	0.9100	0.9370	0.9236	0.8241	0.9117
F1	0.8670	0.8668	0.9672	0.9524	0.9180	0.9412	0.9600	0.9508	0.8889	0.9440
Number of Features	All	All	All	All	All	All	14	5	3	3

Table 1: Results From Evaluation Metrics of Confusion Matrices of All Models

Because our problem was one of classification, it reduced to one of determining how best to separate the points into either malignant or benign groupings. We opted to try KNN for a few reasons. First, there are known abnormalities between cancerous and benign samples based on various characteristics of the cell nuclei [14]. We therefore believed that the KNN method would be useful as we hypothesized that the characteristics between different groupings would be handled well by the model. We also knew there would not be an explicit training step, thus speeding up the process. Moreover, since the only hyperparameter would be the k-value, we could easily tune it to try to optimize the model. For the model building itself, we determined the best k value by iterating over a range of numbers from 1 to 20 and creating a KNN model for each value. We then used our evaluation metrics to determine the optimal k. In order to ensure the validity, we ran the testing with a grid of tuning parameters by using a training control utilizing a repeated cross validation method with 10 resampling iterations and 3 complete sets of repeated k-fold cross-validation. From this, we determined that a KNN model with k value of 11 performed best, as well as had the best overall metrics of every model we looked at.

While this model was successful in predicting the correct diagnosis with high levels of accuracy, there are some concerns with KNN models. For one, while this dataset worked fine, if we were to incorporate more data, the prediction complexity and computational needs could increase. Moreover, if any significant outliers were introduced, there could be significant effects on the predictions. Finally, this model used every feature from the data, and assigned them equal importances, which is not necessarily true. We wanted to see if we could distill our models down even further so that they would only rely on a small number of predictors. Thus, as noted before, we used feature reduction to further improve our model through simplification. With the GAM model, we looked at different combinations of the most important variables and found that using the "worst" area, texture, and concave points, resulted in a model that had a predictive accuracy of greater than 95%. The GAM was set up using the binomial family to specify the logistic model link function and the restricted maximum likelihood method for smoothness selection [15]. We then ran it with a training control that utilized Leave-One-Out-Cross-Validation and generalized cross validation for the prediction error criteria to validate our results.

Results

The results across all of our models can be seen in Table 1. The best performing model was the KNN one; it had the best metrics for accuracy, specificity, Cohen's kappa, and F1 score. Overall, for correctly diagnosing breast cancer, we would want to focus on having the highest specificity and accuracy, as it is critical that correct diagnoses are made especially to avoid wrongly treating patients with invasive followup procedures. However, because these diagnoses would be used in a clinical setting, we want to ensure that it is also easily interpretable. Comparing the GAM and KNN models, while the metrics were slightly lower, they compared very favorably with an average difference of 0.021 between each metric (and equal sensitivities) and the GAM model required only 3 features versus 30.

Our approach looked promising dealing with this dataset, generating high accuracy metrics with a multitude of models, even models requiring only a few important features. However, there are concerns given the size of the dataset about how applicable the methods would extend if more data points were added. Beyond that, all the features in this dataset were calculated using the same image processing methodology. In order to have a system that could completely replace diagnosis by a pathologist, the manner of feature computation would have to be standardized which would require either adoption of the same procedure as used by the team who created this dataset, or a concurrent signal processing structure that could be applied to any set of biopsy images. If this can be developed, we believe our models are well suited to be applied in more clinical settings.

Conclusions and Future Work

In conclusion, our KNN model resulted in the highest performance among all tested models and scored the highest accuracy at 97.66%, while our GAM models using only 3 features reached a comparable accuracy of 95.91%. There are two limitations of our current method that come to mind. First, there is a concern with potential selection bias - samples in this dataset were chosen by experienced pathologists, and it is possible that the ones chosen were mostly suspected malignant cases, thereby influencing the disparities between cell groupings. We could verify this by using random selection of images on a new set of samples and rerunning our models. Second, as mentioned above, there are concerns based on the pre-processing and how easily repeatable it is to use the methodology used for feature extraction. To solve this problem, future steps would include a feature extraction element in the whole data analysis process. Using this, we could create a completely automated system to preprocess the raw biopsy images, compute the features, and predict the malignancy status. Finally, it would be interesting to see if the models used in this study could be applied to other types of cancers in order to ease the burden of prediction diagnosis in those areas as well.

References

[1] Harbeck, N., Penault-Llorca, F., Cortes, J. et al. Breast cancer. Nat Rev Dis Primers 5, 66 (2019). https://doi.org/10.1038/s41572-019-0111-2

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. (2018), Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68: 394-424. https://doi.org/10.3322/caac.21492

[3] Frable WJ. Needle aspiration biopsy: past, present, and future. Hum Pathol 1989;20:504–17.

[4] Wöckel, Achim et al. "The Screening, Diagnosis, Treatment, and Follow-Up of Breast Cancer." Deutsches Arzteblatt international vol. 115,18 (2018): 316-323. doi:10.3238/arztebl.2018.0316

[5] The American Cancer Society medical and editorial content team. (n.d.). Breast cancer statistics: How common is breast cancer? American Cancer Society. Retrieved April 12, 2022, from https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

[6] Bray, F., Ferlay, J., Laversanne, M., Brewster, D., Gombe Mbalawa, C., Kohler, B., Piñeros, M., Steliarova-Foucher, E., Swaminathan, R., Antoni, S., Soerjomataram, I. and Forman, D. (2015), Cancer Incidence in Five Continents: Inclusion criteria, highlights from Volume X and the global status of cancer registration. Int. J. Cancer, 137: 2060-2071. https://doi.org/10.1002/ijc.29670

[7] The American Cancer Society medical and editorial content team. (n.d.). Fine needle aspiration (FNA) of the breast. American Cancer Society. Retrieved April 12, 2022, from https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html#:~:text=During%20a%20fine%20needle%20aspiration,needle%20biopsy%20is%20often%20preferred

[8] Rahul Nadda, Ashish Kumar Sahani & Ramjee Repaka (2021) A Systematic Review of Real-time Fine-needle Aspiration Biopsy Methods for Soft Tissues, IETE Technical Review, DOI: 10.1080/02564602.2021.1955758

[9] Bleyer, A. & Welch, H. G. Effect of three decades of screening mammography on breast-cancer incidence. N. Engl. J. Med. 367, 1998–2005 (2012).

[10] W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Analytical and Quantitative Cytology and Histology, Vol. 17 No. 2, pages 77-87, April 1995.

[11] Iranpour Mobarakeh, Majid. (2007). Breast Cancer Detection from FNA Using SVM and RBF Classifier.

[12] Breast Cancer Wisconsin (Diagnostic) Data Set. UCI Machine Learning Repository: Breast Cancer wisconsin (diagnostic) data set. (n.d.). Retrieved April 12, 2022, from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[13] W.N. Street, W.H. Wolberg and O.L. Mangasarian Nuclear feature extraction for breast tumor diagnosis.IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.707&rep=rep1&type=pdf

[14] Rajesh Kumar, Rajeev Srivastava, Subodh Srivastava, "Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features", Journal of Medical Engineering, vol. 2015, Article ID 457906, 14 pages, 2015. https://doi.org/10.1155/2015/457906

[15] Wood , S. N. (n.d.). Generalized Additive Model Selection. R: Generalized additive model selection. Retrieved April 12, 2022, from https://stat.ethz.ch/R-manual/R-devel/library/mgcv/html/gam.selection.html