Predicting Breast Cancer From FNA Using Machine Learning

Justin Steinman jsteinman@umass.edu

Derek Liu dzliu@umass.edu

Yongye Tan yongyetan@umass.edu

Billy Girard wgirard@umass.edu

1 Introduction

Breast Cancer is among the most commonly diagnosed cancers in women, affecting one in eight women in the United States; while dangerous, 5-year survival rate with early detection is 93% (ear). A machine learning model capable of identifying cases of breast cancer before they are typically detected would prevent countless deaths: as such, the problem this project addresses is quickly and effectively identifying breast cancer. The input to the algorithm consists of the following statistics in vector form, computed from a digitized image of a fine needle aspirate (FNA) of a breast mass:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area - 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension (coastline approximation - 1)

We then use a random forest classifier to predict either a benign or malignant diagnosis.

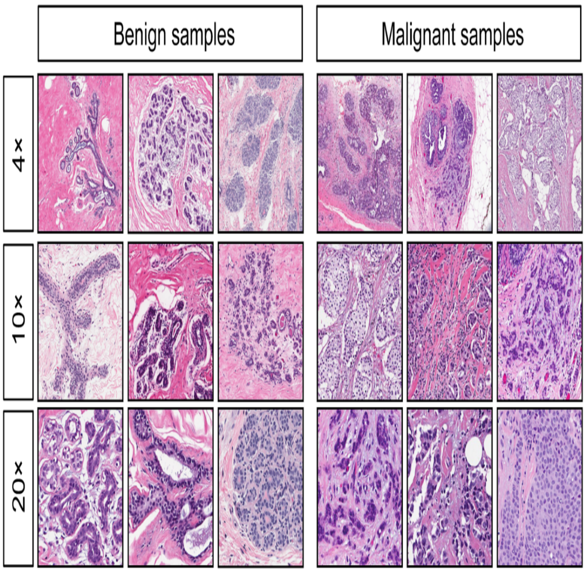
2 Related work

Previous papers have also use the data to train classification models. In the paper "Wolberg et al. (4)," a 97% accuracy was acheieved on the validation data and 100% accuracy on new samples. Other papers use the data set to test the effectiveness of machine learning methods. In the paper "Mert et al. (2)" feature reduction with independent component analysis was tested using this data set. It was shown that 30 features reduction was optimal for k-nearest neighbors.

3 Dataset and Features

Our dataset has **569** patients, which are collected from FNA, and there are 10 features for each patient. Additionally, our data is obtained from the *"Diagnostic Wisconsin Breast Cancer Database,"* and a patient can be represented as (842302, M, 17.99, 10.38, 122.8, 1001, 0.1184, 0.2776, 0.3001, 0.1471, 0.2419, 0.07871). Feature scaling is being used here, before splitting the data set values, to normalize each continuous feature by dividing it by the maximum value of the column as some data from one column is too large compared to other values from the same column.

(Two Types Of Cancer After Zooming-in)



(3)

4 Methods

The first classification model used was k-nearest neighbors, where $k = 5$. In this model, each new data point is compared to the labels of its k nearest neighbors, and the plurality is output.

Next, logistic regression was used. This is a model where each feature is multiplied by some weight, summed, and then that sum is fed into a logistic function (a type of sigmoid function):

$$f(x) = \frac{1}{1 + e^{-x}}$$

The output of this function falls on the range $0 \leq f(x) \leq 1$, which signifies the probability of the output being 0 or 1. If there were more than two possible labels (benign and malignant), multinomial logistic regression could be used to account for this. However, it is not necessary because the output in this case is binary.

The third model used was a decision tree, which is essentially a tree where branches are conditional statements about the features of the tree, and the leaf nodes are the possible labels. Each new data point starts at the root and continues down the tree abiding by its conditionals until it reaches a leaf node and its label is determined.

Finally, a random forest classifier was used, which constructs n decision trees (in this case, $n = 100$) designed to classify a new data point. Once each tree makes a prediction, the plurality of their outputs is returned. The use of many different decision trees allows the model to avoid overfitting to its training set. That is, the classifier will be more consistent and versatile than a single decision tree.

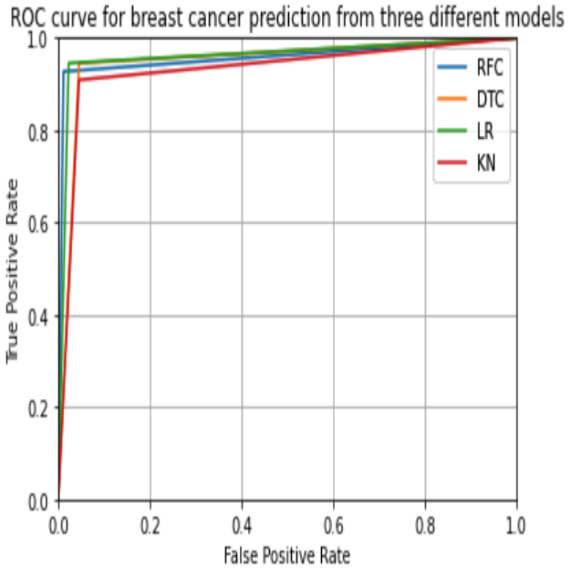
5 Experiments/Results/Discussion

In the models we used to classify the type of breast cancer we tuned the hyper-parameters to optimize the performance of the models. For k-nearest neighbors, we chose $k = 5$ because the model seem to perform well in contrast to when $k = 3$ and $k = 11$. The number of neighbors is important for the performance of the model–too low will increase the effect of noise on the performance, too high and there will be higher bias. We chose an odd number of neighbors because of the chance of a tie with an even number of neighbors. the metrics we used was primarily accuracy. We also calculated the F_1 score, the confusion matrix, and the AUC-ROC curve.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_1 = \frac{precision \cdot recall}{precision + recall}$$



We have found the computed accuracy is 0.944, 0.937, 0.923, and 0.916 and F_1 scores of 0.953, 0.920, 0.917, 0.954 by using random forest, logistic regression, k-nearest neighbor, and decision tree classification, respectively. It is evident that random forest produces the highest accuracy and decision tree classification returns the lowest accuracy.

6 Conclusion

Using only 10 features gathered from three nuclei, this machine learning model was capable of predicting the presence of breast cancer with incredibly high accuracy above 90%. Given more training data and time, this figure could easily increase.

The random forest classification model was concluded to work best with the given training data. k-Nearest neighbors did not perform as well as the other models because with low values of k, a hard decision barrier is formed based off the training data. This causes problems with the testing data because the model fails to generalize the problem. The same goes for just a single decision tree. When many different random trees are generated (i.e., a random forest), the model is more generalized and will work with more data points. Logistic regression and random forest are designed to avoid this problem of over-fitting.

With the use of these classification models, meaningful change in the medical world can be effected to save many lives. While this project focused on breast cancer, one specific disease, machine learning can and will be used for many other ailments to quickly and effectively provide patients with diagnoses.

References

[ear] Early detection is key — carol milgard breast center.

[2] Mert, A., Kılıç, N., Bilgili, E., and Akan, A. (2015). Breast cancer detection with reduced feature set. *Computational and Mathematical Methods in Medicine*, 2015:265138.

[3] Sumbria, S. (2021). Breast cancer diagnostic dataset — eda.

[4] Wolberg, W. H., Street, W., and Mangasarian, O. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, 77:163–171.