Evaluating Benign and Malignant Breast Tumor Cells from Fine-Needle Aspirates

Abstract

Data of 569 Fine-Needle Aspirate (FNA) samples from breast tumors, with records of 10 cell nuclear morphology attributes, was taken from the University of Wisconsin. We aimed to discover whether there were pairs of attributes that could accurately diagnose breast tumor cell aspirates. A conventional decision tree was first constructed to show the four best variables for diagnosing breast tumor cell aspirates. This decision tree had a 91% accuracy and the best four variables were area, perimeter, concave points, and texture. Bivariate decision trees were constructed using all possible pairs of attributes. The best pairs were area and compactness, concave points and texture, and concavity and radius (93%). Finally, we created an interactive visualization for users to experiment with different combinations of cell nuclear morphology attributes.

Background and Significance

According to the National Breast Cancer Foundation, one in eight women in the United States will be diagnosed with breast cancer in her lifetime, making it the second leading cause of cancer death among women. Cancers are malignant forms of tumors and are often defined as being able to spread throughout the body (Lodish et al 2000). Cancer can be diagnosed through multiple methods. One of which is a biopsy. Three common types of biopsies for breast cancer are core-needle biopsy, surgical biopsy, and Fine-Needle Aspiration (FNA). FNA allows for a small sample of the tumor to be extracted in a non-invasive process, making it a safer method than surgical removal.

Wolberg and Mangasarian (1999) developed a technique to use visual characteristics of the cells to identify malignant and benign tumor cells. They took FNA samples and measured the size, shape, and texture cell nuclei. They developed a malignancy analysis algorithm which is 90% correct for a set of 119 FNA samples. Samples from their analysis showed that digital images of cells from FNA samples can be used to assess tumor cells accurately. We took a different approach to study if there are particular nuclei attributes that are more significant than others in determining the malignancy of cells. As malignant nuclei are more deformed than benign cells, we hypothesized that the attributes of area, concavity, texture, and perimeter are significant classifiers because these measurements together characterize the cell nuclei when viewed on a 2D plane and in 3D space.

To address our research question, we used a Classification and Regression Tree (CART) analysis using the rpart package in Rstudio to determine which predictor variables are more significant in determining breast tumor cell malignancy. We also constructed a Shiny App to create an interactive visual on which two factors are the strongest predictor variables for tumor cell malignancy. We hope our Shiny App would make the breast cancer cell data more accessible to others, allowing users to see the exact clusters of malignant and benign cells based on the predictor variables chosen.

Methods

I. Data Description

We obtained our Breast cancer cell data from the University of Wisconsin's Machine Learning Repository website (Lichman 2013). The data set includes measurements from 569 FNA samples of breast tumor cells' nuclei. The explanatory variable is the diagnosis of the tumor cell (malignant or benign) and the predictor variables are radius (μm), texture (grayscale value), perimeter (μm), area(μm^2), smoothness (μm), compactness, concavity (μm), concave points, symmetry (μm), and fractal dimension (μm), all of which are the mean values (table 1). We assume that in each FNA sample, if the majority of the tumor cells are malignant, then the tumor is cancerous.

II. Rationale for Independent Variable

The original breast cancer data set was created by collecting data on ten attributes of cell morphology from breast tumor cell aspirates. The original data set provided the mean, standard error, and the "worst"(the average of the largest three values). For our project, we decided to only use mean values for all attributes. Mean measurements were used by the studies that we referenced (Cui et al. 2007 and Narasimha et al. 2013) and we wanted our final Shiny App to be in accordance with findings by other researchers. The limitation to using mean values is that there may be more overlap of variable values than the "worst" values. The mean takes both the lower and upper end of the sample's measurements into account while the "worst" only accounts for the largest values. It is possible that the difference between mean values for malignant and benign cells would be smaller and harder to detect.

III. CART Analysis

The purpose of implementing decision tree was to predict the malignancy of breast tumor cell aspirates using visible qualities of the cells' morphology. We created a decision tree with mean measurements. By setting a seed for our data, we guaranteed that data could be replicated. We then split the data into two-thirds for training and one-third for testing to prevent the model from being over fit. We used the training data to create a multivariate decision tree and the testing data to verify the accuracy of the the tree by assessing the predictability of the decision tree. Additionally, we created a confusion matrix, which is a table summarizing the actual classifications versus the predicted classifications. The accuracy was calculated by the following formula:

$$Accuracy = (Sum\ of\ true\ positive\ and\ negative) / (Number\ of\ all\ observations)$$

where true positive and negative are the number of correct predictions from the model and total predicted is the sum of both correct and incorrect predictions made with the model for the given testing data . We also created a plot for complexity parameter (CP), which indicates how much a node improves the tree's predictability, minimizing cross-validation errors. We pruned the tree by selecting the CP with the smallest cross-validation error (x-error), and compared the accuracy of the pruned tree with the unpruned tree.

IV. Shiny App

To construct the Shiny App, we first created bivariate decision trees using the training data. Just like how multivariate decision tree was created, the original data was separated into training and testing to build the trees and then assess the accuracy of the trees, respectively. If the decision tree has a poor predictability, then a great number of predictions would be inaccurate. This is a non-conventional approach to using CART because we used the decision trees to determine which two attributes of cell nuclear morphology could construct decision trees that accurately predicted the diagnoses of breast tumor cells. In other words, the Shiny App provides a visual to seeing the classification of breast tumor cells depending on the visual indicators of a cell's nucleus, and was constructed in reference to Dennis Liu's Iris App.

Bivariate decision trees of all possible combinations for the 10 attributes were created and then the accuracy for each tree was evaluated using the testing data. We wanted to see if pruning the trees would simplify the decision trees without sacrificing the tree's accuracy. To know how much to prune each bivariate decision tree, we used the smallest complexity parameter (CP), which corresponds to a tree with the smallest cross validation error. By assigning a CP value to prune the tree, we obtain simpler trees. To test if accuracy was reduced by pruning, we applied testing data to each of the pruned trees.

For the final shiny app, we included an interactive scatterplot in addition to a summary table of the accuracies for all bivariate decision trees and the plots of the decision trees. With the scatterplot users can select the x and y variable from the 10 predictor variables and see the distribution of malignant and benign cells.

Results and Discussion

I. Decision Tree

The multivariate decision tree was created by using all ten variables with diagnosis. The lowest x-error was when CP = 0.015 and the number of nodes was five (fig. 2). The important attributes for the tree turned out to be concave points, area, perimeter, and texture. As illustrated in (fig. 1), Each node showed the predicted probability and percentage of observations in the node. We evaluated the accuracy of the decision tree to be 91.01%. The pruned tree and the original tree had the same accuracy; therefore, we decided not to prune the tree since pruning did not improve the accuracy.

II. Shiny App

A total of forty five bivariate decision trees were created by pairing one the ten predictor variables. We decided not to prune the trees as pruning reduced the accuracy of some models (table 2). In most of the bivariate decision trees, the CP values suggested that the trees did not need to be pruned because the accuracy did not improve with pruning. We evaluated the accuracy of each decision tree by using testing data and found that the accuracy of the decision trees ranged between 70% to 94% (fig. 3**).** The highest accuracy was found in the decision tree using the variables area and compactness, concave points and texture, and concavity and radius, all of which were 93.87% accuracy. The lowest accuracy was found in the decision tree using the variables fractal dimension and smoothness (70.07%), and symmetry and fractal dimension (71.43%). The high accuracy of some decision trees suggest that particular predictor variable pairs are better indicators in diagnosing breast cancer cell malignancy. This is also shown in our app. If users select radius and concavity, they will see that the cluster of malignant and benign cells are less diffused than in a scatterplot using symmetry and fractal dimension (fig. 4).

III. Interpretations and Limitations

The decision tree using the variables fractal dimension and symmetry have the lowest accuracy, suggesting that these two variables are shared by nuclei from cells of both conditions, therefore making it difficult to distinguish malignant and benign cells. The models for compactness and area, and concavity and radius have the highest accuracy, which shows that these variable pairs are clear in distinguishing benign and malignant cells (fig. 4, 5, 6). Malignant cells often have more deformed nuclei than benign cells due to factors such as change in protein expression, loss in nuclear lamina structure, mechanical stress from neighboring cells, which may promote the invasiveness of malignant cells (Webster et al. 2009, Cui et al. 2007, Narasimha et al. 2013, Lodish et al 2000). When comparing cell nuclei by concavity and radius, malignant cells have nuclei with a larger radius and are more concave than benign cells (fig. 5). This may reflect the uncontrollable growth of cancerous cells, causing the nuclei to increase in size without a proper structure. The concave points and texture model show that malignant cells have more concave points than benign cells, which aligns with the fact that malignant cells have deformed nuclei (fig. 6).

A limitation to decision trees is that each tree is only useful for predicting cell malignancy in breast cancer cells from a specific population. The data of breast tumor cell aspirates was collected from a specific population in Wisconsin, so it is possible that breast tumor cells from people of different backgrounds may show different trends in nuclei morphology. We also cannot conclude from these decision trees whether a certain pair of variables are more important than others because it is possible that a more accurate tree can be constructed if more than two variables are used. Furthermore, the trees are not robust compared to a random forest since other features of malignant cells are not taken into account. A greater breadth of FNA samples from people of multiple backgrounds can help widen the scope of understanding cancer cell nuclei morphology.

Conclusion

We conclude from the conventional decision tree that breast tumor cell aspirates can be classified as malignant or benign with 91% accuracy given the cell nuclear morphology attributes of concave points, area, perimeter, and texture. In the bivariate decision tree, the best pairs of attributes are area and compactness, concave points and texture, and concavity and radius. Although the bivariate decision trees oversimplify the morphological attributes of malignant breast cancer cells, it is a useful model for visualizing how breast cancer cells can be distinguished as malignant or benign based on visual observations. For future projects, we are interested in learning advanced techniques such as random forest to develop a more holistic model for classification.

References

Cross, S.S. (1997). Fractals in Pathology. *J. Pathol.* 182, 1–8.

Cui, Y., Koop, E.A., van Diest, P.J., Kandel, R.A., Rohan, T.E. (2007). Nuclear morphometric features in benign breast tissue and risk of subsequent breast cancer. *Breast Cancer Res. Treat.* 104, 103–107.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

Liu, H. Y. (2017). Iris App [Shiny App]. Retrieved from https://dennisliu.shinyapps.io/IrisAPP/ Date accessed Nov 28 2017.

Lodish H, Berk A, Zipursky SL, et al. (2000). Section 24.1, Tumor Cells and the Onset of Cancer. Molecular Cell Biology. 4th edition. New York: W. H. Freeman.

Mouelhi, A., Sayadi, M., Fnaiech, F. (2011). Automatic segmentation of clustered breast cancer cells using watershed and concave vertex graph, in: 2011 International Conference on Communications, Computing and Control Applications (CCCA). Presented at the 2011 International Conference on Communications, Computing and Control Applications (CCCA), 1–6.

Narasimha, A., Vasavi, B., Kumar, H.M. (2013). Significance of nuclear morphometry in benign and malignant breast aspirates. *Int. J. Appl. Basic Med. Res.* 3, 22–26.

Neto, J., 2013. Classification & Regression Trees [WWW Document]. Classification & Regression Trees. URL http://www.di.fc.ul.pt/~jpn/r/tree/tree.html (accessed 12.2.17).

Phinyomark, A., Jitaree, S., Phukpattaranont, P., Boonyapiphat, P. (2012). Texture Analysis of Breast Cancer Cells in Microscopic Images Using Critical Exponent Analysis Method. Procedia Engineering, ISEEC 32, 232–238.

Schmitt, O., Hasse, M. (2008). Radial symmetries based decomposition of cell clusters in binary and gray level images. *Pattern Recognition* 41, 1905–1923.

Webster, M., Witkin, K.L., Cohen-Fix, O. (2009). Sizing up the nucleus: nuclear shape, size and nuclear-envelope assembly. *J. of Cell Sci.* 122, 1477–1486.

Wolberg, W.H., Street, W.N., Mangasarian, O.L. (1999). Importance of Nuclear Morphology in Breast Cancer Prognosis. *Clin. Cancer Res.* 5, 3542–3548.

Appendix

Shiny App:
https://gubnerkevin.shinyapps.io/breast_cancer_prediction_app/

Attributes	**Definition**
Radius	Average distance from cell center to cell perimeter
Texture	Standard deviation of gray-scale values; brightness of pixel (Phinyomark et al. 2012).
Perimeter	Distance around nuclear border
Area	Area of nucleus
Compactness	Perimeter^2/area
Smoothness	Variation in the cell's radial lengths
Concavity	Size of indentations in the nuclear border
Concave points	Number of points on an indented section of the nuclear border (Mouelhi et al. 2011)
Symmetry	Deviation of nuclei shape from ideal (Wolberg, Street, and Mangasarian 1999); measured by finding "the relative difference in length between line segments perpendicular to and on either side of the major axis" (Schmitt and Hasse 2008)
Fractal dimension	Measurement of nuclear border irregularity (Cross 1997)

Table 1. Description of cell nuclear morphology attributes.

Variables	Radius	Texture	Perimeter	Area	Smoothness	Compactness	Concavity	Concave Points	Symmetry	Fractal Dimension
Radius	-	-0.006803	0	0	-0.01361	0	0	0	0	0
Texture	-	-	0	0	0	-0.06122	-0.06802	-0.027210884	0.01361	0.02721
Perimeter	-	-	-	0	0	-0.02041	-0.013605442	-0.020408163	0	0
Area	-	-	-	-	0	0	0	0	0	0.01361
Smoothness	-	-	-	-	-	-0.03401	0	0.006803	0.02041	0
Compactness	-	-	-	-	-	-	0.01361	0.02721	0.03401	0
Concavity	-	-	-	-	-	-	-	-0.01361	-0.02041	0
Concave points	-	-	-	-	-	-	-	-	0	0.01361
Symmetry	-	-	-	-	-	-	-	-	-	0.01361

Table 2. Difference in accuracy of pruned to non-pruned bivariate decision trees.

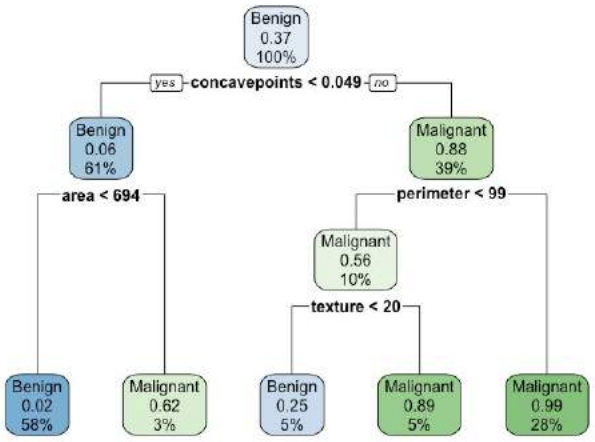

Figure 1. Decision tree for diagnosing breast tumor cells by the four important factors of concave points, area, perimeter, and texture.

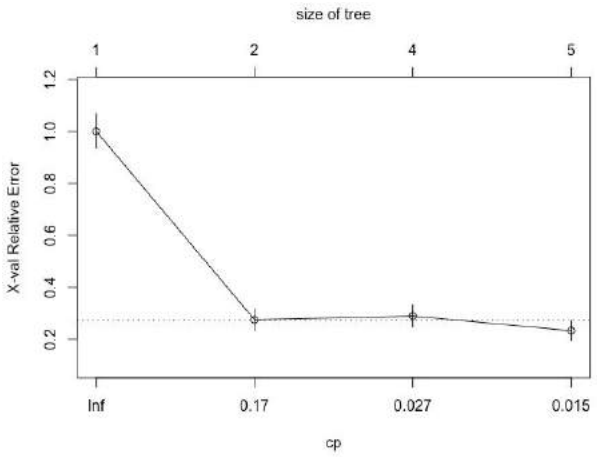

Figure 2. Plot of complexity parameter for decision tree in figure 1.

Scatterplot | Decision Tree | Decision Tree Accuracy Table

Show 11 entries Search:

Variable	Radius	Texture	Perimeter	Area	Smoothness	Compactness	Concavity	Concave_Points	Symmetry	Fractal_
Diagnosis										
Radius		0.9047619	0.8707483	0.9047619	0.8639456	0.9251701	0.9387755	0.9115646	0.8843537	0.884353
Texture			0.8911565	0.9047619	0.7823129	0.8639456	0.9183673	0.9387755	0.7619048	0.7142857
Perimeter				0.8707483	0.8707483	0.8979592	0.9047619	0.9115646	0.8843537	0.884353
Area					0.8775510	0.9387755	0.9115646	0.9183673	0.8911565	0.8911565
Smoothness						0.8571429	0.8639456	0.9047619	0.5938776	0.700680
Compactness							0.8707483	0.8843537	0.7687075	0.8435374
Concavity								0.9251701	0.8707483	0.8775510
Concave Points									0.9115646	0.8979592
Symmetry										0.7142857
Fractal Dimension										

Figure 3: Decision Tree Accuracy Summary table.

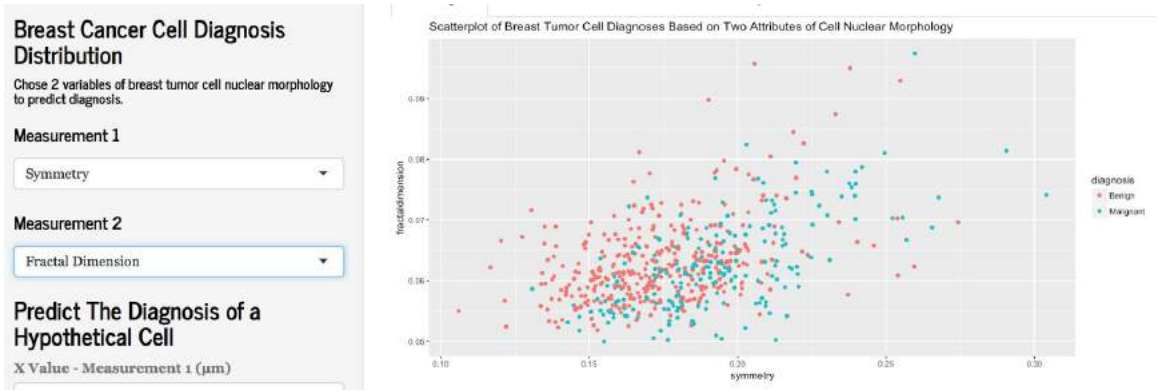

Figure 4. Scatterplot of benign and malignant cell nuclei by symmetry and fractal dimension.

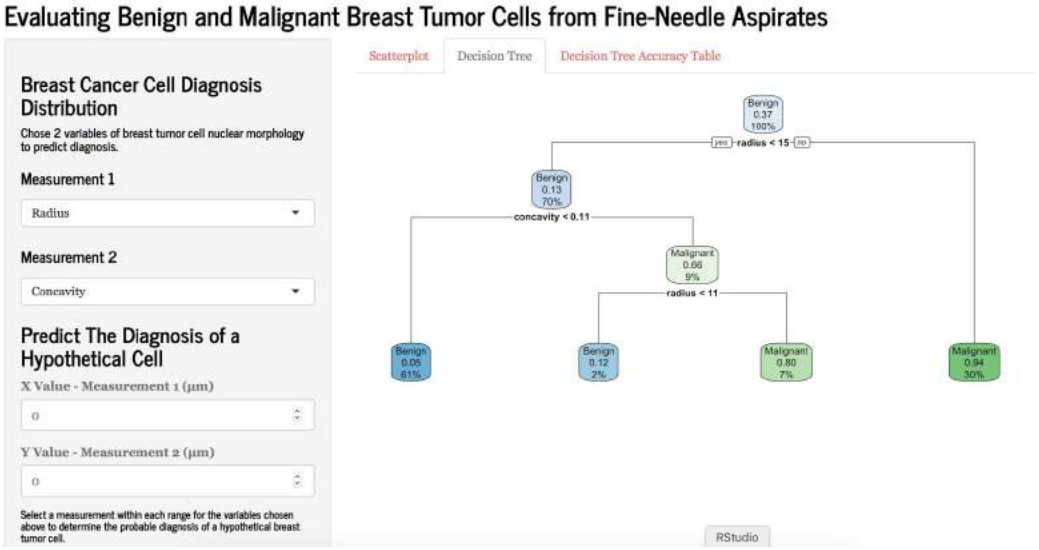

Figure 5. Decision tree predicting breast tumor cell diagnosis by radius and concavity.

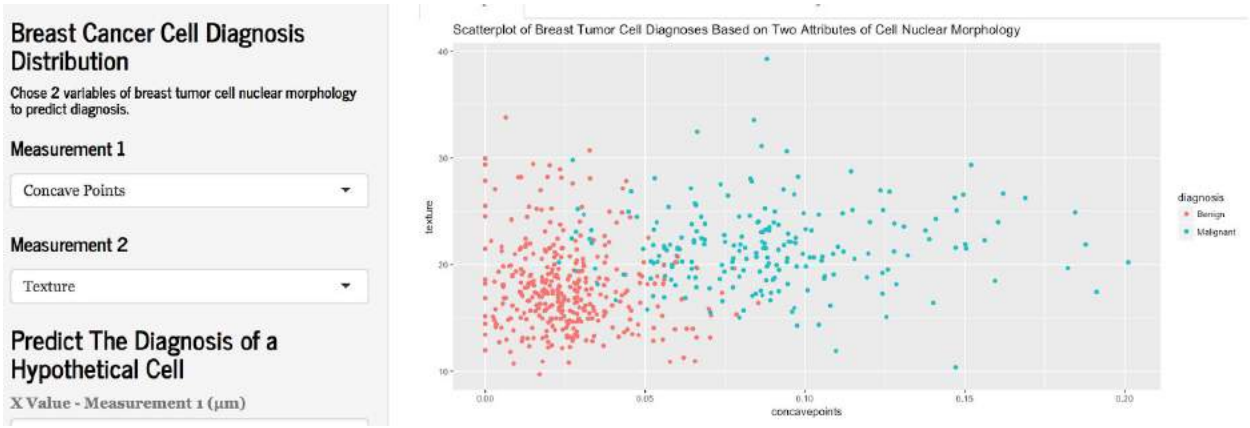

Figure 6. Scatterplot of malignant and benign cells by concave points and texture.

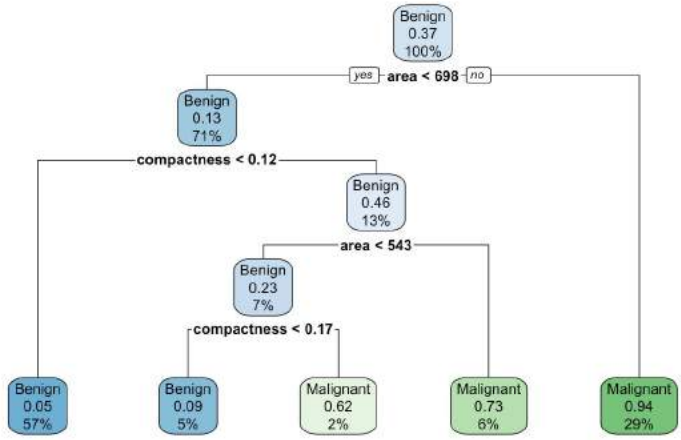

Figure 7: Decision tree for area and compactness. This graphic was taken from the shiny app.

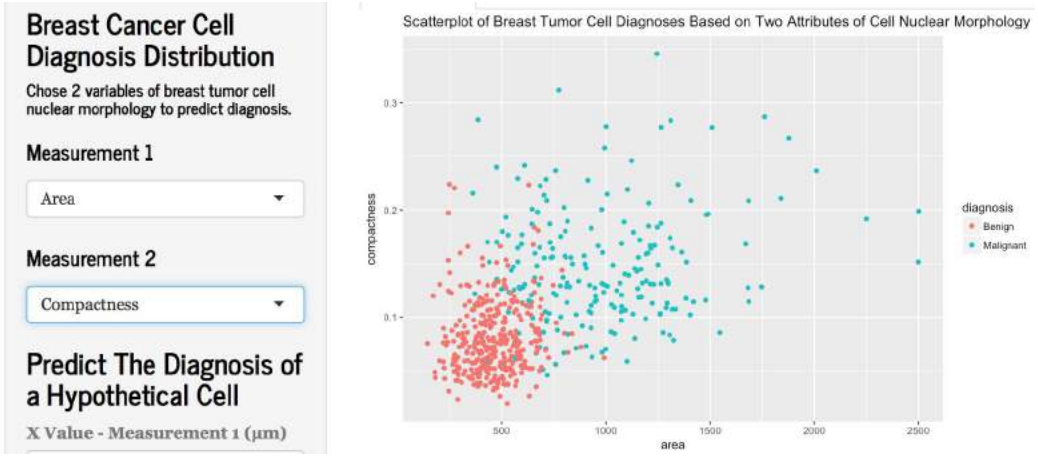

Figure 8: Screenshot of scatter plot from Shiny App. This scatterplot shows the distribution of diagnosis when looking specifically at area and compactness.