# Classifying Breast Cancer Malignancy Based off Cell Type Attributes with Machine Learning

Walker Azam and Maria Partida-Aguilar

**Abstract**

Cancer is a complex disease that is time sensitive and requires early diagnosis for the best prognosis. In nations where there is limited access to extensive screening equipment, we wonder if there is a cost effective and accessible testing, such as Fine Needle Aspirations, that could be effective at diagnosing tumor malignancy early on. Using a dataset provided by Dr. William H Wolberg of the University of Wisconsin-Madison, we explored the possibility of using just tissues samples with ten cell type attributes measured. The classifiers we used to approach this were Nearest Neighbor, Linear Discriminant Analysis and Support Vector Machine. With each classifier we looked for the accuracy of properly classifying a diagnosis for a tissue sample and evaluated the effectiveness of using this type of classifier in the real world.

**Introduction**

In the United States, 1 in 8 women will be diagnosed with breast cancer in their lifetime ("U.S.). Of those women, 30% will develop metastatic breast cancer ("U.S"). In the world, there was a diagnosis of over 2 million new cases of breast cancer in 2018 ("Breast" 2018). Breast cancer has continued to be one of the most diagnosed cancers in women across the world ("Breast" 2018). Breast cancer has been characterized as a western disease with incidences of 89.7 per 100,000 women in Western Europe and 19.3 per 100,000 women in Eastern Africa ("Breast" 2016). Despite those high values, there are well established medical infrastructures put into place in developed countries, resulting in higher survival rates compared to developing nations.

The traditional route of getting a diagnoses of breast cancer in the U.S. can range from a self-breast examination, scheduled mammograms, biopsies of lumps, MRI's for clearer images on location of tumors, invasive surgeries for identification and removal. These are all tests involved in the screening, diagnoses and prognosis of breast cancer. Without health insurance, these tests could get very expensive, and that is just in developed nations. In developing nations there is a lack of access to some of these extensive screening processes which comes from a lack of well-established medical infrastructure. The lack of infrastructure then disadvantages women by depriving them to access to regular check-ups in addition to the extensive screening process that have proven effective in developed nations. This results in lower survival rates in these nations of below 40% because when women do get diagnosed it is mainly late stage breast cancer and once again a lack of access to treatment when there is a diagnosis ("Breast" 2016).

The question becomes how can we better diagnose cancer in these developing nations? Having access to easy, quick and affordable diagnostic testing will allow for earlier diagnosis which could potentially result in higher survival rates. Dr. William H. Wolberg, from the University of Wisconsin-Madison, was asking a similar question of whether a fine needle aspiration (FNA) was enough to get an accurate diagnosis of benign or malignant breast cancer. He obtained FNA samples of 699 patients and measured out nine different cell attributes that were visible and quantifiable underneath a microscope that he considered relevant to diagnosis ("Machine"). This was then comprised into the dataset we obtained from this source: https://www.kaggle.com/roustekbio/breast-cancer-csv.

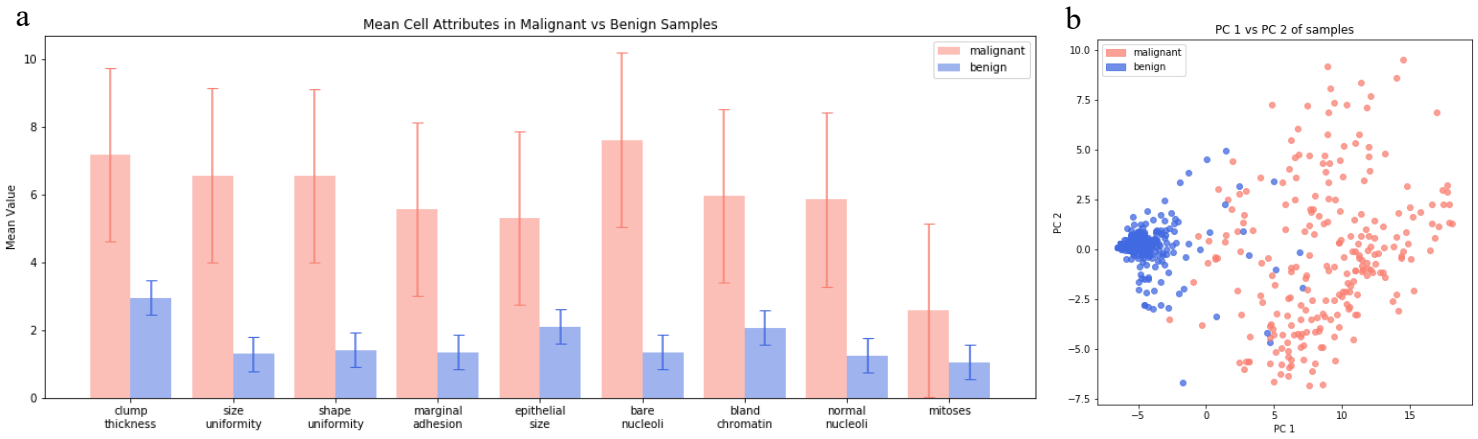


*Figure 1. a) Preliminary visualization of the mean and standard deviation of all the cell type attributes contained within the data set based off classification. Shows that there is a distinguishable difference between benign and malignant tissue samples. b) Graphing of benign and malignant data points based off the PC1 and PC2 values.*

## Methods

As aforementioned, the dataset was made up of 699 tissue samples of FNA and contained 10 identifiers. The identifiers measured were cell attributes such as 'clump thickness', 'marginal adhesion', and etc. The measurements of these identifiers were scaled 0 to 10. The original size of the dataset before cleaning was 699 by 11.

Our coding was done on Python, using the packages pandas, numpy, matplotlib.pyplot, and various scikitlearn functions. We first converted the dataset into a '.csv' file manually by copying and pasting the values into a spread-sheet. The column 'sample ID' was dropped, which was patient ids for each tissue sample. Samples that had missing values, denoted as a '?' were dropped. The 'bare nucleoli' column, which had missing values, was originally a string that had to be converted to integer values. After cleaning the dataset, the final shape of our data was 683 by 10. The 10$^{th}$ column was the 'class' indicator where a 2 classifies as a benign sample and 4 classifies as a malignant sample. We'd use this column as our label throughout our analysis.

We then did a preliminary visualization of the dataset by grouping the samples by their classification. We had 444 benign samples and 239 malignant samples. To visualize the difference between malignant and benign cells, we calculated the average value and standard deviation for each cell attribute, excluding the 'class' column, and plotted this on a bar graph (fig 1a). The graph was separated by benign and malignant samples. From the preliminary visualization, we decided that dimensionality reduction, to generate principal components, would be effective in discerning the two classes.

Principle Component Analysis was used from the sklearn package on python. To determine which principal components should be used in the analysis we plotted the variance of each Principal Component (PC) and plotted the fraction of the total variance explained (fig 2). From the figure, PC 1 and PC 2bexplained most of the variance, with approximately 76% of the total variance explained by those two components. To visualize the data, we plotted PC 1 vs. PC 2 on a scatter plot, distinguishing the malignancy of the samples with color (fig 1b). From the visualization were able to see separation of the samples into two distinct areas of the graph. The Machine Learning classifiers we created would use these first 2 components to determine whether a cell is benign or malignant.

The machine learning classifiers used were Nearest Neighbors Analysis (KNN), Linear

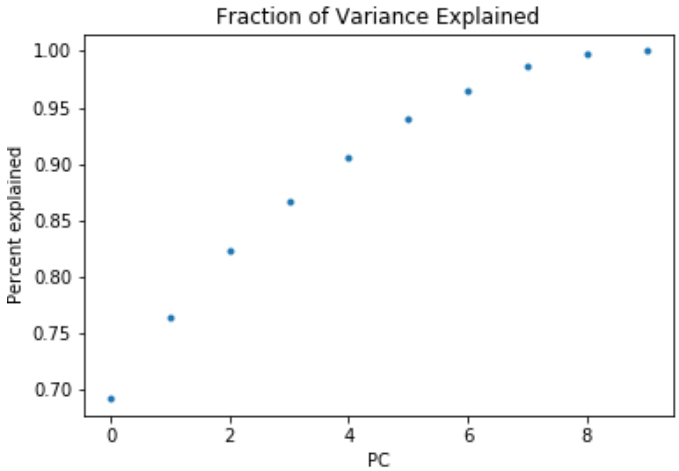


*Figure 2. Plotting of percentage of variance explained by number of principal components used from a PCA of all samples.*

Discriminant Analysis (LDA), and Support Vector Machine with RBF kernel (SVM). These three classifiers were all imported from scikitlearn on python. We implemented a 'split' function that separated and randomized our dataset into two training and testing datasets. The training dataset comprised of 70% of the data, the other 30% was separated into an independent testing dataset. We did this split of our dataset with this ratio and randomization to avoid over-fitting our classifiers. We separated our training dataset into its values and its class labels, running a PCA on the training data values. Our classifiers would use the first 2 PCs to train its classifier on. Since the classifiers were only fit with the training data PCs, we could use the test dataset to generate an accuracy score because it was withheld from the training process. The testing data accuracy was generated using the classifier.score function. To generate the average accuracy score for each of the classifiers, we did 100 iterations and averaged out the accuracy. For each iteration, we randomly split the training and test data, running a PCA with a new set of training data each time. The testing data was separate each time to avoid any interaction with the training of the classifier to guarantee our accuracy has not been tampered with.

Nearest Neighbors uses an approach for classification where it uses the labels of the training dataset to determine classification of new samples, based off how close it is to other labelled data. The classifier generates a decision boundary by calculating the closest distance to a nearby data point called its 'nearest neighbor'. The boundary area is the shortest Euclidean distance any point could be from its closest neighboring data point ("1.6."). The classifier takes in the input of how many nearest neighbors it should consider, determined by its n_neighbors value. For our method we used 2 nearest neighbors because this confers a high accuracy without over-fitting the classifier model to the training data presented. For our data we decided to display the decision boundary of our classifier with our training and testing data points. Our figures display the nearest neighbor's boundary generated from one of the 100 iterations as an example. To generate the boundary, a colormesh had to be constructed where the plot of the decision boundary could be over-laid. Figure 4a displays the test points placed on the decision

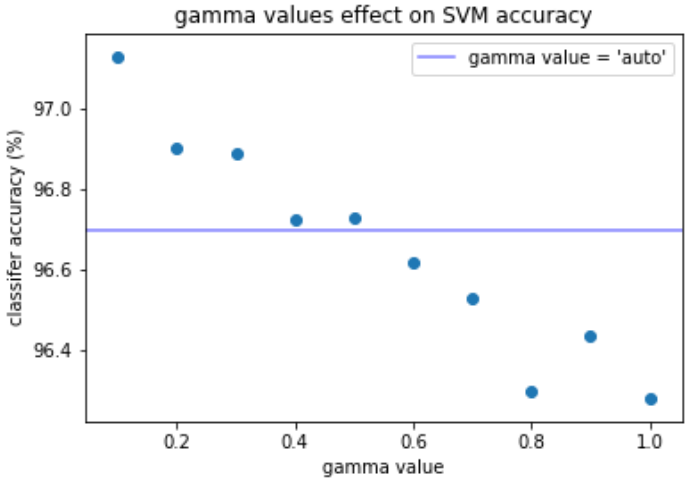


*Figure 3. Plotting of classifier accuracy following a hundred iterations using different gamma values. Horizontal line represents accuracy when using the "auto" gamma value.*

boundary generated. This visualization allows us to observe how the classifier functions with the training values, specifically how it treats 'islands' of points of one classification surrounded by the other.

In order to supplement the visualization of this decision boundary, we decided to also construct a confusion matrix of this iteration that could explicitly display the classification of test data samples (fig 6). The confusion matrix was imported from sklearn, and we decided to use a normalized matrix that could show the accuracy of each classification, and misclassification, as a fraction of 1. The confusion matrix displays the true labels of test points on the y-axis, and the classifier predictions on the x-axis.

The second classifier used was Linear Discriminant Analysis. LDA generates a decision boundary that is linear which avoids islands of classification that KNN has. Whereas Nearest Neighbors plots a decision boundary based of each training points distance, LDA generates a linear divider based off each class's Gaussian Density ("Sklearn"). As displayed in our PC 1 vs. PC 2 plot, our data distribution generally divides the benign and malignant classes into two halves of the graph, however data points in the middle pose a harder challenge. Visualizing LDA boundary offers insight into where certain points are misclassified based on a linear divider. We plotted the decision boundary of the LDA classifier on a generated colormesh, and then overlaid the test data points (fig 4b). We then generated a confusion matrix for this iteration to

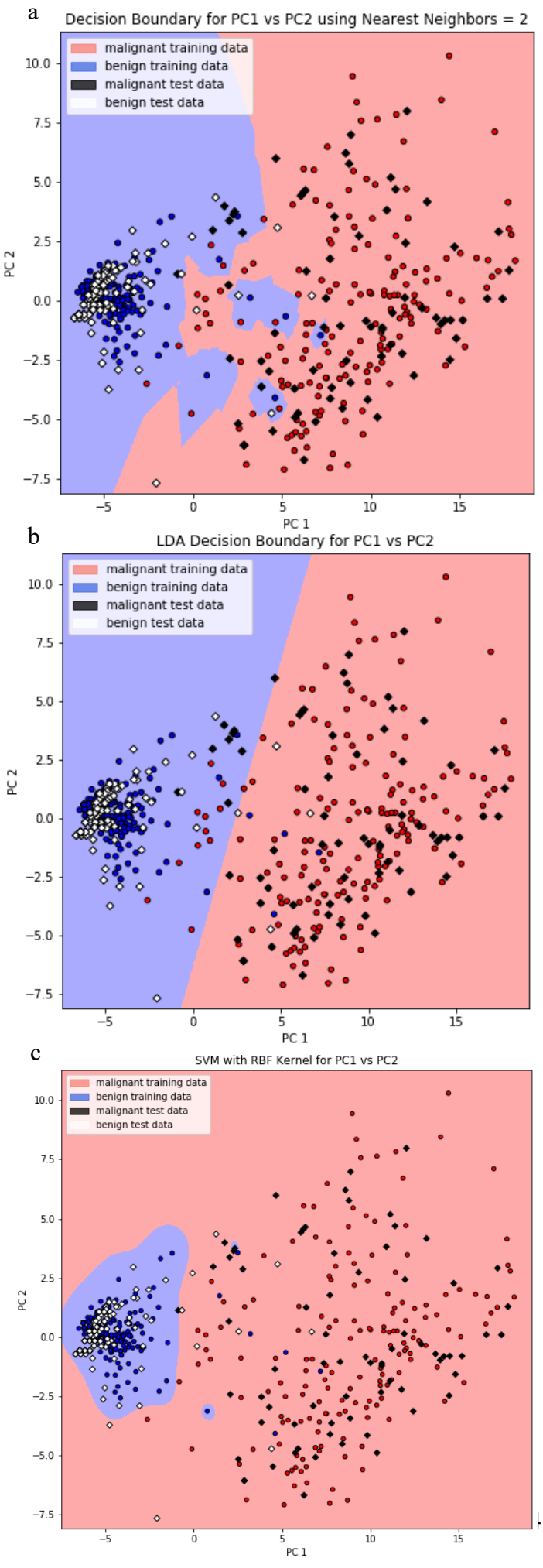


*Figure 4. Plotting of the three classifiers used with their corresponding decision boundaries based off a single iteration of classification. Test points are overlaid the original boundary. a) KNN decision boundary, islands were created from training data that was classified. b) LDA displays a linear boundary. c) SVM with an RBF kernel that produces the main island, and two smaller islands, of benign classification.*

show the normalized score of the LDA classifier (fig 6b).

Support Vector Machine (SVM) is the final classifier we decided to implement. SVM classifier works by generating a decision boundary – called a hyper plane – that maximizes the margin from 'support vectors'. Support vectors are the training data points close to the boundary. These are used to influence the orientation and alignment of the boundary (Gandhi). With support vectors, SVM maximizes margins to create an optimized boundary, used to differentiate classifications. SVM stands out since it is useful for high dimensionality data, however, to reliably compare classifiers we still fit the classifier on the first 2 PCs of the training data. Moreover, SVM is prone to over fitting, which we tried to avoid by comparing the effects of different 'gamma' values from the classifier. Gamma value determines the radius influence each point has as a support vector. We found that gamma value of 0.1 had the highest accuracy, and as gamma value increased, the accuracy of the classifier dropped. We decided to use the 'auto' gamma value. The auto value of gamma scales in terms of the data, and variance, so it is adaptable to any dataset's number of support vectors. This gamma value maintained a high accuracy, yet avoided over fitting, which a 0.1 gamma could have.

SVM also functions with specific 'kernels' that determine the characteristic of its hyper plane boundary. We chose to use an RBF kernel in our classifier. This kernel generates a radius-based boundary around the points and offers a contrast to our other classifier models ("RBF"). We visualized this decision boundary, and again overlaid the test points (fig 4c). For that iteration we also generated a confusion matrix showing the effectiveness of a radial boundary in classification (fig 6c).

After generating a mean accuracy score for each classifier, from 100 iterations, we plotted the results on a histogram (fig 5). The histogram also shows the standard deviation as error bars.

## Results

In the preliminary results, there was a clear distinction visible when looking at the graphed averages of the different cell attributes against each classification of malignant of benign, showing there is a visible trend or difference between the two classifications (fig 1a). This suggests that there is a possibility to effectively classify malignancy just based of cell type attributes obtained from a FNA.

A decision boundary was created for each of the classifiers using the same set of training and testing data to get a comparison of the different classifiers against each other. For each decision boundary the boundary was created using a corresponding classifier on the training data that would create a boundary for classification. In the KNN classifier, the boundary consisted of jagged lines and various islands spread out around the benign classification side (fig 4a). When the testing data was overlaid there was misclassification of benign and malignant samples with points being misplaced (fig 4a). When looking at that same iteration with a confusion matrix it was apparent that a majority of the misclassifications were of malignant samples being misdiagnosed as benign, 14% of the testing malignant samples (fig 6a).

In our LDA figure, there is a single straight line splitting the malignant and benign boundary (fig 4b). With the test data overlaid, there once again was a presence of points being misclassified onto the wrong side of the decision's boundary (fig 4b). When looking at the confusion matrix, there was a similar issue in this one iteration of a greater percentage, 12%, of misclassifications of malignant samples as benign samples (fig 6b).

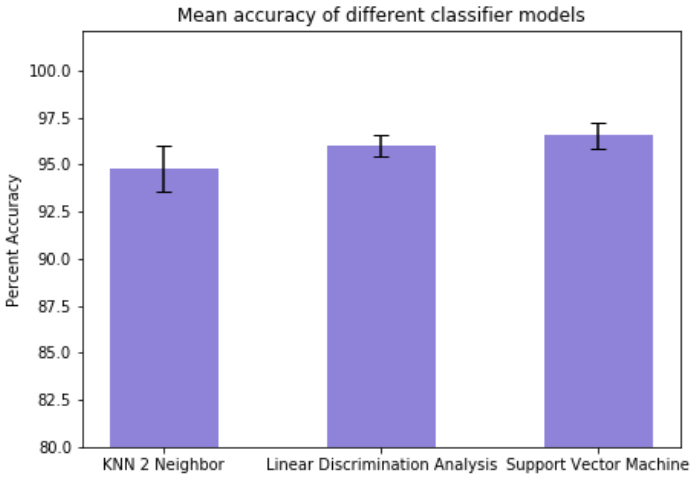


*Figure 5. Graphing of the average accuracy of 100 iterations of each classifier. The standard deviation is included as error bars.*

In our final decision boundary of SVM with a RBF kernel, the decision boundary created formed a large island with two small islands of benign classification (fig 4c). The amount of points misclassified was slightly lower than the other two classifiers. With this single iteration only 4% of the malignant samples were being misclassified as benign (fig 6c). The boundary for malignant classification almost completely encompasses all of the malignant points for both the training and testing data points, this however this could be explained by the majority of the field being labeled as malignant classification. The placement of benign points has more inaccuracies because of the limitation of one main island for classification (fig 4c).

After running a hundred iterations of each classifier and averaging the accuracy and standard deviation, the SVM classifier appears to be the most accurate with a 96.54% accuracy when predicting a samples classification (fig 5). This is then followed

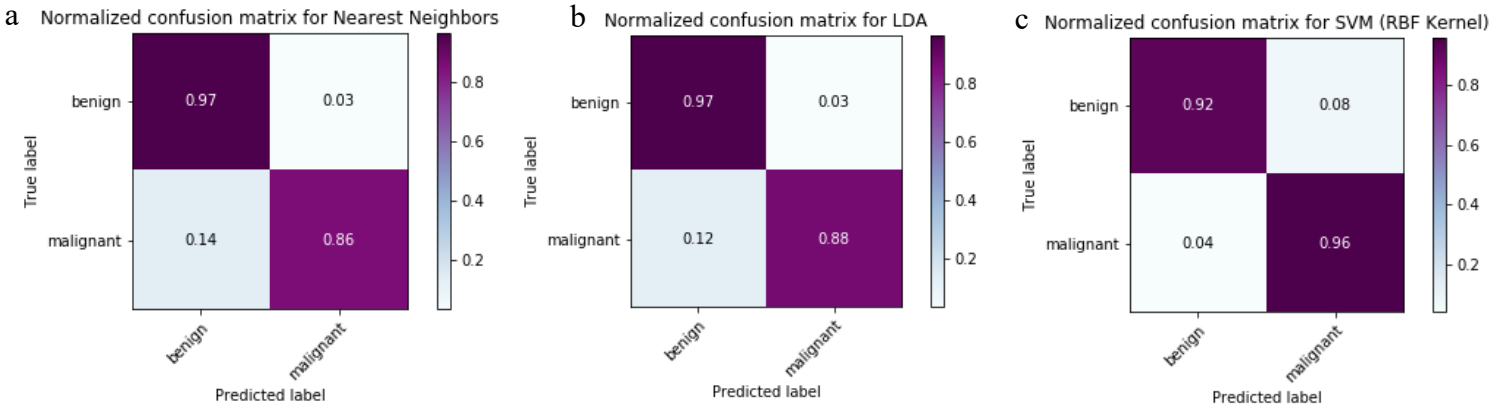


*Figure 6. The confusion matrix of a single iteration of each classifier. It is a normalized matric in which the percentage is displaced, a predicted label that doesn't match the true label is a misclassification.*

by LDA, 96%, and KNN, 84.77%. All three classifiers had an accuracy of at least 94%, showing a confidence in that these cell attributes measured are sufficient in getting a relatively accurate diagnosis.

**Discussion**

The use of Machine Learning Classifiers is effective at distinguishing between potentially malignant of benign tumors with the use of a FNA. In less developed countries were access to equipment such as mammograms and MRIs is limited, there is a potential of implementing FNA for first line detection. With educating women on performing self-exams, concerns over suspicious lumps could be addressed with a FNA and a portable ultrasound and analyzed under a microscope, with sample information being put into the classifier.

Despite the ease of a FNA, there still is a limitation were some nations don't have proper infrastructures for women's health. Additionally, this is just a classifier and cancer is very complex patient to patient, meaning that FNA might only be sufficient in some cases. Tumors have been known to bury themselves deep into the body and a self-examination may not be enough for identification. In order to better support women in developing nations there needs to be better infrastructure which takes time and finances.

**Contributions**

The dataset was acquired through the work of Dr. Willian H Wolberg with the University of Wisconsin-Madison. Feedback was utilized from our peer reviews, Eleanor Lutz, as well as Dr. Bing Brunton and Dr. Kameron Harris.

Walker had found the dataset on Kaggle and Maria had converted it into a ".csv" file. Walker had done the preliminary cleaning and visualization of data. Both of us had done the Project Proposal, while Maria did the Interim Report. Maria had put together the Project Presentation. Walker had done the code for PCA analysis, the graphing of components 1 and 2, the visualization of variance explained by components. Walker defined functions for splitting the data into train and test sets and the function for obtaining the mean accuracy score of 100 iterations of each classifier. Walker did the KNN decision boundary code, Maria had done the LDA decision boundary code, and both of us had done the coding for the SVM with RBF kernel. Walker did the visualization of gamma accuracies, as well as the confusion matrices. Walker wrote the methods portion of the Final Paper, Maria wrote abstract, introduction, results, and discussion portion. Maria formatted the paper with images and wrote the captions. Both of us edited the paper before final submission.

**References**

"1.6. Nearest Neighbors¶." *Scikit*, Scikit-Learn Developers, scikit-learn.org/stable/modules/neighbors.html.

"Breast Cancer Statistics." *World Cancer Research Fund*, World Cancer Research Fund International, 12 Sept. 2018, www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics.

"Breast Cancer: Prevention and Control." *World Health Organization*, World Health Organization, 21 Jan. 2016, www.who.int/cancer/detection/breastcancer/en/index1.html.

"Machine Learning for Cancer Diagnosis and Prognosis." *Machine Learning for Cancer Diagnosis and Prognosis*, pages.cs.wisc.edu/~olvi/uwmp/cancer.html.

"RBF SVM Parameters." *Scikit*, Scikit-Learn Deveopers, scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html.

"Sklearn.discriminant_analysis.LinearDiscriminantAnalysis." *Scikit*, Scikit-Learn Developers, scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html.

"U.S. Breast Cancer Statistics." *Breastcancer.org*, Breastcancer.org, 13 Feb. 2019, 10:47am, www.breastcancer.org/symptoms/understand_bc/statistics.

Gandhi, Rohith. "Support Vector Machine - Introduction to Machine Learning Algorithms." *Towards Data Science*, Towards Data Science, 7 June 2018, towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.