Analysis of Breast Cancer Wisconsin Data Set

Emily Nieves **Ishita Soni** **Induja Nidamarti**

Abstract

In this paper, we analyze the Breast Cancer Wisconsin [1] data set to see if we can classify tumors as malignant or benign from features extracted from breast cancer images and identify the features of a tumor (texture, radius, etc) that best predict cancer. We evaluate several classification models to see which features contribute the most to breast cancer. Our experiments conclude that Random Forest is the best model for feature selection with larger values of radius and concave points attributes correlating the most with malignant tumors while neural networks offer higher accuracy and recall of malignant cases.

1 Introduction

We chose to analyze the Wisconsin Breast Cancer Dataset (imported from Kaggle) for a variety of reasons. One of the primary motivations was how prominent this type of cancer is today, as the most common cancer in women is breast cancer, and the second most common cause of death from cancer in many races of women [2]. About 1 in every three women will develop breast cancer, which is an extremely high likelihood. Ever since 1999, the annual number of new patients diagnosed with breast cancer has increased by about 50,000 total cases [3].

The features from the dataset from Kaggle were collected via a digitized image of a fine needle aspirate of a breast mass. These features describe the characteristics of the nuclei of the cell that are present in the image. One of the categories of features, such as the radius, describe the mean of the distances to various points of the perimeter from the center. Texture is calculated via converting the image into a grayscale image and then computing the standard deviation of the grayscale values. Smoothness is the variation in the lengths of the radius, and concavity/concave points are the number and severity of concave portions in the contour [1]. We want to analyze these numerical values and test for how we can ultimately classify each of the ten features that were given to us to see which features are the most useful, and what values of those features would best diagnose cases as malignant or benign.

2 Data

For our analysis of the Wisconsin Breast Cancer Data Set, we were given a total of thirty features, which included the mean, standard error, and worst measures for the

following ten features of breast cancer tumors: Radius, texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave Points, Symmetry, and Fractal Dimension. The dataset also incorporated 569 total observations, of which 357 were benign cases and 212 were malignant cases.

3 Methods

We combined traditional statistics such as correlation testing and data visualization with newer machine and deep learning methods to analyze the dataset and gain insights into the features that influence a tumor classification as malignant or benign.

In order to identify some of the redundant features amongst the thirty that were initially included in the dataset, we placed these features in a scatterplot matrix. We saw that there was a presence of "multicollinearity" amongst the mean, worst, and standard error of the columns, especially between the "mean" and the "worst" columns. This is because the "worst" columns are a subset of the "mean columns", and the "worst" columns would substantially affect the "mean" columns. For this reason, we decided to process only the features from the "mean" columns in order to avoid redundancy and skewed results. Amongst the columns in the mean, we again ran a scatterplot matrix to see which groups of features were highly correlated to decide which features to use when training the machine learning algorithms.

We evaluated our models using Python in Jupyter and used the sklearn library for the machine learning algorithms. To evaluate our models, we used the train test split method with our data set. When doing so, we used the means of the ten unique attributes in the data set as our x training data and the diagnosis attribute as our y training attribute as it is the target variable for this data set. We used the hold-out model for the training/testing split.

We used several models to evaluate which features of a tumor (texture, radius, smoothness, etc.) best predict breast cancer and to train a model to classify a tumor as malignant or benign. The models we used were logistic regression, decision trees, random forest, linear SVM, and neural networks. To evaluate each of these models we created a method that fit the model with the training data, obtained the prediction data from the model with the ten attributes' testing data, and obtained the accuracy score with the prediction data of the ten attributes as well as the testing data of the diagnosis attribute.

4 Experiments and Results

4.1 Feature Selection

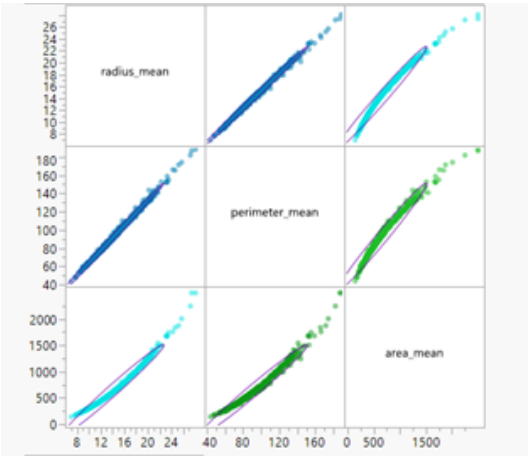
Based on the scatterplot matrix (numerical values) of correlations between the features listed in table 1, we noticed an extremely high correlation amongst radius mean, perimeter mean, and area mean. These three categories had correlations of 0.9874, 0.9979, and 0.9865, indicating that there would be no difference between using any one of the three features. This is most likely because the formula for

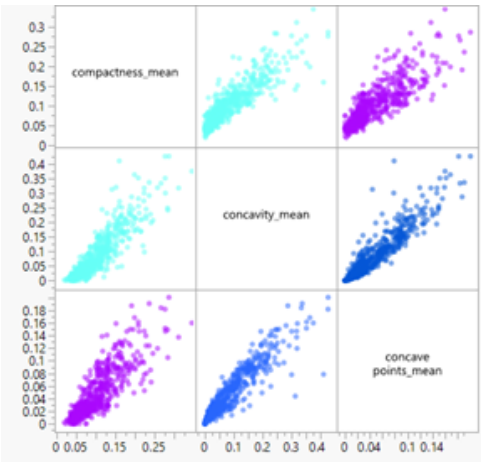
perimeter and area involve the use of radius in some way, as the formulae are $2\pi r$ and πr^2 respectively. Additionally, there was also a high correlation amongst compactness mean, concavity mean, and concave points mean. These categories had correlations of 0.8831, 0.9214, and 0.8311, which are a little less strong compared to the correlations between radius, perimeter, and area, but still very strong overall. From these two groups of features, we decided to choose perimeter and exclude radius and perimeter, and chose concave points mean and excluded compactness mean and concavity mean. Our final set of features was the means of perimeter, texture, smoothness, concave points, symmetry, and fractal dimension.

Table 1. Correlations between the features of the dataset.

Correlations

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
radius_mean	1.0000	0.3238	0.9979	0.9874	0.1706	0.5061	0.6768	0.8225	0.1477	-0.3116
texture_mean	0.3238	1.0000	0.3295	0.3211	-0.0234	0.2367	0.3024	0.2935	0.0714	-0.0764
perimeter_mean	0.9979	0.3295	1.0000	0.9865	0.2073	0.5569	0.7161	0.8510	0.1830	-0.2615
area_mean	0.9874	0.3211	0.9865	1.0000	0.1770	0.4985	0.6860	0.8233	0.1513	-0.2831
smoothness_mean	0.1706	-0.0234	0.2073	0.1770	1.0000	0.6591	0.5220	0.5537	0.5578	0.5848
compactness_mean	0.5061	0.2367	0.5569	0.4985	0.6591	1.0000	0.8831	0.8311	0.6026	0.5654
concavity_mean	0.6768	0.3024	0.7161	0.6860	0.5220	0.8831	1.0000	0.9214	0.5007	0.3368
concave points_mean	0.8225	0.2935	0.8510	0.8233	0.5537	0.8311	0.9214	1.0000	0.4625	0.1669
symmetry_mean	0.1477	0.0714	0.1830	0.1513	0.5578	0.6026	0.5007	0.4625	1.0000	0.4799
fractal_dimension_mean	-0.3116	-0.0764	-0.2615	-0.2831	0.5848	0.5654	0.3368	0.1669	0.4799	1.0000





Figures 1 and 2.

We also examined the correlation of the diagnosis with features using Pearson's correlation coefficient. The correlations helped further with the feature selection task. Only features with high correlations with the endpoint were chosen to train the algorithms. Features related to the concavity or size of the tumor had the highest correlation with the diagnosis while fractal dimension had very little correlation. The results are shown in table 2 below.

Table 2. Pearson's correlation coefficient between features and diagnosis.

Feature	Pearson's Correlation	Feature	Pearson's Correlation
Concavity	0.69636	Area	0.70898
Smoothness	0.35856	Concave Points	0.77661
Compactness	0.59653	Perimeter	0.74264
Texture	0.41519	Fractal Dimension	-0.01284
Radius	0.73003	Symmetry	0.33050

To further examine the correlation of the diagnosis with other features of the tumor, features were chosen and plotted in 2D and 3D graphs and grouped by diagnosis. This allowed for further analysis of both correlation between features and how separable the two classes were. Several features were highly related to the diagnosis and natural clusters could be seen in the feature space.

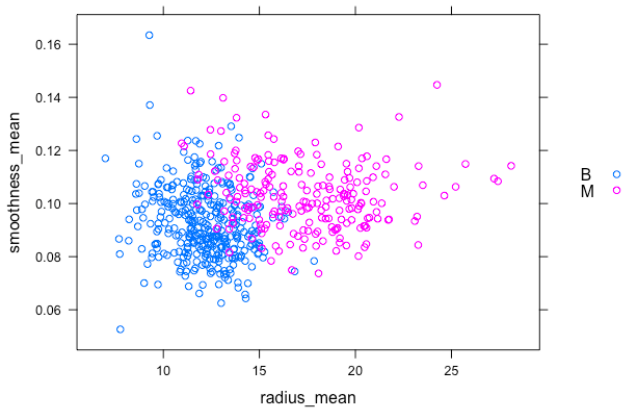


Figure 3. Radius and smoothness plotted and grouped by diagnosis.

After all the correlation analyses were completed, we decided to choose perimeter and exclude radius and perimeter, and chose concave points mean and excluded compactness mean and concavity mean. Our final set of features used for prediction was the means of perimeter, texture, smoothness, concave points, symmetry, and fractal dimension.

4.2 Logistic Regression

Logistic Regression turns linear predictions into probabilities. We ran the Logistic Regression model on three different test sizes, 25%, 50%, and 75%. As the test size

increased, the model produced a slightly lower accuracy each time. The results hovered around ~89%.

4.3 Decision Trees

Decision trees use trees to partition data into different regions and make predictions. It uses a top-down, recursive, divide-and-conquer method. We ran the Decision Tree model on three different test sizes, 25%, 50%, and 75%. As the test size increased, the model produced similar results that hovered around ~90%.

4.4 Random Forest

Random forest is an ensemble learning method for classification. It constructs multiple decision trees at training time. We ran the Random Forest model on three different test sizes, 25%, 50%, and 75%. As the test size increased, the model produced a slightly lower accuracy each time. The results hovered around ~92.5%.

4.5 Linear SVM

Linear support vector machine is an advanced classification technique that searches for a linear optimal decision boundary. Like previous models, we ran the linear SVM model with three different test sizes: 25%, 50%, and 75%. The results peaked at around 94%.

Table 3. Results table

Model	25% Test Size	50% Test Size	75% Test Size
Logistic Regression	91.608%	89.825%	88.290%
Decision Tree	90.909%	88.772%	90.867%
Random Forest	93.007%	92.281%	92.506%
Linear SVM	91.582%	94.002%	93.873%

4.6 Neural Networks

Next, a neural network was used to classify the data to see if any higher accuracy could be achieved. A neural network is a deep learning technique involving connections of nodes based on the physiology of biological neurons. The python library keras [5] was used with tensorflow to create the neural network.

A four layer architecture was constructed consisting of an input layer, two hidden layers with four nodes each, and one output node. RELU activation was used for all the layers and the cross entropy loss function was used. The data was scaled

using the sklearn standard scaler prior to testing and training. 25% of the data was used to test.

The testing accuracy was 94.4%, but the recall score for the malignant cases was only 84% showing that there was some room for improvement.

The effect of various architectures and activation functions was analyzed next. For this procedure, only one hidden layer was used. We changed the size of the hidden layer node, testing 2, 4, 8, and 12 node sizes. We also analyzed the effect of using sigmoid, RELU, and hyperbolic tangent activation functions. The results are summarized in the tables below.

The accuracies did not seem to vary greatly between the different tests. If confidence intervals were calculated, the accuracies would probably overlap. Using the hyperbolic tangent activation did change the malignant recall metric by a significant amount however.

Table 4. Accuracy and recall of the malignant cases with varying number of nodes in the hidden layer of the neural network.

Number of Hidden Layer Nodes	Accuracy	Malignant Recall
2	93%	85%
4	95%	94%
8	94%	92%
12	94%	89%

Table 5. Accuracy and recall of malignant cases with varying activation functions.

Activation Function	Accuracy	Malignant Recall
Sigmoid	92%	85%
RELU	94%	91%
Hyperbolic Tangent	94%	98%

5 Feature Importance

Based off of the model evaluations and their resulting accuracies, we chose to use sklearn's feature importance with Random Forest model to find out the features that contributed the greatest in predicting whether a tumor is malignant or benign. Since

Random Forest is a Decision Tree based model, the feature importance method gave the resulting Gini index values for each of the attributes used in the model evaluations. Concave points_mean and radius_mean had the highest index values meaning these two attributes were the most important when it comes to predicting whether a tumor is malignant or benign.

Table 6. Feature selection Gini Index values using Random Forest

Feature	Gini Index
concave points_mean	**0.398467**
radius_mean	**0.355060**
texture_mean	0.087627
smoothness_mean	0.083694
symmetry_mean	0.041492
fractal_dimension_mean	0.033659

6 Visualizing Data

After determining the most important features in predicting whether a tumor is malignant or benign, we created graphs for each feature. In each graph we plot the measurements and frequencies of malignant and benign tumors to visualize the data and reinforce the results we obtained.

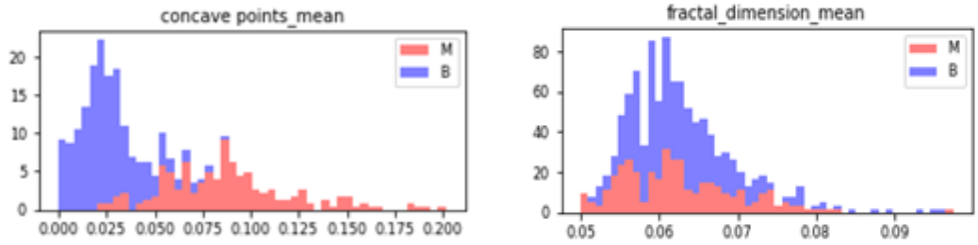


Figure 4. Graphs visualizing data

The graphs reinforce that radius and concave points are the best attributes for predicting breast cancer, and this can be seen by the smaller overlap between benign and malignant tumor histograms. It is also apparent that larger values for these attributes correlate with malignant tumors.

Attributes such as fractal dimension (attribute with the lowest index value) are not biased toward one diagnoses or the other. This is illustrated by the strong overlap between benign and malignant tumor histograms.

7 Conclusion

Our experiments conclude that radius and concave points are two features that yield the highest accuracy when it comes to predicting whether a tumor is benign or malignant. Therefore, physicians should pay extra attention to these characteristics of a tumor when diagnosing a patient.

Neural Networks is the model that yields the highest accuracy as compared to Linear SVM, and Random Forest (close followers), Logistic Regression, and Decision Tree for selection of attributes. However, based on our experiments, we concluded that Neural networks is less optimal for feature selection as it is a relatively new model and feature selection is still difficult to conduct with it. Therefore, we ultimately chose Random Forest for feature importance given its easy interpretability.

References

[1] Wolberg, W., Street, W., Mangasarian, O. (2016). Breast Cancer Wisconsin (Diagnostic) Dataset. Retrieved from https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[2] CDC - Breast Cancer Statistics. (n.d.). Retrieved from https://www.cdc.gov/cancer/breast/statistics/index.htm

[3] USCS Data Visualizations. (n.d.). Retrieved from https://gis.cdc.gov/Cancer/USCS/DataViz.html

[4] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011)

[5] Chollet, F. (2015) keras, GitHub. https://github.com/fchollet/keras