Breast Cancer Biopsy Neural Network Diagnosis

Israel Girón-Palacios

2023-12-06

Abstract

This paper covers the process by which a Neural Network is trained for the purpose of predicting Breast Cancer utilizing the analysis of a Fine Needle Aspiration biopsy

Introduction

Breast cancer is a type of cancer that forms in the cells of the breast[1]. It is one of the most common forms of cancer, second only to skin cancer, accounting for approximately 30% of new cancers in female patients each year[2]. The American Cancer Society's diagnosis estimates for breast cancer for 2023 is around 297,790 cases. They also estimate approximately 43,700 deaths during the same year. Considering these statistics adequate diagnosis from mildly invasive procedures such as Fine Needle Aspiration Biopsy are major consideration.

Data & Methodology

The models trained and chosen in this paper utilize the "Breast Cancer Wisconsin (Diagnostic)"[3] dataset provided by UC Irvine Machine Learning Repository.

This data was created in November 1995 by:

- *Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin*
- *Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin*
- *W. Nick Street, Computer Sciences Dept., University of Wisconsin*

The Data was donated by Nick Street.

Features

The data is composed of 32 variables, two of which are the diagnosis, the response variable, and an ID number.

Predictors are test results one of the following in 3 iterations:

a) *radius (mean of distances from center to points on the perimeter)*

[1] *https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470?utm_source=Google&utm_medium=abstract&utm_content=Breast-cancer&utm_campaign=Knowledge-panel*

[2] https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html#[:~:text=It%20is%20about%2030%25%20(or,(DCIS)%2 //www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html#:~:text=It%20is%20about%2030%25%20(or,(DCIS)%2

[3] *https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic*

b) *texture (standard deviation of gray-scale values)*
c) *perimeter*
d) *area*
e) *smoothness (local variation in radius lengths)*
f) *compactness (perimeter^2 / area - 1.0)*
g) *concavity (severity of concave portions of the contour)*
h) *concave points (number of concave portions of the contour)*
i) *symmetry*
j) *fractal dimension ("coastline approximation" - 1)*

The iterations are in order of appearance, the mean, the standard error(SE) and the mean of the 3 largest values for the specific test. The use of these statistics for biopsy diagnosis is beyond the scope of this paper, as such all features other that ID will be considered for modeling.

The models chosen for training are a Neural Network and Linear eXtreme Gradient Boosting.

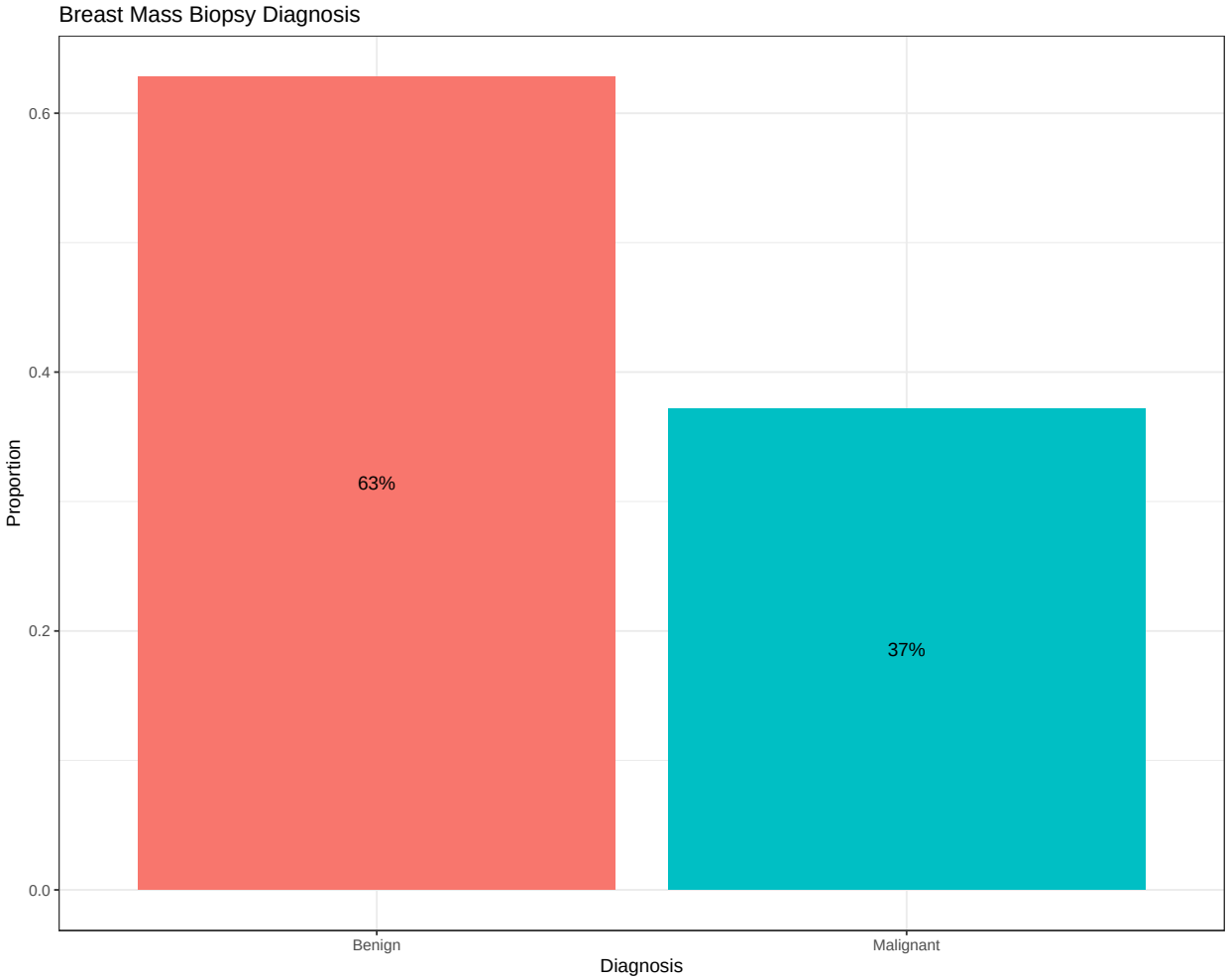
Data Preparation

The model will be chosen by gauging the sensitivity of the predictions on a test set and the model sensitivity will be calculated on the performance on a holdout set. For this purposes the data will be partitioned as 64/26/10.

Data Exploration

Response Variable Distribution

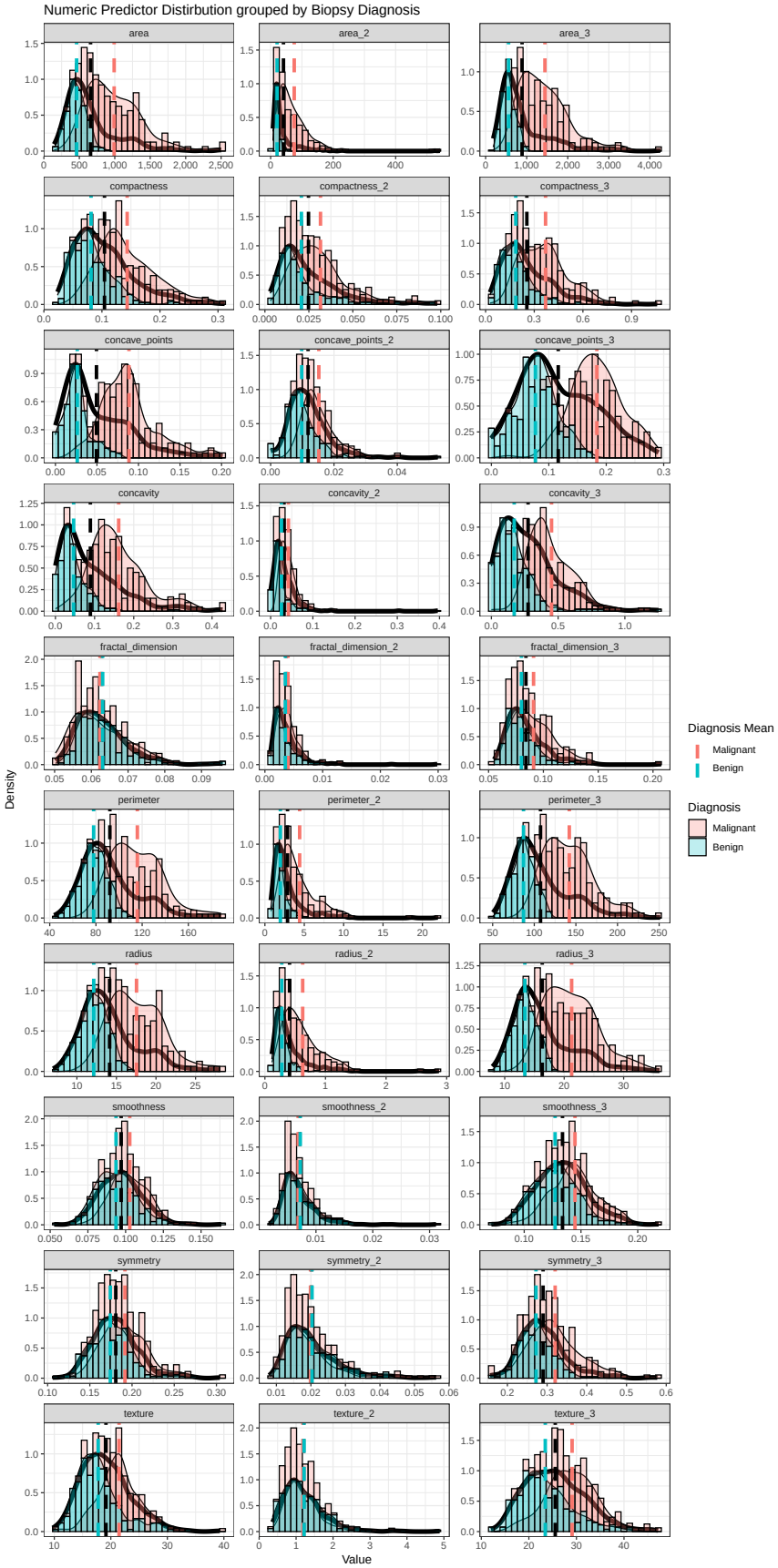
The spread of values for either Benign or Malignant diagnosis is somewhat unbalanced but considering that it is not an extreme difference the data will not undergo balancing.



Diagnosis	N	Proportion
Benign	257	0.6283619
Malignant	152	0.3716381

Predictor Variable Distributions

All of the predictors are numeric therefore the distribution of values in relation to the diagnosis will be explored.

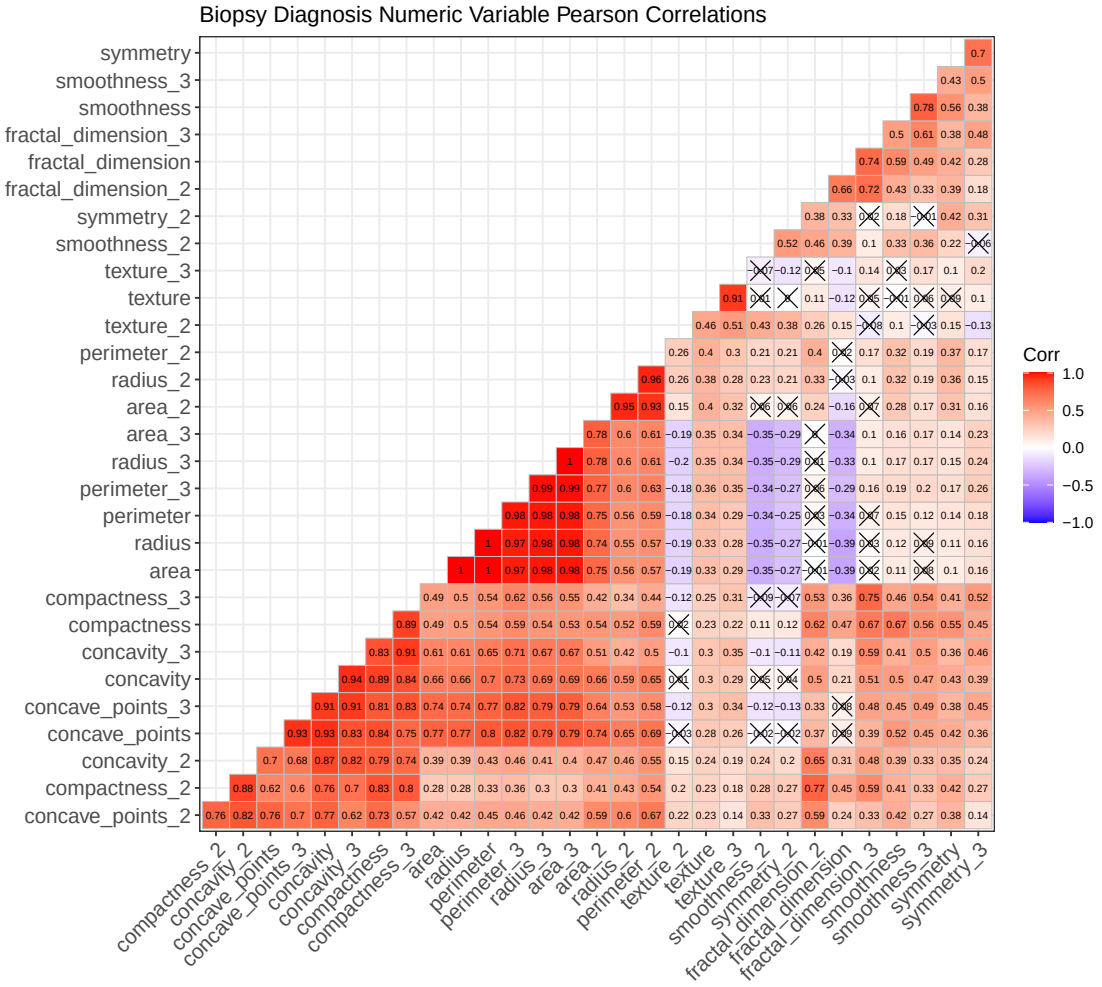

Numeric Predictor Distirbution grouped by Biopsy Diagnosis
area
area_2
area_3
compactness
compactness_2
compactness_3
concave_points
concave_points_2
concave_points_3
concavity
concavity_2
concavity_3
fractal_dimension
fractal_dimension_2
fractal_dimension_3
perimeter
perimeter_2
perimeter_3
radius
radius_2
radius_3
smoothness
smoothness_2
smoothness_3
symmetry
symmetry_2
symmetry_3
texture
texture_2
texture_3
Density
Value
Diagnosis Mean
Malignant
Benign
Diagnosis
Malignant
Benign

There are clear differences in diagnosis being tied to the mean value of a number of predictors. Most notably Area3, which would be the mean of the 3 largest values of area. Analyzing the features further it can be observed that there may be some correlations present between the predictors.

Feature Selection

Correlation

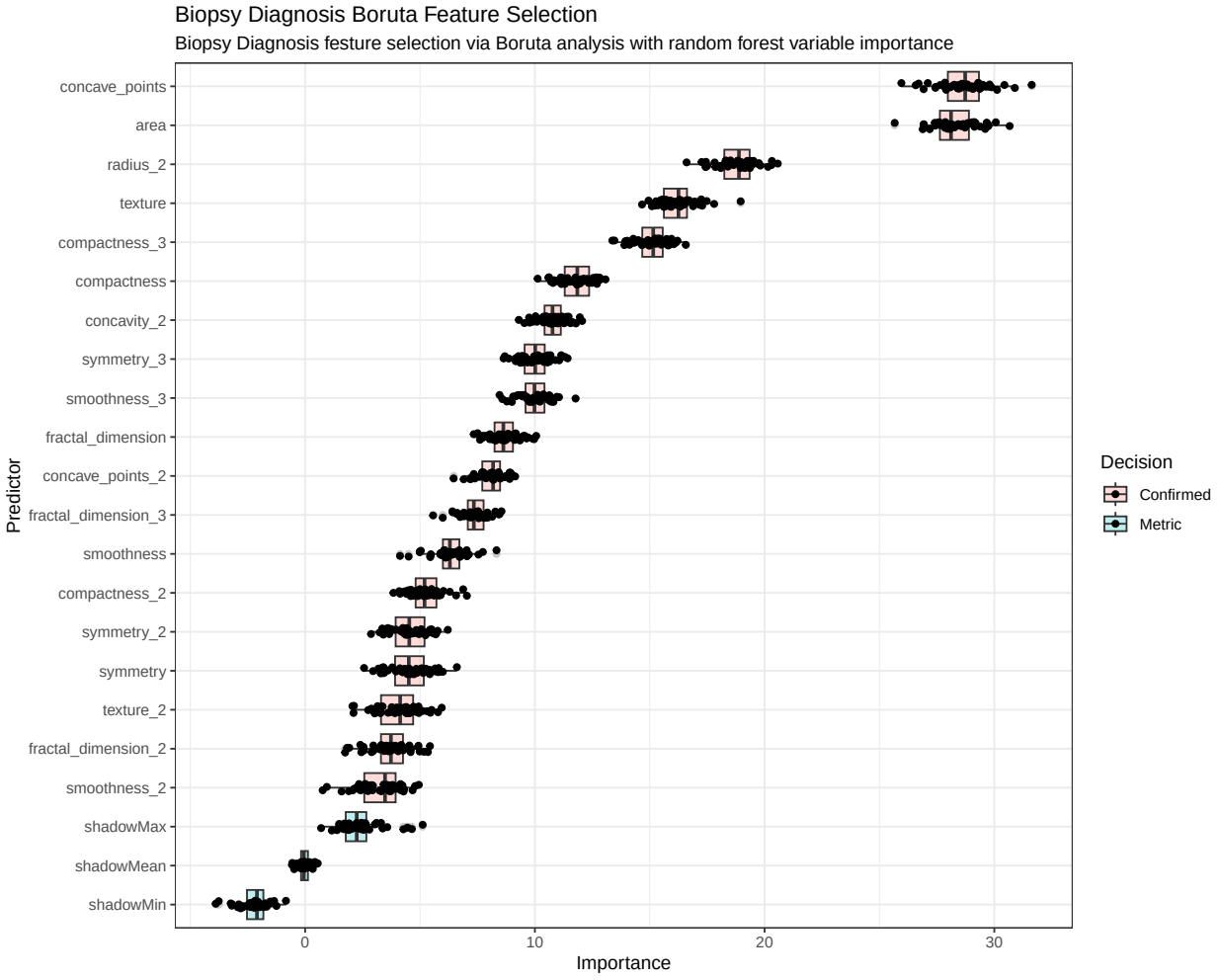
There are a a large amount of correlated predictors, unsurprisingly features such as area and perimeter have a relationship. However, despite the amount of highly correlated values most do not meet the typical cutoff of 0.9.



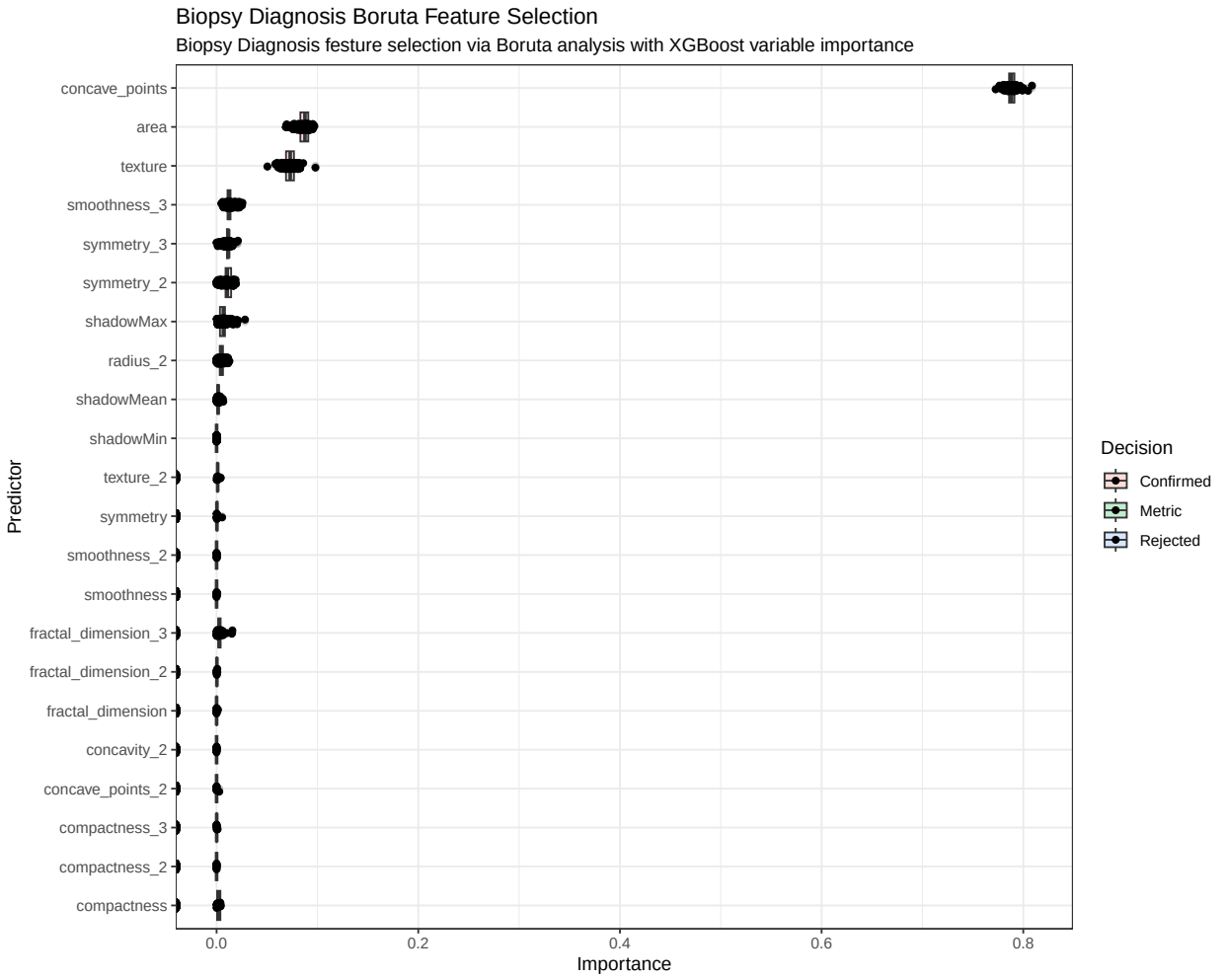
Boruta Feature Selection

As the feature space remains large the Boruta Feasture selection wrapper[4] will be applied in order to consider variable importance when feature interactions are considered.

[4] *https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/*



The standard Boruta Algorithm uses Random forest to calculate feature importance, in this case it considered all the predictors to be important. In order to be certain the alaysis will be run a second time but using the xGBoost algorithm to measure feature importance.



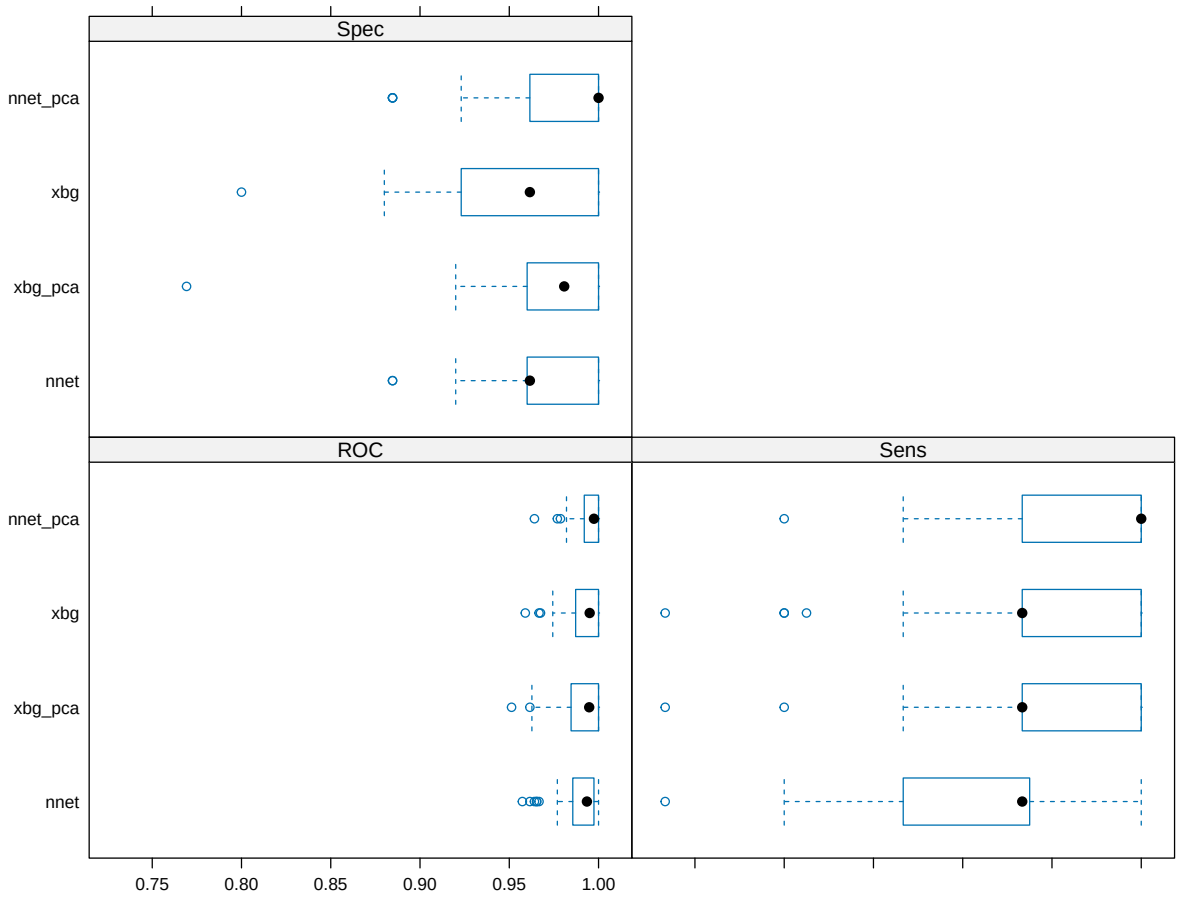
The analysis utilizing the XGBoost algorithm simplifies the feature space significantly and has a clear major contributor to the a predictive model. Both have determined that concave points. Considering these results the features determined by the XGBoost variant will be chosen as model predictors.

Model Training

As previously stated there are two core models to train, a Neural Network and Linear eXtreme Gradient Boosting model. Each of these will be trained on two datasets. One compromised of the selected features in the natural state and another where the features have been centered,, scaled, transformed and have undergone PCA. This will train a total of 4 models with two model methods allowing comparison of the natural and interpretable data to processed data for modeling.

Model Selection

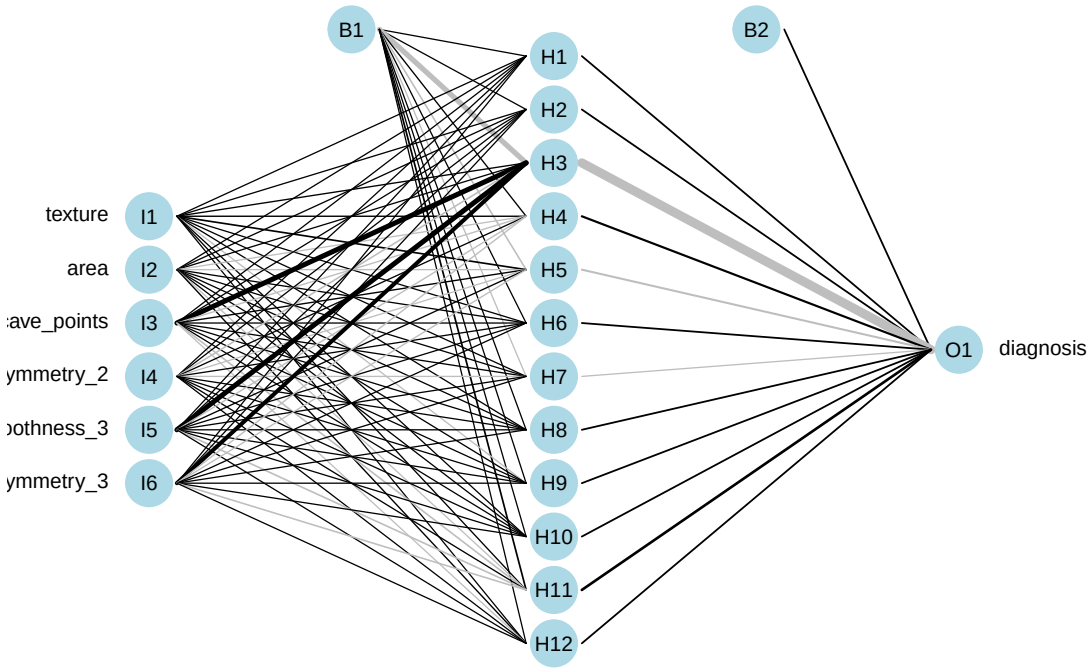
Considering saved training resampled data the model that has the highest expectation is a Neural Network with processed data.



Model	Sensitivity	Specificity
Neural Network	0.9210526	0.984375
Neural Network PCA	0.9210526	0.984375
xGBoost PCA	0.9210526	0.953125
xGBoost	0.8947368	0.968750

Of the four Models the model with the best Sensitivity on the test set is a natural data Neural Network with a sensitivity of 0.9210526 .

The model itself consists of a neural network with 6 predictors, 1 hidden layer, 12 neurons and two bias values.



Holdout Set Model Performance

The confusion matrix of the Neural Network model on the holdout set is:

Metric	Value
Sensitivity	0.8636364
Specificity	0.9722222
Pos Pred Value	0.9500000
Neg Pred Value	0.9210526
Precision	0.9500000
Recall	0.8636364
F1	0.9047619
Prevalence	0.3793103
Detection Rate	0.3275862
Detection Prevalence	0.3448276
Balanced Accuracy	0.9179293

The final model has solid performance on the holdout set with a sensitivity of 0.8636364 and a specificity of 0.9722222 . As such we can be certain that given this training data the model has been accurately trained.

Conclusions

While this type of model is by no mean a substitute for expert medical diagnosis it does highlight the most prominent features to consider when performing a diagnosis. In this instance Texture, Area, number of Concave points, Standard Error of Symmetry and Maximum Mean of Smoothness and Symmetry are major contributors in diagnosis Breast Cancer.