

PREDICTING THE DIAGNOSIS OF BREAST TUMORS

Marta Matosas Fonolleda

1- Introduction

The motivation for this project is that breast cancer is the most frequently diagnosed cancer in women, and the leading cause of death cancer in women worldwide.¹ An estimated over half a million women died due to breast cancer, only in 2016. If breast cancer is detected early, there are more treatment options and a better chance for survival. After a breast mass is detected, a biopsy is performed to classify the type of tumor. There are different types of biopsies (surgical, core needle, fine needle aspirations. One of them is the fine needle aspiration biopsy. This type of biopsy uses a thin needle to remove a small sample of cells. The advantages of this biopsy are that is less invasive and have less risk of infections due to the procedure. The current accuracy of diagnosis by Fine Needle Aspirate biopsy is 85%.²

The aim of this project is to improve the accuracy of the prediction of the diagnosis of a breast tumor to provide a better chance for survival to women worldwide. Since this is a healthcare problem and the importance of early diagnosis has been widely established, I will firstly focus on the reduction of false negatives – that is, malignant tumors classified as benign. Secondly, I will aim to reduce false positives – that is, benign tumors classified and benign. Even though more tests may be performed to confirm the malignancy of a tumor, there are cases where women go under a more aggressive treatment due to the misclassification of the tumor; as suggested by the findings of a study over 240 breast biopsies, 17% were deemed more suspicious or cancer.³ Finally, these patients may also suffer from unnecessary psychological stress.

2- Data gathering

When a fine needle aspirate biopsy is performed, a small drop of fluid is obtained and the sample is evaluated by a pathologist. In this case, the sample is expressed onto a glass slide and stained. The image for digital analysis is generated by a color video camera mounted on top of a microscope. To successfully analyze the digital image, a user selects a set of nuclei and specifies each location of each cell nucleus boundary through a graphical user interface. The system then

¹ https://www.researchgate.net/figure/Global-burden-of-breast-cancer-by-continents-per-100-000-women-per-yearbased-on-Ferlay_fig2_323991514

² <https://www.archivesofpathology.org/doi/pdf/10.5858/arpa.2018-0463-RA>

³ <https://www.cbsnews.com/news/breast-biopsies-often-get-it-wrong/>

analyzes and computes features relative to the size, shape and texture of those nuclei. Specifically, it calculates the mean value, the largest value and the standard error of each feature over the range of cells selected.⁴

3- Description of the dataset

The dataset includes 569 samples and 30 numerical four-digit features.

Those 30 features are divided in 3 sets of the same 10 features. The sets correspond to the before mentioned calculated values: the mean, the largest or worst and the standard error. The 10 extracted features are:

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 ($\text{perimeter}^2 / \text{area} - 1.0$)
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)

Upon inspection of the dataset, the key findings are that there are no missing values and no evidence of noisy or inconsistent data.

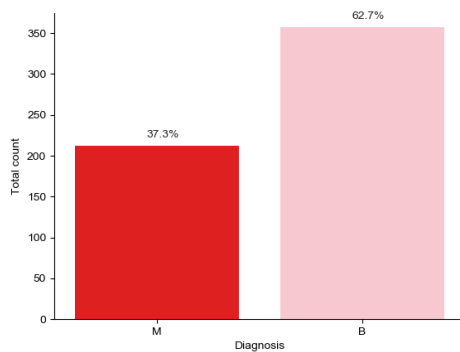
Finally, the target is the diagnosis of the breast tissue: malignant or benign. Since the target is discrete and binary, this is a binary classification problem.

⁴ W. Nick Street, William H. Wollberg and O.L. Mangasarian, "Nuclear Feature Extraction For Breast Tumor Diagnosis"

4- Feature selection

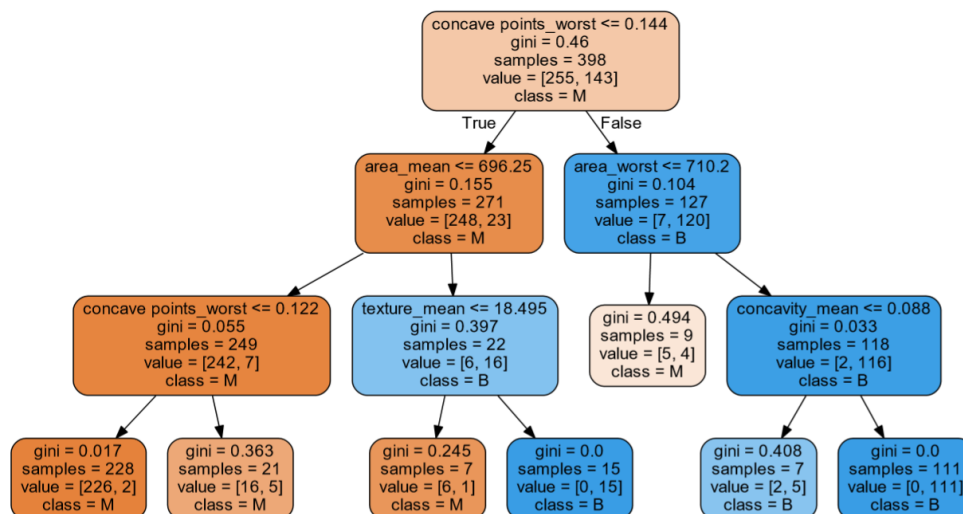
To perform the analysis on the dataset, a Python script was written in PyCharm and the data was then read as a Pandas dataframe.

Upon exploration of the class labels, there is evidence that class label is slightly unbalanced:

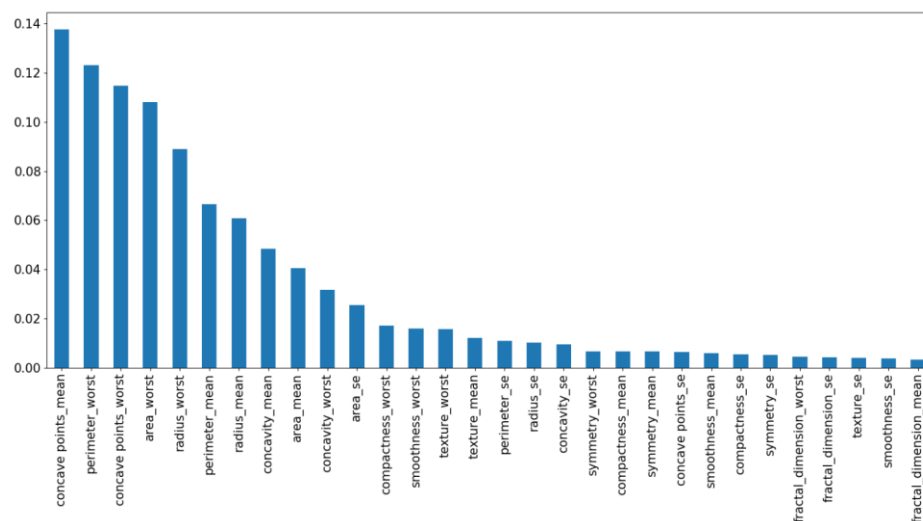


Given the unbalance and the limited data sample, I decided to use stratified k-fold cross-validation as resampling technique because it generally results in a less biased or less optimistic estimate of the model skill than other methods. Specifically, I will use 10-fold following a general recommendation.

Given the number of features present, the first step was to investigate feature importance. I first used the model Decision Tree with criterion Gini and criterion Entropy. The first provided better results for both precision and recall; thus, I noted the features selected (see image below).



I also run Radom Forest to gain insight from a different selection strategy (see image below).



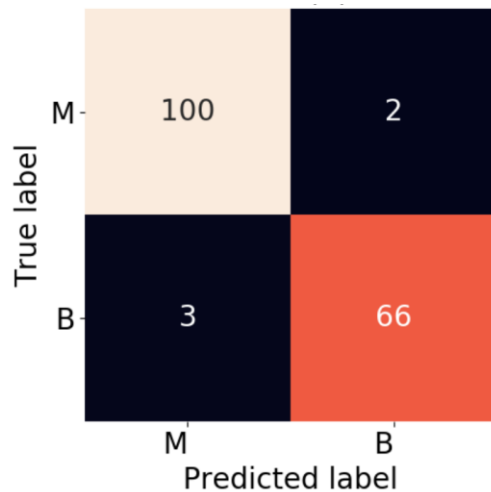
The dataset comes with an accompanying paper from the University of Wisconsin. The team at the University of Wisconsin achieved a 97 accuracy with only 3 features. I took this input as a benchmark. My goal was to find a subset of features that provided a better performance, specially seeking a reduction of false negatives. The goal is to have zero malignant cases misclassified as benign. Please bear in mind that the paper does not provide further details about the classifier and the hyper parameters used so I ran its selection with Random Forest to compare it to my final feature selection as a starting point.

To compare the results of the two feature selection sets, I used the confusion matrix and the ROC AUC score as a measure of the capability of the algorithm to distinguish between the two classes.

Results of the list of features from the paper:

List: [texture_mean, area_worst, smoothness_worst]

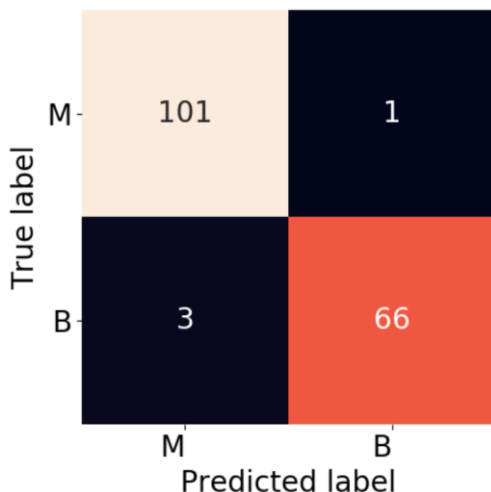
ROC_AUC: 99.60



My custom list of features:

List: [texture_mean, area_worst, smoothness_worst, area_mean, concavity_mean]

ROC_AUC: 99.53



Since the ROC AUC score of my selection is slightly lower than the benchmark and a 1-point improvement on the false negatives (type II error), I decided to proceed to modeling with this feature subset.

5- Modeling

As a next step, I performed hyper parameter tuning with the following classifiers: Logistic Regression, Support Vector Machine (SVM), XGBoost, K-Nearest Neighbors, Random Forest and Decision Tree. This was an iterative exercise: I used previous inputs to change the hyper parameters range and ran again the tuning process until the overall accuracy practically stabilized. The train and test sets were split 70% 30% and the pipeline included standardization of all the features. I set up the GridSearchCV to methodically build and evaluate all the models for each combination of algorithm parameters specified in each of their respective parameter grids. Finally, as mentioned before, I selected stratified 10-fold as the cross-validation method.

These are the best parameters for each model:

SVC

Accuracy: 0.9773717948717948

Best hyper parameters: {'clf__C': 11, 'clf__gamma': 0.01, 'clf__kernel': 'rbf'}

Logistic Regression

Accuracy: 0.9748717948717948

Best hyper parameters: {'clf__C': 10, 'clf__multi_class': 'ovr', 'clf__solver': 'newton-cg'}

XGBoost

Accuracy: 0.9699358974358974

Best hyper parameters: {'clf__eta': 0.1, 'clf__gamma': 0, 'clf__lambda': 1}

Random Forest

Accuracy: 0.9623717948717948

Best hyper parameters: {'clf__max_depth': 5, 'clf__min_samples_leaf': 5, 'clf__min_samples_split': 2, 'clf__n_estimators': 100}

KNN

Accuracy: 0.9598076923076923

Best hyper parameters: {'clf__algorithm': 'auto', 'clf__leaf_size': 1, 'clf__n_neighbors': 11}

Decision Tree:

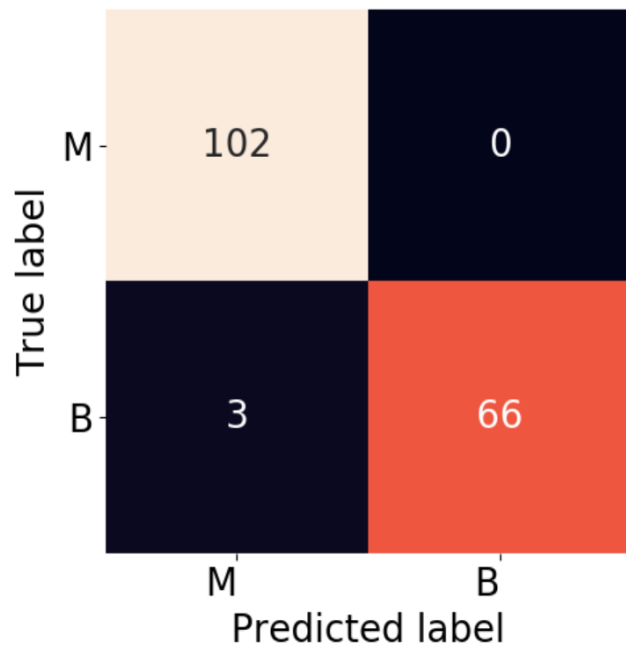
Accuracy: 0.9573717948717949

Best hyper parameters: {'clf__min_samples_leaf': 3, 'clf__min_samples_split': 2}

6- Evaluation

Then I ran each model with their respective best parameters and obtained both the ROC AUC scores and the confusion matrix. The classifier that outperformed is logistic regression:

ROC_AUC Score: 0.9782608695652174



A confusion matrix for breast tumor classification. The y-axis is labeled 'True label' with categories 'M' (Malignant) and 'B' (Benign). The x-axis is labeled 'Predicted label' with categories 'M' and 'B'. The matrix cells contain counts: 102 for (M, M), 0 for (M, B), 3 for (B, M), and 66 for (B, B). The cells are colored: (M, M) is light orange, (M, B) is dark blue, (B, M) is dark blue, and (B, B) is red.

True label	M	B
	102	0
B	3	66
Predicted label		

Even though false negatives are down to zero, there is room for improvement. There still exists false positives, this meaning that women with benign breast tumors can potentially receive a malignant diagnosis. The potential consequences are psychological stress and / or receiving a more aggressive treatment than required.

7- Future work

The future work is to:

1. Further investigate the feature space: as observed during feature selection, some features contribute to decreasing true false outcomes while others are provide improvements in the false positives department.
2. Implement ensemble methods: explore the combination of multiple models to produce a more powerful model.

8 - Bibliography

https://www.researchgate.net/figure/Global-burden-of-breast-cancer-by-continents-per-100-000-women-per-yearbased-on-Ferlay_fig2_323991514

https://www.breastcancer.org/symptoms/understand_bc/statistics

<https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf>

<https://www.ascopost.com/News/60293>

<https://www.ncbi.nlm.nih.gov/books/NBK470268/>

<https://www.archivesofpathology.org/doi/pdf/10.5858/arpa.2018-0463-RA>

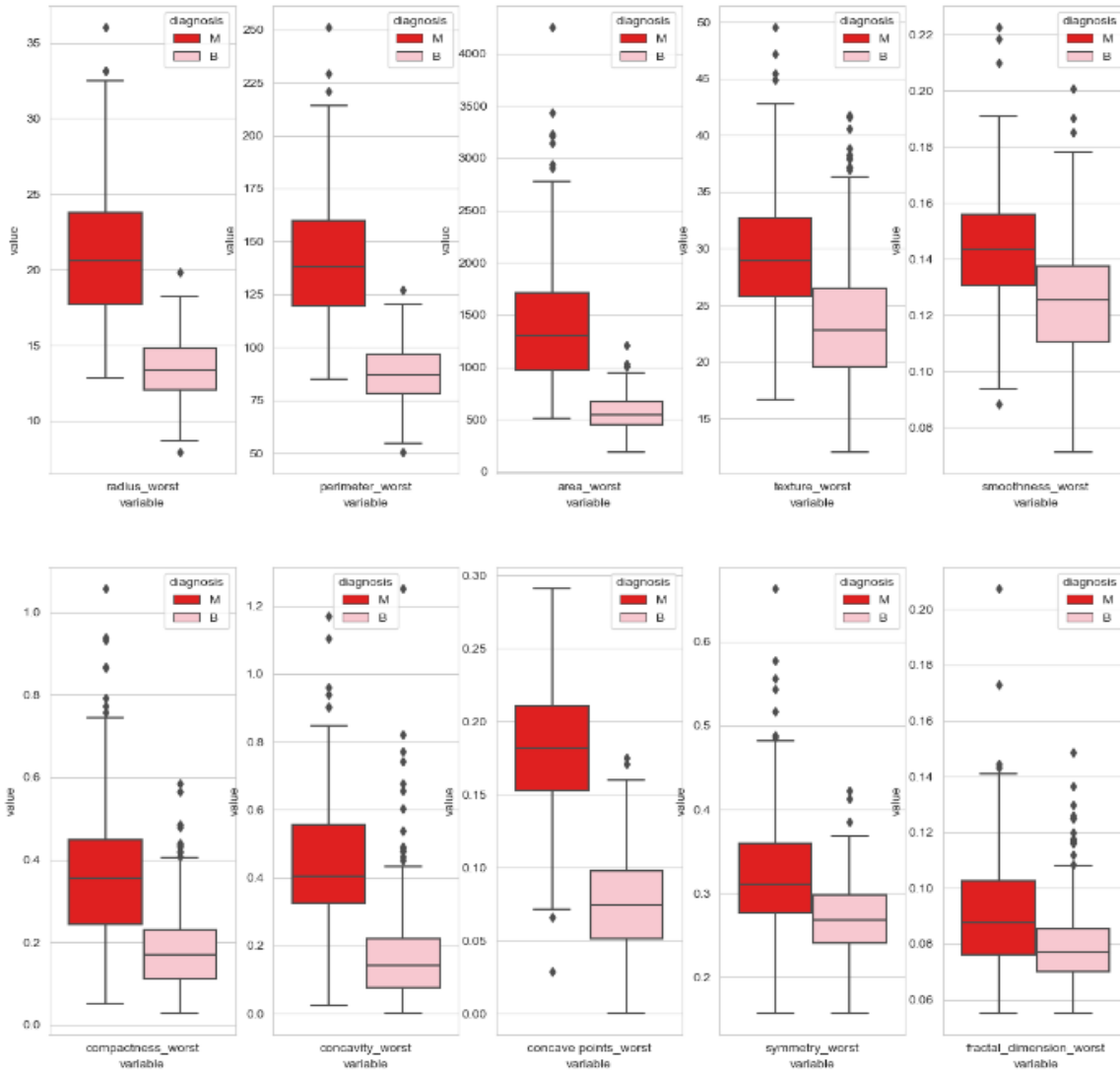
<https://www.europadonna.org/breast-cancer-facs/>

<https://www.cbsnews.com/news/breast-biopsies-often-get-it-wrong/>

W. Nick Street, William H. Wollberg and O.L. Mangasarian, "Nuclear Feature Extraction For Breast Tumor Diagnosis"

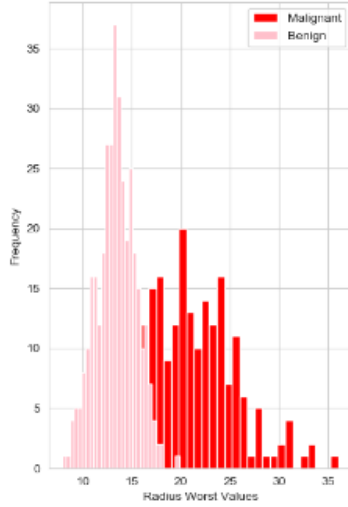
Appendix 1 - Exploratory Data Analysis

Set of features “worst”

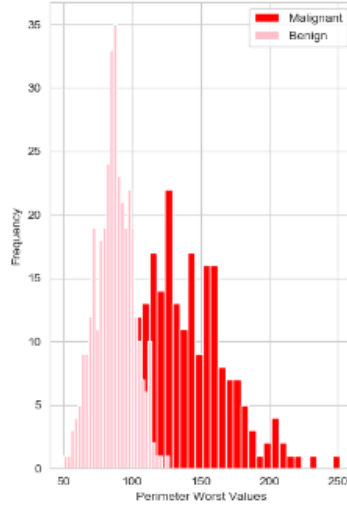


Set of features “worst” (cont.)

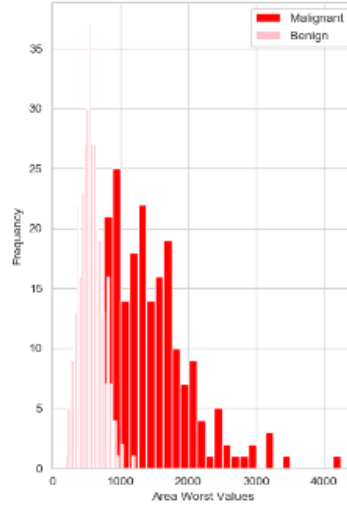
Histogram of Radius Worst for Benign and Malignant Tumors



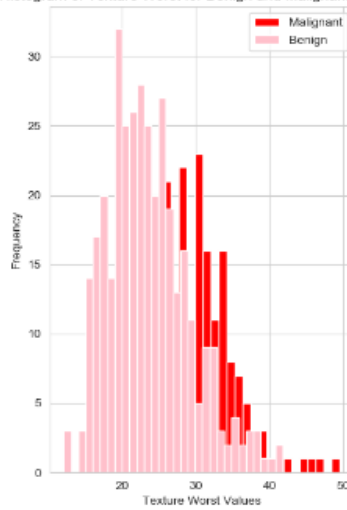
Histogram of Perimeter Worst for Benign and Malignant Tumors



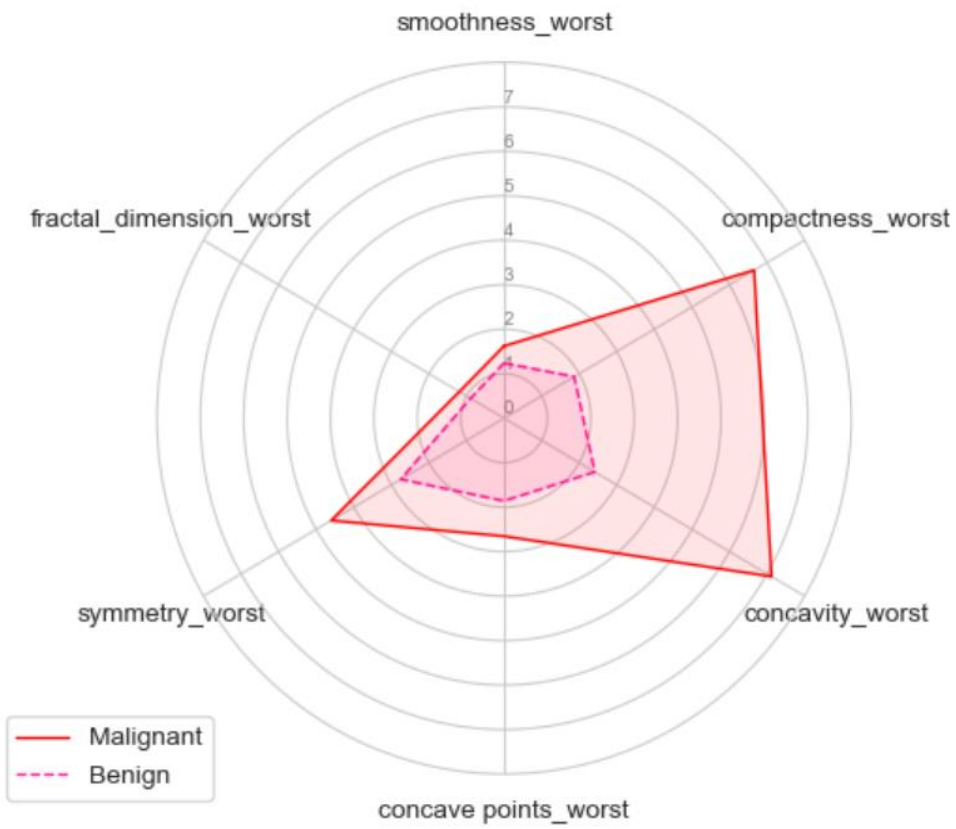
Histogram of Area Worst for Benign and Malignant Tumors



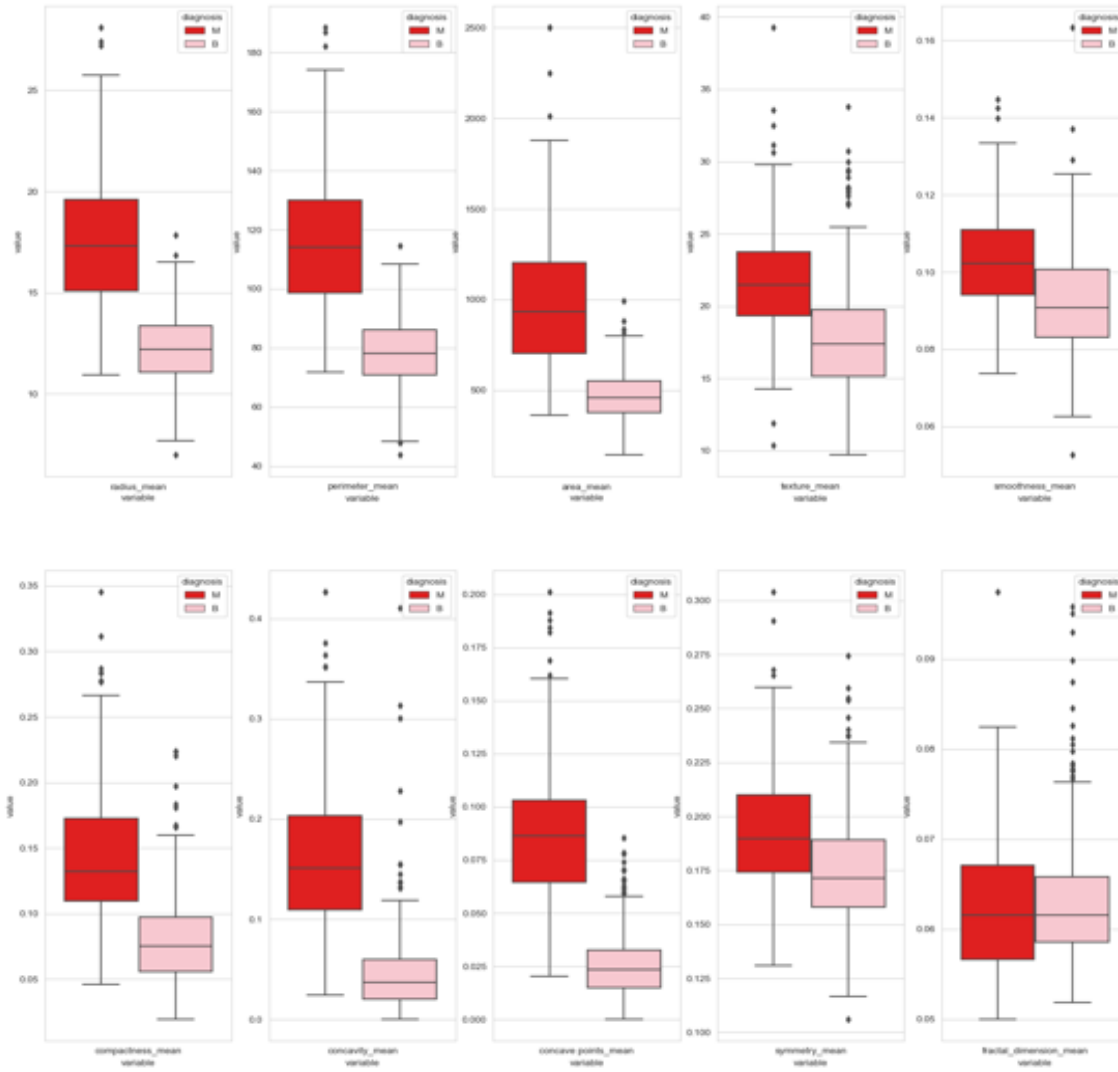
Histogram of Texture Worst for Benign and Malignant Tumors



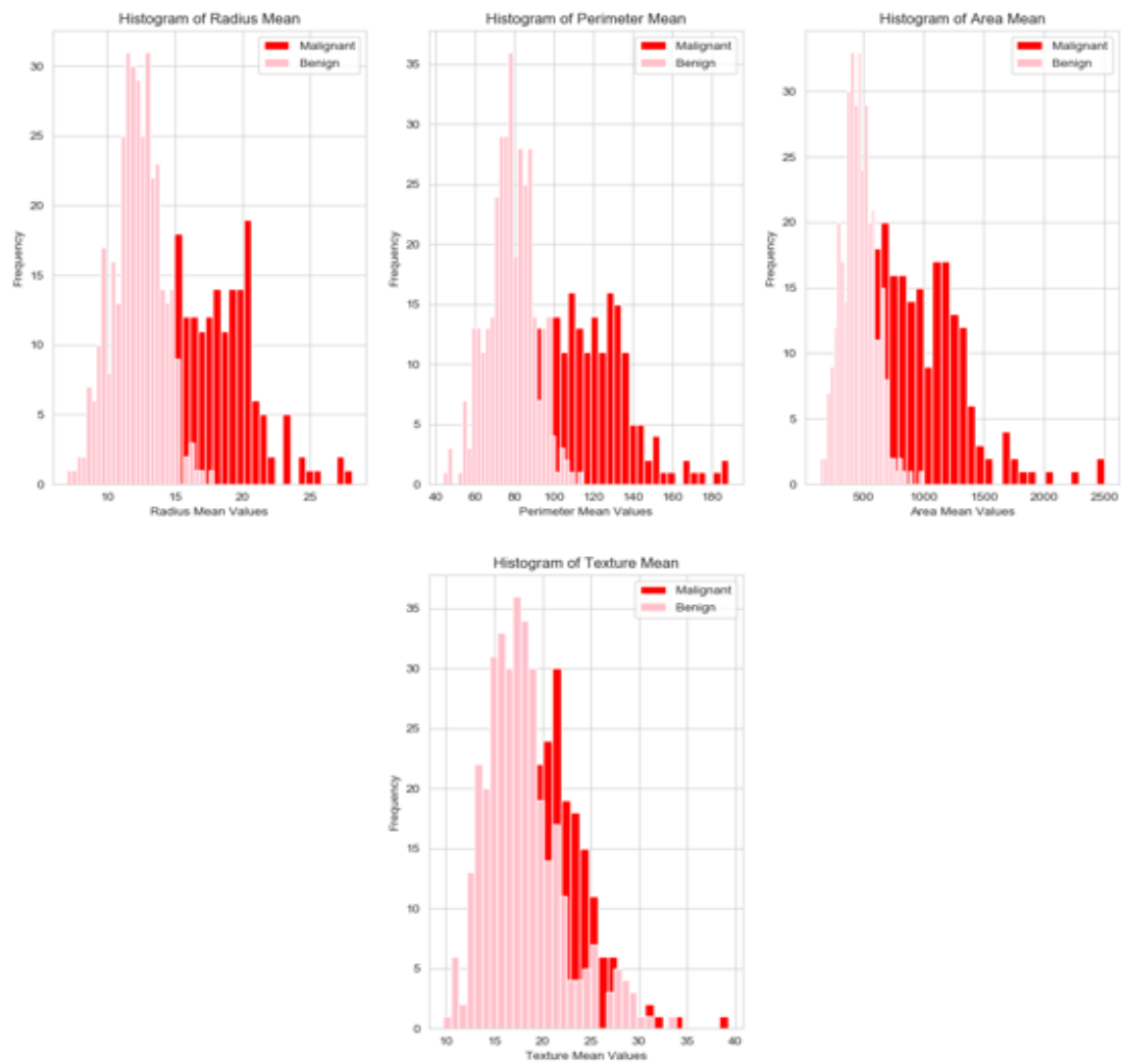
Set of features “worst” (cont.)



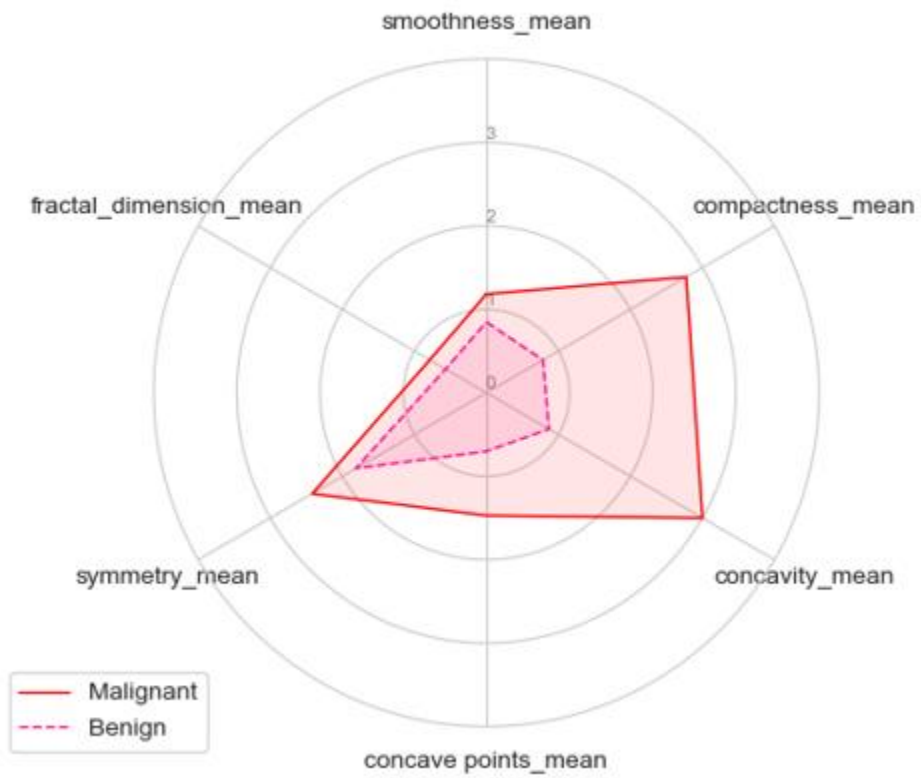
Set of features “mean”



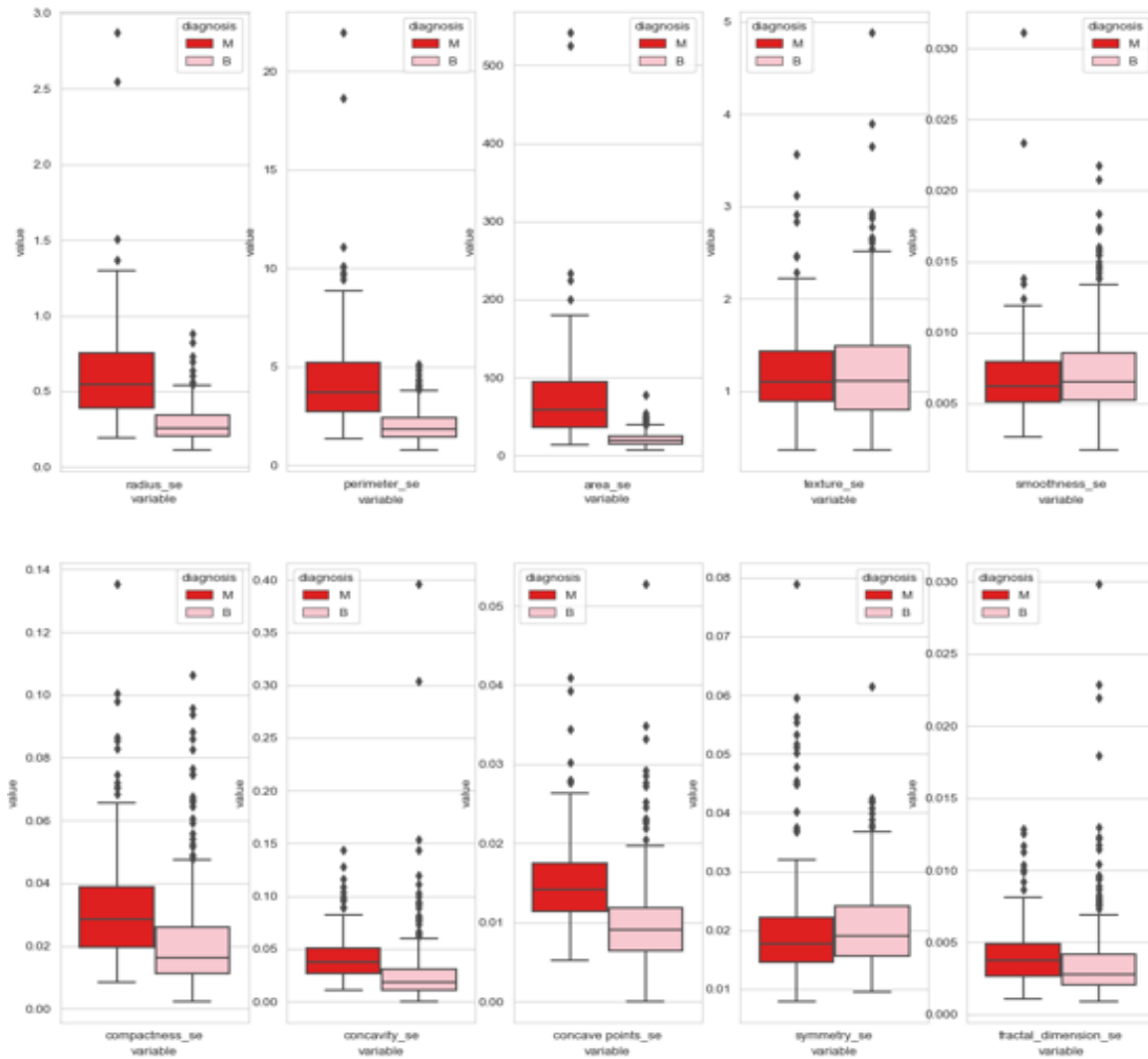
Set of features “mean”



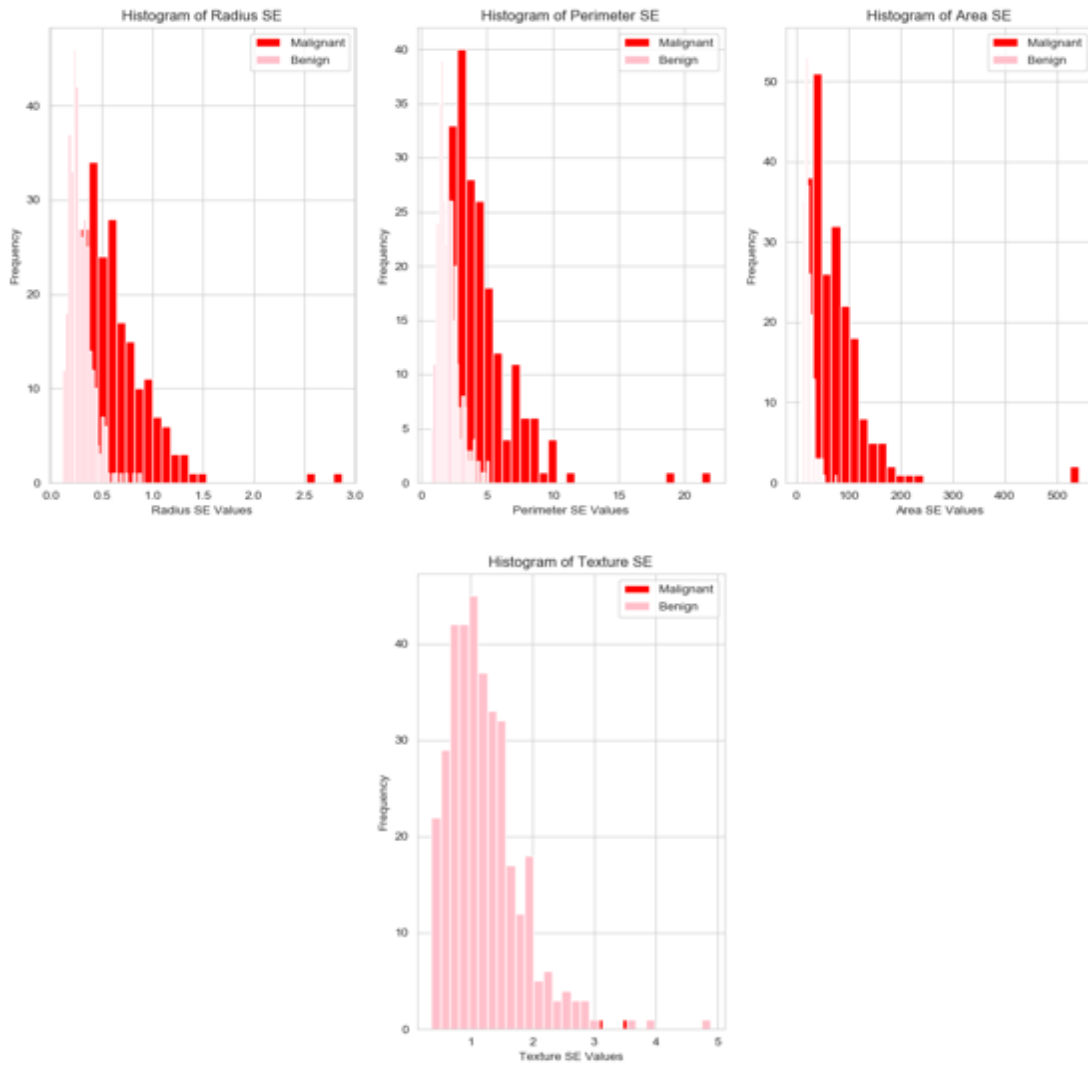
Set of features “mean”



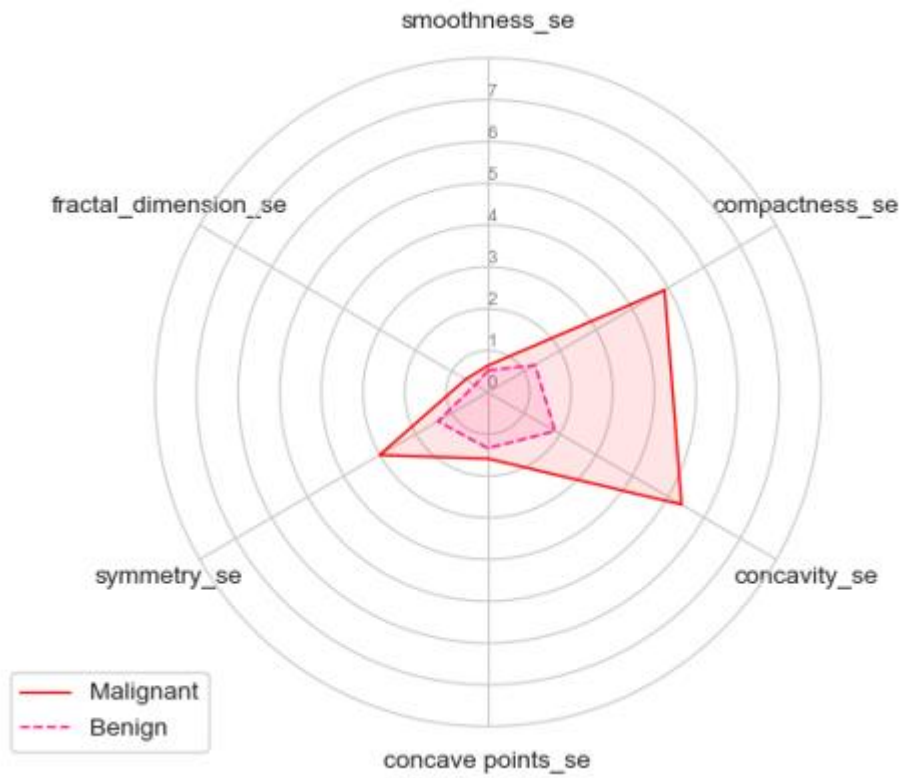
Set of features “standard error”



Set of features “standard error”



Set of features “standard error”



Correlation plots

