RESEARCH

Introduction to Focus Areas in Bioinformatics Project Week 1

Sina Glöckner[*†], Christina Kirschbaum[†], Swetha Rose Maliyakal Sebastian[†] and Gokul Thothathri[†]

[*]Correspondence:
sina.gloeckner@fu-berlin.de
Institute for Informatics, Freie Universität Berlin, Takustr. 9, Berlin, DE
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Goal of the project: Obtain statistical information about the data, two classifiers should be trained to allow classification of the given class. Get the most important features of the individual classifiers.

Main results of the project: We observed that Random Forest classifier and the Support Vector Machine (SVM) results are almost similar for all evaluated values. The statistical significance of this validation set are inconsistent because of the distribution of data.

Personal key learnings:
Sina: feature importance applied to Random Forest
Christina: feature importance works for SVM only with linear kernel
Swetha: Overleaf tool, understanding the statistical significance
Gokul: Fundamental statistical analysis using Python, Overleaf tool

Estimation of the time:
Sina: 10 hours
Christina: 8 hours
Swetha: 6 hours
Gokul: 6 hours

Project evaluation: 3

Number of words: 1687

1 Scientific Background

Breast cancer is currently the most common malignancy among women in developed countries, except for non-melanoma skin tumors [1]. Breast cancer is the second largest cause of cancer-related death for American women, behind lung and bronchial cancer. For cancer detection, a breast exam has moderate Sensitivity and good Specificity, and the predictive value changes by age. Digital mammography is used as a diagnostic tool has a predictive value of this new approach was not shown to be greater to analog; but diagnostic efficacy was found to be higher for women before menopause, who are under the age of 50 and have dense breasts. A pre-operative test called Fine Needle Aspiration Cytology (FNAC) is used to examine breast lumps. It is a cost-effective treatment that can help you avoid unneeded surgery. FNAC has certain advantages over core-needle biopsy in that it uses a finer needle, which reduces the risk of hematoma and other uncommon problems including pneumothorax.

In FNAC, a cytological sample is obtained directly from the tumor with a fine needle. The biopsy sample is then stained and visualized under a microscope to

discover if cancer cells are present. The main step in imaging is done to specify an accurate location of the cell wall and nuclear boundaries and to approximate the initial boundaries of enough nuclei to get a representative sample [2]. The sample is then the subject of microscopic examination. This method is painless; therefore, anesthesia is not required [3]. Since only a small percentage of the information in histological sections is visible to the eye, digital image processing is used to recognize and quantify complex patterns and interactions among elements.

Our hypothesis for this study was that using digital image analysis technologies, visually unquantifiable breast cancer microarchitectural features could be carefully analyzed and tested as prognostic criteria for invasive breast carcinoma of any type.

2 Goal

The main purpose of the project was to gather statistical information about the data. Afterwards, two classifiers were trained and the importance of the features for both was analyzed. Lastly, the classifiers are evaluated and the results were compared.

3 Data and Preprocessing

The data is from Diagnostic Wisconsin Breast Cancer Database, and was obtained using the FNAC method.

The data was imported using the files `breast-cancer-Wisconsin.data` with names from `breast-cancer-Wisconsin.names` from the Breast Cancer Wisconsin (Diagnostic) Data Set [4]. First, missing values were replaced with the attribute's mean value. In this case, this only applies to the null values in Bare Nuclei. For the target, values for benign tumors (2) were replaced with 0 and for malignant tumor (4) with 1.

Then, a statistical analysis was performed to get an overview of the data. Histograms, the p-value, correlation and co-variance were identified for this.

3.1 Histograms

In statistical practice, the histogram is one of the most essential graphical objects. With only a few assumptions, the histogram provides a consistent approximation of any density function [5]. It illustrates the frequency of each value and represents the distribution of variables. The frequency is here defined as the number of times each value appears in the data set.

Using the frequency of Mitoses, Marginal Adhesion, Uniformity of Cell Size, Clump thickness, Single Epithelial Cell Size and Bland Chromatin, histograms (Figure 1) were created for each of the features. This shows the distribution of these six features. None of the chosen attributes are in a normal distribution which renders the mean of a single variable (red line in Figure 1) meaningless.

3.2 p-values

The p-values indicate the likelihood of an event occurring in the data by chance.

It can be observed that the combinations of qualities have a higher possibility to occur in the hypothesis testing with the null hypothesis stating that the combinations chosen are not related or false relationships with each other.

3.3 Co-variance

The tendency of two variables to fluctuate together is measured by co-variance. When two vectors are identical, co-variance is maximal.

The mean of the Mitoses, Marginal Adhesion, Uniformity of Cell Size, Clump thickness, Single Epithelial Cell Size and Bland Chromatin were found and further used for the calculations of co-variance between the attributes. Co-variance was found between the three pairs Mitoses and Marginal Adhesion mean, Uniformity of Cell Size and Clump thickness mean, Single Epithelial Cell Size and Bland Chromatin mean.

3.4 Correlation

Correlation is the relationship strength between two variables. The correlation between all features was computed and a correlation heat map (Figure 2) was generated.

4 Methods

After preprocessing the data and getting a general overview, the data set was split into training data and validation data. Then, we applied two different classifiers, the random forest model (subsection 4.1) and a support vector machine (subsection 4.2), to the training data. The trained models were further used to predict the classification of the validation data. This way, the classifiers could be compared. The implementation used the python package sklearn [6].

4.1 Random Forest

Random forest models are an approach to supervised learning, because of the enormous number of decision trees involved in the process, it is utilized for classification and regression. This classifier is made up of numerous random decision trees that are less sensitive to the training data. The name Random Forest originates from the fact that we employ a multitude of trees. Using Bootstrapping, the algorithm selects random rows from the data set to construct a fresh training data set. We created such a model and trained with the training data set. Further, we calculated the training score and the validation score. This model was then used to validate the data and analyze the feature importance. A plot for the permutation feature importance and drop column importance. The F1 score for Random forest was obtained. The evaluation of training data gave us the ROC curve.

4.2 Support Vector Machine

Support vector machine (SVM) is another typical classification approach combined with regression analysis for given data. The classifier is a supervised learning approach to evaluate data by categorizing it into two groups. The result of this technique is a layout map in which the sorted data set with the categories' boundaries is as large as possible [7]. A model was constructed for testing data, and then used to classify the validation data. Furthermore, the feature importance was investigated with three different methods. First, the sklearn-internal calculation of feature importance was applied. Then, a plot of the permutation Feature Importance and drop column importance was obtained, as well as the confusion matrix in form of a heat map.

5 Results

For the Random Forest classifier, no feature proved significantly more important than another. The results with different methods contradict each other (Figure 3). However, the classification had a high Accuracy with 0.9571. Sensitivity, also called Recall, and Specificity were high with values at 0.9333 and 0.9684 respectively. This is reflected in the ROC curve (Figure 6) as well. Additionally, the Precision of the model was at high 0.9333. The F1 score supports this as well, since it is very high at 0.93. Most individuals in the validation data set were classified correctly (Figure 5 (A)).

The computations from SVM bear similar results. In this case, the Accuracy lies at 0.9643. While the Sensitivity is slightly lower at 0.9111, the Specificity, Precision, and F1 score are even higher at 0.9895, 0.9762, and 0.94 each. The importance of the features is not clear once again.

A better comparison between both models can be seen in Table 1.

Table 1: All results for Random Forest and Support Vector Machine.

	Accuracy (%)	Sensitivity / Recall (%)	Specificity (%)	Precision (%)	AUC score	F1 score
Random Forest	95.71	93.33	96.84	93.33	0.99	0.93
SVM	96.43	91.11	98.95	97.62	1.00	0.94

6 Discussion

In conclusion, both models acted similarly. This is reflected in their confusion matrices, and therefore, the ROC-Curves, and other evaluation methods.

A clear difference is shown in the feature importance though. However, this difference seems mostly dependent on the method, how feature importance is calculated. For instance, the drop column importance always rates the Bare Nuclei attribute high. There is no consensus among the various calculations on the importance of any feature. The feature importance implemented with sklearn results in the highest values, but the values are also close to each other. The same can be said about the permutation importance, which computes lower, but also closely clustered, values.

The project incorporated statistics in the analysis of the data, data wrangling, data visualization in various ways and some Machine learning with the classifiers. These are all core skills of a data scientist.

On a more personal note, this project helped to further deepen our understanding on classifying data. It highlighted the importance of the distribution of the data. Since the parameters were not in a normal distribution, some classical statistical methods for visualizing and analyzing the data, such as the mean, were not meaningful. Additionally, the correlation between two attributes emphasized the possible dependence of two attributes on each other. The comparison of two classifiers on one data set was more exact as well. While the feature importance was very contradictory in our example, it also shone a light on more methods to evaluate a classifier with special regard to the data.

Competing interests
The authors declare that they have no competing interests.

Author's contributions

Swetha Rose Maliyakal Sebastian informed about the scientific background and wrote about this, the goal, data and preprocessing, and methods. Gokul Thothathri implemented the statistical analysis. Sina Glöckner implemented the random forest model and feature analysis. Christina Kirschbaum implemented the SVM model and the respective feature analysis. Both wrote the results and discussion, and created the figures.

References

1. Nazário, A., Facina, G., Filassi, J.: Breast cancer: News in diagnosis and treatment. Revista da Associação Médica Brasileira **61**, 543–552 (2015). doi:10.1590/1806-9282.61.06.543
2. Street, N., Wolberg, W., Mangasarian, O.: Nuclear feature extraction for breast tumor diagnosis. Proc. Soc. Photo-Opt. Inst. Eng. **1993** (1999). doi:10.1117/12.148698
3. Kowal, M., Filipczuk, P., Obuchowicz, A., Korbicz, J., Monczak, R.: Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. Computers in biology and medicine **43**, 1563–72 (2013). doi:10.1016/j.compbiomed.2013.08.003
4. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 Accessed 2021-11-06
5. Scott, D.: Histograms: Theory and Practice, pp. 47–94 (2008). doi:10.1002/9780470316849.ch3
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
7. Hariharan, K., Vigneshwar, W.S., Sivaramakrishnan, N., V, S.: A comparative study on heart disease analysis using classification techniques. International Journal of Pure and Applied Mathematics **119**, 13357–13365 (2018)

Figures

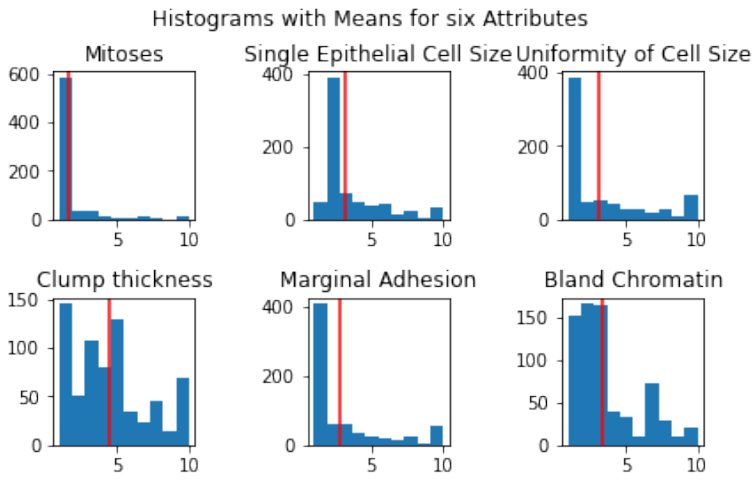

Figure 1: **Histograms** for six different attributes. The red line shows the mean-value of the attribute.

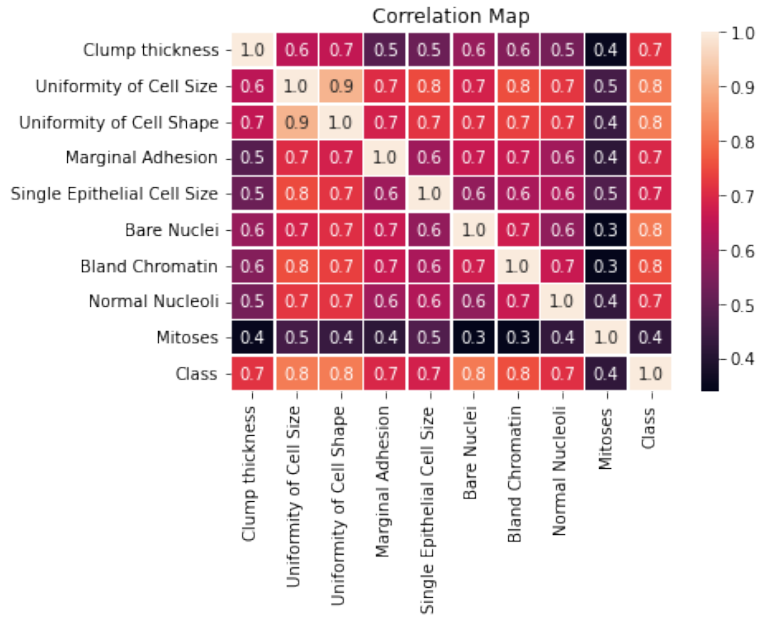

Figure 2: **Correlation map** shows the correlation between every attribute in the dataset.

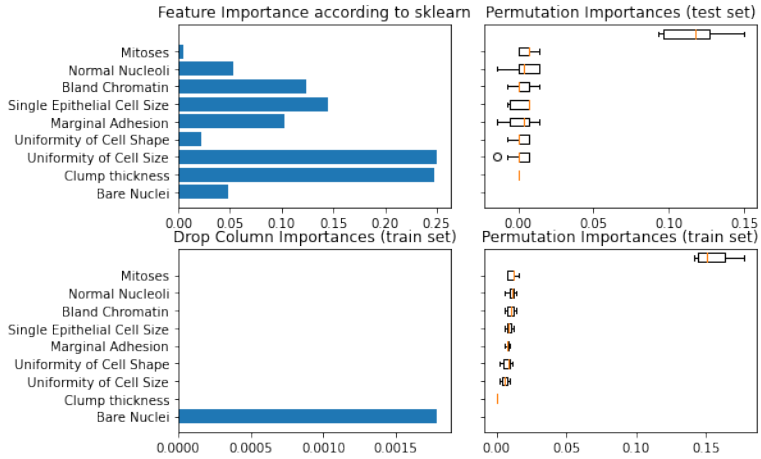

Figure 3: **Feature Importance for Random Forest** Here are four different measures for feature importance applied to the Random Forest model.

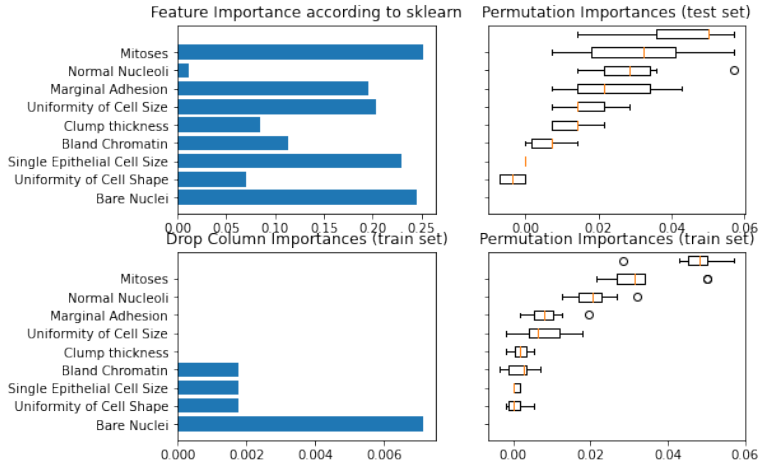

Figure 4: **Feature Importance for SVM** Here are four different measures for feature importance applied to the Support Vector Machine.

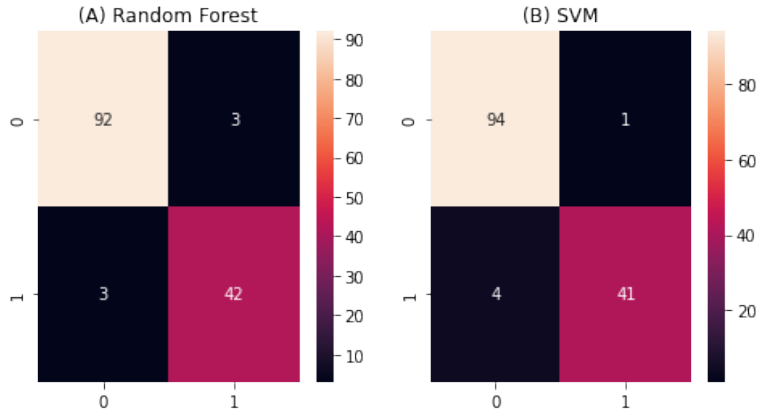

Figure 5: **Confusion Matrices** This show the confusion matrices for predictions with both classifiers, random forest (A) and SVM (B).

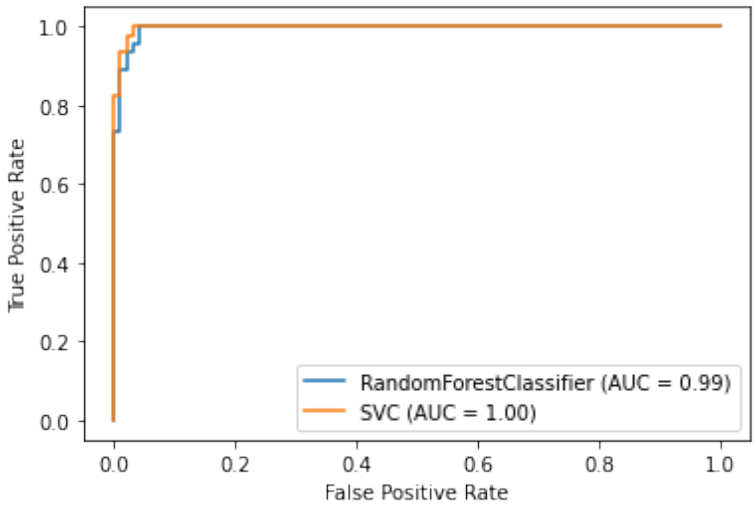

Figure 6: **ROC-Curve**