The analysis of Random Forests and Artificial Neural Networks in the application of breast mammography

Joel Kischkel

July 16, 2020

Abstract

With around 20% reduction of mortality, mammography is a useful method for the early prediction of women breast cancer. But this approach leads to around 70% of unnecessary biopsies. A possibility of reduction is using machine learning algorithm for the prediction of the severity of a mammography outcome. Based on Mammography Mass Database we have chosen Random Forest and Artifical Neural Networks to perform classification. Models were trained with and without cross validation. To measure the performance the auc (area under the curve) was computed. Both approaches were able to predict the severity. Artifical Neural Networks perform slightly better with an auc of 0.869 compared to Random Forest with 0.867.

1 Introduction

For 2020 the American cancer society estimates 279,100 new cases of breast cancer in the US. [1] A crucial point is the early detection of breast cancer. Cases with an early detection of breast cancer diagnosis leads to a higher chance of recovery. [2] Different screening methods are used for the early detection. The idea behind screening is that this method can detect breast cancer before any symptoms show up. One of the most effective methods is the mammography. This method can reduce the mortality in a range between 20% and 25%. In case of a positive screening result a biopsy is performed to confirm the result. The disadvantage is that the interpretation of x-ray images of mammography is difficult and leads to around 70% of unnecessary biopsies. To support the decision if a biopsy in necessary or not different computer aided approaches were developed. Elter Schulz-Wendtland and Wittenberg showed that computer based approach is able to predict the outcome of a mammography. During their research they built up their own mammography database. [3] [4] Based on their database the aim of this paper is to predict by the application of Random Forests and Artifical Neural Networks the need of performing a biopsy. At the same time this paper compares the chosen approaches and analyses accuracy of the prediction.

2 Dataset and Methods

2.1 Description of the dataset and preprocessing

The Mammographic Mass Data Set was originally created by Schulz-Wendtland and donated by Elter. The data set is puplicy available at the UCI Machine Learning Repository. [5] The original data set contains 961 values with 6 different features and is used for the prediction of breast cancer. These features are BI.RADS Assignment, Age, Shape, Margin, Density and Severity. The term BI.RADS refers to the word Breast Imaging-Reporting and Data System and was developed by the American College of Radiology. [6] Breast cancer diagnosis are categorized in 7 different categories. The category of 0 means the assessment was incomplete and the likelihood of cancer is not available. Category 1 and 2 label a 0% likelihood of cancer. In category 3 the likelihood is in the interval between 0% and 2%. Category 4 and 5 describes a likelihood smaller than 95% respectively greater than 95%. Last but not least category 6 means an already positive biopsy that confirmed breast cancer. [7] The feature Age is not explaining. The Shape can take 4 different values on nominal scale. The shapes can be round, oval, lobular and irregular with respectively assigned values 1, 2, 3 and 4. The same scale is applied to the feature Density where 1 is assigned to a high, 2 to a iso, 3 to low and 4 to a fat-containing density. The last feature that we will use in the application is the margin mass. The margin mass can be described with 5 different values. An integer of 1 describes circumscribed, while a 2 means microlobulated and a 3 obscured. 4 and 5 describe ill-defined spiculated respectively. The Severity is used to predict the outcome which is binary. An outcome of 0 means that no biopsy is necessary, while 1 means that a biopsy should be performed. The dataset is delivered with some missing values. The following table 1 shows the exact amount for each feature. [5] For preprocessing we decide to eliminate all NA values inside the data set. We assume that missing values are randomly distributed in the data set. This allows us to delete all rows which containing missing values. We can check if this changed the distribution of the Severity variable. If we plot before and after a histogram of this feature, we see that the data is still balanced. The mean of the feature is now 0.4855 which tells us that nearly 50% of the data has an outcome of 1. After this procedure 830 values are left in the data set. In the next step we are checking the data for outliers. The only we find is in the BI.RADS column. The value of the outlier is 55. By definition of the Bi.RADS the value need to be between 0 and 6. We can conclude that this is caused by a typo error, since 5 and 55 are from a typing point of view are close to each other. Based on this conclusion we manually fix the value by replacing it with 5.

BI-RADS	Age	Shape	Margin	Density	Severity
2	5	31	48	76	0

Table 1: Missing values

2.2 Methods

For analysing the data, we use Random Forests and Artifical Neural Networks. We describe shortly the concept of each method and define hyperparameters. We train different models with different hyperparameters. For a robust estimate of the error we are using cross validation. Each model is tested on a test data set. The test is done by predicting the outcome of the Severity variable. To measure the prediction performance for each mode the area under the curve (auc) is computed. The computation of the algorithms is done in R and depends heavily on the H2O R package. The reason therefore is that H2O comes with many useful tools for the analysis of algorithms.

3 Random Forest

3.1 Introduction

Decision trees have been proved to be successful to represent decision processes in medical applications. An very simple example is to ask for the patient age and using the answer for further guidance depending on the result. [3] Decision Trees are easy to implement and interpret but tend to overfit. To avoid this problem, we focus on Random Forests. These models are based on a collection of Decision Trees where each tree generated by a random subsample of our sample space. To predict the label a majority vote of all trees is done. [8] The Figure 1 shows the decision tree for the data. If the result of the BI.RADS is greater than 5 the tree predicts the outcome of 1 with a probability of 0.91. 40% of the data is used in this leave. It becomes more complex if the result of the BI.RADS is smaller than 5. Then it depends on the next step from the shape. If the variable is smaller than 4 in the next node the person is younger than 65, the tree predicts an outcome of 0.

3.2 Random Forest

To keep the complexity of the Random Forest algorithm low we are using just one hyperparameter. Other hyperparameters are defined through the default condition of the H2O package. This means max and minimal nodes are fixed as well as maximum and minimum number of terminal nodes. The

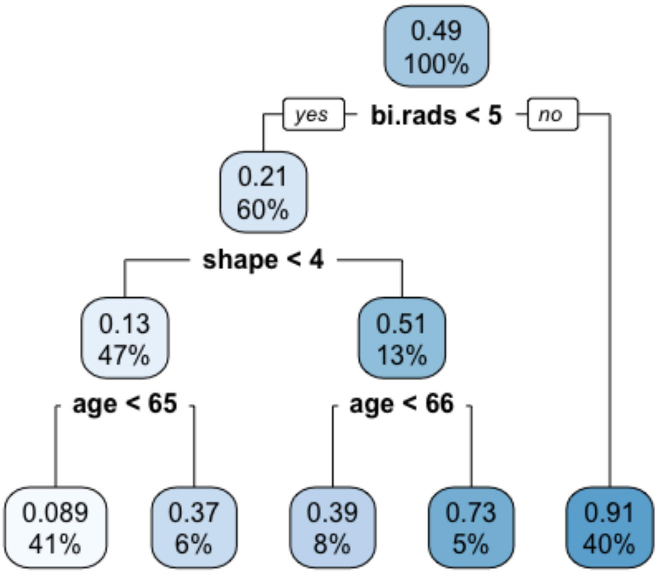

Figure 1: Decision tree

number of optimal predictors is computed with the following formula:

$$m \approx \sqrt{p}$$

M is the number of predictors used in each bootstrap and p are the number of features in the dataset. In this case p is equal to 5. Based on this formula each random generated tree is using two features. Back to our hyperparameter the question is how many trees should we use? To know the right number of trees we train a random forest with 150 trees. We expect that this Random Forest will not perform well. If we check the classification error in figure 2, we see that in the beginning additional trees reduces the classification error. But if we add to many trees the classification error starts to increases. We now build models with different number of trees and compare them.

3.3 Results

We see in Table 2 that the model with 11 trees achieves an auc of 0.8600413. We compare this with the default model of 50 trees. The auc of this model is slightly higher. In Table 2 we see that the lowest classification error do not implies that the auc has to be the highest. The Random Forest with 50 trees achieves a higher auc then the model with 11 trees. In Figure 3 we see that the highest training auc is caused by an model with 31 trees. If we add more trees we cannot improve the auc anymore. Table 2 shows us that the training auc of the Random Forest with 100 trees is close to the model with just 50 trees. We see as well that the auc achieved on the test data in Table 3 not increases as well. We train all the models again with cross validation. The Table 2 shows us that the cross validation auc of the Random Forest with 11 trees is higher than the training auc. The model with 31 trees has an cross validation auc close to the training auc. The model with 50 trees has a higher training auc than cross validation auc. This means these model perform really good on the training data but on the cross validation set they misclassify data. Of course the decrease is minimal in our case. We are now using the test dataset to measure the performance of prediction. As seen in Table 3 the best result is caused by the model with 50 trees. Surprisingly the model with 31 trees, which achieved the highest training auc, performs worse. A reason therefore can be that the model overfits. Because the auc on the test set of the model with 50 trees is the highest, we will choose this one for further analysis. If we analyze the variable importance we see in Table 4 the most important variables are BI.RADS and Age. Followed by Margin and Shape. A not very important variable is the Density. The result is just depending with around 2.2% on it.

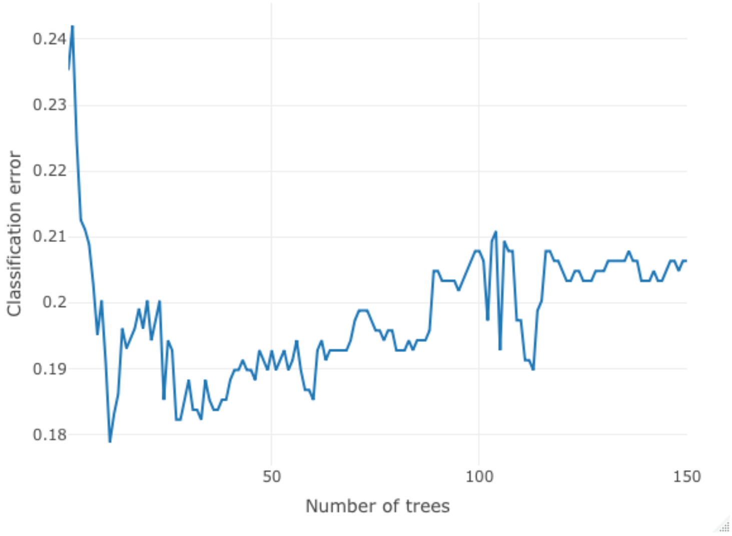

Figure 2: Random forest classification error

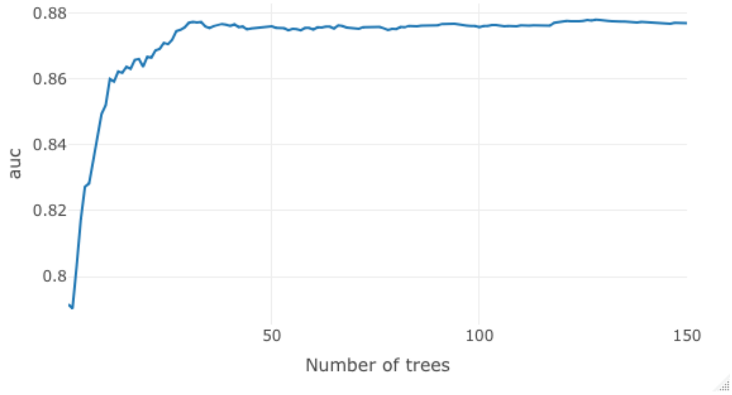

Figure 3: Random Forest auc

Model	Training AUC	CV AUC
11 trees	0.8600413	-
31 trees	0.8774041	-
50 trees	0.8760517	-
100 trees	0.8757477	-
11 trees and CV = 10	0.8600413	0.8749989
31 trees and CV = 10	0.8774041	0.8775811
50 trees and CV = 10	0.8760517	0.8759927

Table 2: Random forest training and cross validation auc

Model	AUC
11 trees	0.8664773
31 trees	0.8692453
50 trees	0.8746358
11 trees and CV = 10	0.8664773
31 trees and CV = 10	0.8692453
50 trees and CV = 10	0.8746358

Table 3: Random forest test auc

4 Artificial Neural Networks

4.1 Introduction

The second type of model we want to use to predict the outcome are Artifical Neural Networks (ANN). An Artifical Neural Network contains about at least one neuron. Several neurons connected together build then a network. [8] Figure 4 shows a neural network. Red lines are associated with negative weights and black lines with positive weights. The network has five input nodes, four hidden nodes and two output nodes. We have two output nodes because we do not predict exactly 0 or 1. Our model predicts the probability to be 0 or 1. We compute two outputs with for example 0: 0.2, 1: 0.8. This means that this output is equal to one with probability of 0.8.

4.2 ANN with H2o Deep Learning

We are using a three-layer neural network. This means it has one input, one hidden and one output layer. We are referring to the universal approximation theorem that just one hidden layer is necessary for the prediction of the outcome by a continuous function. [9] With this we define that the hyperparameter is the question of how many nodes are necessary in the hidden layer. For simplicity we will stick to just this hyperparameter and letting the rest defined by the default constellation of the deeplearning function of H2O. We build different networks and compare the performance of them. As loss function we use the cross entropy. Schulz-Wendtland and Elter proposed four hidden nodes in their research. [4]. The second ann contains 50 nodes in the hidden layer. The training is done by backpropagation. As activation function we use rectifier and the epoch is 10. This means that the neural network uses each datapoint 10 times for training. Again we are using the auc for measuring the performance. All of the models will be also trained with 10-fold cross validation.

4.3 Results

In Table 5 we see that the training auc of the ann with 4 hidden nodes is smaller compared with a model with 50 hidden nodes. If we test the models on the test set wee see in Table 6 that the neural network with 50 nodes performs just slightly better than the network with 4 nodes. If we are using cross validation both models cross validation auc is smaller than the training auc. But this difference is in the four nodes network smaller. Also in the test data the auc achieved by the models with cross validation is smaller. In fact, we have to emphasize here that the cross validation reduces the bias in

Variable	Importance percentage
BI.RADS	0.413796
AGE	0.233692
Margin	0.183511
Shape	0.146750
Density	0.022250

Table 4: Variable importance

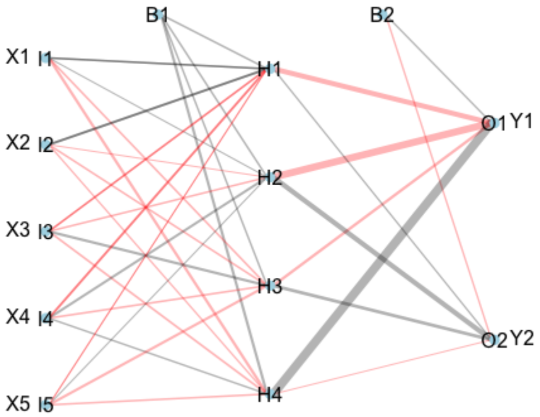

Figure 4: ANN

the model. The larger network can predict the training data better but in the validation the difference is bigger.

Model	Training Auc	CV Auc
ANN with 4 nodes in hidden layer	0.8947901	-
ANN with 50 nodes in hidden layer	0.9204084	-
ANN with 4 nodes in hidden layer and 10 fold CV	0.8936283	0.8852961
ANN with 50 nodes in hidden layer and 10 fold CV	0.9215566	0.9045428

Table 5: Training and cross validation auc of artificial neural network

5 Comparison

For the comparison we are choosing the Random Forest with 50 trees versus the Neural Network with 4 nodes in the hidden layer. We have chosen this Random Forest because it caused the highest aucin the test data. The reason for the Neural Network is, that this one is smaller and faster to compute. Based on the auc the ANN performs slightly better than the Random Forest. The Random Forest was able to achieve an auc of 0.8746358, while the ANN achieved 0.8758013. Both results are close to each other. To see the differences the confusion matrix is computed. We see this confusion matrix in Table 7. The total error rate of both models are close to each other. Inside the matrix we find the differences. The Random Forest can predict the Severity equal to 0 better than the ANN. The neural

Model	Prediction Auc
ANN with 4 nodes in hidden layer	0.8758013
ANN with 50 nodes in hidden layer	0.8999854
ANN with 4 nodes in hidden layer and 10 fold CV	0.8737617
ANN with 50 nodes in hidden layer and 10 fold CV	0.8932838

Table 6: Prediction auc of artificial neural network

network is better to predict the Severity equal to 1. A big difference is that the ANN predicts just in 5 cases the Severity equal to 0 when the true value is 1. Here the Random Forest classifies more than twice higher the Severity equal to 0 even the true value is 1. If we plot the roc cruve 5 we see that the ANN, the blue line, has a steeper slope than the Random Forest. The False Positive Rate of the Random Forest is slightly higher than of the ANN. We have seen that the cross validation often reduced the auc. We want to check now how good the models are with different data splits. To do this we build a loop and train the models with different splits of the data. After this we are using a bootstrap to simulate the distribution of the test auc. The ANN has a mean auc of 0.869317. This is close to the auc we computed before. The confidence intervals are for a 95% level are (0.8624652, 0.8758515). Figure]6 shows the distribution of the auc values. For the Random Forest we compute a mean of 0.8677496 with a confidence interval of (0.8640693, 0.8714809). Also here the confidence level is 95%. The Figure 7 shows the histogram of the Random Forest.

Result	ANN 0	ANN 1	ANN Error	ANN Rate	RF 0	RF 1	RF Error	RF Rate
0	64	24	0.272727	=24/88	71	17	0.193182	=17/88
1	5	73	0.064103	=5/78	15	63	0.192308	=15/78
Total	69	97	0.174699	=29/166	86	80	0.192771	=32/166

Table 7: Confusion matrix of ANN and RF

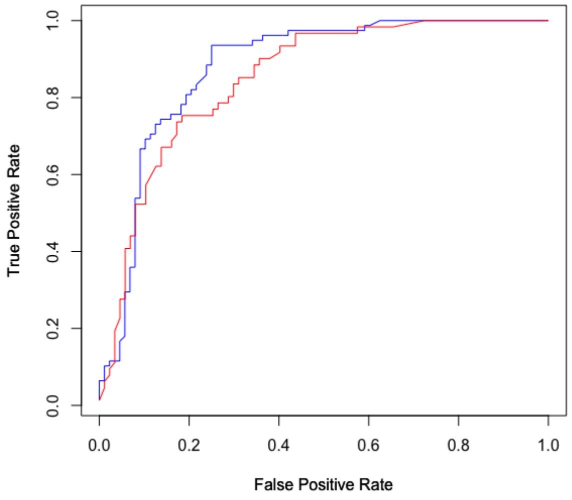

Figure 5: ROC Curve

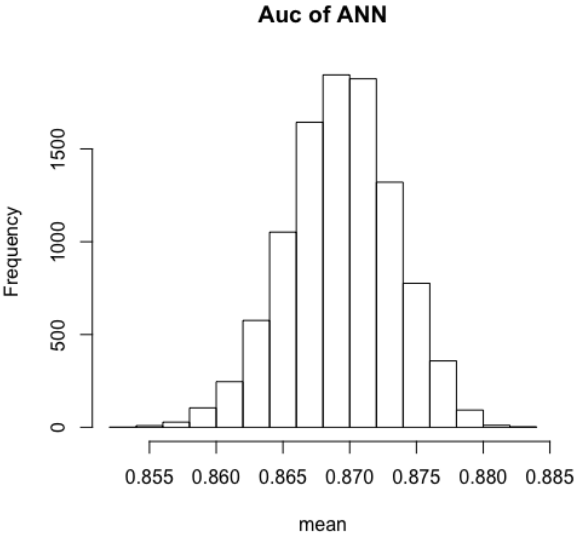

Figure 6: Histogram of the auc of the ANN

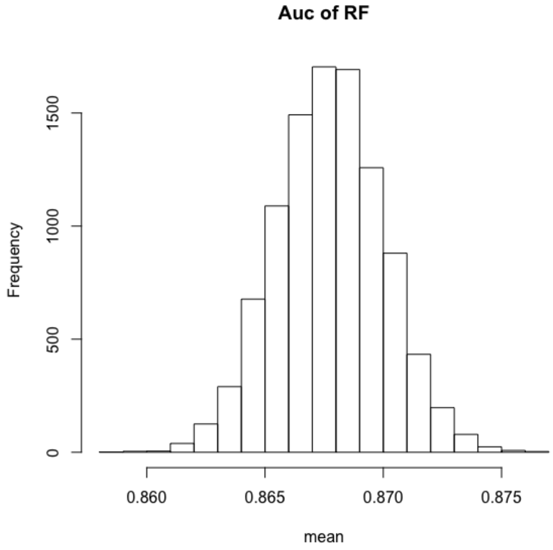

Figure 7: Histogram of the auc of the RF

6 Conclusion

To put it all into a nutshell. Both models can predict the severity. We discovered that the ANN performs slightly better than the Random Forest. We proved this by computing the mean and confidence intervals. Another important result is that the data split and preprocessing have a huge effect on the result. In experiments where no outliers were cleaned from the data, the ANN performs very bad. With different data splits we could influence the result as well. We have to add on this point that we have two open points that would may improve the result of the neural network. First, we have just few data available. More data would allow as to train and test networks with more data in each round. Second, we did not make a tuning of all hyperparameters. Here is further research necessary, since this can positively influence the outcome of a network.

References

[1] American Cancer Society. Breast Cancer Statistics. https://cancerstatisticscenter.cancer.org/?_ga=2.8685175.354318906.1591549652-887627775.1591549652#!/cancer-site/Breast, Unknown (accessed: 2020-06-23).

[2] N.E. Day, D.R.R. Williams and K.T. Khaw. Breast cancer screening programmes: the development of a monitoring and evaluation system. *British Journal of Cancer*, 59:954–958, 1989.

[3] J. A. López-Vallverdú, D. Riaño, J. A. Bohada. Improving medical decision trees by combining relevant health-care criteria. *Expert Systems with Applications*, 39(14):11782–11791, 2012.

[4] M. Elter, R. Schulz-Wendtland, T. Wittenberg. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical Physics*, 34(11):4164 – 4172, 2007.

[5] R. Schulz-Wendtland, M. Elter. Mammographic Mass Data Set. http://archive.ics.uci.edu/ml/datasets/mammographic+mass, 2007 (accessed: 2020-06-01.

[6] I.T. Nakano, H.R. Schelin, V. Denyak, S. Paschuk, S. Tacara, J.A.P. Setti. Mammographic breast density distribution and the fifth-edition BI-RADS criteria. *Radiation Physics and Chemistry*, 167, 2020.

[7] American College of Radiology. ACR BI-RADS® ATLAS — MAMMOGRAPHY. https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/Mammography-Reporting.pdf, Unknown (accessed : 2020-06-15).

[8] S. S. Shai, B.D. Shai. *Understanding Machine Learning - From Theory To Algorithms.* Cambridge University Press, 2014.

[9] Kurt Hornik. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, 4(2):251–257, 1991.