Classifying breast tumor using Neural Network

Matej Petrinović

1. INTRODUCTION

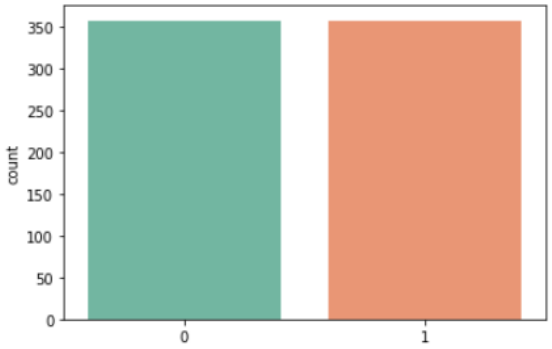
Breast cancer is the most common malignancy among women, accounting for nearly 1 in 3 cancers diagnosed among women in the United States, and it is the second leading cause of cancer death among women. Breast Cancer occurs as a result of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. A tumor does not mean cancer - tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as MRI, mammogram, ultrasound and biopsy are commonly used to diagnose breast cancer performed

In this paper we will discuss a diagnosis technique that uses the FNA (Fine Needle Aspiration), which is a quick and simple procedure to perform, which removes some fluid or cells from a breast lesion or
cyst with a fine needle like a blood sample needle.

This is an analysis of the Breast Cancer Wisconsin (Diagnostic) Dataset, obtained from UCI. This data set was created by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA. To create the dataset, data were taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan. The program uses an algorithm, to compute ten features from each one of the cells in the sample, then it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector.

2. DATA

As we have 30 features in the data, we will not display several rows of data for clarity, nor will we process each feature separately.
The aim of the study is to classify the tumor as malignant (label = 1) or benign (label = 0), which is our target variable.
The data consist of 714 observations, which we oversampled using **SMOTE** algorithm to get more data and thus equalized in tumor classes.



Picture 1. Target variable distribution

First, as a data processing, we will perform an analysis of correlations between features and graph it.

Picture 1. Target variable distribution

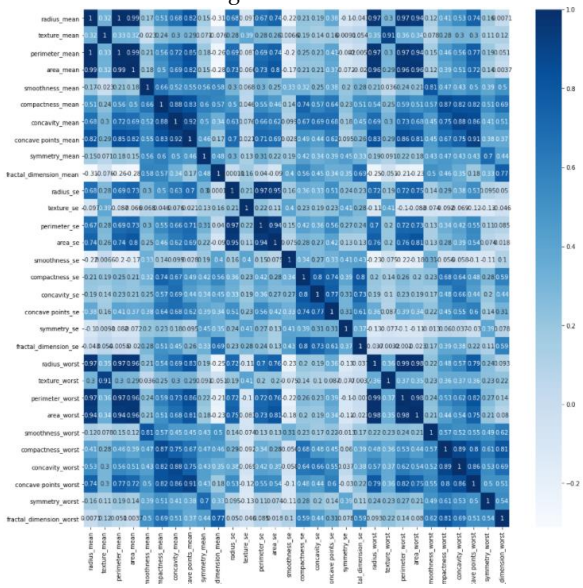


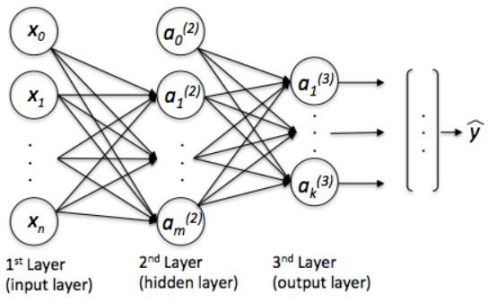
Table 1. Correlation matrix

As we see there are some very correlated features, with correlation greater than 0.9. This could make some problem in modelling step. So, these features will be removed.
The features that will be removed are
`perimeter mean, area mean, perimeter se, area se, radius worst, perimeter worst, area worst.`

Next, we'll scale data to (0,1) interval for reducing large data ranges, using *Minmax* Scaler. Now we have all our data ready and we're ready for modeling.

3. MODELING

For modeling we will use concept of Neural network for classification of a tumor, i.e., to calculate the probability of having malignant type of tumor. Artificial neural networks, usually simply called neural networks, are computing systems vaguely inspired by the biological neural networks that constitute animal brains.



Picture 2. Neural network

Our model will have the following structure:

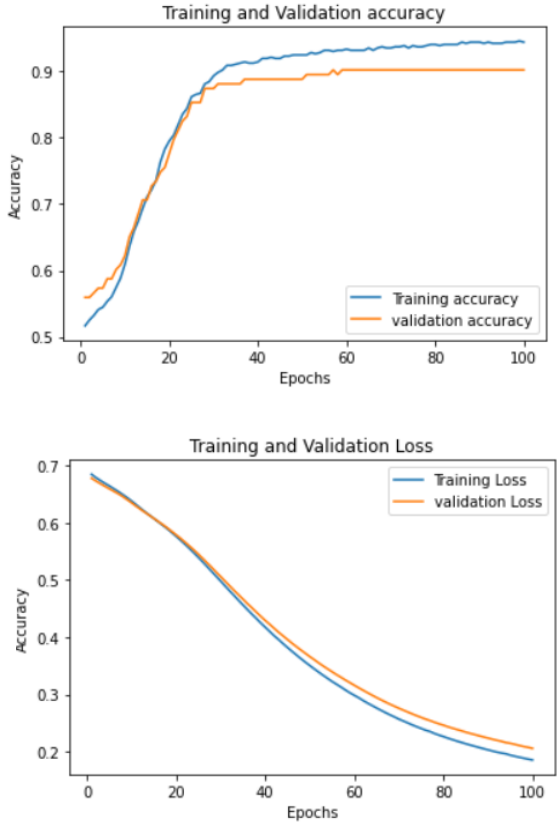
# Neurons Hidden Layer	Activation function
23	ReLu(x)
20	ReLu(x)
8	ReLu(x)
1	Sigmoid(x)

Table 2. Model structure

For optimizing algorithm Adam will be used with a learning rate μ = 0.0001. Training will be in 100 epochs with a train-test spit 80:20%.

4. RESULTS

For showing the results firstly we'll show a graph that shows model accuracy and model loss.



Picture 3. Model accuracy and loss

As it can be seen model accuracy improving improves through the epochs, and loss decreases. Next, we'll look at the confusion matrix which shows as the number of classification and misclassification.

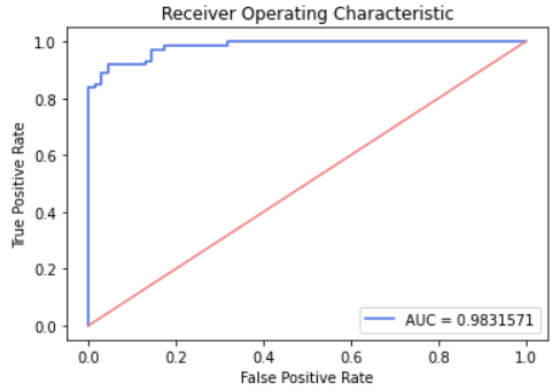
Actual/Predicted	*0*	*1*
0	59	10
1	4	70

Table 3. Confusion matrix

	Precision	*Recall*	*F1*
0	0.94	0.86	0.84
1	0.88	0.95	0.91
Accuracy			0.90
Macro avg	0.91	0.90	0.90
Wei. avg	0.90	0.90	0.90

Table 4. Classification report

As we can see from report table, all metrics are very high which is good. Next, *ROC* curve will be shown. *ROC* is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.



Picture 4. *ROC* curve

ROC is almost perfect. *AUC*, which is area under the *ROC*, is 0.9831. It can be concluded that trained model is good classificatory for tumor.

5. CONCLUSION

In this work we have trained a classification model for tumor types, malignant or benign. For model we have chosen neural network. Firstly, data cleaning and scaling were performed. Dataset was split on train and test part, 80%, 20% respectively. After training, model was tested on testing part of data and showed excellent results. For this analysis and model diagnostic it can be used for further uses.