# Analysis of tumor classification algorithms for breast cancer prediction by machine learning methods

https://doi.org/10.31713/MCIT.2021.07

Nataliya Boyko
department of artificial intelligence
Lviv Polytechnic National University
Lviv, Ukraine
Nataliya.i.boyko@lpnu.ua

Olena Kulchytska
department of artificial intelligence
Lviv Polytechnic National University
Lviv, Ukraine
olenakulchutska3@gmail.com

*Annotations* — **The paper presents machine learning methods for classification and forecasting problems. Tumor classification algorithms based on the Random Forest method are considered. To understand the distribution of classified data, a 3D graph of the three attributes of the date set was implemented. For a better understanding, graphs were constructed, namely the ROC curve and the RP curve. The AUC value for the model was also determined. The results of the graphs and AUC values were compared with the NoSkill model, ie the model without skills. High quality of the received models is offered.**

*Keywords* — **machine learning; tumor; random forest; foresight; classification.**

## I. INTRODUCTION

Breast cancer is the most common cancer among women in the world. It accounts for 25% of all cancers, and in 2015 alone more than 2.1 million women were affected. Early diagnosis significantly increases the chances of survival [2, 3]. The main problems facing its detection are how to classify tumors into malignant (cancerous) or benign (cancerous). A tumor is considered malignant if the cells can grow into surrounding tissues or spread to distant parts of the body. A benign tumor does not penetrate the surrounding tissues and does not spread to other parts of the body, like cancerous tumors. But benign tumors can be serious if they put pressure on vital structures such as blood vessels or nerves [1, 8].

The aim of the work is to analyze tumor classification algorithms based on the Random Forest method. Achieving this goal will be implemented using methods of data mining and machine learning and requires specific tasks:

- establish general principles of classification;
- clarify its characteristics;
- building a process model according to the dataset.

## II. ANALYSIS OF MATERIALS AND METHODS

Random Forest is a controlled learning algorithm. It can be used for both classification and regression. It is also the most flexible and easy to use algorithm. The forest consists of trees. It is said that the more trees in him, the stronger the forest. Random forests create decision trees on randomly selected data samples, obtain forecasts from each tree, and select the best solution by voting. It also provides a pretty good indicator of the importance of the function.

The primary task of machine learning is classification, to determine the class (group) of affiliation of observation. Data science provides many classification algorithms, such as logistic regression, vector support machine, naive Bayesian classifier, and decision tree. But at the top of the classifier hierarchy is a random forest classifier.

Random forest is implemented in various programs in various fields, such as recommendation mechanisms, image classification and feature selection. In this study, it should be used to classify and predict diseases. It is the basis of Borut's algorithm, which selects important functions of the data set [6, 10].

A random forest consists of a large number of individual decision trees that act as an ensemble. Each individual tree in a random forest gives a prediction of the class, and the class with the most votes becomes the forecast of our model.

To begin with it is necessary to analyze the offered algorithm. Technically, this is an ensemble method (based on a "divide and conquer" approach) of decision trees formed on a randomly divided data set. This collection of decision tree classifiers is also known as forest. Individual decision trees are generated using an attribute selection indicator, such as information increment, gain, and Gini index for each attribute [4, 6].

Each tree depends on an independent random sample. In the classification problem, each tree votes, and the end result is the most popular class. In the case of regression, the average value of all tree outputs is considered the end result. It is simpler and more powerful compared to other nonlinear classification algorithms [5, 7].

The basic concept behind the random forest is a simple but powerful concept – the wisdom of the crowd. In data science, the reason that the random forest model works so well is because of the large number of relatively uncorrelated models (trees) that act as a committee, surpassing any of the individual components of the model [9,11].

The algorithm works in four steps:

1. Select random samples from a given data set.

2. Build a decision tree for each sample and get the result of forecasting from each decision tree.

3. We vote for each expected result.

4. Choose the result of forecasting with the largest number of votes as the final forecast.

Low correlation between models is a key factor. The reason for this wonderful effect is that the trees protect each other from their individual mistakes (if they are not always all wrong in one direction). While some trees may be wrong, many other trees will be right, so as a group, trees can move in the right direction. Thus, the preconditions for an arbitrary forest to work well are:

- Functions must have some actual signal for models built using them to work better than random guesses.
- Predictions (and therefore errors) made by individual trees should be low correlated with each other.

## III. THE PROBLEM OF CANCER IN MACHINE LEARNING

Machine learning techniques can significantly increase the level of breast cancer diagnosis. Studies show that experienced doctors can detect cancer with an accuracy of 79%, while machine learning can achieve an accuracy of 91% (sometimes up to 97%) [4, 10].

The random forest method was chosen to classify tumors into benign and malignant. The Random Forest method occupies one of the leading places in modern machine learning. Random Forest is a controlled learning algorithm. It can be used for both classification and regression. It is also the most flexible and easy to use algorithm. The forest consists of trees. Random forests create decision trees on randomly selected data samples, obtain forecasts from each tree, and select the best solution by voting [3, 7].

## IV. EXPERIMENTS

### *A. Analysis and data processing*

Before creating a classification model, a very important step is to pre-process and clean the data. The data contains information about a specific tumor. Tumor parameters are ‘mean distance from center to points around the perimeter’, ‘standard deviation of gray scale values’, ‘average tumor nucleus size’, ‘mean of local changes in radius lengths’, ‘mean weight of concave contours’ and others. Among all the signs there is one target categorical variable, which acquires the values "M" or "B" (M - malignant - malignant, B - benign - benign tumor). Working with categorical data is not very suitable for the classifier, so the change of the type of the objective function from categorical to binary is implemented. The serial number of the study of a particular tumor does not carry any useful information for prediction, so it is unnecessary. An important step in pre-processing the data is to check the data for null values and duplicates and clear the dataset from them.

Visualization of data classification in 3D space is part of data analysis. Figure 1 shows that the smaller the values of the signs, namely the average number of concave parts of the contour, the largest average value for the number of concave parts of the contour and the largest perimeter of the tumor, the greater the probability of benign tumor.

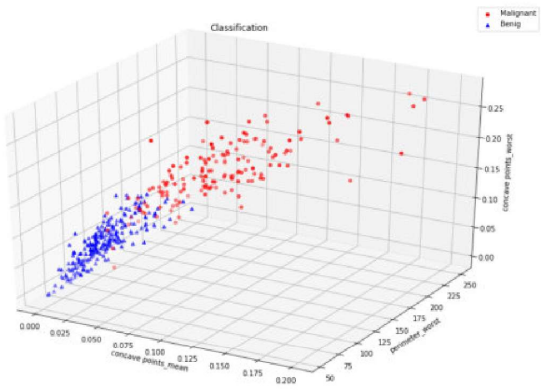


Figure 1. Data visualization

To build a good model, you should make a feature selection to clear the data from noise. To do this, we visualize the correlation matrix. The linear correlation coefficient is estimated between two random variables. Because of this, it is often called the pairwise correlation coefficient. Pearson's correlation coefficient - in statistics, the correlation index (linear dependence) between two variables X and Y, which takes values from −1 to +1 inclusive. It is widely used in science to measure the degree of linear dependence between two variables. That is, constructing a matrix, we obtain data on the linear relationship between all features.

The requirement for features is a high correlation with the target variable. However, independent variables that have a very high correlation with each other should be avoided. In this case, one of these features is removed.

Analyzing the obtained correlation matrix, it is seen that some features have a low correlation (Fig. 2) with the objective function, ie not a high linear relationship. From this we can conclude that this data can make noise and degrade the accuracy of the model. So we can remove one of them to improve accuracy.

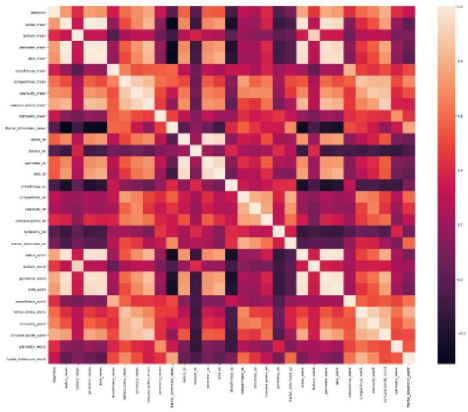

Figure 2. Correlation matrix

*B. Building a model*

After processing and analyzing the data, you can create and train a random forest model. Each random forest consists of a number of decision trees.

All committee trees are built independently according to a certain procedure: the classification of objects is done by voting: each committee tree refers to an object that is classified to one of the classes, and wins the class for which the largest number of trees voted.

The optimal number of trees is selected so as to minimize the classifier error in the test sample. In our case it is 10 trees (fig. 3).

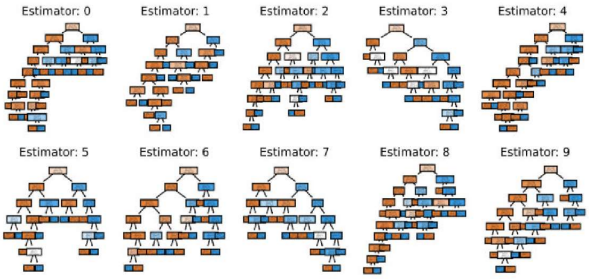


Figure 3. Visualization of a random forest model

*C. Quality assessment of the model*

It is not enough to consider one parameter to evaluate the model, as it is not objective. Therefore, it is necessary to determine several parameters of the model that will determine an adequate assessment of the model. These parameters include AUC (area bounded by the ROC curve) and F1-score (accuracy), R2-score (clarity) and Accuracy-score. All parameters are contained in table 1.

TABLE 1. MODEL EVALUATIONS

| **Model** | **Parameters** | | | |
|---|---|---|---|---|
| | ***AUC*** | ***F1score*** | ***R2 score*** | ***Accurancy score*** |
| Random Forest | 0.995 | 0.959 | 0.874 | 0.971 |

AUC is the area bounded by the ROC curve and the axis of the proportion of false positive classifications. The higher the AUC, the better the classifier, with a value of 0.5 indicates the unsuitability of the selected classification method (corresponds to conventional guessing). The AUC of our model is 0.995, which indicates the high quality of the classifier.

The F-measure is one of the measures of test accuracy. It is calculated through the accuracy and completeness of the test, where accuracy is the number of correctly determined positive results divided by the number of all positive results, including incorrectly defined, and completeness is the number of correctly determined positive results divided by the number of all samples to be defined as positive. The F-measure of our model is 0.959, which indicates a very good accuracy of the classifier.

R-squared is a measure of the clarity of the model and expresses the proportion of variance of the model. In the presented study of the model, the estimate is 0.874, which means a high degree of certainty of the model.

Accuracy score is an assessment of the accuracy classification. In a multi-label classification, this function calculates the accuracy of the subset: the set of labels provided for the sample must exactly match the set of labels from the test or training sample. In our case it is equal to 0.971 on the test data. This means that about 97% of the data is classified correctly, which is a very good result and proves the high accuracy of the classifier.

*D. ROC and PRC curves*

ROC-curve – a graph that allows you to assess the quality of binary classification, shows the ratio between the share of objects from the total number of carriers of the feature, correctly classified to the total number of objects that do not carry features, erroneously classified as having a feature. Also known as the error curve.

The shape of the curve contains a lot of information, including what worries us most about the problem, the expected false-positive and false-negative indicators.

A classifier without skills is one that cannot distinguish classes and assumes a random or permanent class in all cases. At point (0.5, 0.5) the model without skills is presented. The model, which does not have skills at each threshold, is represented by a diagonal line from bottom to left in the graph at the top right and has an AUC of 0.5. In fig. 4 unskilled model is shown by a dotted blue line.

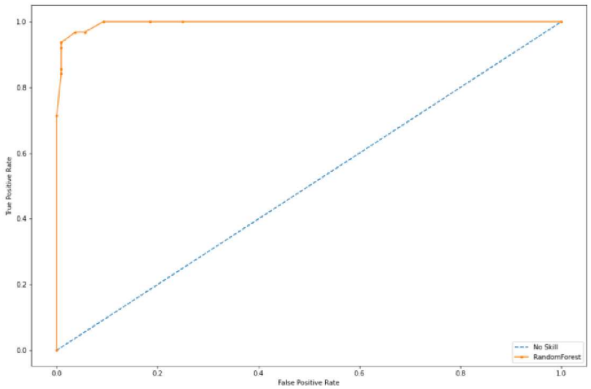


Figure 4. ROC-curve

A good model assigns a higher probability to a randomly selected real positive phenomenon than a negative one on average. This is what is meant when it is emphasized that the model has good skills. As a rule, good models are represented by bends that lean to the upper left corner of the site. In this case, you can see the bends in the upper left corner.

Quantitative interpretation of ROC is given by the indicator - the area bounded by the ROC-curve and the

axis of the share of false positive classifications. The area under the curve (AUC) can be used as a result of the skill of working with the model. An ideal model when the entire area is filled with color, but the proposed model is quite accurate, because the area under the ROC curve occupies almost the entire area of the graph.

PRC or Precision Recall Curve is a curve that shows the changes in Precision and Recall values depending on the threshold. Normally, the threshold is set to zero because the result will be a probability sign on the object. PRC will make it possible to understand how our model distributed probabilities across elements of different classes and how well it did so. This curve can be used only if the probabilities are in the range from 0 to 1, and in our case it is so.

This approach measures precision and recall. Accuracy describes how well the model predicts a positive class. Accuracy is called positive prognostic value. The recall is calculated as the ratio of the number of true positives divided by the sum of true positives and false negatives. Response is the same as sensitivity.

A classifier without skills is one that cannot distinguish classes and assumes a random or permanent class in all cases. The line of lack of skills changes based on the distribution of positive to negative classes. This is a horizontal line with the value of the ratio of positive cases in the data set. For a balanced data set, this is 0.5. The figure shows a blue dotted line (see Figure 5).

The model with good skill is depicted as a point in (1,1). A good model is represented by a curve (yellow) that slopes to (1.1) above a straight line, a model that has no skill (blue dotted).

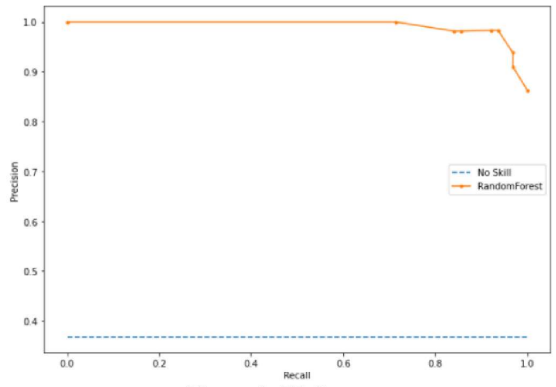


Figure 5. PRC curve

## V. CONCLUSION

Summarizing all the above, we can conclude about the work done. During the study, the classification of tumors by Random Forest method was presented.

When applying the algorithm to the data set, it was seen that it had a fairly high accuracy and the quality metrics were very good. Therefore, the data is cleaned and prepared correctly, the type of classifier is well chosen.

For the analysis of the decisive forest, the visualization of individual forest trees as its separate parts was realized.

So, let's summarize the advantages of the Random Forest algorithm over other classifiers:

a) It has the ability to efficiently process data with a large number of features and classes.

b) The decision forest is insensitive to the scaling of feature values.

c) It handles both continuous and discrete features equally well.

Equally important for large volumes of data is that this classifier is characterized by high parallelism and scalability.

## REFERENCES

[1] N. Montañez-Godínez, A.C. Martínez-Olguín, O. Deeb, R. Garduño-Juárez, and G. RamírezGalicia, "QSAR/QSPR as an Application of Artificial Neural Networks", Journal Artificial Neural Networks, Vol. 1260, 2015, pp. 319–333.

[2] M. Nekoei, M. Mohammadhosseini, and E. Pourbasheer, "QSAR study of VEGFR-2 inhibitors by using genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm-support vector machine (GA-SVM): a comparative approach", Journal of Medicinal Chemistry Research, Vol. 24, No. 7, 2015, pp. 3037–3046.

[3] O. Ivanciuc, "Drug Design with Artificial Neural Networks", Encyclopedia of Complexity and Systems Science, 2009, pp. 2139–2159.

[4] A. Khajeh, H. Modarress, and H. Zeinoddini-Meymand, "Modified particle swarm optimization method for variable selection in QSAR/QSPR studies", Journal of Structural Chemistry, Vol. 24, No. 5, 2013, pp. 1401–1409.

[5] M. Goodarzi, M.P. Freitas, and R. Jensen, "Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl) uracil derivatives using MLR, PLS and SVM regressions", Journal of Chemometrics and Intelligent Laboratory Systems, Vol. 98, No. 2., 2009, pp. 123–129.

[6] L. Blum, and J. Reymond, "970 million drug like small molecules for virtual screening in the chemical universe database GDB -13", Journal of American Chemical Society, Vol. 131 (25), 2009, pp. 8732–8733.

[7] A. Nonell-Canals, and J. Mestres, "In silico target profiling of one billion molecules", Molecular Informatics, Vol. 30 (5), 2011, pp. 405–409.

[8] M. Shahlaie, A. Fassihi, A. Pourhossein, E. Arkan, "Statistically validated QSAR study of some antagonists of the human CCR5 receptor using least square support vector machine based on the genetic algorithm and factor analysis", Journal of Medicinal Chemistry Research, Vol. 22, No. 3, 2013, pp. 1399–1414.

[9] X.B. Zhou, W.J. Han, J. Chen, and X.Q. Lu, "QSAR study on the interactions between antibiotic compounds and DNA by a hybrid genetic-based support vector machine", Monatshefte für Chemie – Chemical Monthly, Vol. 142, No. 9, 2011, pp. 949–959.

[10] B. Sprague, Q. Shi, M.T. Kim, L. Zhang, S.Sedykh, E. Ichiishi, H. Tokuda, K. Lee, and H. Zhu, "Design, synthesis and experimental validation of novel potential chemopreventive agents using random forest and support vector machine binary classifiers", Journal of Computer-Aided Molecular Design, Vol. 28, No. 6, 2014, pp 631–646.

[11] N. Boyko, "Application of mathematical models for improvement of "cloud" data processes organization", in Mathematical Modeling and Computing, Vol. 3(2), 2016, pp. 111–119. DOI: https://doi.org/10.23939/mmc2016.02.111