INTERNATIONAL SCIENTIFIC AND VOCATIONAL JOURNAL (ISVOS JOURNAL)

Vol.: 2 Issue: 2 Date: December 2018 Received: 25.10.2018 Accepted: 07.12.2018 Final Version: 25.12.2018

ISVOS Journal 2018, 2(2):48-56

Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers

Ahmet Saygılı[a]

[a]Namık Kemal University, Çorlu Engineering Faculty, Department of Computer Engineering, Tekirdağ, Turkey

Abstract

Cancer is one of the leading causes of human death in the world and has caused the death of approximately 9.6 million people in 2018. Breast cancer is the most common cause of cancer deaths in women. However, breast cancer is a type of cancer that can be treated when diagnosed early. The aim of this study is to identify cancer early in life by using machine learning methods. The characteristics of the people included in the Wisconsin Diagnostic Breast Cancer (WDBC) dataset were classified by support vector machines (SVM), k-nearest neighborhood, Naive Bayes, J48, random forest and multilayer perceptron methods. The preprocessing step was applied to the dataset prior to classification. After the preprocessing stage, six different classifiers were applied to the data using 10-fold cross-validation method. Accuracy, sensitivity, specificity, ROC area values, and confusion matrices were used to measure the success of the methods. As a result of the application, it was found that random forest was the most successful method with 98.77 % accuracy value. The second most successful method was the multilayer perceptron method with an accuracy value of 98.41%. When the results obtained from feature selection are evaluated, it is seen that feature selection and other preprocessing methods increase the success of the system. It can be said that the success achieved in comparison with previous studies is at a good level.

Keywords: "Breast Cancer, WDBC, Support Vector Machines, Gain Ratio, k-NN, Random Forest"

1. Introduction

Cancer is a general term for a large group of diseases that can affect any part of the body. Other terms are malignant tumors and neoplasms [1]. Cancer is characterized by the rapid spread of abnormal cells that go beyond their normal limits and then invade adjacent parts of the body and can spread to other organs. This process is called metastasis [1]. Metastases are the main cause of cancer-related deaths. Cancer is a worldwide fatal disease. In 2018, approximately 9.6 million people died due to cancer [1]. Globally, one out of six deaths is caused by cancer. Approximately 70% of deaths from cancer occur in low- and middle-income countries. The causes of deaths from cancer include body mass index, low fruit and vegetable consumption, lack of physical activity, tobacco use and alcohol use. Tobacco use is the most important risk factor for cancer and accounts for about 22% of cancer deaths [2]. Case and mortality rates for cancer types are shown in Table 1 [1];

Table 1. Number of cases and deaths of the most common cancer types worldwide [1]

Cancer Type	Case (#)	Death (#)
Lung	2.09 million	1.76 million
Breast	2.09 million	627.000
Colorectal	1.80 million	862.000
Stomach	1.03 million	783.000

Cancer is a type of disease that is caused by an uncontrolled growth of cells in the body. It is often referred to by the name of the structure in which the cancer disease is effective in the body. Breast cancer in women is a type of cancer with a very high mortality rate. Rapidly dividing cells form breast masses in breast cancer. These masses are called tumors. Tumors are divided into two groups as benign and malignant. Malignant tumors penetrate healthy body tissues and damage them. Harmful cells inside the tumor can spread to different organs of the body and damage them. Breast cancer means a malignant tumor placed in the breast.

Breast cancer is the most dangerous cancer that causes death among women aged 40-55. According to the World Health Organization, 2.09 million people are diagnosed with breast cancer every year [1]. Therefore, many studies have been carried out

Corresponding Author. Tel.: +90-282-250-2376 ; fax: +90-282-250-9924.
E-mail address: asaygili@nku.edu.tr

for the early diagnosis of cancer, which causes such harmful effects on humans. In this study, it has been tried to be diagnosed with cancer using Wisconsin Diagnostic Breast Cancer (WDBC) breast cancer data [3, 4].

There are many studies that have been performed on the WDBC breast cancer dataset and their success is quite high. Quinlan et al. carried out the first of these studies. In the study, C4.5 decision tree was used for classification and a success of 94.74% was achieved [5]. The fuzzy genetic algorithm was used in a study by Pena Reyes and a success of 97.36% was obtained [6]. In another study, Nauck and Kruse achieved a 95% success using blurred neurons [7]. In Setiono's study using feedforward neural networks, there was a success of 98.1% [8]. In a study by Albrecht et al., the perceptron neural network method was used and a success rate of 98.8% was obtained [9]. In a study using a fuzzy clustering method by Abonyi and his friend, a success of 95.57% was achieved [10]. Kiyan et al. using generalized regression neural networks achieved a success of 98.8% [11]. Polat and Güneş's study achieved a success rate of 98% [12]. In 2007, multilayer perceptron neural network (MLPNN), combined neural network (CNN), probabilistic neural network (PNN), recurrent neural network (RNN) and support vector machine (SVM) was used by Übeyli. In this study, the highest success was obtained using support vector machines with 99.54% [13]. In the study that Akay used together with feature selection and support vector machines, a success of 99.5% was achieved [14]. Peng et al. they achieved a success rate of 99.50% using the filter and wrapper methods [15]. In 2012, Salama et al. performed by the support vector machines, a diagnostic success of 97.71% were obtained [16].

Remaining parts of this article is organized as follows. Section 2 introduces the materials and methods used for the proposed diagnosis system. Section 3 presents the experimental results. Finally, Section 4 concludes with the contributions of the study and discussions.

2. Materials and Methods

The dataset used in this study was taken from the Irvine Machine Learning Repository of the University of California (UCI). The dataset is from the University Hospital of California created by Wolberg [17]. The UCI Machine Learning repository is an open-source repository with many datasets that can be used for experimental analysis of machine learning algorithms [3].

The dataset used in this study is the WDBC dataset consisting of 569 samples and 32 features in the UCI Machine Learning store. Some of the features included in the dataset are; properties such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal size for each cell nucleus. 212 were malignant (Malignant) and 357 were benign (Benign) of the 569 breast cancer data in the dataset. Figures 1 and 2 show examples of benign and malignant cancer cells in the dataset. The distribution of benign cancer cells is more uniform and structural malignancies are found in malignant cancer cells as shown in these figures.

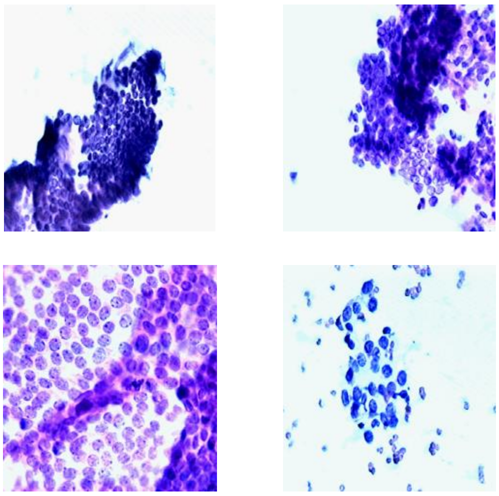

Fig 1. Benign cancer cell samples [18, 19]

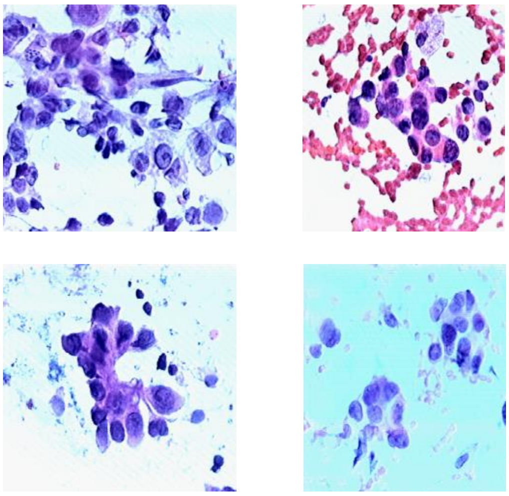

Fig 2. Malignant cancer cell samples [18, 19].

Some of the features in the datasets are more selective and decisive than other features. Moreover, the determination of these features significantly increases the success of the system. Feature selection methods are used to determine these features. Gain Ratio feature selection method was used in this study. In order to understand this method, it is necessary to mention briefly about decision trees and information gain. The most important step in the decision tree algorithm is to decide which feature to select in each node in the tree. Decision-tree-based algorithms use the Entropy measure of information to search from the features that give the valuable information to create the decision tree [20, 21]. The measure of the value of the feature is determined by a statistical value called information gain. Entropy characterizes the uncertainty of the samples. The entropy value of the S set is shown in Equation 1 [22].

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{1}$$

where p_i is the ratio of the number of instances of the ith class in the S to the number of all samples in the S set. c is the number of classes. If all the samples in the S set are in the same class, then the Entropy value is zero. If the number of classes in the S set is equal, the Entropy value takes a value with a maximum, that is the uncertainty. The efficiency measure of a feature is used with the term information gain. The gain of knowledge of feature A in Equation 2 is defined as Gain (S, A).

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

where Values (A) is a set of all possible values of feature A; S_v is a set of instances with v value of A feature in the S set. The C4.5 decision tree algorithm uses the value in Equation 3 for the gain ratio by normalizing the information gain [23];

$$SplitInfo_s(S) = -\sum_{i=1}^{v} \left(\frac{|S_i|}{|S|}\right) log_2(\frac{|S_i|}{|S|}) \tag{3}$$

With the help of Equation 3, the gain ratio is calculated in Equation 4;

$$Gain\ Ratio(A) = Gain(S, A)/SplitInfo_A(S) \tag{4}$$

The features selected by the gain ratio feature selection method are modeled with support vector machines, k nearest neighborhood, Naive Bayes method, J48 decision tree method, random forest, and multilayer perceptron method. Confusion matrix, accuracy, sensitivity, specificity and ROC area (AUC) metrics were used to measure the classification success of the methods. Equations 5, 6 and 7 show how these metrics are obtained. The confusion matrix is the matrix that represents the actual classes with the classes that are estimated in a classification system. Table 2 shows this matrix;

Table 2. Confusion matrices

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	TP	FN
	Negative	FP	TN

- True Positive (TP): Data that is actually sick and labeled as a patient.
- True Negative (TN): Data that is actually not sick and labeled as non-patient.
- False Positive (FP): Data that is actually sick and labeled as non-patient.
- False Negative (FN): Data that is actually not sick and labeled as patients.

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN} \tag{5}$$

$$\text{Specificity}=\frac{TN}{FP+TN} \tag{6}$$

$$\text{Sensitivity}=\frac{TP}{TP+FN} \tag{7}$$

Another measure of success is the value of the area under the receiver operating characteristic (ROC) curve. This value is drawn according to the true positive rate (TPR) and the false positive rate (FPR). TPR is a synonym for sensitivity in Equation 7. FPR is 1-Specificity (Eq. 6). A ROC curve is plotted to the TPR and FPR values of different classification methods as in Figure 3. The gray area below the broken lines shows the area under the ROC curve (AUC) value. The AUC provides a total performance measurement across all possible classification thresholds.

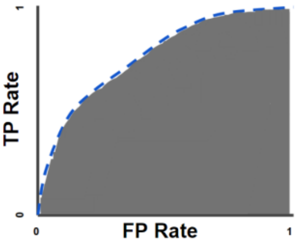

Fig 3. ROC curve [24]

3. Experimental Results

In our study, the WEKA platform was used for the preprocessing and classification of the WDBC dataset [25]. WEKA is an open source platform designed with JAVA programming language. WEKA supports various data mining functions such as data preprocessing, classification, clustering, association, regression, feature selection. Data points can be nominal, numerical, normal and other types of features.

The WDBC dataset used in our study consists of 569 samples and 32 features. One of the features is a class tag and one is an ID variable. First, the ID feature within 32 properties has been manually removed. Because ID is just a sequence number used to show examples. It is not a feature that can be used to evaluate data. Then, the Gain Ratio feature selection method is applied to the dataset. In this method, the feature with the highest gain ratio is chosen as the separation feature [26]. Following the determination of the gain ratios, fractal_dimension3, symmetry2, symmetry1, fractal_dimension2, smoothness2, fractal_dimension1, and texture2 properties, which have a gain value of less than 0.1, were excluded from the dataset. Thus, the number of features decreased to 24. After the feature selection, a random sampling preprocess was performed. The classification performance is often insufficient when learning from datasets in which the class distribution is unbalanced [27]. An unbalanced distribution in the WDBC dataset was attempted to be resolved using the random sampling strategy which increases the success rate of the classification method.

Many machine-learning methods give results that are more successful when data attributes are at the same scale. For this reason, the dataset has been normalized to the range [0-1]. After these preprocessing stages, the classification process was started. Six different methods were used in the classification stage: SVM, kNN (IBk in Weka), Naive Bayes, J48, random forest, and multilayer perceptron. For SVM, libsvm library was used on Weka platform [28]. 10-fold cross-validation method was used in the classification process. The flow chart of the operations performed is as shown in Figure 4. First, the raw dataset was classified without any preprocessing stages. In this way, it is aimed to analyze the effect of feature selection and preprocessing stage. The classification success rates for the raw dataset are shown in Table 3. Table 4 shows the results obtained after the preprocessing stages.

Table 3. Success rates without preprocessing stages

Method	Correctly Classified Samples (#)	Misclassified Samples (#)
SVM-Linear	544	25
SVM-Polynomial	370	199
SVM-Radial Basis	357	212
1-NN	545	24
3-NN	551	18
5-NN	552	17
Naive Bayes	530	39
J48	535	34
Random Forest	546	23
Multilayer Perceptron	552	17

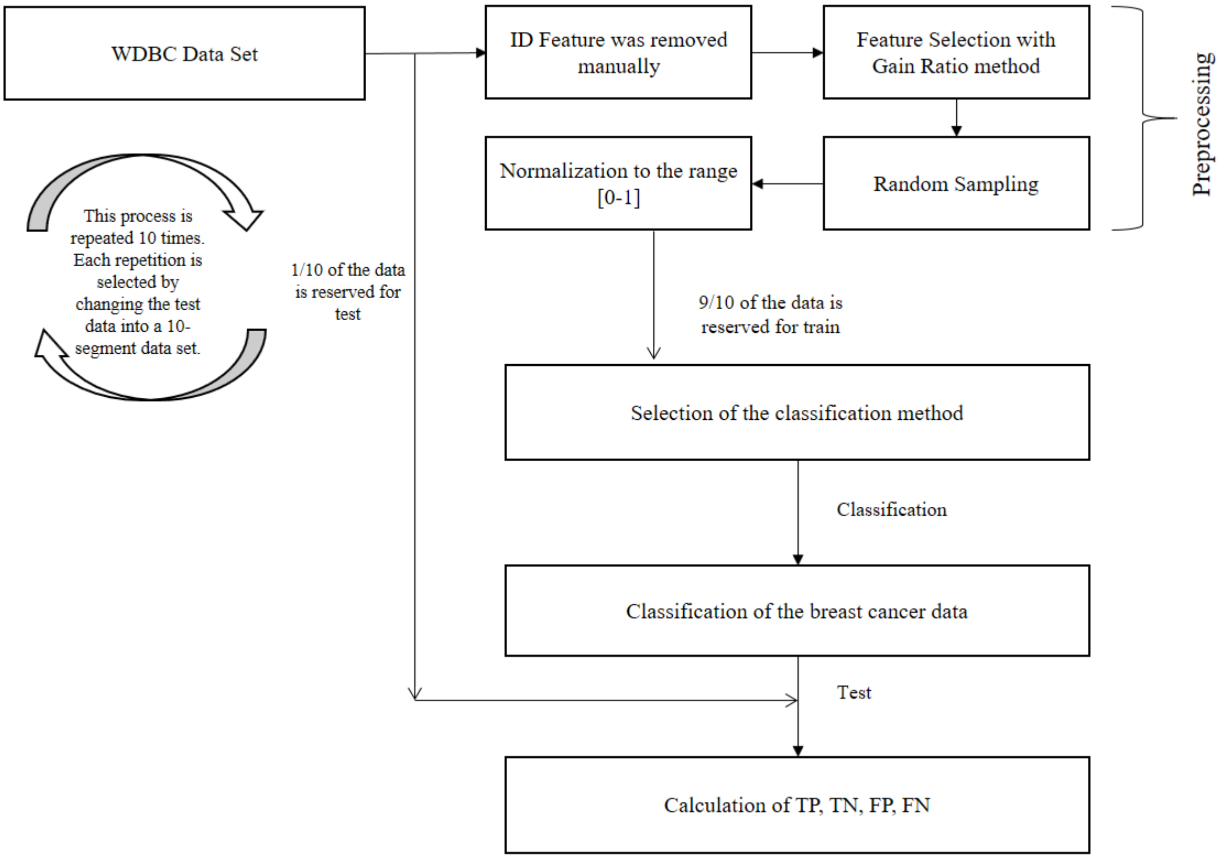
WDBC Data Set
ID Feature was removed manually
Feature Selection with Gain Ratio method
Random Sampling
Normalization to the range [0-1]
Preprocessing
This process is repeated 10 times. Each repetition is selected by changing the test data into a 10-segment data set.
1/10 of the data is reserved for test
9/10 of the data is reserved for train
Selection of the classification method
Classification
Classification of the breast cancer data
Test
Calculation of TP, TN, FP, FN

Fig 4. Flowchart for the classification system

When Table 3 and 4 are examined, it is seen that the preprocessing phases affect the success of classification. Especially in the random forest method, which gives the most successful results, it was observed that the selection of features changed the success considerably. In addition, the implementation of random sampling process has put forward the success of random forest. Because it shows the success of random forest in the datasets that are properly distributed. In support vector machines, three different kernel functions are used: linear, polynomial and radial based as seen in Table 4.

Table 4. Classification results with feature selection and preprocessing phases

Method	Correctly Classified Samples (#)	Misclassified Samples (#)
SVM-Linear	558	11
SVM-Polynomial	386	183
SVM-Radial Basis	555	14
1-NN	553	16
3-NN	554	15
5-NN	552	17
Naive Bayes	541	28
J48	558	11
Random Forest	562	7
Multilayer Perceptron	560	9

In Table 5, the confusion matrices are shown for each classification method, which is obtained because of the classification process after the selection of the features and preprocessing operations.

Table 5. Confusion matrices for classification methods

Method	Predicted Class		
	M (Malignant)	B (Benign)	Actual Class
SVM-Linear	205	7	M
	4	353	B
SVM-Polynomial	29	183	M
	0	357	B
SVM-Radial Basis	201	11	M
	3	354	B
1-NN	204	8	M
	8	349	B
3-NN	206	6	M
	9	348	B
5-NN	204	8	M
	9	348	B
Naive Bayes	197	15	M
	13	344	B
J48	208	4	M
	7	350	B
Random Forest	209	3	M
	4	353	B
Multilayer Perceptron	206	6	M
	3	354	B

Breast cancer data were classified by six different classification methods. The accuracy, sensitivity, specificity and AUC values obtained as a result of this classification process are shown in Table 6.

Table 6. Evaluation of the classification methods

	Accuracy (%)	Sensitivity	Specificity	AUC
SVM (linear)	98.07	0.981	0.975	0.978
k-NN (k=3)	97.36	0.974	0.973	0.991
Naive Bayes	95.08	0.951	0.942	0.986
J48	98.07	0.981	0.981	0.983
Random Forest	98.77	0.988	0.987	0.999
Multilayer Perceptron	98.41	0.984	0.979	0.998

4. Conclusions and Discussions

In this study, we used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to investigate the most successful breast cancer classification model. Support vector machines, k nearest neighborhood, Naive Bayes, J48 decision tree, random forest and multilayer perceptron methods were used in the classification. For a general comparison of success between methods, it is seen that the random forest method is the most successful method with a value of 0.999 when we evaluate it according to the preferred AUC value. This is followed by the multilayer perceptron method with a value of 0.998. The third method is the k-NN method with 0.991.

The performance criterion values of the models with the highest achievements for each method are shown in Table 6. The studies carried out from past to present, the methods they used and their successes are seen in Table 7.

Table 7. Success rates of the studies in the literature

Author(s)	Method	Success Rate (%)
Quinlan [5]	C4.5	94.74
Pena-Reyes and Sipper [6]	Fuzzy-GA	97.36
Nauck and Kruse [7]	NEFCLASS	95.06
Setiono [8]	Neuro-rule	98.10
Albrecht et al. [9]	Perceptron	98.80
Abonyi and Szeifert [10]	Fuzzy Clustering	95.57
Kiyan et al. [11]	Statistical Neural Networks	98.80
Polat and Gunes [12]	LS-SVM	98.53
Übeyli et al. [13]	SVM	99.54
Akay [14]	F-score + SVM	99.51
Peng et al. [15]	Filter and wrapper methods	99.50
Salama et al. [16]	SVM	97.71
This study	Random Forest	98.77

The comparison of Table 3 and 4 is important. Because in Table 3, the classification process is applied without any preprocessing steps. In addition, the correct and incorrect classified sample numbers are given in this table. In Table 4, after the many preprocessing stages, the classification was carried out. There is a significant increase in the success achieved. In this way, it can be clearly seen how the correct shaping of the dataset changes the success in such studies.

References

[1] O. WH. (2018, 10.01.2018). Cancer. Available: http://www.who.int/en/news-room/fact-sheets/detail/cancer

[2] C. Fitzmaurice, C. Allen, and R. Barber, "A systematic analysis for the Global Burden of Disease Study," JAMA Oncol, vol. 3, pp. 524-548, 2017.

[3] A. Asuncion and D. Newman, "UCI machine learning repository," ed, 2007.

[4] (10.01.2018). Repository UML. Breast Cancer Wisconsin (Diagnostic) Data Set. Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[5] J. R. Quinlan, "Improved use of continuous attributes in C4. 5," Journal of artificial intelligence research, vol. 4, pp. 77-90, 1996.

[6] C. A. Pena-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis," Artificial intelligence in medicine, vol. 17, pp. 131-155, 1999.

[7] D. Nauck and R. Kruse, "Obtaining interpretable fuzzy classification rules from medical data," Artificial intelligence in medicine, vol. 16, pp. 149-169, 1999.

[8] R. Setiono, "Generating concise and accurate classification rules for breast cancer diagnosis," Artificial Intelligence in medicine, vol. 18, pp. 205-219, 2000.

[9] A. A. Albrecht, G. Lappas, S. A. Vinterbo, C. Wong, and L. Ohno-Machado, "Two applications of the LSA machine," in Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on, 2002, pp. 184-189.

[10] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers," Pattern Recognition Letters, vol. 24, pp. 2195-2207, 2003.

[11] T. Kiyan and T. Yildirim, "Breast cancer diagnosis using statistical neural networks," IU-Journal of Electrical & Electronics Engineering, vol. 4, pp. 1149-1153, 2004.

[12] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," Digital signal processing, vol. 17, pp. 694-701, 2007.

[13] E. D. Übeyli, "Implementing automated diagnostic systems for breast cancer detection," Expert systems with Applications, vol. 33, pp. 1054-1062, 2007.

[14] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," Expert systems with applications, vol. 36, pp. 3240-3247, 2009.

[15] Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification," Journal of Biomedical Informatics, vol. 43, pp. 15-23, 2010.

[16] G. I. Salama, M. Abdelhalim, and M. A.-e. Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers," Breast Cancer (WDBC), vol. 32, p. 2, 2012.

[17] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cancer Wisconsin (diagnostic) data set," UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/], 1992.

[18] W. Wolberg. (1993). Cancer Images. Available: ftp://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer/cancer_images/

[19] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cytology diagnosis via digital image analysis," Analytical and Quantitative Cytology and Histology, vol. 15, pp. 396-404, 1993.

[20] J. R. Quinlan, "Induction of decision trees," Machine learning, vol. 1, pp. 81-106, 1986.

[21] R. E. Blahut, Principles and practice of information theory: Addison-Wesley Longman Publishing Co., Inc., 1987.

[22] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," International Journal of Information Technology and Knowledge Management, vol. 2, pp. 271-277, 2010.

[23] J. R. Quinlan, "Bagging, boosting, and C4. 5," in AAAI/IAAI, Vol. 1, 1996, pp. 725-730.

[24] Classification: ROC and AUC. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, pp. 10-18, 2009.

[26] J. Han and M. Kamber, "Data mining concepts and techniques San Francisco Moraga Kaufman," 2001.

[27] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," Journal of Artificial Intelligence Research, vol. 19, pp. 315-354, 2003.

[28] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," ACM transactions on intelligent systems and technology (TIST), vol. 2, p. 27, 2011.