AN EFFECTIVE HYBRID CLASSIFIER FOR BREAST TUMOR CLASSIFICATION

Sannasi Chakravarthy S. R. and Harikumar Rajaguru
Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam, India
E-Mail: elektroniqz@gmail.com

ABSTRACT

Even though the world is well-advanced and fast-enough, cancer is still a life-threatening disease for every living being. The global mortality rate due to cancer is steadily increasing all days. In particular, breast cancer is the one which plays a major role in affecting human lives. Thus, a proper automated and computer-aided diagnosis tool is essential for the prediction of breast tumours. The paper aims to propose a hybridized algorithm that integrates Quadratic Discriminant Analysis (QDA) with Multi-Layer Perceptron (MLP) techniques. The idea behind the proposed algorithm is that the output of Quadratic Discriminant Analysis is cascaded with the Multi-Layer Perceptron network for the automatic classification of breast tumours. The paper utilizes the standard benchmark breast cancer dataset, Breast cancer wisconsin (diagnostic) dataset. The evaluated results are compared against the support vector machine, random forest, adaboost and Gaussian process classification algorithms. These comparisons are done through the calculation of confusion matrix elements. By using the elements of confusion matrix, several performance metrics are derived and used for the comparison of proposed classification algorithm with the existing ones. The paper exactly assesses the classification of tumour severity of breast cancer i.e. benign and malignant ones. The evaluated results show that the proposed hybrid algorithm is better in classifying the benign and malignant inputs than the existing algorithms.

Keywords: benign, malignant, breast cancer, hybrid, neural network, wdbc.

INTRODUCTION

Among all the types of cancer, breast cancer is the invasive one found in women. And after lung cancer, it is being the second highest form of cancer amongst women. A cancer which origins in the breast cells make the breast cancer. Advance methodologies in both screening and treatment will improve the recovery rates radically since 1989 [1]. This type of cancer can be found in higher rates for women and very rare for men. The primary signs of breast cancer are the formation of a lump around the breast region; changes occur in the texture or shape of the entire breast or in nipple part and discharge of blood dots occur in the nipple [1].

In general, the breast cancer starts to affect either in the ducts or lobules part of the breasts. Here, the lobules represent the glands which are responsible for producing milk whereas ducts refer to the pathways which are responsible for carrying the milk to the nipple from the glands of breast [2]. This type of cancer might also affect the fibrous connective or the fatty tissue around the breast regions. The uncontrolled and aggressive cancer affected cells habitually invade the nearby healthy cells in the breast and can have the ability to spread to the lymph nodes that are usually found under the arms [2]. From this we can say that these lymph nodes can act as a primary pathway which supports the transportation of cancer cells to all other regions of the breast.

In the early stage of breast cancer, it may not show us any type of symptoms. And in several cases, a tumour might be very small and it can't be able to felt physically by the cancer affected women. If any woman can have the ability to feel the tumours, then the initial sign is generally a newer lump found in the breast regions [3]. On the other hand, all the lumps are not decided as cancer. This controversy makes the screening of breast cancer very harder. Thus, the diagnosis for the primary stages of breast cancer should always be a tougher one for any physicians.

The laboratory tests which are used for breast cancer screening include the use of mammograms and ultrasound imaging. The most common approach for checking or screening of breast for breast cancer is with an imaging method, mammography approach. Several mid-aged and older-aged women have undergone annual mammogram check-up for breast cancer screening [4]. If any clinician suspects that if anyone has a tumour or suspicious lump, then they recommend preparing for a mammogram. And if any abnormality lump or spot is found on the obtained mammograms, then the clinicians might request further tests for breast cancer [3]. A breast ultrasound is also an imaging technique, which uses acoustic waves for providing an image of the deeper breast tissues. And this can help the clinicians to make a distinction between a tumour (solid mass), and a benign cyst [4].

If the clinicians suspects cancer in the breast, then they might recommend for both mammogram and ultrasound scanning of breast. If these two tests are not able to help the clinicians to make decisions on breast cancer, then the clinicians should recommend for taking breast biopsy test. During this biopsy test, the clinicians will take a sample of tissue from the suspicious spot of breast [5]. There are various forms of breast biopsies. In some of the biopsy tests, the clinicians make use of a needle to collect the sample of breast tissue. And in other biopsy tests, the clinicians will make an incision in breast area and then the sample is collected [6]. This paper utilizes the standard benchmark breast cancer dataset, Breast cancer wisconsin (diagnostic) dataset. In this dataset, the features were extracted from the breast biopsy test that makes use of fine needle.

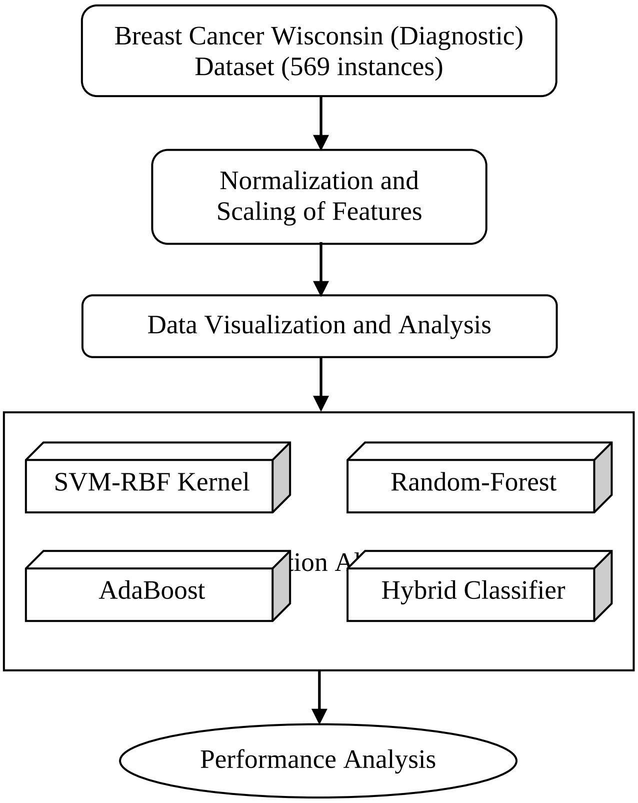
Breast Cancer Wisconsin (Diagnostic) Dataset (569 instances)
Normalization and Scaling of Features
Data Visualization and Analysis
SVM-RBF Kernel
Random-Forest
ion A
AdaBoost
Hybrid Classifier
Performance Analysis

Figure-1. Work-Flow.

Figure-1 illustrates the flow of work to be followed in this paper for an effective classification of breast tumour. As shown in Figure-1, the paper employs Breast Cancer Wisconsin (Diagnostic) dataset that contains a total of 569 instances. The normalization and scaling of features is done for making the decision making easier. The features are then visualized for its further classification analysis. After visualization, four distinct classification algorithms are utilized for classifying the benign and malignant inputs. Finally, the results are compared using standard metrics for finding its effectiveness.

MATERIALS AND METHODS

This section discusses about the materials and methods used for classifying the benign and malignant inputs.

Dataset and its Visualization

As shown in Figure-1, the paper utilizes the Breast Cancer Wisconsin (Diagnostic) data corpus for its evaluation. The dataset is shortly and popularly termed as WDBC dataset. The WDBC dataset has the characteristics of having 569 instances with thirty numeric and predictive attributes [7]. The attributes of the dataset includes radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. In this, the radius refers to the mean of distances measured from centre to different points noted on the perimeter values. The texture attribute refers to the standard deviation (SD) measurement of gray scale values. During the biopsy test, the local variation measurement in radius lengths is referred as smoothness attribute. The compactness attribute is calculated based on the formula of,

$$Compactness \ = \frac{perimeter^2}{area-1} \tag{1}$$

The concave attribute refers to the severity of identified concave portions related to its contour. The amount of concave portions related to the contour is referred as concave points. And the complementary of coastline approximation is referred as fractal dimension. The standard error, mean, and the largest or worst of these attributes were determined for each obtained image, yields a total of thirty features. By this way, the dataset was created by Mr Nick Street, Dr. William Wolberg, and Olvi Mangasarian [7]. Out of the total 569 instances, the dataset has no missing values and this makes this dataset as a more popular one among the medical researchers. The output class or labels of this dataset is denoted as B (benign) and M (malignant). And this class distribution is given in Figure-2.

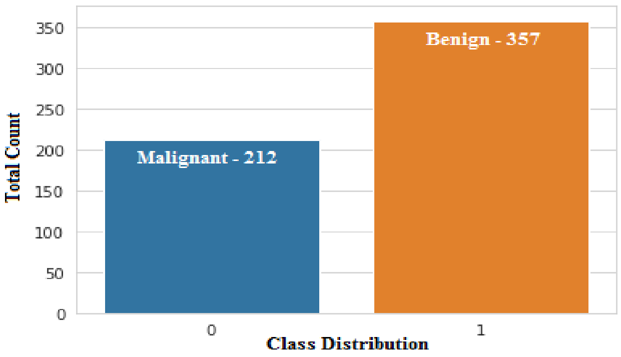

Figure-2. Class distribution of WDBC dataset.

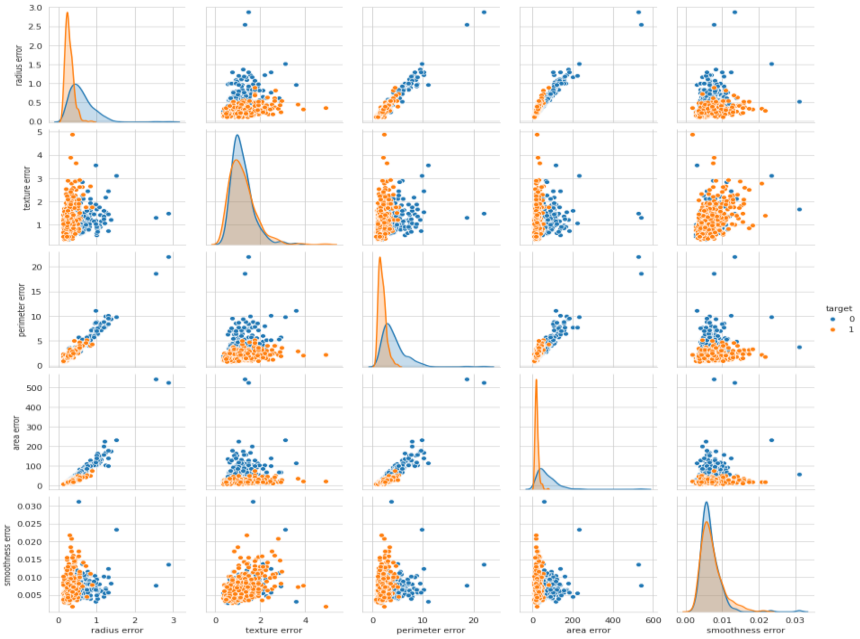

Figure-3. Visualization of features of WDBC dataset.

Figure-3 shows the visualization of pair plot of some error features such as radius error, texture error, perimeter error, area error, and smoothness error with respect to the output class targets. As in Figure 3, all the features are skewed to different range and thus normalization and scaling of features is necessary to make the further classification easier. In this Figure-3, the blue and orange colours represent the malignant and benign classes respectively.

Normalization and Scaling

After splitting the dataset randomly into training (80%) and testing (20%) data, the normalization and scaling is done. This refers to the way of rescaling all the attributes to lie in the range → 0 to 1. This makes the dataset with the maximum value of all attributes as 1 and the least value of all attributes as 0. The paper follows the idea of min-max normalization to scale the features of WDBC dataset.

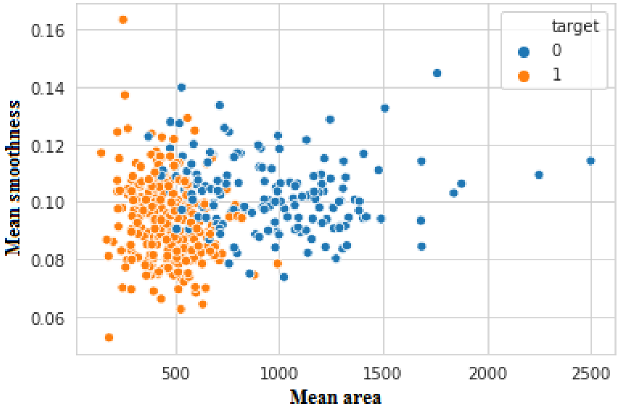

Figure-4. Sample features before normalization.

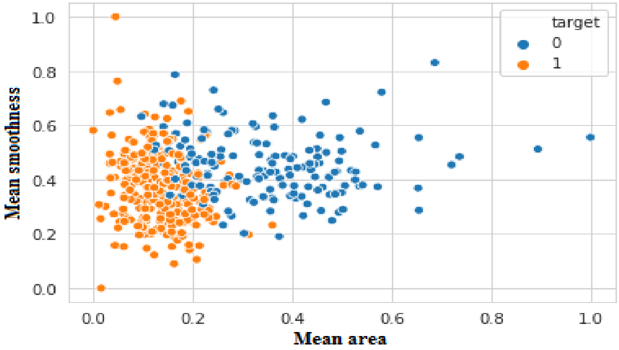

Figure-5. Sample features after normalization.

The relative difference between the features before and after normalization is shown in Figure 4 and in Figure-5. For data visualization, we have considered the mean area and mean smoothness features. In these figures, these two features are plotted against its output target - benign and malignant. From Figure-5, the features are confined to the range of 0 and 1.

CLASSIFICATION ALGORITHMS

Support Vector Machine (SVM)

In machine-learning, SVM is a supervised learning model that can be applied for both classification and regression problems [8]. But, SVM is more popularly used for solving classification related applications. In the SVM model, we can plot every data attribute as a point in n −dimensional space. For our problem, the n value is 569 since the dataset contains 569 instances. In this plot, the each feature value is represented as the value of a specific coordinate. Afterwards, classification is carried out by computing the hyper-plane. Thus, SVM performs the task of classification by plotting the hyper-plane for the considered problem. This hyper-plane is used for distinguishing the benign and malignant output classes [8]. Here, it is easier for the SVM to provide a linear hyper-plane between the output targets. Can we use this SVM model for non-linear classification problems? For this, SVM is having a kernel trick to solve for non-linear classification problems. The kernel used in SVM is a function which considers lower-dimensional input space and then transforms to a different i.e. higher-dimensional space. This implies that SVM converts non-linear (non-separable) problems to linear (separable) one by using the kernel trick. From Figures 3, 4, and 5, the input dataset is found to be non-linear. Thus, the paper utilizes the SMVM model with radial basis function (RBF) kernel.

Random Forest (RF) Algorithm

Ensemble models are those which make use of more than one models of either same or different type for classification problems. Random forest (RF) classifier is a type of ensemble tree-based learning model. The RF algorithm consists of a group of decision trees which are taken from the random subset of training data [9]. The RF classifier simply aggregates all the votes resulted from various decision trees to conclude the final output class of the test data. The algorithm of random forest is summarized as [9]:

a) At initial step, the algorithm starts with the random selection of sample data from the input dataset.
b) Then, RF model will start to create a decision tree for every input. And at the same time, the algorithm will acquire the prediction result from these decision trees.
c) Now, voting process will be done for the above predicted results.
d) Finally, the algorithm will select and decide the most-voted prediction result as the final output.

AdaBoost (AB) Algorithm

The previously discussed random forest algorithm is an ensemble learning model that works by bagging technique. AdaBoost model is also an ensemble learning model but works by boosting technique [10]. The AdaBoost classifier is one of the first models which use boosting technique adapted for solving practices. Simply, the Adaboost algorithm works by combining several “weak classifier” into a single “strong classifier”. This makes the algorithm to be named as Adaptive-Boosting (AdaBoost). The algorithm of AdaBoost is summarized as [10]:

a) Initialize the sample (identical) weights for the taken input.
b) Construct a decision tree along with each attribute, classification of data is carried out and then the result is evaluated.
c) Compute the significance of all the trees obtained in the final classification. The significance can be calculated using:

$$significance = \frac{1}{2} log\left(\frac{1-total\ error}{total\ error}\right) \qquad (2)$$

Here the sum of all the weights corresponds to the incorrectly classified inputs is referred as total error.

d) Updation of sample weights is done and so the subsequent decision tree will consider the error made by the former decision trees into account and construct a new data.
e) Repeat the steps (a) to (d) until the iteration count will become same as the number of estimators as specified in the hyperparameter.

Hybrid Classification Algorithm

For any machine learning approach being used in solving classification tasks of medical data, it is expected to provide some desired things such as better performance, better transparency of algorithm, better ability to decide predictions and better ability of the model to reduce the classification error to acquire reliable results. To obtain better classification results, the classifier model has to provide higher classification accuracy even on new input cases. Nowadays to obtain better performance, different learning models are employed and being tested on the newer dataset.

Among these, the best ones (either one or two algorithms) will be chosen for implementation of a robust classification framework [11]. In literature, the artificial neural network (ANN) always try to provide a better performance in medical classification problems but still its performance is needed to enhance its classification accuracy [11]. In order to achieve this, the work combines the Quadratic Discriminant Analysis (QDA) with the Multi-layer Perceptron (MLP) neural network model. Let’s see the advantages in combining the classifier models to get better classification results in the next section.

RESULTS AND DISCUSSIONS

The input dataset (WDBC) taken is splitted randomly into training (80%) and testing (20%) sets. In these sets, the algorithms are implemented and their results are evaluated. All the work carried in the paper is implemented using Python 3.6 accompanied with Intel core i5 vPro, 8 GB RAM, 2 TB hard-drive and with Windows 7 operating system. The attained classification results are evaluated using standard performance metrics - Sensitivity (Se), Specificity (Sp), Accuracy (Ac), Precision (Pr), F1 Score (F1) and Matthews Correlation Coefficient (MCC). These metrics are derived after constructing the confusion matrix for the classification algorithms.

Table-1. Confusion matrix for classification algorithms.

Classifiers	Confusion Matrix			
	TP	FN	FP	TN
SVM-RBF	158	54	53	304
Random-Forest	164	48	49	308
AdaBoost	181	31	38	319
Hybrid Classifier	196	16	26	331

Table-1 provides the confusion matrix results of four classifiers taken for our classification task. As from Table-1, the number of false predictions is more for SVM algorithm and the number of true predictions is more for the proposed hybrid classifier. These predictions are calculated and obtained for both benign and malignant cases.

Table-2 shows the calculated performance metrics that are computed using the confusion matrix as shown in Table-1. As in Table-2, six different performance measures are considered for the comparison of performance of the different algorithms taken for our problem.

Table-2. Performance measures of different algorithms.

Classifiers	Performance Measures					
	Se	Sp	Acc	Pr	F1	MCC
SVM-RBF	74.5	85.1	81.2	74.8	74.7	59.7
Random-Forest	77.3	86.2	82.9	77	77.1	63.5
AdaBoost	85.3	89.3	87.8	82.6	83.9	74.2
Hybrid Classifier	92.4	92.7	92.6	88.2	90.3	84.4

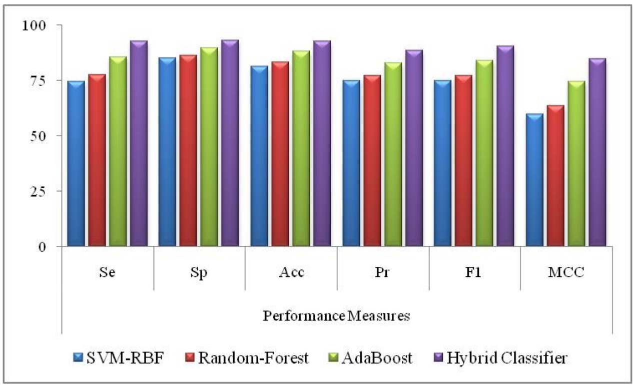

Figure-6. Graphical performance analysis.

As shown in Table-2, the performance analysis of the classifiers is compared graphically in Figure-6. From Table-2, the maximum classification accuracy is yielded for the proposed hybrid classifier i.e. the output of Quadratic Discriminant Analysis (QDA) is cascaded to the input of Multi-layer Perceptron (MLP) neural network model. Here, the highest classification accuracy of 92.6% is obtained for this hybrid classifier. Even though, the standard SVM algorithm employs radial basis function kernel, it provides a maximum of 81.2% of classification accuracy. The ensemble learning models - random forest and adaboost classifiers are in the position of giving 82.9% and 87.8% of classification accuracies for the task of breast cancer classification. These performance measures shown in Table-2 are graphically plotted in Figure-6. As from both Table-2 and Figure-6, the performance of adaboost classifier is significantly greater than the random forest classifier. As from Figure-6, the hybrid classifier provides a maximum classification performance for distinguishing the benign and malignant inputs. In other words, the hybrid classifier provides a highest sensitivity, specificity, accuracy, precision, F1 score and MCC values as compared with the SVM algorithm with RBF kernel, random forest algorithm and with the AdaBoost classifier models. The decision boundary is plotted for different classifier is shown in Figure-7. In Figure-7, it is noted that the decision boundary between benign and malignant input features is plotted in a better way for the proposed hybrid classifier and this will make the hybrid classifier to perform well than the others.

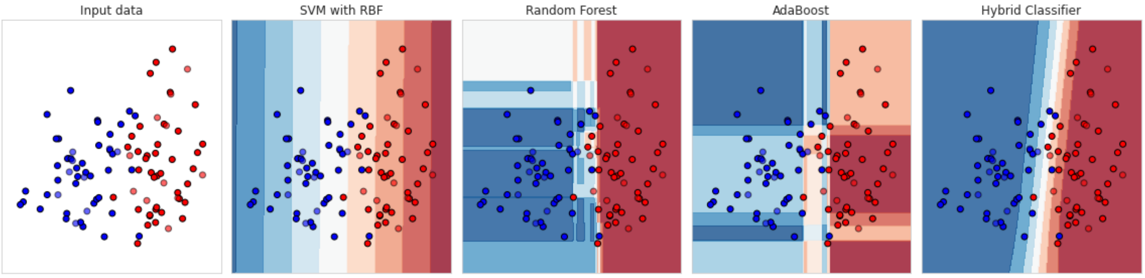

Figure-7. Plot of decision boundaries for different classifiers.

CONCLUSIONS

The computer-aided and automatic classification framework is proposed in this paper for the classification of breast cancer (benign or malignant one). For distinguishing between the benign and malignant input features, the performance of the proposed hybrid classifier is calculated and compared with the support vector machine algorithm, random forest algorithm and with the AdaBoost classifier algorithm. Here, for solving the non-linear classification, the SVM is implemented with the radial basis function kernel. Rather than giving raw features to the classification algorithms, the input features are analysed and normalized using a standard scaling method. This will make the further classification process easier and will influence the classification results of the algorithms. Thus, our proposed hybrid classifier provides a maximum performance over others. The future work of this paper involves the implementation of the hybrid classifier for different datasets and for different severities.

REFERENCES

[1] DeSantis C. E., Ma J., Gaudet M. M., Newman L. A., Miller K. D., Goding Sauer A., Jemal A. and Siegel R. L. 2019. Breast cancer statistics, 2019. CA: a cancer journal for clinicians. 69(6):.438-451.

[2] Northouse L. L., Templin T., Mood D. and Oberst M. 1998. Couples' adjustment to breast cancer and benign breast disease: A longitudinal analysis. Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer. 7(1): 37-48.

[3] Evans A., Whelehan P., Thomson K., McLean D., Brauer K., Purdie C., Jordan L., Baker L. and Thompson A. 2010. Quantitative shear wave ultrasound elastography: initial experience in solid breast masses. Breast cancer research. 12(6): R104.

[4] Chaurasia V., Pal S. and Tiwari B. B. 2018. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology. 12(2): 119-126.

[5] Abirami C., Harikumar R. and Chakravarthy S. S. 2016. March. Performance analysis and detection of micro calcification in digital mammograms using wavelet features. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 2327-2331). IEEE.

[6] Sannasi Chakravarthy S. R. and Rajaguru H. 2020. Detection and classification of microcalcification from digital mammograms with firefly algorithm, extreme learning machine and non-linear regression models: A comparison. International Journal of Imaging Systems and Technology. 30(1): 126-146.

[7] Blake C. L. and Merz C. J. 1998. UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences.

[8] Huang S., Cai N., Pacheco P. P., Narrandes S., Wang Y. and Xu W. 2018. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics-Proteomics. 15(1): 41-51.

[9] Biau G. and Scornet E. 2016. A random forest guided tour. Test. 25(2): 197-227.

[10] Wyner A. J., Olson M., Bleich J. and Mease D. 2017. Explaining the success of adaboost and random forests as interpolating classifiers. The Journal of Machine Learning Research. 18(1): 1558-1590.

[11] Chen C., Zhang G., Yang J. and Milton J.C. 2016. An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. Accident Analysis & Prevention. 90: 95-107.