Classifying Breast Cancer Types Based on Fine Needle Aspiration Biopsy Data Using Random Forest Classifier

Farzana Kabir Ahmad
Computational Intelligence Research Cluster,
School of Computing, College of Arts and Sciences
Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
e-mail: farzana58@uum.edu.my

Nooraini Yusoff
Computational Intelligence Research Cluster,
School of Computing, College of Arts and Sciences
Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
e-mail: nooraini@uum.edu.my

Abstract **-** Breast cancer is a complex and heterogeneous disease due to its diverse morphological features, as well as different clinical outcome. As a result, breast cancer patients may response to different therapeutic options. Currently, difficulties in recognizing the breast cancer types lead to inefficient treatments. Generally, there are two types of breast cancer, known as malignant and benign. Therefore it is necessary to devise a clinically meaningful classification of the disease that can accurately classify breast cancer tissues into relevant classes. This study aims to classify breast cancer lesions which have been obtained from fine needle aspiration (FNA) procedure using random forest. Random forest is a classifier built based on the combination of decision trees and has been identified to perform well in comparison to other machine learning techniques. This method has been tested on approximately 700 data, which consists of 458 instances from benign cases and 241 instances belong to malignant cases. The performance of proposed method is measured based on sensitivity, specificity and accuracy. The experimental results show that, random forest achieved sensitivity of 75%, specificity of 70% and accuracy about 72%. Thus, it can be concluded that random forest can accurately classify breast cancer types given a small number of features and it works as a promising tool to differentiate malignant from benign tumor at early stage.

Keywords: *Breast cancer, fine needle aspiration, random forest*

1.0 INTRODUCTION

Breast cancer is a heterogeneous disease that can be appeared in several clinical and histological forms. Due to this nature, most of breast cancer patients with same clinical and diagnostic profile may develop different clinical outcome [1]. As a result, breast cancer clinical progression is very difficult to predict using the existing prognostic factors. Furthermore, difficulties in recognizing the breast cancer types accuracy also lead to inefficient treatments. Generally, there are two types of breast cancer, known as malignant and benign. Benign tumors are non-cancerous but it is the areas that involve extreme cell growth although it can also happen at sluggish rate. This type of tumor does not invade to another part of body and usually is harmless. In contrast, malignant tumors are associated with cancer. The main characteristic of malignant tumors is it ability to metastasis. This means that malignant tumor experiences several mutations and can be spread throughout the body. Once they adhere to another part of body, malignant tumors have the ability to interfere in normal tasks by infecting healthy tissues and consequently transform the cells into cancerous one.

Currently, various kinds of techniques have been used by physicians in order to identify the type of breast cancer, which includes imaging technique such as mammogram and CT- scan. Mammogram has been employed as one of the standard screening method to diagnose breast cancer. Although this technique capable of visualizing breast masses smaller than 0.5 cm, it is prone to several errors. Patients who undergone mammograms test are mainly exposed to mutagenic effect of radiation. Moreover, interpreting mammograms result is a complex task as it may very similar to normal glandular tissues. Even though, tools for computer-aided diagnosis have been developed by using advanced method in image processing field that can permit an easy visualization of mammograms, the interpretation of mammograms results are basically rely on the experiences and expertise of radiologists. Even if the radiologists can differentiate the malignant and benign lesion, present computer aided diagnose suffers from high number of false positive detection. In this regard, tissues cell nuclei which is obtain from fine needle aspiration (FNA) biopsy procedure has become an alternative way to detect breast cancer lesion.

Fine needle aspiration (FNA) biopsy is a diagnostic procedure used to investigate the breast lesion. In this technique, a thin, hollow needle is inserted into the mass for sampling of cells that will be examined under a microscope as shown in Figure 1. Although this technique is safer and involve minor surgical procedure, the incompleteness and uncertainty of the information

contained in the histo-pathological image are the major problem. In addition, the imperfection of the data acquisition process in the form of noise, chromatic distortion and deformity of histo-pathological material caused by its preparation also has increased the problem complexity. Besides that, most of pathologist examines the abnormalities manually under the microscope, in which their experience may influence the examination results and lead to false positive cases. Furthermore, examining manually a large amount of FNA biopsy data is also a time consuming task. To accelerate the diagnosing procedure in ensuring better treatments can be prescribed to patients, various kinds of computational intelligence methods have been developed by researchers using different type of data.

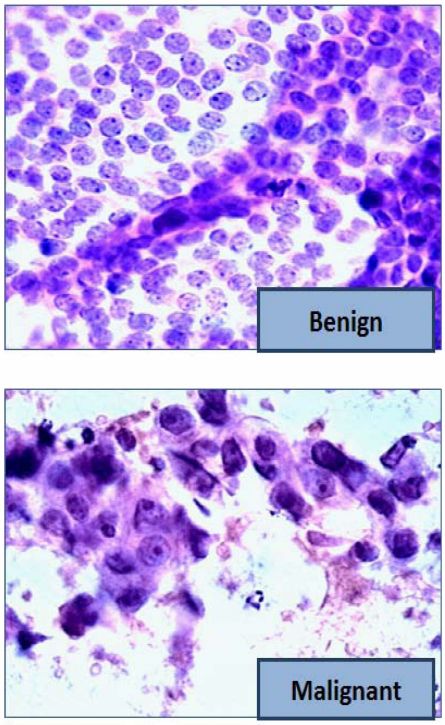

Figure 1: FNA results for benign and malignant tumor under the microscope

Given the growing rate of mortality due to breast cancer and the importance of accurate diagnosing assessment, a number of machine learning technique have been used over the past several years. The major types of algorithms include 1) artificial neural networks (ANN); 2) decision trees (DT); 3) genetic algorithms; 4) linear discriminant analysis (LDA) methods; 5) *k*-nearest neighbor algorithms [2]. Although the used of computational intelligence techniques are prominent in the medical domain, the success of machine learning algorithms are not always guaranteed. A good understanding of problem and limitation of data is important criterion to design successful algorithm. In addition, choosing a set of variables is also crucial to classify novel data. Therefore, this paper attempt to explore Random Forest (RF) in classifying different types of breast lesion based on FNA biopsy data.

The remainder of this paper is organized as follows. Section 2 described the RF algorithm and data used to classify breast cancer lesion. Section 3 on the other hand presents the results and discussion. A detail comparison with other techniques will also discuss in this section. Finally, Section 4 offers concluding and future direction remarks.

2.0 METHOD

This section is divided into two main parts; a) the description of Random Forest method and b) dataset that has been used.

2.1 Random Forest

The Random Forests classifier was introduced by Breiman in 1999. It is built based on the combination of decision trees $\{DT(x,\Theta_k)\}_{k=1}^{T}$, where x is the input vector and Θ_k denoted to random split independent vector with equal distribution of trees in the forest, $\Theta_1,\ldots,\Theta_{k-1}$. Meanwhile, T is the ensemble bootstrap sample drawn from training data. Each tree is built on a different bootstrap sample, consisting of N samples drawn at random, and with replacement from the N samples of the training set. At each node a number m of the total number of predictors M is chosen at random. The best split among these m is used to split the node. The value of m is held constant throughout the forest. The random forest algorithm is as follows (adapted from [3]):

For bootstrap, b = 1,…, T; create a bootstrap sample $\acute{L}_b$ by randomly building N samples with replacement from the N samples in the learning set Ĺ. At this point, the tree, T is build by using $\acute{L}_b$ as given below:

1. At node n, randomly sample m of the M predictor variables.
2. For each of the m sampled variables v_k, whereby k = 1,…, m find the best split s_k among all possible splits.
3. Then, select the best split s* among the k = 1,…,m splits s_k in order to split the node. This variable v_{best} is identified on which cut point c* is used to split the node.
4. Split all the data entries i = 1,…, n that is present in the parent node, by sending the observations with $vi_{best} < c^*$ to the left descendant node and all observations vibest ≥c* to the right descendant node.

5. Repeat steps 1- 4 on all descendant nodes to grow a maximally sized tree T_b.

Given C classes, the Gini criterion for each n node, G(n) is defined as shown below. This criterion is used to select the split with the lowest impurity at each node. For each tree in the forest, the predicted class for each observation is obtained. The class with the maximum number of votes among the T trees in the forest is the predicted class of an observation.

$$G(n) = 1 - \sum_{k=1}^{C} p^2(k \mid n) \qquad (1)$$

Where p2(k|n) for k = 1…C are the estimations of class probabilities after the node split. This ratio also represents the entries belong to class k to the total entries on one side of the split.

RF have been identified to have an outstanding performance in comparison to other machine learning techniques for example neural network (NN), support vector machine (SVM) and *k*-nearest neighbour [4-6]. In addition, most machine learning methods, are useful for classifying but do not give any insight as to what variables are most important with respect to the derived classifier. RF is also highly tolerated to noisy data, which make it appropriate for this study.

2.2 Description of Dataset

The FNA biopsy data which is used in this study has been obtained from public UCI Machine Learning Repository. Approximately 700 data has been used in this research, which has been collected from six cohorts of patients in 1989 to 1991. These data includes 458 instances from benign cases and 241 instances belong to malignant cases. For each patient ten features are measured to determine the type of lesion. Description of these features is listed as given in Table 1. Prior to the construction of RF model, the data is pre-processed by using *k*-NN to address the issues of missing values.

Table 1: Ten feature measure in the FNA biopsy data

No.	Features	Description
1.	Clump thickness	measurement of thickness of clustered mass tissues
2.	Uniformity of cell size	degree of consistent cell size
3.	Uniformity of cell shape	having one form of shape
4.	Marginal Adhesion	the stable joining of parts to one another, which may occur abnormally
5.	Single Epithelial Cell Size	number of layers present in epithelium
6.	Bare Nuclei	having sufficient nucleus
7.	Bland Chromatin	unperturbed genetic
8.	Normal Nucleoli	normal round granular body composed of protein and RNA in the nucleus of a cell
9.	Mitoses	the entire process of cell division including division of the nucleus and the cytoplasm.
10.	Class:	Class for benign or malignant

2.3 Evaluation Method

In order to evaluate the performance of RF model in classifying breast cancer types, this study has used receiver operating characteristics (ROC). ROC is a graph technique for visualizing, organizing and selecting classifiers based on their performance. ROC graphs are commonly used in medical decision making, and in recent years have been used increasingly in machine learning and data mining research. As a result, current years have seen an increase in the use of ROC graphs in the machine learning community, due in part to the realization that simple classification accuracy is often a poor metric for measuring performance. In addition to being a generally useful performance graphing method, they have properties that make them especially useful for domains with skewed class distribution and unequal classification error costs. These characteristics have become increasingly important as research continues into the areas of cost-sensitive learning and learning in the presence of unbalanced classes.

The performance of the RF classification algorithm is evaluated by computing the percentages of sensitivity, specificity and accuracy, the respective definitions are as follows:

$$\textit{Sensitivity} = TP/(TP+FN)*100 \qquad (1)$$
$$\textit{Specificity} = TN/(TN+TP)*100 \qquad (2)$$
$$\textit{Accuracy} = (TP+TN)/(TN+TP+FN+FP)*100 \qquad (3)$$

Where TP is the number of true positives, TN is the number of true negatives; FN is the number of false negatives, and FP is the number of false positives. Since this study aims to estimate the performance of RF classifier based on the classification of benign and malignant breast using FNA data, the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are defined appropriately as shown below:

FP: Predicts benign as malignant.
TP: Predicts malignant as malignant.
FN: Predicts malignant as benign
TN: Predicts benign as begin.

Specificity and sensitivity are another two terms that has been frequently used in ROC analysis. The ROC space is defined by false positive rate (FPR) and true positive rate (TPR) as *x* and *y* axes respectively as illustrated in Figure 2. It also represents and depicts related trade-off between specificity and sensitivity since TPR is equivalent with sensitivity and FPR indicates the value of 1-specificity. Figure 2 shows an example of ROC space, whereby the result of method *D* is significantly outperformed against other methods, with the point approximately reaching (0, 1), while result for method *E* is worst in comparison to others, namely *A, B,* and *C*.

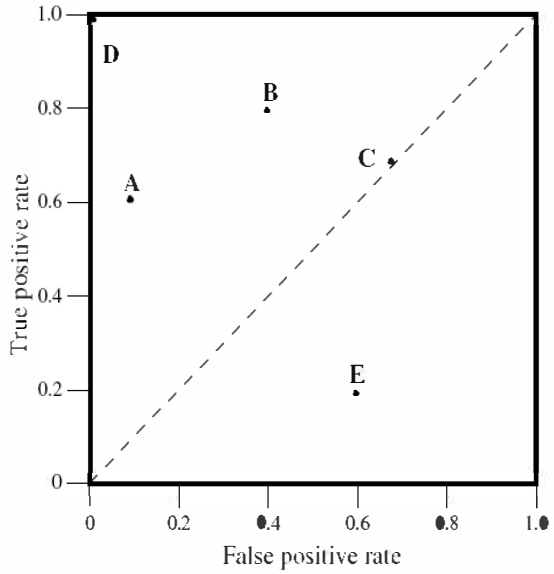

Figure 2: The ROC space: TPR (sensitivity) vs. FPR (1-specificity)

3.0 EXPERIMENTAL RESULTS AND DISCUSSION

To assess the effectiveness of RF in classifying breast cancer types, three experiments are conducted. The first experiment is focusing on variable importance by importing all features into RF; meanwhile the second empirical study aims in searching the most important features that denotes to different types of breast cancer. Finally this study examine the specificity and sensitivity of RF algorithm and reports the accuracy results for all instances that have been described in Section 2.2.

3.1 Features Importance Result

The features importance which has been obtained by the training samples using the RF algorithm is illustrated in Figure 3. This figure displayed the result for each feature when all features are used as input in the RF. In this study the feature importance is determined by the mean decrease permutation accuracy. The result shows that mitoses, bland chromatin, single epithelial cell size and uniformity of cell sizes appear to be the most relevant features. The result also indicates that genetic alterations play a significant factor in determining the presence of cancerous cells. This finding is align with result that have been reported by [7].

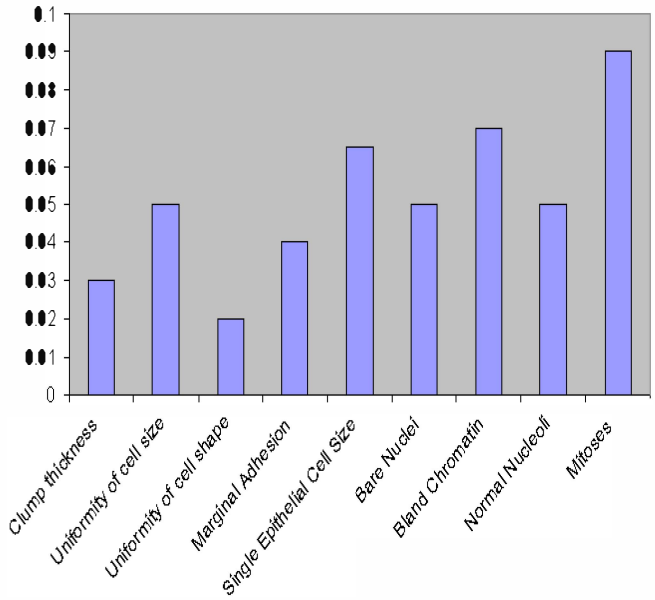

Figure 3: Feature importance based on RF algorithm

3.2 Specificity and Sensitivity Result

Specificity and sensitivity are two terms that have been widely used in accessing the performance of classifier in medical domain. For such reason, this performance metric has been used in this study. Sensitivity is the ability of a test that correctly classifies the patient that has a breast cancer (malignant) into respective class. On the other hand, specificity is the ability of a test that correctly determined cancerous free patient (benign group). The experiment has shows that RF can achieved sensitivity of 75% while obtained specificity of 70% as illustrated in Figure 4. The results demonstrate that the algorithm can easily classify patients with malignant cells but slightly underperform to differentiate between malignant and benign type. It is denotes to the fact that breast cancer is a

heterogeneous disease and therefore required various kind of data to be examined.

In addition, accuracy as been calculated to examine the performance of proposed method. The result for testing data is given in Table 2, with accuracy of 72%.

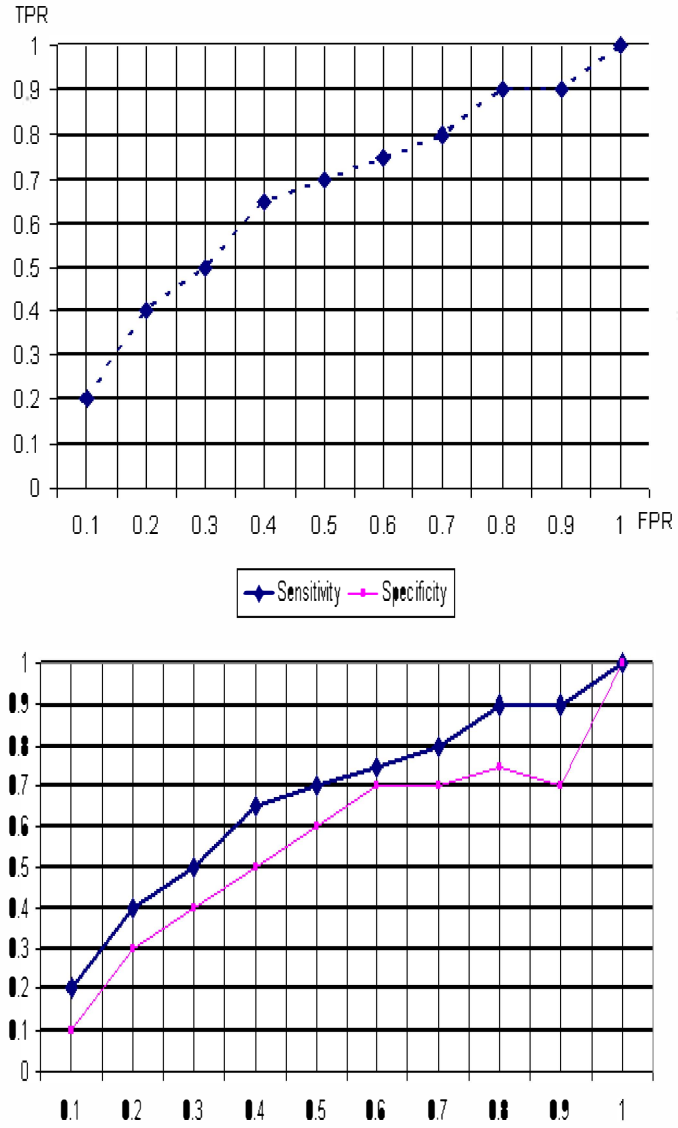

Figure 4: Specificity and sensitivity results

Table 2: The results after training the RF algorithm

No of testing cases	Sensitivity	Specificity	Accuracy
250	75%	70%	72%

4.0 CONCLUSION AND FUTURE REMARK

Breast cancer is a common disease among women worldwide. Generally there are two types of breast cancer known as malignant and benign tumor. Classification of these tumor types into relevant classes is important to help physician in prescribing the appropriate treatments. Therefore, this paper has presented a study on classifying breast cancer types using RF algorithm. The algorithm has been tested on a secondary FNA data that contained several features. Approximately 700 dataset has been used in this study and the performance of proposed method has been measured. The obtained accuracy of proposed method was 72%, whereas the sensitivity and specificity were found 75% and 70% respectively. It can be concluded that RF can accurately classify breast cancer tumors, however further research is required to analyzing and determine the best split at a node. Furthermore, for the purpose of future work, this study attempts to compare various types of methods in determining the accuracy and effectiveness of RF algorithm as well as verify the results that have been obtained.

5.0 REFERENCES

[1] P. Boyle and B. Levin, "World Cancer Report," World Health Organization, Geneva, Switzerland 2008.

[2] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Informatics*, vol. 2, pp. 59–78, 2006.

[3] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, pp. 2249 – 2260, 2008.

[4] L. Breiman, "Statistical modeling: the two cultures," *Statistical Science*, vol. 16, pp. 199–231, 2001.

[5] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, pp. 169–186, 2003.

[6] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern Recognition*, 2010.

[7] M. J. v. d. Vijver, "Genetic alterations in breast cancer," *Current Diagnostic Pathology*, vol. 6, pp. 271-281, 2000.