Using Five Machine Learning for Breast Cancer Biopsy Predictions Based on Mammographic Diagnosis

David Oyewola*‡; Danladi Hakimi*; Kayode Adeboye*, Musa Danjuma Shehu*

*Department of Mathematics, Federal University of Technology, Minna, Nigeria

(davidakaprof01@yahoo.com)

‡ Department of Mathematics, Federal University of Technology, Minna, Nigeria , davidakaprof01@yahoo.com

Abstract- Breast cancer is one of the causes of female death in the world. Mammography is commonly used for distinguishing malignant tumors from benign ones. In this research, a mammographic diagnostic method is presented for breast cancer biopsy outcome predictions using five machine learning which includes: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Random Forest (RF) and Support Vector Machine (SVM) classification. The testing results showed that SVM learning classification performs better than other with accuracy of 95.8% in diagnosing both malignant and benign breast cancer, a sensitivity of 98.4% in diagnosing disease, a specificity of 94.4%. Furthermore, an estimated area of the receiver operating characteristic (ROC) curve analysis for Support vector machine (SVM) was 99.9% for breast cancer outcome predictions, outperformed the diagnostic accuracies of Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Random Forest (RF) methods. Therefore, Support Vector Machine (SVM) learning classification with mammography can provide highly accurate and consistent diagnoses in distinguishing malignant and benign cases for breast cancer predictions.

Keywords Logistic regression, Linear discriminant analysis, Random forest, Quantitative discriminant analysis, Support vector machine, Breast cancer.

1. Introduction

Breast cancer is a malignant tumor that starts in the cells of the breast. A malignant tumor is a set of cancer cells that grow into surrounding tissues or spread to distance areas of the body. Breast cancer can occur in both men and women, although it is much more common in women. [1] .Breast cancer was rated second highest among women in the United States. Some women are at higher risk for breast cancer than others because of their personal or family medical history or because of certain changes in their genes [2].A patients using mammograms regularly can lower the risk of dying from breast cancer. Preventive Services Task Force in the United States recommends that average-risk women who are 50 to 74 years old should have a screening mammogram every two years**.** Average-risk women who are 40 to 49 years old should talk to their doctor about when to start and how often to get a screening mammogram.

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning have been used in cancer detection and diagnosis for a score [4-6]. Nowadays machine learning techniques are being used in a wide range of applications ranging from detecting and classifying tumors via X-ray and CRT images [7-8] to the classification of malignancies from proteomic and genomic (microarray) assays [9-10]. According to the latest PubMed statistics,more than 1500 papers have been published on the subject of machine learning and cancer. However, the vast majority of these papers are concerned with using machine learning methods to identify, classify, detect, or distinguish tumors and other malignancies. In other words machine learning has been used primarily as an aid to cancer diagnosis and detection [12]. Breast Cancer data can be useful to discover the genetic behaviour of tumors and to predict the outcome of some diseases. There are many techniques to predict and classify breast cancer pattern. This paper compares performance of five machine learning techniques classifiers.

2. Materials and Methods

In this study, the Wisconsin Breast Cancer Database an UCI Machine Learning Repository was analysed which was located in breast-cancer Wisconsin sub-directory, filenames root: breast-cancer-Wisconsin having 568 instances, 2 classes (malignant and benign), and 32 attributes (ID, diagnosis, 30 real-valued input features) (see Table 1). Our methodology involves use of machine learning techniques such as; Logistic regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Random Forest (RF) and Support Vector Machine (SVM).

Table 1. Wisconsin Diagnostic Breast Cancer Attributes

Number	Attributes
1	ID number
2	Diagnosis (M = malignant, B = benign)
3-32	ten real-valued features are computed for each cell nucleus:
a)	Radius (mean of distances from center to points on the perimeter)
b)	Texture (standard deviation of gray-scale values)
c)	Perimeter
d)	Area
e)	Smoothness (local variation in radius lengths)
f)	Compactness (perimeter^2 / area - 1.0)
g)	Concavity (severity of concave portions of the contour)
h)	Concave points (number of concave portions of the contour)
i)	Symmetry
j)	Fractal dimension ("coastline approximation" -1)

2.1. Logistic Regression (LR)

Logistic regression is a generalized linear model that can be binomial or multinomial. Binomial or binary logistic regression can have only two possible outcomes: for example, "chronic disease" vs. "non-existence of chronic disease". The outcome is usually coded as "0" or "1", as this leads to the most straightforward interpretation. If possible outcome is success then it is coded as "1" and the contrary outcome referred as a failure is coded as "0". Logistic regression is used to predict the odds of a case based on the values of the independent variables (predictors). The odds are the probability that a particular outcome occuring divided by the probability that it is not occuring.

2.2. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a technique developed by Roland Fisher. It can also be called Fisher Discriminant Analysis (FDA). The main objective of LDA is to separate samples of distinct groups. Essentially, it transforms data to a different space which optimally distinguishes classes which can be referred to as the "between class" and "within class".

2.3. Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA) is much like Linear discriminant analysis. QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution and plugging estimates for the parameters in order to perform prediction. QDA assumes that each class has its own covariance matrix which leads to the number of parameters increases significantly.

2.4. Random Forest (RF)

Random forest provide an improvement over bagged trees by way of small tweak that decorates the trees.

In Breiman's approach, each tree in the collection is formed by first selecting at random, at each node, a small group of input coordinates to split on and, secondly, by calculating the best split based on these features in the training set. The tree is grown using CART methodology (Breiman et al., 1984) to maximum size, without pruning. This subspace randomization scheme is blended with bagging (Breiman, 1996; Buhlmann and Yu, 2002; Buja and Stuetzle, 2006; Biau et al., 2010) to resample, with replacement, the training data set each time a new individual tree is grown.

2.5. Support Vector Machine (SVM)

Support vector machine (SVM) is a powerful machine learning technique for classification. SVM is becoming popular in pattern recognition in bioinformatics, cancer diagnosis, and more. SVM is a maximum margin classification algorithm rooted in both machine and statistical learning theory. It is the method for classifying both linear and non-linear data. Basically the method involves finding a hyper plane that separates the examples of different outcomes. Being primarily designed for two-class problems, it find a hyper plane with a maximum distance to the closest point of the two classes; such a hyper plane is called the optimal hyper plane. A set of instances that is closest to the optimal hyper plane is called a support vector.

In this study, logistic regression, linear discriminant analysis, quadratic discriminant analysis, random forest and support vector machine algorithm can be assessed by confusion matrix which is shown in Table 2 below. Confusion matrix provides a detailed layout which represents the performance of the two algorithm. The row of this matrix represents the predicted class instances while each of column of the matrix represents the actual class instances as shown below. This matrix is also used to show the correct and incorrect instances.

Table 2. Confusion Matrix

Predicted Class		Actual Class	
		Positive(P)	Negative(N)
	True(T)	True Positive(TP)	True Negative (TN)
	False(F)	False Positive (FP)	False Negative (FN)

True Positive(TP): This instance indicates benign samples that were classify as benign.
True Negative(TN): This instance indicates malignant samples that were classify as malignant.
False Positive(FP): This instance indicates benign samples that were classify as malignant.
False Negative: It indicates malignant samples that were classify as benign.

2.6. Performance Metrics

Performance metrics such as accuracy, sensitivity and specificity is the most widely used medicine and biology. The performance metrics are presented in Table 3.
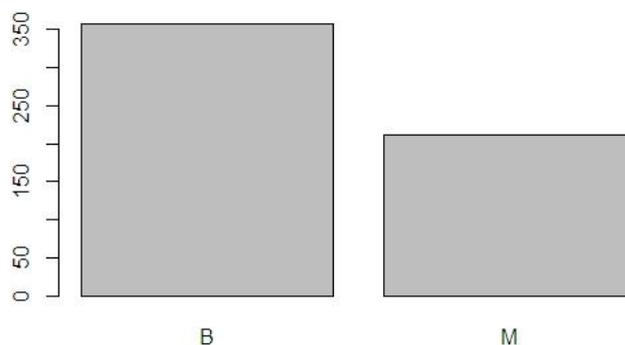
Table 3. Performance Metrics

Measure	Formula
Accuracy	$\dfrac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	$\dfrac{TP}{TP + FN}$
Specificity	$\dfrac{TN}{TN + FP}$

3. Result and Discussion

The experimental results of the breast cancer disease for prediction using Logistic regression, linear discriminant analysis, quadratic discriminant analysis, random forest and support vector machine are analysed in this section. The data related to breast cancer diseases are collected from 568 patients who are provided by National Cancer Institute.

In order to visually compare profiles from the two groups such as benign and malignant cancer patients. The Figure 1 below consists of patients that have benign cancer which is represented as B and malignant patients is represented by M as displayed below.

**Fig. 1.** Benign and Malignant cancer patients.

The correctly classified data for diagnosis of breast cancer has been observed and its accuracy is calculated for the five machine learning are shown in Table 4 above. After completing the training of the five machine learning classification model. Using 568 clinical instances of the mammographic mass dataset. From Table 4 above, the testing results shows that Support Vector Machine (SVM) in terms of accuracy performs better than other remaining four machine learning.

Figure 2 is called a Receiver Operating Characteristic curve (or ROC curve) it is a useful technique for organizing classifiers and visualizing their performance. ROC graphs are two-dimensional graphs in which true positive rate is plotted on the Y axis and false positive rate is plotted on the X axis. Figure 2 displays an ROC graph with five classifiers that were used in this paper which was plotted on the same

Table 4. Results for diagnosing of breast cancer

Techniques	LR	LDA	QDA	RF	SVM
TN	346	349	346	341	354
TP	191	182	180	194	190
FP	20	29	31	17	21
FN	11	8	11	16	3
TN+TP+FP+FN	568	568	568	568	568
TP+FN	202	190	191	210	193
TN+FP	366	378	377	358	375
Accuracy	94.5	93.5	92.6	94.2	95.8
Sensitivity	94.6	95.8	94.2	92.4	98.4
Specificity	94.5	92.3	91.8	95.3	94.4

graph. The diagonal line, from (0,0) to (1,1), is an indicative of an independent variable that discriminates no different from guessing (50/50 chance). ROC space is better than another if it is to the northwest (TP rate is higher, FP rate is lower, or both) of the first. From the Figure 2 above the perfect curve was obtained from SVM since it is closer to the northwest of True positive rate.
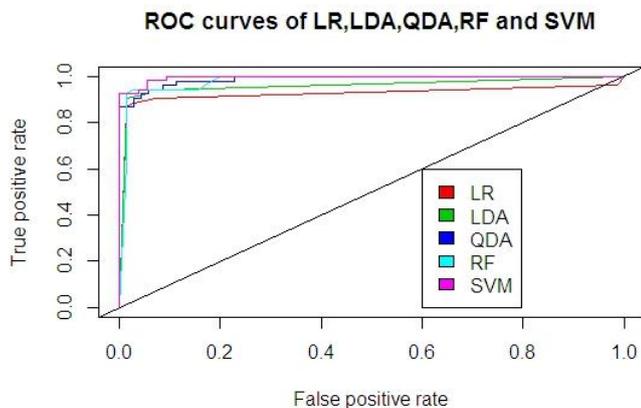
Fig. 2. ROC curves of LR, LDA, QDA, RF and SVM.

The AUC is a measure of the discriminality of a pair of classes. Table 5 above shows the AUC results obtained from ROC curve. From the table 5 above SVM has the highest predicted value.

Table 5. Area under the curve (AUC)

Techniques	Area Under the curve (AUC)(%)
SVM	99.9%
RF	98.07%
QDA	98.89%
LDA	96.06%
LR	92.51%

References

[1] Department of Health and Human Services Centers for Disease Control and Prevention, World Cancer Day, February 3, 2015.

[2] Department of Health and Human Services Centers for Disease Control and Prevention, United States Cancer Statistics, Technical Notes 2007.

[3] American Cancer Society, Cancer Facts & Figures 2016, Atlanta, Georgia, American Cancer Society, pp. 1–63, 2016.

[4] Simes RJ. Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer. J Chronic Dis, 38:171-86, 1985.

[5] Maclin PS, Dempsey J, Brooks J, et al. Using neural networks to diagnose cancer J Med Syst, 15:11-9, 1991.

[6] Cicchetti DV. Neural networks and diagnosis in the clinical laboratory: state of the art. Clin Chem, 38:9-10, 1992.

[7] Petricoin EF, Liotta LA. SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. Curr Opin Biotechnol, 15:24-30, 2004.

[8] Bocchi L, Coppini G, Nori J, Valli G. Detection of single and clustered micro calcifications in mammograms using fractals models and neural networks. Med Eng Phys, 26:303-12, 2004.

[9] Zhou X, Liu KY, Wong ST. Cancer classification and prediction using logistic regression with Bayesian gene selection. J Biomed Inform, 37:249-59, 2004.

[10] Dettling M. Bag Boosting for tumor classification with gene expression data. Bioinformatics, 20:3583-93, 2004.

[11] Wang JX, Zhang B, Yu JK, et al. Application of serum protein finger printing coupled with artificial neural network model in diagnosis of hepatocellular carcinoma. Chin Med J (Engl), 118:1278-84, 2005.

[12] McCarthy JF, Marx KA, Hoffman PE, et al. Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. Ann N Y Acad Sci, 1020:239-62, 2004.

[13] L. Breiman. Bagging predictors. Machine Learning, 24:123–140, 1996.

[14] P. Buhlmann and B. Yu. Analyzing bagging, The Annals of Statistics, 30:927–961, 2002.

[15] A. Buja and W. Stuetzle. Observations on bagging. Statistica Sinica, 16:323–352, 2006.

[16] G. Biau, F. Cerou, and A. Guyader. On the rate of convergence of the bagged nearest neighbor estimate. Journal of Machine Learning Research, 11:687–712, 2010.