Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis

Dana Bazazeh[1] and Raed Shubair [1,2]
[1]Electrical & Computer Engineering Department, Khalifa University, UAE
[2]Research Laboratory of Electronics, Massachusetts Institute of Technology, USA
Email: dana.bazazeh.ae@ieee.org; raed.shubair@kustar.ac.ae; rshubair@mit.edu

***Abstract*—Breast cancer is one of the most widespread diseases among women in the UAE and worldwide. Correct and early diagnosis is an extremely important step in rehabilitation and treatment. However, it is not an easy one due to several uncertainties in detection using mammograms. Machine Learning (ML) techniques can be used to develop tools for physicians that can be used as an effective mechanism for early detection and diagnosis of breast cancer which will greatly enhance the survival rate of patients. This paper compares three of the most popular ML techniques commonly used for breast cancer detection and diagnosis, namely Support Vector Machine (SVM), Random Forest (RF) and Bayesian Networks (BN). The Wisconsin original breast cancer data set was used as a training set to evaluate and compare the performance of the three ML classifiers in terms of key parameters such as accuracy, recall, precision and area of ROC. The results obtained in this paper provide an overview of the state of art ML techniques for breast cancer detection.**

I. Introduction

Cancer is a heterogeneous disease that can be divided into several types. According to [1], 25% of the females in the US are diagnosed with breasts cancer at some stage in their life. In UAE, 43% of female cancer patients are diagnosed with breast cancer [2]. Accurately predicting a cancerous tumor remains a challenging task for many physicians. The emergence of new medical technologies and the enormous amount of patient data have motivated the path for the development of new strategies in the prediction and detection of cancer. Although data assessment that is collected from the patient and a physicians intake greatly contributes to the diagnostic process, supportive tools could be added to help facilitate accurate diagnoses. These tools aim to eliminate possible diagnostic errors and provide a fast way for analyzing large chunks of data.

Machine Learning (ML), is a subfield of Artificial Intelligence (AI) that allows machines to learn without explicit programming by exposing them to sets of data allowing them to learn a specific task through experience [3] . Over the last few decades, ML methods have been widespread in the development of predictive models in order to support effective decision-making. In cancer research, these techniques could be used to identify different patterns in a data set and consequently predict whether a cancer is malignant or benign. The performance of such techniques can be evaluated based on the accuracy of the classification, recall, precision, and the area under the ROC [4].

In this paper, three popular ML techniquess applied to a breast cancer data set are investiagted and compared. These techniques are Random Forest (RF), Support Vector Machine (SVM), Bayesian Networks (BN). The rest of the paper is organized as follows: Section II explains the fundamental concept of the three ML methods being investigated. Section III describes the simulation setup that has been used for carrying out the comparative study. Simulation results and interpretations are provided in Section IV. Finally, conclusions are provided in Section V.

II. Machine Learning Techniques

The learning process in ML techniques can be divided into two main categories, supervised and unsupervised learning. In supervised learning, a set of data instances are used to train the machine and are labeled to give the correct result. However, in unsupervised learning, there are no pre-determined data sets and no notion of the expected outcome, which means that the goal is harder to achieve.

Classification is among the most common methods that goes under supervised learning. It uses historical labeled data to develop a model that is then used for future predictions. In the medical field, clinics and hospitals maintain large databases that contain records of patients with their symptoms and diagnosis. Therefore, researchers make use of this knowledge to develop classification models that can make inference based on historical cases. Medical inference has therefore become a much simpler task with machine-based support using the sheer amount of medical data that is available today. It is useful to note that all of the techniques used in this paper fall under classification models.

A. Support Vector Machine (SVM)

SVM is one of the supervised ML classification techniques that is widely applied in the field of cancer diagnosis and prognosis. SVM functions by selecting critical samples from all classes known as support vectors and separating the classes by generating a linear function that divides them as broadly as possible using these support vectors. Therefore, it can be said that a mapping between an input vector to a high dimensionality space is made using SVM that aims to find the most suitable hyperplane that divides the data set into classes [5]. This linear classifier aims to maximize the distance between the decision hyperplane and the nearest data point, which is called the marginal distance, by finding the best suited hyperplane [6].

978-1-5090-5306-3/16/$31.00 ©2016 IEEE

Fig 1 shows a scatter plot of two classes with two properties. A linear hyperplane is defined as $ax_1 + bx_2$ and the aim is to find a,b and c such that $ax_1 + bx_2 \leqslant c$ for class 1 and that $ax_1 + bx_2 > c$ for class 2 [5][7]. Different from other techniques, SVM depends on the support vectors, which are the data sets closest to the decision boundary, in their algorithms. This is because removing other data points that are further away from the decision hyperplane will not change the boundary as much as if the support vectors were removed.

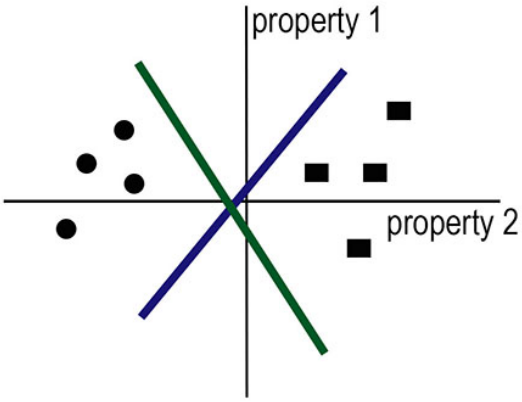

Fig. 1. SVM generated hyper-planes.

B. Random Forest (RF)

Similar to how a jury of people is used to make a court decision, RF brings together many decision trees to ensemble a forest of trees. The argument used is that having a single decision tree can either provide a simple model or a very specific one [4]. Using RF results in increased stability as compared to using single decision trees. This indicates that RF is insensitive to the noise of the input data set. One of the primary reasons behind using RF in cancer detection is its ability to handle data minorities. For example, a tumor can be classified as either benign or malignant, despite the latter class is only 10% of the input data set.

The RF method is based on a recursive approach in which every iteration involves picking one random sample of size N from the data set with replacement, and another random sample from the predictors without replacement. Then the data obtained is partitioned. The out-of-bag data is then dropped and the above steps repeated many times depending on how many trees are needed. Finally, a count is made over the trees that classify the observation in one category and in the other. Cases are then classified based on a majority vote over the decision trees [8].

C. Bayesian Networks (BN)

BN is a subfield of probabilistic graphical models that are used for prediction and knowledge representation of uncertain domains [9]. BN correspond to a widely used structure in machine learning called the directed acyclic graph (DAG). This graph consists of several nodes, each corresponding

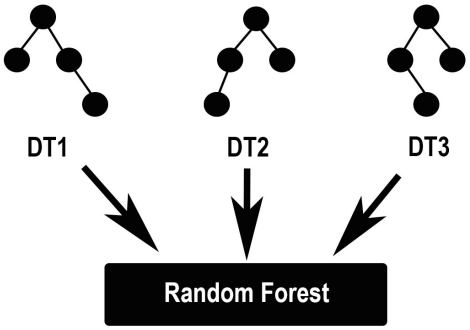

Fig. 2. A visual of how random forests works.

to a random variable and the node edges represents direct dependence among the corresponding nodes in the graph. An edge between X_1 and X_2 represents a direct dependence between these two nodes. Statistical methods are often used to obtain an estimate for these dependencies. Every variable must have a conditional probability table that shows its probability distribution knowing its direct antecedents. Also, all variables are conditionally independent of their non-descendants given their instant predecessors in the nodal frame. To compute the joint probability of the values $(x_1, x_2, , x_n)$ assigned to the network variables $(X_1, X_2, , X_N)$, the following equation is used.

$$P(x1, x2, , xn) = \prod_{j=1}^{n} P(xi|Parents(xi)) \qquad (1)$$

The parents of a node represent their direct predecessors in the network and the conditional probability is obtained from the probability table associated with each node.

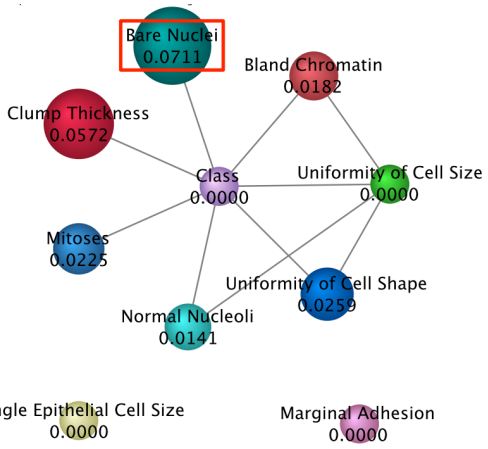

Fig. 3. DAG model combining breast cancer attributes [10]

III. Simulation Setup

A. Data Set

Our investigation is based on the Original Wisconsin Breast Cancer Data set that is obtained from the UCI Machine

Learning Repository, an online open source repository [11]. This data set was collected periodically over three years by Dr. William H. Wolberg from the University of Wisconsin Hospitals and consists of 669 instances, where the cases are classified as either malignant or benign. 458 of the cases are benign and 241 are malignant. The 10 attributes are:

- Clump Thickness
- Cell Size Uniformity
- Cell Shape Uniformity
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nuclei
- Mitoses
- Class

All of the above attributes except for the class are of numerical type, and their values range between 1 and 10. The class has a value of 2 for benign and 4 for malignant.

B. Training Set

The classifier will be tested using the $k - fold$ cross-validation method. This validation technique will randomly separate the training set into k subsets where 1 of the $k - 1$ subsets will be used for testing and the rest for training. 10-fold cross-validation is the preferred k value used in most validation in ML and will be used in this paper [12][14]. This means 9 subsets will be used for training of the classifier and the remaining 1 for the testing. This technique is used to avoid over fitting of the training set, which is likely to occur in small data sets and large number of attributes.

Fig. 4. 10-k cross validation method.

C. Simulation Software

In this paper, the Waikato Environment for Knowledge Analysis (WEKA) software was used as a ML tool. WEKA is a Java based open source tool that was first released to the public in 2006 under the GNU General Public License [13]. This tool provides several ML techniques and algorithms including the classification techniques that are being investigated in this paper. Other features include data preprocessing, clustering, feature selection evaluation and rule discovery algorithms. Data sets are accepted in several formats, such as CSV and ARFF. Beside being an open source tool, WEKA is also attractive due to its portability and ease of use GUI.

IV. Results and Discussion

This sections describes the parameters and presents the results that assists the three classifiers that are being investigated in this paper.

A. Accuracy

The classifier accuracy is a measure of how well the classifier can correctly predict cases into their correct category. It is the number of correct predictions divided by the total number of instances in the data set. It is worth noting that the accuracy is highly dependent on the threshold chosen by the classifier and can therefore change for different testing sets. Therefore, it is not the optimum method to compare different classifiers but may give an overview of the class. Hence, accuracy can be calculated using the following equation:

$$Accuracy = \left(\frac{K_{TP} + K_{TN}}{K_P + K_N}\right) \times 100\% \quad (2)$$

Where TP and TN represent the True Positive and True Negative values, respectively. Similarly P and N represent the Positive and Negative population of Malignant and Benign cases, respectively. The results show an accuracy rate of 97%.

B. Recall

Recall, also commonly known as sensitivity, is the rate of the positive observations that are correctly predicted as positive [12]. This measure is desirable, especially in the medical field because how many of the observations are correctly diagnosed. In this study, it is more important to correctly identify a malignant tumor than it is to incorrectly identify a benign one.

$$Recall = \left(\frac{K_{TP}}{K_P}\right) \times 100\% \quad (3)$$

The recall values for all three techniques are shown in Table I:

TABLE I
Recall Values.

	SVM	RF	BN
Benign	97.4%	96.9%	96.5%
Malignant	96.3%	95.9%	98.3%
Average	97.0%	96.6%	97.1%

C. Precision

Precision, also commonly known as confidence, is the rate of both true positives and true negatives that have been identified as true positives. This shows how well the classifier handles the positive observations but does not say much about the negative ones.

$$Precision = \left(\frac{K_{TP}}{K_{TP} + K_{TN}}\right) \times 100\% \quad (4)$$

The precision values for all three techniques are shown in Table II:

TABLE II
PRECISION VALUES.

	SVM	RF	BN
Benign	98.0%	97.8%	99.1%
Malignant	95.1%	94.3%	93.7%
Average	97.0%	96.6%	97.2%

D. ROC Area

A receiver operating characteristics (ROC) graph is a way to visualize a classifiers performance by showing the tradeoff between the cost and benefit of that classifier. ROC is one of the most common and useful performance measure for data mining techniques. This 2-D graph plots the TP rate (benefit) on the y-axis and the FP rate on the x-axis (cost) [15]. The domains of both axis stretch between 0 and 1. The graph is plotted by obtaining the TP and FP rates for every possible threshold value of the classifier. A point on the ROC is deemed more preferable if it is in the upper left quadrant of the plot where there is a high TP rate and low FP rate. A point on the diagonal line $y = x$ reflect a poor classifier that is based on random guessing, where the number of TPs and FPs are the same. The Area under a ROC graph reflects the performance of the classifier. This is obtained by dividing the area under the plot with the total area of the graph. Values that are closer to 1 show a higher performance of the classifier. Percentage values for the ROC area of all three techniques are shown in Table III:

TABLE III
AREA UNDER ROC VALUES

	SVM	RF	BN
Benign	96.4%	99.8%	99.0%
Malignant	96.8%	99.9%	99.2%
Average	96.6%	99.9%	99.1%

E. Discussion

The results presented in Tables I & II show that Bayesian Network (BN) has the best performance in terms of recall and precision. On the other hand, Table III shows that Random Forest (RF) technique has the optimum ROC performance when compared to the two other techniques. This implies that RF has a higher chance of discriminating between malignant and benign cases.

V. CONCLUSION

ML techniques have been widely used in the medical field and have served as a useful diagnostic tool that helps physicians in analyzing the available data as well as designing medical expert systems. This paper presented three of the most popular ML techniques commonly used for breast cancer detection and diagnosis, namely Support Vector Machine (SVM), Random Forest (RF) and Bayesian Networks (BN). The main features and methadolgy of each of the three ML techniques was described. Performance comparison of the investigated techniques has been carried out using the Original Wisconsin Breast Cancer Data set.

Simulation results obtained has proved that classification performance varies based on the method that is selected. Results have showed that SVMs have the highest performance in terms of accuracy, specificity and precision. However, RFs have the highest probability of correctly classifying tumor.

REFERENCES

[1] WHO — Breast Cancer: Prevention and Control (2015) Retrieved 20 Jan 2015, from WHO — World Health Organization. http://www.who.int/cancer/detection/breastcancer/en/index1.html

[2] Y. Elobaid, T.-C. Aw, J. N. W. Lim, S. Hamid, and M. Grivna, "Breast cancer presentation delays among Arab and national women in the UAE, a qualitative study," *SSM - Popul. Heal.*, Mar. 2016.

[3] E. D. Michie, D. J. Spiegelhalter, and C. C. Taylor, "Machine Learning , Neural and Statistical Classification," *Proceeding*, 1994.

[4] I. Kononenko, "Machine learning for medical diagnosis: history , state of the art and perspective," vol. 23, 2001.

[5] G. Williams, "Descriptive and Predictive Analytics", *Data Min. with Ratt. R Art Excav. Data Knowl. Discov. Use R*, pp. 193-203, 2011.

[6] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol.* J., vol. 13, pp. 8-17, 2015.

[7] T. J. Cleophas and A. H. Zwinderman, "Machine Learning in Medicine," pp. 1-271, 2013.

[8] Y. Yasui and X. Wang, *Statistical Learning from a Regression Perspective by BERK, R. A.*, vol. 65, no. 4. 2009.

[9] B. Networks, F. Faltin, and R. Kenett, "Bayesian Networks," *Encycl. Stat. Qual. Reliab.*, vol. 1, no. 1, p. 4, 2007.

[10] S. Conrady and L. Jouffe, Bayesian Node Analysis. 2013.

[11] M. Lichman, UCI Machine Learning Repositry, 2013. [Online]. Available: https://archive.ics.uci.edu/.

[12] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Stat. Comput.*, vol. 21, no. 2, pp. 137-146, 2011.

[13] K. J. Edwards and M. M. Gaber, *Astronomy and Big Data.*2014

[14] D. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37-63, 2011.

[15] L. F. Carvalho, G. Fernandes, M. V. O. De Assis, J. J. P. C. Rodrigues, and M. Lemes Proença, "Digital signature of network segment for healthcare environments support," *Irbm*, vol. 35, no. 6, pp. 299-309, 2014.