SURVEY ON ANALYSIS OF BREAST CANCER PREDICTION

Lingaraj N[1], Krishna Kumar S[1], Banu Priya K[1], R.M. Shiny[2]

[1]*UG Student, Department of Computer Science and Engineering, Agni College of Technology, Tamil Nadu, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Agni College of Technology, Tamil Nadu, India*

---***---

Abstract - *Breast cancer is one of the major threats in middle aged women throughout the world. In today's world, this is the second most threatening cause of death in women. Early diagnosis can significantly reduce the chances of death. But it is not an easy due to several uncertainties in detection. Machine Learning techniques are used to develop a tool which is used as an effective mechanism for early detection and diagnosis of breast cancer. Early diagnosis is helpful for physicians which will greatly enhance the survival rate of cancer patients. This paper compares three of the most popular ML techniques which are used for breast cancer detection and diagnosis. The techniques are Support Vector Machine (SVM), Random Forest (RF) and Naive Bayes (NB). The probability will be calculated for each of these techniques and the algorithm with highest probability provides the more accurate results.*

Key Words: Breast Cancer, Machine Learning, Detection, Diagnosis, Prediction, Analysis.

1. INTRODUCTION

Cancer is a heterogeneous disease that may be divided into many types. According to World Health Organization [1], twenty five percent of the females are diagnosed with breasts cancer at some stage in their life. In UAE[2], forty third of feminine cancer patients are diagnosed with breast cancer. Accurately predicting a cancerous growth remains a difficult task for several physicians. The emergence of latest medical technologies and therefore the monumental quantity of patient information have impelled the trail for the event of latest methods within the prediction and detection of cancer. Recurrent breast cancer is the one which comes back in the same or opposite breast or chest wall after a time period when the cancer couldn't be detected. Though information assessment that is collected from the patient and a physicians intake greatly contributes to the diagnostic method, supportive tools could be superimposed to assist facilitate proper diagnoses. These tools aim to eliminate possible diagnostic errors and supply a quick method for analyzing the large chunks of Data.

1.1 Machine Learning:

Machine Learning (ML), is a subfield of Artificial Intelligence (AI) that permits machines to learn without explicit programming by exposing them to sets of information permitting them to learn a selected task through expertise. Over the previous few decades, machine learning strategies have been widespread within the development of predictive models so as to support effective decision-making. In cancer analysis, these techniques could be used to determine completely different patterns during a information set and consequently predict whether or not a cancer is malignant or benign. The performance of such techniques will be evaluated supported the accuracy of the classification, recall, precision, and therefore the space underneath ROC.

1.2 Data Preprocessing

Recently, data processing has become a well-liked economical tool for data discovery and extracting hidden patterns from massive datasets. It involves the employment of refined knowledge manipulation tools to get antecedently unknown, valid patterns and relationships in massive dataset. We tend to apply 3 robust data processing classification algorithms i.e. SVM, Random forest and Naive Bayes, a medium sized knowledge set that contained thirty five attributes and 198 cancer patient data.

2. MACHINE LEARNING TECHNIQUES

The learning method in ML techniques are often divided into 2 main categories, supervised and unsupervised learning. In supervised learning, a set of data instances are used to train the machine and are labeled to grant the right result. However, in unsupervised learning, there are no pre-determined knowledge sets and no notion of the expected outcome, which suggests that the goal is tougher to achieve.

2.1. Support Vector Machine:

Support Vector Machine is one of the supervised machine learning classification techniques that is widely applied in the field of cancer identification and prognosis. SVM is functioned by choosing the critical samples from all categories. These samples are called support vectors. These classes are separated by generating a linear function that divides them broadly as possible using these support vectors.

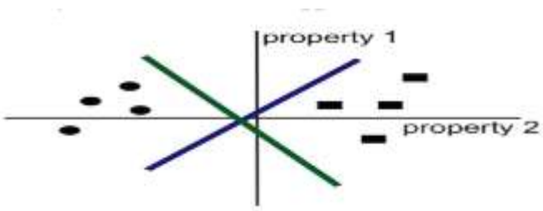


Fig -1: SVM generated hyper-planes

2.2. Random Forest:

RF brings along many decision trees to ensemble a forest of trees. The RF methodology relies on an algorithmic approach within which each iteration involves choosing 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 other data is then dropped. These steps are repeated certain times depending on how many trees are needed. Finally, a count is made over the trees that classify the observation in one category. Cases are then classified based on a majority vote over the decision tree.

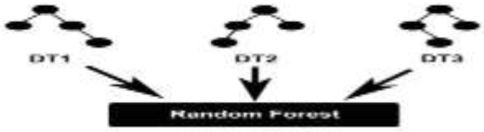


Fig -2: Working of Random Forest

2.3. Naive Bayes:

Naive Bayes is a subfield of probabilistic graphical models which is used for prediction [3] and representation of knowledge in uncertain domains. Na�e Bayes corresponds to a widely used structure in machine learning known as the directed acyclic graph (DAG). This graph consists of many nodes, each equivalent to a variable and the node edges represents direct dependence among the corresponding nodes in the graph.

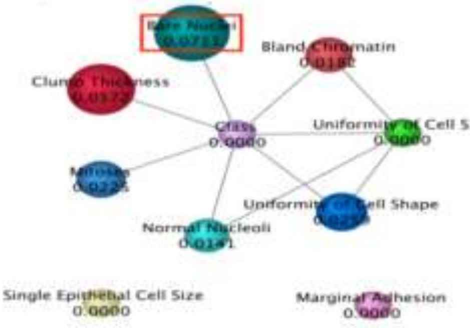


Fig -3: Breast cancer attributes – DAG Model

3. SIMULATION SETUP:

3.1. Data Set:

Our investigation is based on the initial Wisconsin breast cancer data set that's obtained from the UCI Machine Learning Repository [5] , an online open source repository. This data set was collected over 3 years from the University of Wisconsin Hospitals by Dr. William H. Wolberg and it consists of 669 instances, where the classification results are either malignant or benign. There were 458 benign cases and 241 was malignant. The ten attributes used are: • Clump Thickness • Cell Size Uniformity • Cell shape Uniformity • Marginal Adhesion • Single epithelial cell Size • Bare Nuclei • Bland chromatin • normal Nuclei • Mitoses • class.

3.2. Training set:

The classifiers are tested using the k – fold cross validation methodology. This validation technique can randomly separate the training set into k subsets where one of the k-1 subsets are used for testing and the rest for training. 10-fold cross-validation is the preferred k value utilized in most validation in ML and will be used in this paper. This implies nine subsets will be used for training of the classifier and the remaining one for the testing. This system is used to avoid over fitting of the training set, which is likely to occur in small data sets and large number of attributes.

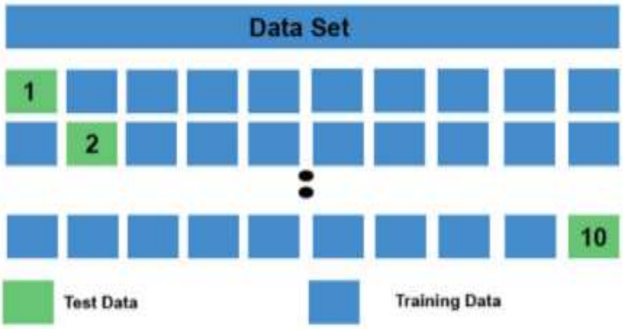


Fig -3: 10- k cross validation method

3.3. Simulation software:

Here, WEKA - Waikato Environment for Knowledge Analysis [4] software is used as a Machine Learning tool. WEKA is a Java based open supply tool that was initially released to the public in 2006 under the GNU General Public License. This tool provides many ML techniques and algorithms as well as the classification techniques that are being investigated in this paper. Alternative features include data preprocessing, clustering, feature selection evaluation and rule discovery algorithms. Datasets are accepted in many formats, like CSV and ARFF. Besides being an open source tool, WEKA is additionally attractive due to its portability and ease of use GUI.

4. RESULT:

This section describes the parameters and presents the results that assist the 3 classifiers that are being investigated in this paper.

4.1. Accuracy:

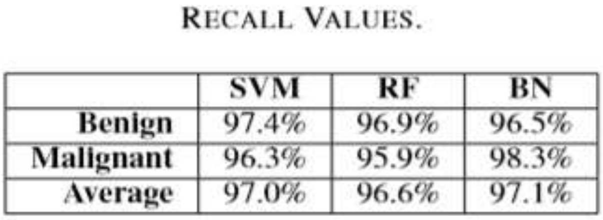
The classifier accuracy is a measure of how well the classifier will properly predict cases into their correct category. It is the number of correct predictions divided by the total number of instances within the data set. It is worth noting that the accuracy is highly dependent on the threshold chosen by the classifier and may therefore change for various testing sets. Hence, accuracy may be calculated using the subsequent equation:

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

4.2. Recall:

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

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



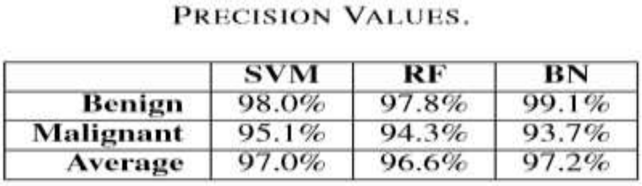
RECALL VALUES.

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

Table-1: Recall values comparison

4.3. Precision:

Precision, also commonly called confidence, is the rate of both true positives and true negatives that are identified as true positives. This shows how well the classifier handles the positive observations but doesn't say much regarding the negative ones. The precision values for all 3 techniques are shown in Table-2:



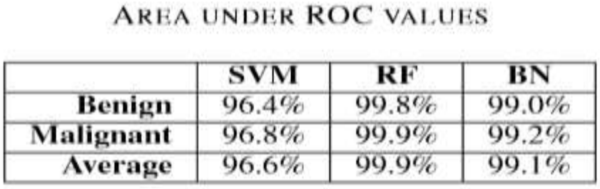
PRECISION VALUES.

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

Table-2: Precision values comparison

4.4. ROC area:

A receiver operating characteristics (ROC) graph is a way to visualize a classifiers performance by showing the trade-off between the price and advantage of that classifier. ROC is one amongst the foremost common and useful performance measure for data processing techniques. Percentage values for the roc area of all 3 techniques are shown in Table-3:



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%

Table-3: ROC values

5. CONCLUSION

ML techniques are 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 says about three of the most common ML techniques which are normally used for breast cancer detection and diagnosis. They are Support Vector Machine (SVM), Random Forest (RF) and Na�e Bayes (NB). The main features and methodology of each of the 3 ML techniques was represented. Performance comparison of the investigated techniques [6] has been applied using the original Wisconsin breast cancer data set. Simulation results obtained has proved that classification performance varies based on the strategy that is selected. Results show that SVM have the highest performance in terms of accuracy, specificity and precision. But, RF has the highest probability of properly classifying tumor.

REFERENCES

1. World Health Organization - Breast Cancer: Prevention and Control, 20 Jan 2015. http:// www.who.int/cancer/detection/breastcancer/en/inde x1.html.

2. “Breast cancer presentation delays among Arab and national women in the UAE, a qualitative study,” Y. Elobaid, T.-C. Aw, J. N. W. Lim, S. Hamid in Mar. 2016.

3. Bayesian Node Analysis, S. Conrady and L. Jouffe, 2013.

4. Weka 3.5.6, an open source data mining software tool developed at university of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/weka/ 2009.

5. "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Wolberg, William H., and Olvi L. Mangasarian.

6. "Analysis of feature selection algorithms on classification: a survey", Vanaja, S., and K. Ramesh Kumar. International Journal of Computer Applications (2014).