March - April 2020
ISSN: 0193-4120 Page No. 9293 - 9300

Diagnosis of Breast Cancer Using Supervised Machine Learning Techniques

Dr. S. Mohana[1], Dr. S. A. Sahaaya Arul Mary[2]
[1]Associate Professor, [2]Professor and Head
[1,2]Department Of Computer Science and Engineering ,Saranathan College of Engineering,TamilNadu, India
mohana-cse@saranathan.ac.in[1], mary-cse@saranathan.ac.in[2]

Article Info
Volume 83
Page Number: 9293 - 9300
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 09 April 2020*

Abstract

Breast cancer is a kind of disease which is most common among women. Next to lung cancer it is the second foremostcause of cancer death in women. In order to increase the survival rate of patient who are suffering from breast cancer, a prediction of breast cancer recurrence required. With the help of machine learning techniques and advanced technologies, a cancer can be diagnosed, accuracy can be detected and we can improve the performance.Machine Learning is a statistical model and it is an application of artificial intelligence (AI) .Artificial Intelligence makes the system the to learn automatically and learn from experience without any explicit program. This paper provides comparison of most popular machine learning techniques that are used for breast cancer detection and diagnosis on Wisconsin Breast Cancer Dataset(WBCD). Comparison is performed on both classification and regression categories of Supervised learning-Support Vector Machine, Random Forest, Decision Tree, Multilayer perception, Linear regression. The result shows Support Vector Machine provides high accuracy under Classification algorithm whereas Multilayer Perception regressor gives less error under Regression algorithm.

Keywords; *Machine Learning; Classification; Regression; Support Vector Machine; Decision Tree; Multilayer Perception; linear regression; Logistic Regression- Breast cancer; Wisconsin Dataset*

I. INTRODUCTION

Breast cancer is a type of disease after skin cancer, and is the most common cancer in women. It may be diagnosed in children, too. Breast cancer occurs as breast tissue cells mutate, and continue to reproduce. In these breast tissue, the abnormal cells typically grow together to form a tumour. Tumor is a cancer when the abnormal cells that are present in the tissue enter other parts of the breast or spread through the lymphatic system to other areas of the body. Breast cancer that is created in the lobules or tube-shaped ducts carrying milk to the nipple. Breast cancer is caused in the breast cancer cells by genetic mutation in DNA[7].

In the year 2018, 266,120 women with invasive breast cancer were identified in the United States and 63,960 people were diagnosed with situ breast cancer. The prevalence of breast cancer in India is comparatively lower, but continually growing. Indian women are exposed to low rates of life with breast cancer with a prevalence of disease of 66.1 per cent between 2010 and 2014. At least 17,97,900 women in India may have breast cancer by 2020, the estimate suggests. It is important to identify breast cancer at an early stage in order to provide the appropriate treatment. Diseases can be predicted by translating the information pointing to the disease from the observed data. A comparative study of Machine Learning approaches to increase the predictive rate of breast cancer is conducted in this article.

Machine learning involves multiple ways of learning

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 9293 - 9300

that they are supervised, unsupervised, semi-supervised and validated. They each have different approaches but all adopt the same basic mechanism and philosophy. In medical science huge quantities of data are generated everyday from doctors, hospitals, research and development (R&D). In medical science, the use of machine learning includes drug discovery / manufacturing, identification / diagnosis of cancer, radiology and radiotherapy, etc. Machine learning is growing rapidly in the medical domain thanks to an effective approach in prediction and classification[1].

Predicting breast cancer will be applied in this paper using Supervised Machine learning methods. This paper uses machine learning classifiers and regressors to predict breast cancer in terms of accuracy, accuracy and recall, mae, rmse and r2 scores.

II. LITERATURE SURVEY

Machine Learning plays a vital role as a tool for classification in medical and many emerging applications. In this paper ,various important work which has already made its mark in this field of study is discussedTo identify the respiratory discomforts using pulmonary acoustic signals ,the authors[7] have employed kNN machine learning algorithm and Support Vector Machine (SVM). Also kNN classifier is proved to yield greater results when compared with Support Vector Machine.ThekNN accuracy was found out to be 98.26% as compared to 92.19% of SVM. In [5], the authors , made thorough study on diabetes disease diagnosis using neural networks on a pima-diabetes dataset. It is proved that multilayer neural networks with Levenberg–Marquardt (LM) algorithm provide better performance compared to neural network based classifiers.

A comparative study is made[6] regarding Parkinson's Disease diagnosis on 197 data collected from patients suffered from Parkinson's disease.Almost 22 features were compared on DMNeural, Neural Network, Regression, and Decision tree classification models and neural networks proved to provide 92.9% classification accuracy compared to other algorithm

In [2], To perform disease detection the authors made use of many algorithm like Decision trees J48 , Support vector machines,NaiveBayesand Radial Basis Function (RBF) kernel, Radial basis neural networks, and simple CART and it is proven that SVM RBF kernel method performed well than other classifier techniques.

A survey conducted in 2018 by National Institute of Cancer Prevention and Research in the name of India against cancer says that breast cancer accounts for 14% of all cancers in women. Globocan data 2018 says that newly registered breast cancer patients are 1,62,468 and the deaths is of 87,090.

Registry	Total	#	%	R
Mumbai	18528	5620	30.3	1
Bangalore	13125	2052	15.6	2
Chennai	17499	3921	22.4	2
Thi'puram	18809	5354	28.5	1
Dibrugarh	2276	336	14.8	1
Guwahati	4679	674	14.4	2
Chandigarh	2092	341	16.3	2

Table 1Female Breast cancer (Icd-10:C50)

The above table has relative proportion of breast cancer in females varied from 14.4% in Guwahati to 30.3% in Mumbai. In this table, Number is represented in '#' symbol, Relative proportion(%), Rank(R). In India, around 1 in 28 women is likely to develop breast cancer in her lifetime. Fig:1 Graphical representation of Breast cancer Statistics

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 9293 - 9300

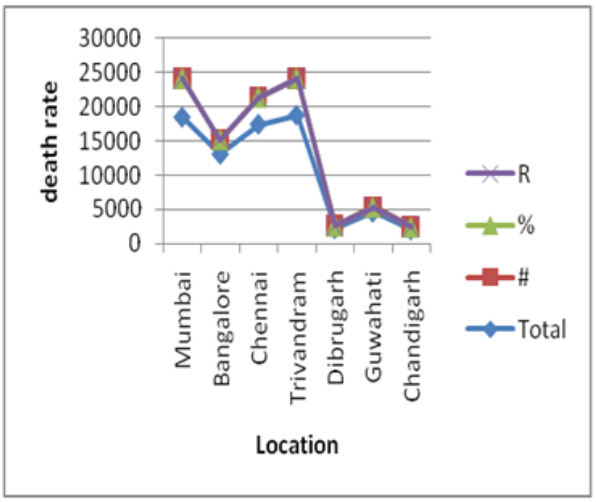


Fig:1 Graphical representation of Breast cancer Statistics

The survey that was conducted by Indian Council ofMedical Research during 1982 to 2005 in metropolitan cities says that the incidence of breast cancer has doubled from earlier survey.

III. METHODS

For forecasting the recurrence of breast cancer, a comparative study of commonly used supervised machine learning methods is conducted on the Wisconsin Breast Cancer Database (WBCD). Supervised learning is a learning process in which we train the computer using data that is well defined (i.e) Data has already been labelled with the correct answer. After knowing the data, algorithms learn from labeled data, computer is presented with new sample set, then algorithms decide the label should be assigned to the data based on the pattern and compare the patterns with unlabeled data. Supervised learning can be categorized into two types : Classification and Regression algorithms.

Classification model aims to reflect any finding from observed values. Either they forecast categorical class labels and values in the attributes given one or more inputs to the classification model, and then use it to classify new data. Other type of supervised learning is regression.

The key distinction between regression and classification is that the regression produces a number rather than a class. Regression is therefore helpful in forecasting number-based problems such as temperature for a given day, stock market values, etc. Unsupervised learning is teaching the system with knowledge that is neither marked nor labelled only input data is included in the examples, but complex patterns can be found hidden within the data without any labeling.

Methods used in this analysis involve:

- Support Vector machine
- Random Forest
- Decision Tree
- Logistic Regression
- Linear Regression
- Multilayer Perception

Support Vector machine:

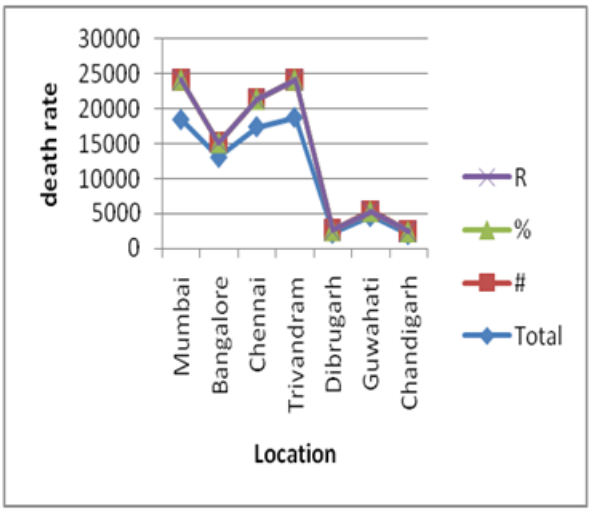
Support Vector Machine is a supervised model of learning which analyzes data used for the analysis of classification and regression. It is a discriminative classifier, identified by hyperplane separation. Each data object is plotted in this algorithm as a point in n-dimensional form (where n is the number of features we have) with the value taken as each function is the value of the particular coordinate. SVM builds a hyperplane or group of hyperplans that are used for grouping, regression and identification of outliers. Hyperplane has the largest distance to any class' neighboring training data points (functional margins), and typically the margin reduces the classifier's generalization error. It also supports regression of vectors close to grouping, but has some small differences[1].

SVR specified the loss function within a certain distance of true value. This form of function is known as the intensive-loss epsilon function.

Generally, hyper plane is defined as ax_1+bx_2we have to find a,b and c such that $ax_1+bx_2 \leq c$ for class1 and $ax_1+bx_2<c$ for class2.

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 9293 - 9300



Decision Tree:

Decision tree is the most popular and powerful classification and prediction tool. It follows a hoerarchical structure in which each internal node represents s test on attribute, each branch represents the test result and each leaf node represents a class name. The path from root node to leaf node follows classification rules. Decision tree creation is achieved by separating the source set into subsets depending on the value check of the attributes. On each derived subset this process is repeated recursively also known as recursive partitioning. Decision tree can handle data of large dimensions and gives reasonable precision[9].

Random Forest:

Random forest is a supervised learning methodology that creates multiple tree decisions and blends them to produce more accurate results and consistent prediction. Most of the decision trees time group are educated using "bagging" method. Random forest's big advantage is that it can be used for problems of grouping and regression. In Random Forest instead of looking for most important features when splitting a node, the dataset searches for the best feature among a random subset of features. One can accurately calculate the relative output of each function of it. It prevents overfitting by building a random subset of features and using these subsets to create small trees[4].

Linear Regression:

Linear regression is a paradigm for exploring the relationship between linear regression There are two different types. The model is called simple linear regression if it uses one independent variable and the method is called multiple linear regression for more than one independent variable. The main objective of linear regression is to achieve a line that best fits the data where the best line of fit is line where the cumulative errors in estimation are as small as possible. Error is the distance from that point to the line of regression[18].

Logistic Regression:

Logistic Regression is a widely used statistical model, it uses a logistic function or logistic curve to model a binary dependent variable. In regression analysis, it is a form of binomial regression. It is used when the dependent variable is categorical i.e binary or dichotomous. For instance, to predict whether an email is spam(1) or not (0). The main objective of logistic regression is to find the best fitting model that describes the relationship between dependent variable and a set of independent (predictors) variables. This algorithm has three different types they are binomial or binary logistic regression, multinomial logistic regression and ordinal logistic regression.

$$Hypothesis => Z = WX + B$$

$$h(\theta) = sigmoid(Z)$$

The estimated probability is the output value of hypothesis which infers how the predicted value is equivalent to actual value when an input X is given[13].

Multilayer Perception:

Multilayer perception is a network of artificial neurons which are usually interconnected in a feed-forward way whereas feed-forward network is type of artificial neural network where connection between the nodes do not form any cycle. Neurons present in each layer has directed connection to the

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 9293 - 9300

neurons of subsequent layer. MLP uses different learning techniques, one of the most popular technique is back propagation where the output values are compared to the actual values in order to predict the pre-defined error function. It consist of three or more layer an input layer, an output layer with one or more hidden layer.

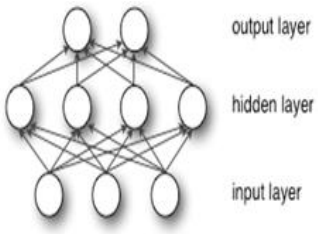


Fig:2 MLP with Single Hidden Layer

MLP function of hidden layer is f: RD□RL , where D is the size of input vectorx and L is the size of output vector f(x). It palys a vital role in diverse fields such a speech recognition, image recognition and machine translation software but have strong competition with support vector machine[20].

IV. EXPERIMENTAL DESIGN

Main aim of this paper is to predict the performance of various Machine Learning techniques. For classfication algorithms the performance is measured in terms of Accuracy, Precision, Recall, f1-score and for regression performance is measured with the help of MAE, RMSE and R2. Each term is defined as follows,

Precision: Precision is the ratio between correct predictiona and total predictions.It is also called positive predictive model.

$$\text{Precision}= \frac{True\ positive}{True\ positive+False\ positive}$$

Recall: Ratio of correct predictions and the total number of correct items in the set. It is expressed as % of total positive items correctly predicted by the model.

$$\text{Recall}=\frac{True\ positive}{True\ positive+False\ Negative}$$

Accuracy: Accuracy is the ratio of number of predictions to the total number of input instances. It works well if each class has equal number of samples.

F1-score: F1-score is also called F-measure or F-score which is a measure of test accuracy. It takes the value of both precision and recall of test set to compute score.

$$\text{F1-Score}=2\frac{precision*recall}{precision+recall}$$

MAE: Mean Absolute Error is an average of difference between total values and predictions. It is used to penalize huge errors than MSE also it is not much sensitive to outliers.

RMSE: Root Mean Square Error is a square root of mean square error, mean square error calculates the difference between predictions and target then the values is squared and then average for those value will be taken. RMSE is used to scale the errors to be same as the scale of targets.

R2: R-Square is almost similar to MSE but it is scale free, if R2 is negative then the model is worse than predicting mean. It always ranges between -∞ to 1.

V. DATASET

Wisconsin Breast Cancer Dataset(WBCD) is used for this research work, this dataset is collected from reputed UCI Repository which is a collection of database and data generators used for analysis of machine learning algorithms. The dataset contains 569 instances and 32 attributes. There is no missing values present in the dataset.

Features of the dataset are determined from a digitized image of Fine Needle Aspirant of breast lump. Fine Needle Aspirant (FNA) is a biopsy procedure , a thin needle is inserted into lump for sampling of cells which helps us to make diagnosis of cancer. The dataset contains 2 classes one is malignant and other is benign and class distribution are 212 malignant and 357 benign.

Dataset contain Ten real valued features that are

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 9293 - 9300

computed for each cell nucleus:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area - 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension ("coastline approximation" -1)

The Mean, Standard Error and Worst values are computed for these 10 features, which results in 30 features and the remaining two attributes are "id" and " Diagnosis (M = malignant, B = benign)".

After importing the packages into the working environment the dataset is separated into training and test set in the ratio of 8:2

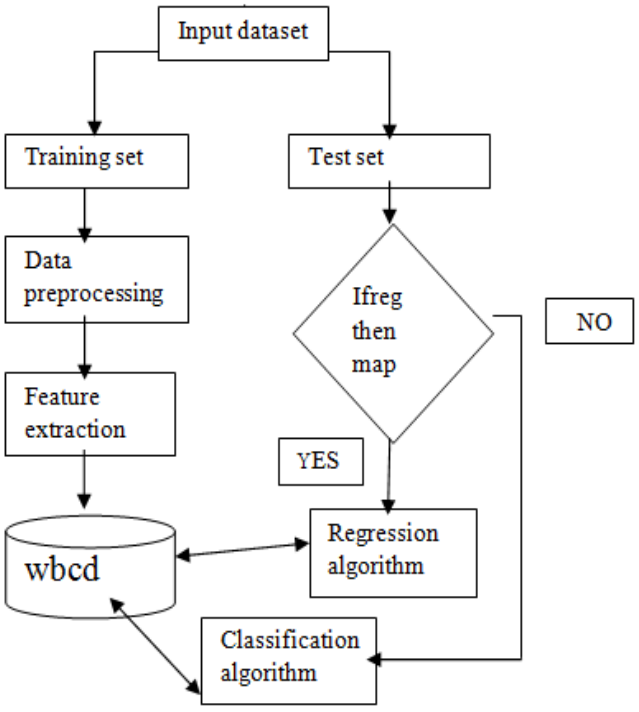


Fig 3: Architectural flow of the work

VI. EXPERIMENTAL RESULTS

A set of experiments have been performed using Anaconda 3.7 Jupyter Notebook with python programming language. To diagnosis breast cancer, different types of machine learning techniques (SVM, DT, RF, MLP, logistic and linear regression) are executed. The comparative study of these supervised machine learning techniques are performed using accuracy, precision and recall in classification and root mean square, R-square error value in regression as mentioned in Table 2 and Table 3. Above performance measures are implemented in all attributes.

Classification	SVM	DT	RF	MLP	Logistic reg
Accuracy	98.24	95.61	96.49	97.36	97.36
Precision	98%	96%	97%	97%	97%
Recall	98%	96%	96%	97%	97%

Table 2: Performance Measures Of Ml Techniques-Logistic regression

Regression	SVR	DTR	RFR	MLP	Linear regression
Root mean square	0.43	0.26	0.18	0.11	0.25
R-Square	0.81	0.71	0.82	0.94	0.68
Mean absolute error	0.12	0.06	0.07	0.02	0.19

Table 3 :Performance Measures Of Ml Techniques-Linear regression

According to Table 2,

-In classification, SVM performances is better than other model when all predicators are involved in testing and training set.

According to Table 3,

- MLP performances is better than other Regression model when all predicators are applied in training and hidden layer (7,2).

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 9293 - 9300

Accuracy of SVM (98.24) is higher than Decision tree (95.61), Random forest (96.29) and logistic regression (97.36). Precision and Recall are also greater in SVM than DT, RF, Logistic regression.Comparatively SVM performs better than other classification techniques.

As a result of regression, MLP performs better compared to other model because it has the least error.

Fig 4 represents the performance comparison of classification techniques. Here X-axis represents the algorithm being used and Y-axis represents the accuracy of ML techniques.

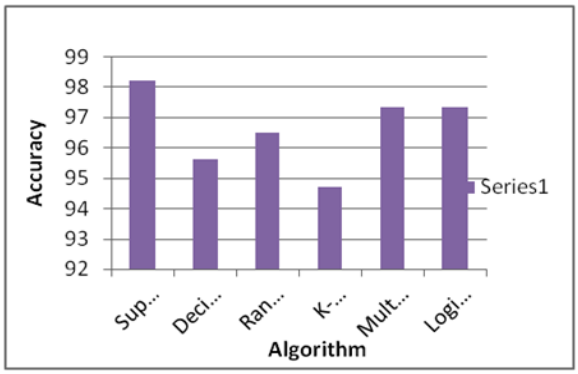


Fig 4: Performance measures of ML techniques (classification)

Fig. 5 represents the performance comparison of regression techniques. Here X-axis represents the ML model and Y-axis error values (Root mean square, R-square and Mean absolute error).

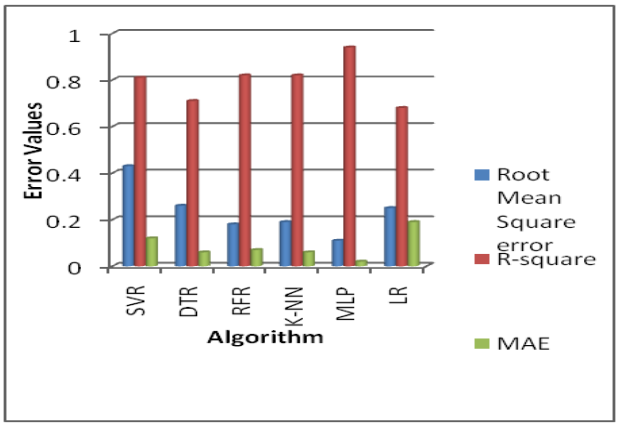


Fig 5: Performance measures of ML techniques (Regression)

VII. CONCLUSION

The main aim of this work is to improve the prediction of breast cancer diagnosis by selecting efficient ML techniques through which performance measures of the method can be increased.The comparative analysis by supervised machine learning techniques of both Classification and Regression are Support vector machine (SVM), Decision tree (DT), Random forest (RF) , Multilayer percepton (MLP), Logistic regression and Linear regression are executed. Compared to other classification model SVM performance gives classification accuracy of 98% ,precision and recall of 98% which is found to be better than DT,RF,MLP.and in case of Regression, MLP techniques has better performance than other ML techniques in terms of Root mean square value(0.11),R-square value(0.94),Mean absolute error(0.02).

REFERENCES

[1] American Cancer Society, “Detailed Guide: Breast Cancer”, cancer.org, 2014 [Online]. Available: www.cancer.org/Cancer/BreastCancer/DetailedGuide/index. [Accessed: Sept. 10,2016].

[2] S. Aruna, Dr S.P. Rajagopalan, L.V. Nandakishore, “An empirical comparison of supervised learning algorithms in Disease Detection”, International Journal of Information Technology Convergence and Services (IJITCS), Vol .1 No. 4, August2011.

[3] IndraKantaMaitra, Sanjay Nag, Samir Kumar Bandyopadhyay, "Identification of Abnormal Masses inDigital Mammography Images", International Journal of Computer Graphics, vol. 2, no. 1, 2011.

[4] S. Aruna, Dr S.P. Rajagopalan, L.V. Nandakishore, “An empirical comparison of supervised learning algorithms in Disease Detection”, International Journal of Information Technology Convergence and Services (IJITCS), Vol .1 No. 4, August2011.

Published by: The Mattingley Publishing Co., Inc.

[5] R. Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease." Expert Systems with Applications, Vol. 37 No .2, pp. 1568- 1572, 2010.

[6] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines", Machine Learning, vol. 46, pp. 389-422, 2002.

[7] H. Temurtas, N. Yumusak, and E. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks." Expert Systems with applications Vol. 36 No. 4, pp. 8610-8615,2009.

[8] Mehul P Sampat, Mia K Markey, Alan C Bovik et al., "Computer-aided detection and diagnosis in mammography", Handbook of image and video processing, vol. 2, no. 1, pp. 1195-1217, 2005.

[9] S. Beucher S., C. Lenteuejoul., "Use of watersheds in contour detection", Proceedings of the International Workshop on Image Processing: Real-Time Edge and Motion Detection/Estimation, 1979, pp. 2.1-2.12.

[10] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers." Machine learning Vol. 29 No. 2-3, 1997,pp.131-163.

[11] R.Palaniappan, K.Sunderaj, S. Sundaraj, “A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals”, BMC Bioinformatics, 15.1, pp. 1-8,2014.

[12] Fear, E. C. and M. A. Stuchly, "Microwave detection of breast cancer," IEEE Transactions on Microwave Theory and Techniques, Vol. 48, 1854-1863, 2000.