

A Classifier Model For Diagnosis of Benign/Malignant Breast Cancer Tumors

Vikhyat Srivastava^a, Ritika Choudhury^a, L.Shanmukhi^a, Jyotirmayee Dash^a

ritu0696@gmail.com

^aSRM Institute of Science and Technology, Kattankulathur, Tamilnadu-603203, India.

Abstract— Diagnosis of breast cancer is still a challenging task. Even though recent development in technology has made detection easier, improving its accuracy is an ongoing process. This paper aims at diagnosis of breast cancer and identifies the type of tumor (benign or malignant). A set of data from Wisconsin Breast Cancer Dataset (WBCD) that contains various real valued features such as radius, texture, area, perimeter, compactness, concavity, concave points, smoothness, symmetry and fractal dimension of the breast tumor cell nucleus has been obtained. Data analysis has been carried out using MATLAB software and a threshold value for each type of tumor (benign or malignant) has been determined. Various combinations of features have been selected and correlation has been observed between them and the tumor type. Suitable features have been selected and used as predictors for a classification model based on Quadratic Support Vector Machine (SVM) algorithm. The model was trained using different training dataset until accuracy of prediction was maximum. This model was used for testing an entirely new set of data and hence predict the breast cancer stages.

Index Terms— Breast Cancer, Classification, Feature selection, Machine learning, Support Vector Machine (SVM)

I. INTRODUCTION

Breast cancer has been affecting a majority of women across the globe. Although the causes of breast cancer remain unknown, symptoms include lump in the breast, discharge from the nipple, inverted nipple, discomfort in the breast, swollen lymph nodes, persistent fever, change in the shape and texture of breast or nipple that indicate the presence of breast cancer. If any of the symptoms are identified, physical examination, mammogram images, biopsy or Position emission tomography-computed tomography (PET-CT) scan is generally used for diagnosis. Depending upon the stage of cancer, people undergo various treatments such as radiation, chemotherapy, hormone therapy and surgery. Several researches have been carried out to propose methodologies for early diagnosis of breast cancer, to reduce the mortality rates. With the advent of recent technologies, it has become possible to extract features from given mammogram images. Kemal Polat^[1] used the Least Square Support Vector Machine (SVM) algorithm in the

dataset containing 683 samples, out of which 444 samples belong to benign class and the remaining were malignant. He considered nine features: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nuclei and mitoses to diagnose breast cancer. An accuracy of 94.44% was obtained in 80-20% training-testing pattern. Mehmet Fatih Akay^[2] used the same dataset and obtained a statistical analysis using the F1 score. The SVM parameters were modified by grid search. Using 80-20% training-testing pattern, an accuracy of 99.51% was obtained. Murat Karabatak^[3] used automatic diagnosis system for detecting breast cancer based on association rules and neural network and gained an accuracy of 95.6%^[3]. U.Rajendran Acharya^[4] used thermal imaging as a potential tool for detecting breast cancer thermal patterns that were recorded using Infrared camera. Hui-Ling Chen^[5] used Rough Set based Support Vector Machine (RS SVM) as a feature selection tool to remove redundant features and improved accuracy by SVM. An accuracy of 99.3% was obtained. A.Marciano-Cederio^[6], used Artificial Metaplasticity Multilayer Perceptron (AMMLP) for simulating the biological property of metaplasticity on MLP with back propagation. Bichen Zheng^[7] used K-means to identify hidden patterns of benign and malignant tumor and used them as a new feature in training the classifier model. Support Vector Machine (SVM) algorithm was used to differentiate between benign/malignant tumor type. He selected six features from thirty two original features consisting of mean, standard error and worst/largest value of the following parameters: radius, perimeter, texture, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. An accuracy of 97.38% was obtained based on 10-fold cross validation. Chien-Hsing Chen^[8] presented a hybrid intelligence model that used cluster analysis techniques with feature selection for analyzing clinical breast cancer diagnosis.

In this paper, a unique method for feature selection has been proposed. A decision making model based on Support Vector Machine (SVM) combined with feature selection was used for diagnosis of breast cancer.

II.. METHODOLOGY

The data set present in Wisconsin Breast Cancer Dataset (WBCD) consisted of three datasets. The first dataset

represented the chronological grouping of data. The second dataset included cases having invasive breast cancer and no evidence of distinct metastases. The third dataset were computed from a digitized image of a fine needle aspirate of a breast mass that describe characteristics of the cell nuclei present in the image. The features mentioned in first two datasets were not obtained from mammogram images and the number of malignant and benign cases was not sufficient for prediction. Hence, we have selected the third dataset for prediction. It consists of 30 parameters. Out of 569 cases, 358 were benign and the remaining were malignant. The mean, standard error and largest/worst values of the parameters such as radius, perimeter, texture, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension were considered. The threshold values of the two stages against each parameter have been determined. This enabled data reduction, classification and increased the amount of percentage of accuracy in the true class and predicted class of malignant and benign stages.

III. RESULTS AND DISCUSSION

A. Feature selection

Determining a suitable combination of features is essential for obtaining high precision and accuracy as too many features pose the problem of overfitting the model. Using only few features reduces the complexity of models, requires less time and is easier to understand [9]. Feature selection was divided into three steps: box plot, correlation coefficient and parallel coordinate plot.

Box plot is used to display the first quartile, median and third quartile for benign and malignant stage of each parameter. The parameters which have distinguished values of median under both benign and malignant stages are chosen.

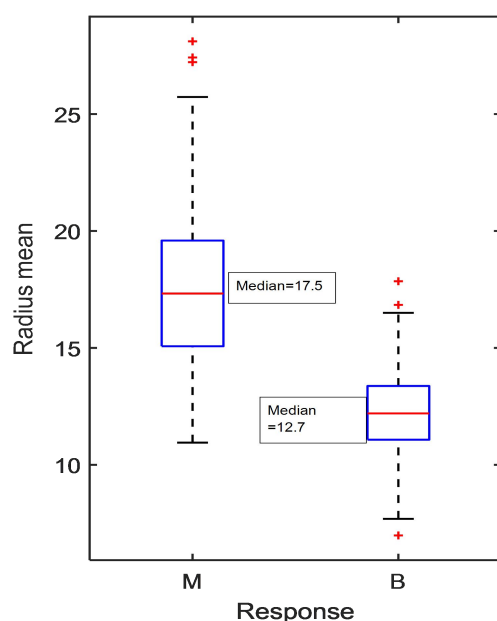


Fig. 1. This figure displays the box plot for radius mean. M represents malignant stage and B represents benign stage. The median value for malignant stage is 17.5 and benign stage is 12.7. As these values are distinguishable from each other, this parameter is taken into consideration.

Further, correlation coefficients are used to determine the relationship between various parameters. The correlation coefficient for any two random variables A and B is given by [10]:

$$R = \begin{bmatrix} 1 & \rho(A,B) \\ \rho(B,A) & 1 \end{bmatrix}$$

$$\text{Where, } \rho(A,B) = \frac{\text{Cov}(A,B)}{\sigma(A)\sigma(B)}$$

The correlation coefficient can range between -1 to 1. Positive value indicates that the two parameters increase or decrease together. Negative value indicates that, when one parameter increases, the other decreases. Hence, only positive values of correlation coefficient are taken into consideration. Further, a model has been created a model using machine learning. Different combinations of parameters have been selected and correlation was observed between them using parallel coordinate plot. In this plot, the X and Y axis lie parallel to each other. It is ideal for comparing multiple variables together and observing the relationship between them. The eight parameters chosen from the previous steps are: Radius mean, perimeter mean, area mean, concave points mean, radius worst, perimeter worst, area worst, concave points worst. The parallel plot of these parameters is easily distinguishable in terms of benign and malignant stage.

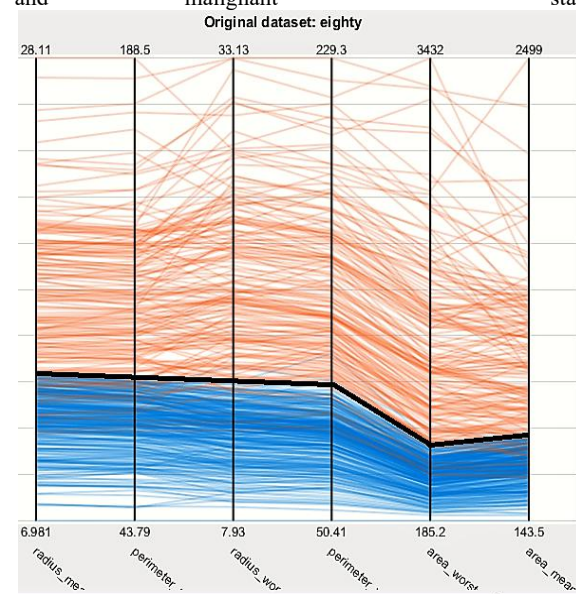


Fig. 2. This figure displays the parallel coordinate plot of the selected features. Blue lines represent benign stage and red lines represent malignant stage. Hence, it is clear that the two stages are easily separable.

B. Training and testing

The accuracy of prediction has been improved by supervised training on the model using Support Vector Machine (SVM). The model has been trained using 50-50%, 70-30%, 80-20% training-testing pattern. An accuracy of 98.2% in the 80-20% training-testing pattern has been gained using the Quadratic SVM model.

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression. In this algorithm, each data is plotted as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes very well^[11].

A quadratic equation is of the form $(x-a)(x-b)$, linear separation is achieved by mapping the data to a higher dimensional space. Thus, the quadratic SVM model was obtained. In this, the quadratic equation is in the form of a parabola in the two-dimensional space.

The following table displays the percentage of dataset used for training-testing pattern against the accuracy obtained in each case.

TABLE 1

Training-testing (%)	Accuracy (%)
50-50	91.9
70-30	95.2
80-20	98.2

To describe the performance of the model in the form of a table, confusion matrix has been used. It allows visualization of the performance of an algorithm. It also allows easy identification of confusion between classes. For example, one class is commonly mislabeled as the other^[12].

True class	B	283	2
	M	5	165
		B	M
		Predicted class	

Fig. 3 This figure represents the confusion matrix for validation with 80% data.

Out of 455 cases, 283 cases were benign and predicted correctly. The remaining 5 cases were benign but could not be predicted accurately as malignant. 165 cases were malignant and predicted correctly. The remaining 2 cases were malignant, but could not be predicted accurately as benign. Overall, 98.2% of the predictions are correct remaining 1.8% could not be predicted accurately.

IV. CONCLUSION

A decision making model based on Support Vector Machine (SVM) combined with feature selection has been used for early diagnosis of breast cancer. Various machine learning algorithms have been tried and tested. The proposed method yields prediction accuracies of the trained classifier model as 98.2% for 80-20%, 95.2% for 70-30% and 91.9% for 50-50% training-testing pattern. Hence, highest accuracy of 98.2% has been obtained for training with 80% data and testing with 20% data. This has been verified using the confusion matrix. Further work can be done on larger amount of cases and linking the classification model to the features extracted by image processing on mammogram images.

ACKNOWLEDGMENT

This work has been supported by Electronics and Instrumentation Department of SRM Institute of Science and Technology.

REFERENCES

- [1] Kemal Polat, Salith Gunes, "Breast cancer diagnosis using least square support vector machine" *Elsevier. [Online]*, vol. 17, no. 4, pp. 694-701, Jul. 2007.
- [2] Mehmet Fatih Akay, "Support vector machines combined with feature selection for breast cancer diagnosis. *Elsevier. [Online]*, vol. 36, no. 2, pp. 3240-3247, Mar. 2008.
- [3] Murat Karabatak, "An expert system for detection of breast cancer based on association rules and neural network" *Elsevier. [Online]*, vol. 36, no.2, pp. 3465-3469, Mar 2009.
- [4] U.Rajendran Acharya, "Thermography based breast cancer detection using texture features and Support Vector Machine. *Springer. [Online]*, vol. 36, pp. 1503-1510, Oct. 2010.
- [5] Hui-Lign Chen, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis" *Elsevier. [Online]*, vol. 38, no. 7, pp. 9014-9022, Jul. 2011.
- [6] A.Marcario-Cederio, "WBCD breast cancer database classification applying artificial metaplasticity neural network" *Elsevier. [Online]*, vol. 38, no. 8, pp. 9573-9579, Aug. 2011.
- [7] Bichen Zheng, Sang Won Yoon and Sarah S.Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms" *Elsevier. [Online]*, vol. 4, no. 4, pp. 1476-1482, Mar 2014.
- [8] Chien-Hsing Chen, "A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection" *Elsevier. [Online]*, vol 20, pp. 4-14, Jul. 2014.
- [9] www.shirringithub.com
- [10] <https://in.mathworks.com/help/matlab/ref/corrcoef.html>
- [11] <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>
- [12] <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>

