International Journal of Applied Engineering Research ISSN 0973-4562 Volume 16, Number 4 (2021) pp. 316-320

Prediction of Breast Cancer Using SVM Algorithm

Deepika S* and Kapilaa Ramanathan**

Student
Department of Information Technology
Sri Venkateswara College of Engineering,
Sriperumbudur, Tamilnadu, India – 602117

Devi N***

Assistant Professor
Department of Information Technology
Sri Venkateswara College of Engineering,
Sriperumbudur, Tamilnadu, India - 602117

Abstract

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society. As the diagnosis of this disease manually takes long hours and the lesser availability of systems, there is a need to develop the automatic diagnosis system for early detection of cancer. In recent years, machine learning has been widely used in detection and achieved favorable performance. In this paper, we have proposed a system that predicts the stages of the breast cancer. Without directly applying the machine learning techniques, we first perform K fold cross validation in the dataset to find which technique is more suitable for this dataset. The Classification and Regression Tress(CART), Linear Support Vector Machines(SVM), Gaussian Naïve Bayes(NB) and k-Nearest Neighbors (KNN). techniques are validated and found that SVM is better for the breast cancer dataset. Our proposed model is then constructed using SVM. Using machine learning methods for diagnostic can significantly increase processing speed and on a big scale can make the diagnostic significantly cheaper. Our proposed system has an accuracy of 99%.

Keywords— Breast Cancer, Machine Learning, SVM, Cross validation, PCA

1. INTRODUCTION

The most common cancer found amongst women is the breast cancer that formulates to almost 25% of the overall cancer cases found. It has affected close to 2.5 million people from 2017-2019. It has become one of the most found cancer among women, contributing to the rise in cancer deaths. The chances of survival from breast cancer has seen an increase remarkably when diagnosed in the early stages. The classification of the tumor into benign (non-cancerous) or malignant(cancerous) proves to be one of the important challenge against the detection of the tumor. This challenge is overcome by using Machine Learning technique that can dramatically improve the level of diagnosis in breast cancer. Research shows that experienced physicians can detect cancer by 79% accuracy, while a 97% accuracy can be achieved using Machine Learning techniques. The main aim of the paper is the classification of the tumor detected, into benign(non-cancerous) or malignant(cancerous) using Machine learning techniques.

2. RELATED WORKS

Artificial Intelligence(AI) techniques were the most used approach for breast cancer diagnosis in the works from the literature survey. These techniques were used to improve the time efficiency and classification methods. Arpit B and Aruna T. [1] proposed a neural network that is genetically optimized(GONN) for the classification of the breast cancer. New mutation operators and new crossovers were introduced in the network. Classical back propagation and classical model were compared with the accuracy, specificity, confusion matrix, ROC curves, AUC under ROC curves of GONN and sensitivity from the classification made using WBCD. The approach presented a good accuracy classification. On the other hand, a larger dataset could be used to improve the accuracy rather than WBCD, feature extraction to make GONN more effective and efficient for diagnosis in real time. Ashraf O. I. and Siti M. S. [2] A computer-based method was proposed to classify the breast cancer in an automated method. Neural network multilayer perceptron(MLP) that is relied on non-dominated sorting genetic algorithm(NSGA-II) is used to boost the accuracy and the network structure. Improvement in the classification accuracy was found in this method. A major drawback of MLP is the local minima. Na L. et al. [3] A hybrid feature selection approach model for the diagnosis of breast cancer was proposed: To remove redundant and irrelevant feature from CSSVM (cost sensitive support vector machine) gain directed simulated annealing genetic algorithm wrapper (IGSAGAW). Improvement in the accuracy and reduction in the computational cost was found in this model. The method was applied on Wisconsin Original Breast Cancer (WBC) and WBCD to verify its effectiveness. A better performance and decrease in the calculation complexity was observed. Nawel Z. et al. [4] Computer Assisted Detection (CAD) used for the classification of mammogram images was implemented. To reduce the dimensionality of the feature vector and semi supervised support vector machine (S3VM) GA-based features selection algorithm was implemented. Digital Database for Screening Mammography (DDSM) dataset was used for validation. Improvement in accuracy was detected. Abdulkader H. et al. [5] Breast tissue classification was automated by the

proposed system. Machine learning algorithms: Feed forward neural network using the back propagation learning algorithm (BPNN) and radial basis function network (RBFN) were used. The classification of the cancerous tissue consisted of 6 types Carcinoma, Fibro-adenoma, Mastopathy, Glandular, Connective, and Adipose tissue. Electrical impedance spectroscopy (EIS) approach was applied to acquire the data. Back propagation network was surpassed by the Radial basis function network with regard to the accuracy, maximum epochs, minimum error and training time in classification. The automated system showed advancement in accuracy and efficiency in training time. The drawback of the automated system was the decrepitation in generalization and local optima. Haifeng W. et al. [6] A learning model based on SVM was proposed for the diagnosis of breast cancer. The model consists of SVM structures which include a C-SVM and a-SVM, along with six different kernel functions. A Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE) mechanism is recommended for model hybridization for various base classifiers. The evaluation was with regard to different datasets: the Wisconsin Breast Cancer (WBC) dataset the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, the Surveillance, Epidemiology, and End Results (SEER) dataset. The designed model shows a drastic escalation in the diagnosis accuracy when measured up to other works based on a single SVM. The drawback detected is with regards to the method being lavish and deficient in training time. In this paper, we propose a predictive analytics model to diagnose breast cancer stages of patients.

3. PROPOSED SYSTEM

The proposed system describes the breast cancer model and draws the execution of the Support Vector Machine (SVM) algorithm to classify breast cancer tumor in to benign or malignant

The outcomes of SVM consist of accuracy and precision. It is used for classification, which trains models to categorize cancer patients according to their diagnosis. The system also focuses on providing the, detailed description of the tumor for each patient diagnosed.

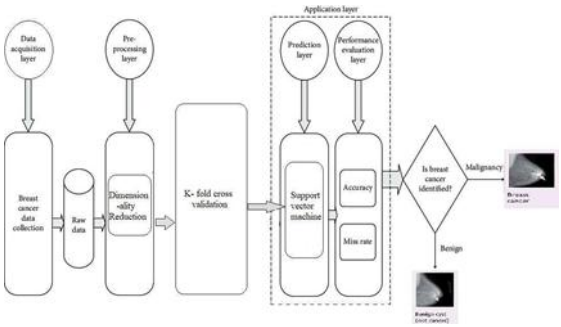

Figure 1. Proposed Architecture For Breast Cancer Prediction

Figure 1 depicts the overall architecture of the system. The data collection consists of the parameters regarding the tumor, from the lab results. The data is pre-processed and cleaned using Dimensionality reduction techniques. The relevant datasets are then used in further steps like classification or regression. The pre-processed data undergoes K-fold cross validation, where the accuracy of various machine learning algorithms on the data is tested. From, the results of the cross validation, the data is trained using SVM algorithm for the analysis of the tumor as cancerous(malignant) or non- cancerous(benign). The data is split into 80% training and 20% test data. The SVM algorithm is then tested for its accuracy and miss rate in the performance evaluation layer. The tested data is then used to identify the breast cancer and classify it based on the SVM algorithm.

3.1 Pre-Processing

The preprocessing of data module includes the implementation of the dimensionality reduction technique. Dimensionality Reduction is a process in which the number of independent variables is reduced to a set of principle variables by removing those which are less significant in predicting the outcome. Dimensionality Reduction is used to get two dimensional data for better visualization of machine learning models. It is done by plotting the prediction regions and the prediction boundary for each model.

PCA is the next step in the layer which is depicted in figure 2. It is an unsupervised linear dimensionality reduction algorithm which is used to find the strongest features based on the covariance matrix of the dataset. It flattens large number of dimensions to 2 or 3 dimensions. It is used when needed to tackle the curse of dimensionality among data with linear relationships.

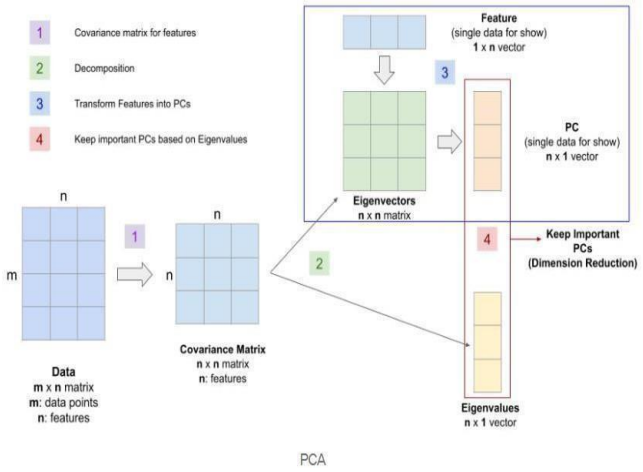

Figure 2. Principal Component Analysis

3.2 K-Fold Cross Validation

The data is analysed and a model is built to predict if the tumor is cancerous or not. It is a binary classification problem, and a few algorithms appropriate are used, to test the data for accuracy. To find the most efficient algorithm , a test is performed on the machine learning algorithms with a default setting to get an early indication of their performance. The 10

fold cross validation process is used for the test. The steps for the cross validation is given below:

Step 1. Shuffle the dataset randomly.

Step 2. Split the dataset into k groups

Step 3. For each unique group:

- Take the group as a hold out or test data set
- Take the remaining groups as a training data set
- Fit a model on the training set and evaluate it on the test set
- Retain the evaluation score and discard the model

Step 4. Summarize the skill of the model using the sample of model evaluation scores

The following non-linear machine learning algorithms are tested, namely: Classification and Regression Tress(CART), Linear Support Vector Machines(SVM), Gaussian Naïve Bayes(NB) and k-Nearest Neighbours (KNN). The results of the K-fold cross validation is presented in figure 3. From the figure 3, it is evident that SVM works better for classification problem. Therefore, we chose SVM for constructive the predictive analysis model.

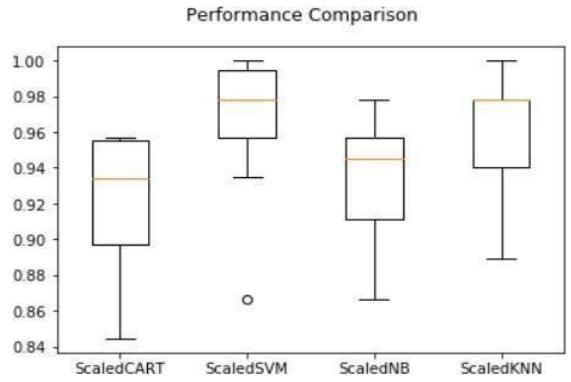

Figure 3. Performance Comparison using K-Fold Cross Validation

3.3 Construction of SVM Classifier

A Support Vector Machine (SVM) is a binary linear classification whose decision boundary is explicitly constructed to minimize generalization error as shown in figure 4. It is a very powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression and even outlier detection. SVM is well suited for classification of complex but small or medium sized datasets.

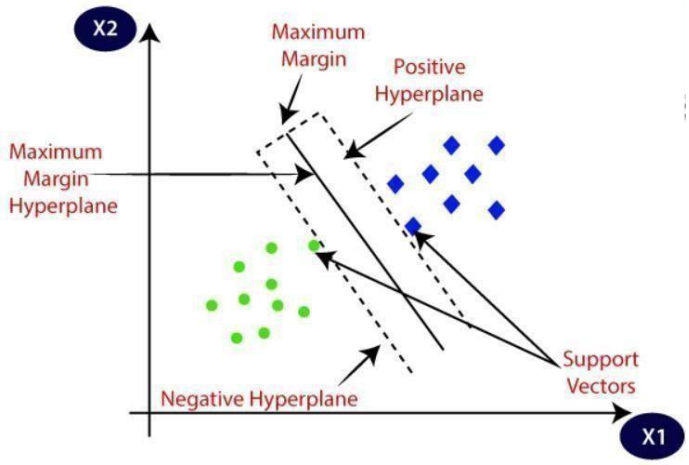

Figure 4 SVM Classifier Model

3.3.1 Model Training

From the dataset, the target and predictor matrix are created. "y" is the feature used to predict (Output). In this case the aim is to predict whether the "target" is cancerous (Malignant) or not (Benign) using the "target" feature here. "X" is the predictors which are the remaining columns consist the mean of perimeter, radius etc. The python library that splits the dataset into training and testing data is imported.

- Training data = the subset of the data used to train the model.
- Testing data = the subset of the data model used for performance evaluation.

A confusion matrix is created for better evaluation of the classifier's performance on the test dataset.

The SVM classifier achieves an accuracy of 99.0% on the held-out test dataset. From the confusion matrix, there is only 1 case of mis-classification. The performance of this algorithm is expected to be high given the symptoms for breast cancer should exhibit certain clear patterns. The classifier then predicts the tumor into malignant (cancerous) or benign (non-cancerous).

4. EXPERIMENTAL RESULTS

Our proposed system is trained and tested with the dataset downloaded from UCI machine learning repository. The dataset has 10 real valued attributes namely radius, texture, perimeter, area, smoothness, compactness, concavity, concave, concave points, fractal dimension. These attributes are computed from the cell nucleus of the breast mammogram. The statistical relationship between two variables is referred to as their correlation. The correlation present in the dataset is shown in the figure 5.

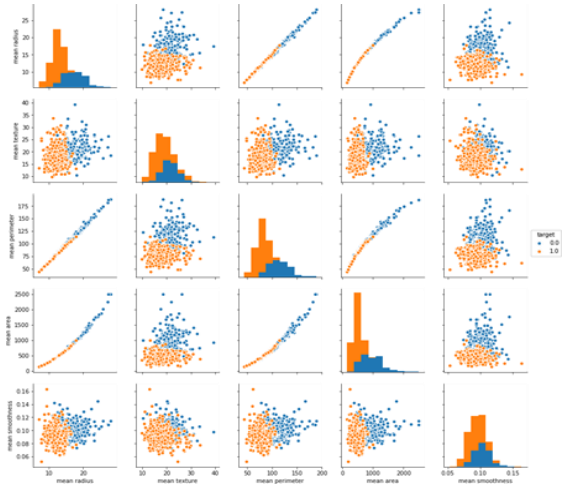

Figure 5. Correlation of Parameters Present in the dataset

The data parameters include features of the tumour that are considered to predict the stage of the breast cancer. Variables within a dataset can be related for lot of reasons.

For example:

- One variable could cause or depend on the values of another variable.
- One variable could be lightly associated with another variable.
- Two variables could depend on a third unknown variable.
- It can be useful in data analytics and modelling to understand the relationships between variables.

The correlations between the data attributes are checked for better analysis and prediction. The red around the diagonal suggests that attributes are correlated with each other. The yellow and green patches suggest some moderate correlation and the blue boxes show negative correlations in figure 6.

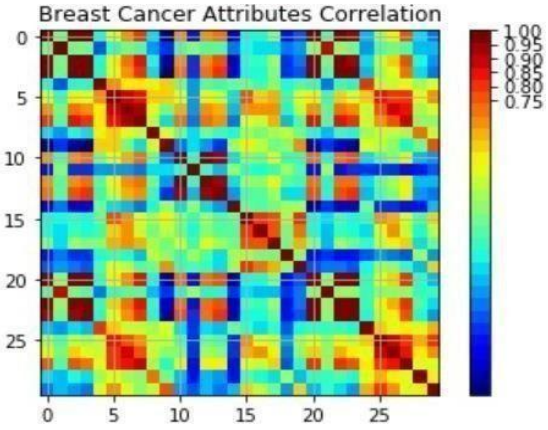

Figure 6. Breast cancer attributes correlation

A strong correlation between mean radius and mean perimeter, as well as mean area and mean perimeter is shown in figure 7.

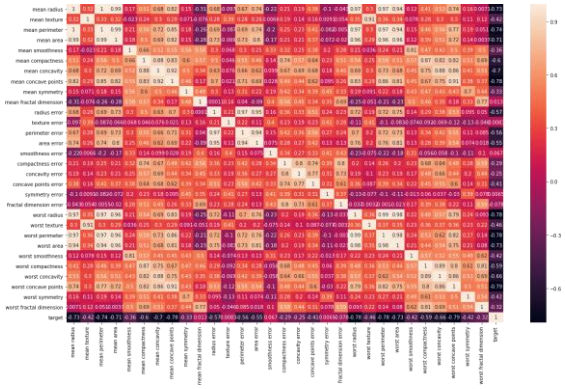

Figure 7. Breast cancer data attributes correlation

The data model is split into 80% for training and the 20% for testing and the model is built using SVM classifier for prediction. A confusion matrix is introduced into the data set to evaluate the performance of the classifier better for efficient prediction which is shown in the figure 8.

```
[[74  1]
 [ 0 39]]
```

Figure 8. Confusion Matrix of our Proposed Approach

The confusion matrix resulted in providing better accuracy and performance of the SVM classifier on the data model. The SVM classifier thus resulted in an 99.0% accuracy rate as shown in figure 9.

	precision	recall	f1-score	support
0	1.00	0.99	0.99	75
1	0.97	1.00	0.99	39
avg / total	0.99	0.99	0.99	114

Figure 9. SVM accuracy evaluation result

The SVM classifier being proved more efficient and accurate from figure 9 for the prediction and classification of the tumor into malignant (cancerous) and benign (non-cancerous). The efficient performance of the SVM classification in prediction and classification is shown in figure 10.

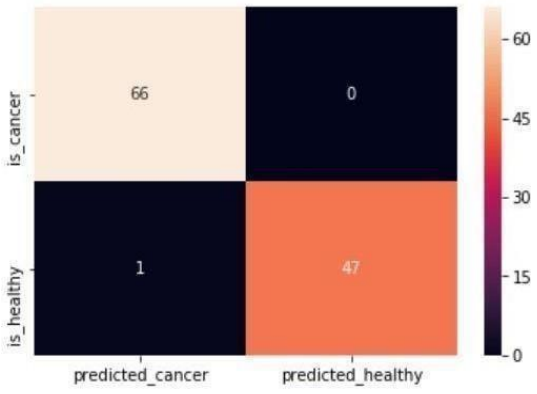

Figure 10. Prediction and Classification of the Tumor

5. CONCLUSION AND FUTUREWORKS

A suitable breast cancer diagnosis model can assist scientific practitioners with accurate and dependable prediction effects. Early detection of breast cancer makes it treatable which may significantly increase the chances of survival. The proposed system classifies the tumors into malignant or benign using features obtained from cell images with 99% accuracy. Hence it can be used as an efficient application for detection and prevention of breast cancer.

The integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide eminent tools for inference in this domain. Further research in this field can be carried out to improve the performance of the classification techniques to predict with different variables. The chosen approach can be developed into a prospective pragmatic model, facilitating doctors with swift consultation in diagnosing and identifying the breast cancer.

REFERENCES

[1] Arpit, B., Aruna, T.: Genetically Optimized Neural Network Model for Breast Cancer diagonosis. Expert Systems with Applications, Elsevier, pp. 15-30(2017)

[2] Ashraf, O. I. and Siti, M. S.: Breast cancer diagnosis using multilayer perceptron neural network. International Journal of Computer Aided Engineering and Technology, Vol. 13, No. 3, pp. 443-447(2017)

[3] Na, L., Qi, E., Gui-Qiu, L.: Breast cancer diagnosis- a novel approach. Information Processing and Management, Elsevier, pp. 710-745(2018)

[4] Nawel, Z., Nabiha, A and Mokhtar, S.: Adaptive Semi Supervised Support Vector Machine. Journal of Medical Imaging and Health Informatics, American scientific publisher,pp. 63-73(2017)

[5] Abdulkader, H., John.:c Classification of breast tissue using machine learning. In: 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW, pp. 511-521. Procedia Computer Science, Elsevier. Budapest, Hungary (2018)

[6] Haifeng, W., Bichen, Z., Sang, W.Y., Hoo,: Ensemble Algorithm for Breast Cancer Diagnosis. European Journal of Operational Research, Elsevier, pp. 44-88 (2015).