# Machine Learning Approaches for Breast Cancer Diagnosis and Prognosis

Ayush Sharma
Department of Computer Science
Jaypee Inst. of Information Technology
Uttar Pradesh, India
contact.ayush95@gmail.com

Sudhanshu Kulshrestha
Department of Computer Science
Jaypee Inst. of Information Technology
Uttar Pradesh, India
sudhanshu.kulshrestha@jiit.ac.in

Sibi Daniel
Department of Computer Science
Jaypee Int. of Information Technology
Uttar Pradesh, India
5181daniel@gmail.com

***Abstract* — For breast cancer diagnosis in patients, radiologists conduct Fine Needle Aspirate (FNA) procedure of breast tumor. This procedure reveal features such as tumor radius, concavity, texture and fractal dimensions. These features are further studied by medical experts to classify tumor as Benign or Malignant. The cardinal aim of this paper is to predict breast cancer as benign or malignant using data set from Wisconsin Breast Cancer Data using sophisticated classifiers such as Logistic Regression, Nearest Neighbor, Support Vector Machines. Furthermore, using Wisconsin Prognostic data set, probability of recurrence in affected patients in calculated. As a result, a concrete relationship between precision, recall and the number of features in the data set is achieved, which is shown graphically.**

***Keywords—Supervised Machine Learning; Breast Cancer; Fine Needle Aspirate; Feature Scaling; Decision Boundary***

## I. INTRODUCTION

Recent years have witnessed a shift in computation intensive methods being used to analyze biomedical signals. The crux of these methods is machine learning and artificial intelligence. One of numerous applications of machine learning involves creating a classifier that can separate subjects into two or more classes based on the labels (features) measured in each of the subjects. There has been an exponential increase in the number of breast cancer cases across the globe. According to [1], breast cancer commences when breast cells grow intractably. The respective cells usually end up forming a tumor that can be detected via an X-Ray or felt as a lump via clinical examination by the doctor followed by Fine Needle Aspirate (FNA) of the same. If the cells are able to pervade in the surrounding tissue or metastasize to distant areas of the body, then the tumor is ascribed as a malignant one. While there may be other symptoms as well, the most eccentric feature of breast cancer is the presence of a lump or mass in the tissue. Not all lumps are indicative of breast cancer, they are benign. If the cancerous cells metastasized to the lymph system, the cancer progresses to terminal stage and invariably becomes difficult to treat. There are various arcane causes of cancer, out of which the most conspicuous ones can help us get insight. These include age (most invasive breast cancers are found in women above 55 years of age [1]), sex (predominantly a female disease, with 10 times more probable to happen in females than in males.) and presence of certain genes (5% to 10% of cancer cases are a direct result of genetic mutation, resulting due to gene mutations passed on from parents). The research here revolves around classifying the correct labels for each of the entries by modeling data. There are three types of inputs to such models, which are; training data, the dependent variables and the independent variables. Once a classification model is built, it can be used to classify unidentified objects as having benign or malignant breast cancer.

Various supervised classification algorithms are available and the ones used here are Logistic Regression, Nearest Neighbor and Support Vector Machine. Both Support Vector Machine and Logistic Regression are proven to produce good classification accuracy performance, however the results are inconsistent depending on the data sets. Due to inconsistent results obtained, the primary aim of this paper is to further substantiate the performance of Support Vector Machines (SVM), k-nearest neighbor classifier (KNN) and Logistic Regression on different breast cancer data sets; namely Wisconsin Diagnostic and Prognostic Breast Cancer data set, which is made publicly available at UCI Machine learning repository [2]. Furthermore, it also aims to validate the relationship between number of features in a data set and model performance. The paper is organized as follows: Section 2 provides the related studies carried out on breast cancer classification using various classifiers and basic concept of the same. In section 3, the dataset is explained in detail. Section 4 described the methodology of this study. In section 5, the results and discussion are summarized as tables to show the performance and the comparison of applied classifiers. Finally, the paper is concluded in section 6.

## II. LITERATURE REVIEW

### *A. Related Work*

The research using machine learning techniques in the medical domain has been prevalent for a long time, especially in the diagnosis of breast cancer. Various researches have been conducted, using an approach similar to the one delineated here. The data used in these researches has been collected from the UCI Machine Learning [2] repository as well. One of them includes the research by [3] which used classifiers such as Naïve Bayes with an accuracy of 95.99% and a multilayer Perceptron with an accuracy of 95.29% on the same data set.

A similar approach proposed by S. Kharya [4] states that artificial intelligence and neural networks are one of the most widely used predictive techniques in breast cancer prediction. Despite of their good accuracy, these approaches are difficult to comprehend. The same paper elucidates the benefits and the limitations of the techniques such as Naïve Bayes, neural networks and decision trees.

Another research conducted by [5] summarizes the applications of machine learning in the field of breast cancer prediction. It also encapsulates the basics of Artificial neural networks, establishes the simplifications required in implementing a neural network approach and lays out the research work carried out by various other researches in the same. Consequently, it evaluates various models such as Relevance Vector Machines (RVM), Support Vector Machines (SVM) and then compares them on the basis of performance and other metrics such as specificity and sensitivity. The most cogent research was pursued by H. Yusuf [6], in which diagnosis of breast cancer from mammograms was complemented using logistic regression. The data set was not from the UCI Machine Learning repository [2], but instead collected from a survey of questions completed by a radiologist during his observations with cancer patients. From a sample of 130 patients the mammogram result accuracy was 91.5% while accuracy same compared to 46 test samples of validation test, the accuracy obtained was 67.4%. The author reached the conclusion that presence of mass calcifications, skin thickening, and distortions as a result of mammography had high odds of cancer being malignant.

### *B. Logistic Regression*

Logistic Regression is a predictive analysis technique when the output label to be predicted is dichotomous, that means it is binary. Here the output label being predicted is Benign or Malignant (B or M). Like all other regression models, it is used to delineate data relationship between dependent and independent variable. It maps the input variable space on a straight line, and then predicts unseen data by visualizing where the unseen data lies on the line. The logistic function is also called the Sigmoid function and is calculated by using the cost function below [8]:

$$f_i(\theta) = p\,(y_i = 1|\mathrm{x}) = \frac{1}{1+exp\,(-\theta^T x)} \qquad (1)$$

### *C. Nearest Neighbor Classifier*

The nearest neighbor classifier predicts the class of a data point to be the most common class among that point's neighbors [9]. Suppose we have n training data points, where the $i^{th}$ point has both a vector of features $\boldsymbol{x_i}$ and class label $\boldsymbol{y_i}$. For a new point $\boldsymbol{x^*}$, the nearest neighbor classifier first finds the set of neighbors of $\boldsymbol{x^*}$, denoted $\mathbf{N}(\boldsymbol{x^*})$. The class label for $\boldsymbol{x^*}$ is then predicted to be:

$$y^* = max \sum_{i \in N(x^*)} (y_i = c) \qquad (2)$$

where the indicator function *I()* is 1 if the argument is true, and 0 otherwise.

### *D. Support Vector Machine Classifier*

Support Vector Machines can be used for regression or classification purposes which perform by fitting a hyperplane in high-dimensional space. A good separation is one that is able to maximize the functional margin, i.e. the largest distance from the plane to the nearest data training data. Thus, higher the functional margin, lower will be the generalization error of the respective classifier [10]. The type of SVM used here is linear SVM. Given a feature set, *xi*, and a label, $yi \in \{0,1\}$, the loss function is minimized by the support vector machine using the equation below:

$$f_i(\theta) = max\,(1 - \theta^{\mathrm{T}}\mathrm{x}, 0) \qquad (3.1)$$

In contrast to other classifiers, there is an additional intercept term that is added to entries by adding the number 1 using the below formula:

$$\min_{\theta} \lambda \sum_n^1 fi(\theta) + \|\theta\| \qquad (3.2)$$

where λ is the penalty parameter that tells the SVM optimization to avoid misclassifying each training example in the data set. For a larger value of this parameter, the optimizer would choose a smaller margin for the hyperplane for correct classification whereas a smaller value of the same would lead to choosing a larger margin for the hyperplane, inadvertently leading to misclassification.

## III. EXPERIMENTAL DATA

The data sets used in this research were procured from the open-source Machine Learning repository of University of California, Irvine [2], which is available as a comma-separated (CSV) downloadable file. The name of the data set for the research conducted is Wisconsin Breast Cancer dataset. The dataset was created by Dr. W.H. Wolberg out of the desire to diagnose breast cancer as benign or malignant solely based on the observations of FNA [11]. Further, in collaboration with Prof. Managasarian and two of his graduate students, namely Rudy Sentiono and Kristin Bennett, a classifier was generated [13] that used multi-surface method of pattern separation one the nine features to finally diagnose 97% of new cases, ultimately producing the Wisconsin Breast cancer data set [14].The nine numeric features are ranged between 1-10, with the ‘class’ attribute having a discrete value of either 2 (benign) or 4 (malignant), and the sample code number being a unique patient identification number. The summary of the data sets is provided below.

The diagnostic data set has 458 entries with benign tumor and 241 with malignant tumor. It has the following features:

TABLE I. WISCONSIN DIAGNOSIC BREAST CANCER DATASET DESCRIPTION

| *S no.* | *Attribute/Feature* | *Range* |
|---|---|---|
| 1. | ID Number | Identification number for patients |
| 2. | Diagnosis | 2: Benign, 4: Malignant |
| 3. | Radius | 11-27 |
| 4. | Area | 360-2300 |
| 5. | Perimeter | 71-82 |
| 6. | Texture | 11-40 |
| 7. | Smoothness | 0.05-0.2 |
| 8. | Compactness | 0.04-0.45 |
| 9. | Concavity | 0.02-0.5 |
| 10. | Concave Points | 0.02- |
| 11. | Symmetry | 0.1-0.3 |
| 12. | Fractal Dimension | 0.05-0.1 |

Similarly, we preprocess the diagnostic data set in such a way that it can be used to map the patients with recurring cancer. For this purpose, we only extract the patients treated with malignant cancer followed by addition of two new attributes: tumor size and lymph node status. The tumor size and the spread in the lymph node helps identify the cancer stage in recurring patients for better understanding and results. The prognostic data set has 47 recurring patients and 143 non-recurring patients.

TABLE II. WISCONSIN PROGNOSTIC BREAST CANCER DATASET DESCRIPTION

| *S no.* | *Attribute/Feature* | *Range* |
|---|---|---|
| 1. | ID Number | Identification number for patients |
| 2. | Diagnosis | 2: Benign, 4: Malignant |
| 3. | Time | 1-125 |
| 4. | Radius | 11-27 |
| 5. | Area | 360-2300 |
| 6. | Perimeter | 71-82 |
| 7. | Texture | 11-40 |
| 8. | Smoothness | 0.05-0.2 |
| 9. | Compactness | 0.04-0.45 |
| 10. | Concavity | 0.02-0.5 |
| 11. | Concave Points | 0.02- |
| 12. | Symmetry | 0.1-0.3 |
| 13. | Fractal Dimension | 0/05-0/1 |
| 14. | Tumor Size | 0.4-10 |
| 15. | Lymph Node Status | 0-27 |

The diagnostic data set comprises of 699 entries, whereas prognostic comprises of 199 entries only. The training and testing split for modelling of data is performed by keeping 70% data for training and 30% for testing, in both the cases.

## IV. METHODOLOGY

Diagnosis of breast cancer is an arduous task. The first step is the clinical examination to detect the tumor/lump in the breast by the General Surgeon or by using imaging techniques such as Mammography. This is followed by FNA on the lump detected. The FNA procedure provides with attributes such as tumor size, radius, area etc. These can be used as features for modelling the data into a classifier. Finally, models such as SVM, k-nearest neighbor and Logistic Regression are trained using the obtained data to make predictions. The general methodology of modelling after data is obtained can be divided into four general steps which can be shown in the figure below. No matter what machine learning algorithm is being implemented, the below mentioned steps are key to any algorithm.

### *A. Data Preprocessing*

This is the first step before modelling data. At times data obtained may be incomplete, like lacking values in certain rows. It can also be noisy, replete with errors and outliers, and inconsistent as well, with arrant discrepancies. Thus, in order to eliminate all of the above, we preprocess the data before modelling. Here, we use min max normalization, also known as feature scaling, since the entries inside our data are primarily numeric. It can be described by the following formula [16]:

$$y = \frac{x\text{-}min(x)}{max(x)\text{-}min(x)} \tag{4}$$

where x is the initial value and min and max are the highest and lowest value respectively.

### *B. Data Visualization*

Data Visualization is the process of discerning data for patterns by presenting it in a pictorial or graphical format. The prime importance of data visualization lies in identification of patterns or trends in data, how to weigh individual features, and how to identify outliers in data. Furthermore, it ultimately helps in selection of right model for modelling of data. Before visualizing data, it is important to understand the size as well as the cardinality of data. High cardinality means there's a larger percentage of unique values, whereas converse is true for data with low cardinality. Secondly, it is important to determine what you're trying to visualize.

There are various ways in which we can visualize data. The most common ones are bar graphs, line charts, box plots, and heat maps etc. The same will be discussed further in the paper.

### *C. Model Selection and Implementation*

The choice of selecting the right model is entirely subjective. It depends strongly on the type of data, and what is the primary aim of the author. If primary aim is accuracy, the

best bet is to test data on number of models and then select the best one using cross validation. However, when the need is of a good enough model, there are certain things to be kept in mind. Firstly, the size of training set plays an integral role. For a training set with low number of training points, the high bias or low variance classifiers have a slight advantage over the low bias or high variance classifiers, as latter tend to over fit. Each model has its own disadvantages and advantages. For example, using Logistic Regression offers many ways to regularize the model, and eliminates the constraint of features being correlated. The data set is then segregated into two disparate sets: training and test set. The split is usually done using the ratio of 80/20, which means 80% of data is used for modelling and 20% of data is used for evaluation and prediction. Thus, we select here four models as discussed earlier, and divide both data sets into training and testing sets.

### *D. Model Evaluation*

Once the training data is modelled, we evaluate the test data and predict the outcome. The labels of test data are recorded and the incorrectly predicted labels are counted, giving us the simplest form of evaluation of model. The primary conundrum that is to be solved while using machine learning techniques, specially classification, is to predict labels or classes of unseen data accurately. The assumption here is that all sample are drawn from the same probability distribution and are completely independent from each other.

The performance of the classifiers was evaluated using metrics such as accuracy, sensitivity, specificity, and area under curve (AUC) etc. Sensitivity which is also defined as True Positive Rate (TPR) is the percentage of benign tumors data classified as benign by the classifier. The classifier that can correctly classify benign tumors will have a higher result in sensitivity. Sensitivity is defined as follows [15, 17]:

$$\text{Sensitivity}\ (\%) = \text{TPR} = \frac{\text{TP}}{\text{FN} + \text{TP}} \text{X100} \qquad (5.1)$$

Specificity is the percentage of malignant tumors data classified as malignant by the classifiers. The classifier that can correctly classify malignant tumors will have a better result in specificity. It is calculated as follows [18, 19]:

$$\text{Specificity}\ (\%) = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \text{X100} \qquad (5.2)$$

Accuracy is the most crucial metric for model evaluation. It invariably combines specificity and sensitivity for whole of the data combined. It is given by [9,11,13,14]:

$$\text{Accuracy}\ (\%) = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}} \text{X100} \qquad (5.3)$$

Area Under Curve (AUC) is a disposition of sensitivity and specificity over all possible thresholds. The AUC value of 100% represents perfect discrimination (the classifier can classify the tumors correctly), whereas an AUC value of 50% is equivalent to random model. AUC was calculated as follows [15]:

$$\text{AUC}\ (\%) = \frac{1}{2}\left(\frac{\text{TN}}{\text{TN+FP}} + \frac{\text{TP}}{\text{TP+FN}}\right) \text{X}\ 100 \qquad (5.4)$$

Although there are other metrics for model evaluation available as well, classification accuracy is the one most pertinent to our research.

## V. Results and Discussion

Throughout the research, 3 models were closely scrutinized and evaluated for two different datasets using the aforementioned metrics such as sensitivity, specificity, accuracy, and area under curve. The confusion matrix for each model is formulated in order to evaluate the models with ease.

For the Wisconsin breast cancer diagnostic data set, the results were as follows:

TABLE III. CLASSIFIER RESULTS ON WDBC

| *Metric.* | *Logistic Regression* | *k-nearest neighbor* | *Support Vector Machine* |
|---|---|---|---|
| Specificity | 99.89 | 94.7 | 84.9 |
| Sensitivity | 90.06 | 90.09 | 88.2 |
| Accuracy | 96.89 | 93.06 | 89.6 |
| AUC | 95.245 | 92.39 | 86.55 |

TABLE IV. CLASSIFIER RESULTS ON WPBC

| *Metric.* | *Logistic Regression* | *k-nearest neighbor* | *Support Vector Machine* |
|---|---|---|---|
| Specificity | 75.3 | 61.2 | 79.7 |
| Sensitivity | 42.9 | 40.89 | 41.2 |
| Accuracy | 88.6 | 82.56 | 89.73 |
| AUC | 59.1 | 51.045 | 60.45 |

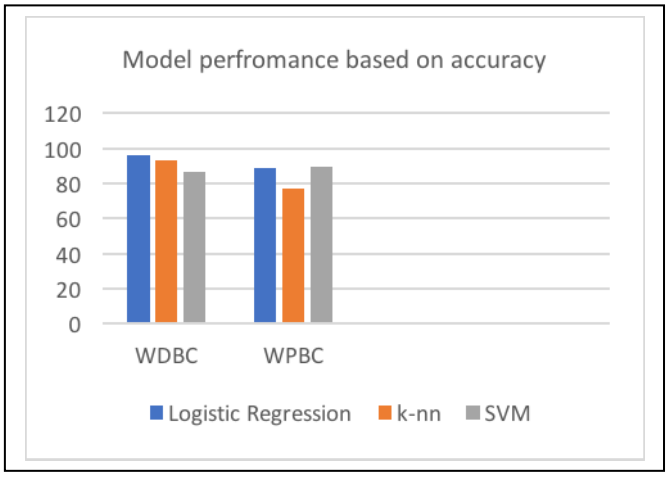


Fig. 2 Model performance based on accuracy of classifiers

We see that keeping accuracy as the sole criterion for evaluation, logistic regression is the best classifier model for diagnostic data set whereas support vector machine is the best classifier for prognostic data set. The reason is straightforward; small data size in case of prognostic data set means accuracy would be higher for a low bias/high variance classifier and converse for the diagnostic data set. Accuracy itself cannot be used to decide the model to be chosen. At times, the relationship between number of features and accuracy needs to substantiated in order to assess whether increasing the features would in turn increase the accuracy of the classifier to what extent. The same is elucidated using three cases in logistic regression of feature sets in data in the table 5 below.

While we see that the logistic regression classifier performs better for the particular data set, there are some differences in the models which should be kept in mind:

- **Decision boundary**: Logistic regression learns a linear classifier, while nearest neighbors can learn non-linear boundaries as well.
- **Predicted values**: Logistic regression predicts probabilities, which are a measure of the confidence of prediction. nearest neighbors predict just the labels.

TABLE V. NO. OF FEATURES VS MODEL ACCURACY EVALUATION

| ***Metric.*** | ***Single featured model*** | ***10-featured model*** | ***34-featured model*** |
|---|---|---|---|
| Accuracy | 88.1 | 96.3 | 96.89 |
| Precision | 87.2 | 95.01 | 96.1 |
| Recall | 80.4 | 90.09 | 90.06 |

## VI. CONCLUSION

While previous researches conducted on breast cancer show that sophisticated algorithms give an accuracy ranging from 90-94% on the same dataset of Wisconsin Breast Cancer, the study proposed here has achieved training accuracies ranging from 93-97% from trivial models such as Logistic Regression, Support Vector Machines, and Nearest Neighbor classifier.

For any model, accuracy alone is not the only metric to be taken into account. The following metrics must be evaluated in order to do a complete evaluation of the model; specificity, sensitivity, area under curve and model accuracy. Furthermore, there may be other metrics as well such as F-score and ROC, which in turn scrutinize the evaluation process.

Eventually, we can extend these techniques to hospitals, and maintain a repository of patients with their current diagnosis, and every other detail, which helps in accounting for new cases and immediately checking the database for similar patients using the same Machine Learning techniques. The use of machine learning techniques cannot be precluded and thus has become increasingly popular, however in a field such as biomedical sciences which does not allow even the slightest of perturbations in the result, there is a need evaluate these techniques rigorously before using them commercially.

## REFERENCES


[1] American Cancer Society, *Breast Cancer – A detailed Guide* http://www.sas.com/en_us/insights/analytics/machine-learning.html, Accesed on March, 2016.

[2] UCI Machine Learning Repository, *Wisconsin Breast cancer data set* https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original), accesed on August, 2016.

[3] Gouda I. Salama1, M.B. Abdelhalim, and Magdy Abd-elghany Zeid, Breast Cancer Diagnosis on Three different datasets using Multi-classifiers, International Journal of Computer and Information Technology (2277 – 0764) Volume 01– Issue 01, September 2012.

[4] S. Kharya, D. Dubey, and S. Soni, Predictive Machine Learning Techniques for Breast Cancer Detection, International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013, 1023-1028.

[5] H. Yusuff, N. Mohamad, U.K. Ngah & A.S. Yahaya, Breast Cancer analysis using Logistic Regression, IJRRAS 10 (1), January 2012.

[6] I.S. Jacobs David A. Freedman (2009). *Statistical Models: Theory and Practice*. Cambridge University Press. p. 128.

[7] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning, Logistic Regression, Chapter 4, pg. 119.

[8] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". *The American Statistician*. **46** (3): 175–185.

[9] Cortes, C.; Vapnik, V. (1995). "Support-vector networks". *Machine Learning*. 20 (3): 273–297.

[10] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[11] William H. Wolberg and O.L. Mangasarian*: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.*

[12] O. L. Mangasarian, R. Setiono, and W.H. Wolberg*: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.*

[13] K. P. Bennett & O. L. Mangasarian*: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).*

[14] Keyvanfard F., M. A. Shoorehdeli, and M. Teshnehlab. 2011. Feature Selection and Classification of Breast Cancer on Dynamic Magnetic Resonance Imaging using ANN and SVM. American Journal of Biomedical Engineering. 1: 20–25.

[15] Subashini T. S., V. Ramalingam, and S. Palanivel. 2009. Breast Mass Classification Based on Cytological Patterns using RBFNN and SVM. Expert Systems with Applications. 36: 5284-5290.

[16] Murat C., E. Mehmet, Z. B. Erkan, and Y. A. Ziya. 2009. Early Prostate Cancer Diagnosis by using Artificial Neural Networks and Support Vector Machines. Expert Systems with Applications. 36: 6357–6361.

[17] Ren J. 2012. ANN vs. SVM: Which One Performs Better in Classification of MCCs in Mammogram Imaging. Knowledge-Based Systems. 26: 144–153.

[18] Vapnik, V. 1995. The Nature of Statistical Learning Theory. New York: Springer.