2018, IEEE Third International Conference on Circuits, Control, Communication and Computing

# *Using Machine Learning algorithms for breast cancer risk prediction and diagnosis*

Anusha Bharat
*Dept of Telecommunication Engineering*
*Ramaiah Institute of Technology*
Bangalore, India
anush.bharat@gmail.com

Pooja N
*Dept of Telecommunication Engineering*
*Ramaiah Institute of Technology*
Bangalore, India
pooja97.natarajan@gmail.com

R Anishka Reddy
*Dept of Telecommunication Engineering*
*Ramaiah Institute of Technology*
Bangalore, India
ranishka1997@gmail.com

***Abstract*—Machine learning is frequently used in medical applications such as detection of the type of cancerous cells. Breast cancer represents one of the diseases that causes a high number of deaths every year. It is the most common type of cancer and the main cause of women's deaths worldwide. The cancerous cells are classified as Benign (B) or Malignant (M). There are many algorithms for classification and prediction of breast cancer: Support Vector Machine (SVM), Decision Tree (CART), Naive Bayes (NB) and k Nearest Neighbours (kNN). In this project, Support Vector Machine (SVM) on the Wisconsin Breast Cancer dataset is used. The dataset is also trained with the other algorithms: KNN, Naives Bayes and CART and the accuracy of prediction for each algorithm is compared.**

***Keywords*—Breast Cancer, knn, naives bayes, CART, SVM**

## I. INTRODUCTION

Breast cancer is a type of cancer that occurs mostly in females and is the leading cause of women's deaths. These deaths can be reduced by early detection of the cancerous cells. Cancerous cells are detected by performing various tests like MRI, mammogram, ultrasound and biopsy. The dataset used in this project contains features that are computed from a digitized image of a fine needle aspiration (FNA) biopsy of a breast mass. They describe characteristics of the cell nuclei present in the image. Diagnosis of breast cancer is done by classifying the tumour. Tumours can be either benign or malignant. Malignant tumours are more harmful than the benign. Unfortunately, not all physicians are expert in distinguishing between the benign and malignant tumours and the classification of tumour cells may take up to 2 days. Machine learning algorithms are used to predict the type of cancerous cells efficiently and accurately. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The different algorithms used are: Support Vector Machine (SVM), Decision Tree (CART), Naive Bayes (NB) and k Nearest Neighbours (k-NN).

### *A. K – Nearest Neighbour (KNN)*

KNN makes predictions using the training dataset directly. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbours) and summarizing the output variable for those K instances. For regression this might be the mean output variable, in classification this might be the mode (or most common) class value. To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance. Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (xi) across all input attributes j.

Euclidean Distance(x, xi) = sqrt( sum( (xj – xij)^2 ) )

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point.

### *B. Naives Bayes*

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Bayes' Theorem is stated as: P(h|d) = (P(d|h) * P(h)) / P(d)

Naive Bayes is a classification algorithm for binary (twoclass) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. Assume that we have a dataset with two classes of data inside. We have an equation for the probability of a piece of data belonging to Class 1:p1(h,d), We have an equation for the class belonging to Class 2:p2(h,d). To classify a new measurement with features (h,d), we use the following rules:If p1(h,d) > p2(h,d), then the class is 1.If p2(h,d) > p1(h,d), then the class is 2.

### *C. Classification and Regression Trees (CART)*

A Classification and Regression Tree (CART), is a predictive model, which explains how an outcome variable's values can be predicted based on other values. A CART output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable. The representation for the CART model is a binary tree. Each root node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric). The leaf nodes of the tree contain an output variable (y) which is used to make a prediction.

### *D. Suppport Vector Machine (SVM)*

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well. Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples. It is worth noting that probabilistic outputs can also be obtained for SVMs figure below illustrates how an SVM might work in order to classify tumours among benign and malignant based on their size and patients' age. The identified hyperplane can be thought as a decision boundary between the two clusters. Obviously, the existence of a decision boundary allows for the detection of any misclassification produced by the method.

## II. LITERATURE SURVEY

From [1] the Wisconsin Breast Cancer dataset was obtained. In [2] Authors proposed a Support Vector Machines (SVMs) based classifier in comparison with Bayesian classifiers and Artificial Neural Networks for the prognosis and diagnosis of breast cancer disease. The paper provides the implementation details along with the corresponding results for all the assessed classifiers. A SVM model is implemented for the breast cancer diagnosis and prognosis problem using the Wisconsin Diagnostic Breast Cancer (WDBC) as well as the Wisconsin Prognostic Breast Cancer (WPBC) datasets. The optimized SVM algorithm performed excellently, exhibiting high values of accuracy (up to 96.91%), specificity (up 97.67%) and sensitivity (up to 97.84%).

[3] states that artificial neural network have been the most widely used predictive technique in medical prediction, though its structure is difficult to understand. The paper lists out the benefits and limitations among various machine learning techniques such as Decision trees, Naïve Bayes, neural network and SVM.

In [4], each algorithm performs in a different way depending on the dataset and the parameter selection. For overall methodology, KNN technique has given the best results. Naive Bayes and logistic regression have also performed well in diagnosis of breast cancer.But SVM is the most suited technique for recurrence/non-recurrence prediction of breast cancer.

## III. MATERIALS AND METHODOLOGY

Materials that we have used include: Python software for coding and breast cancer data from UCI depository. Our methodology involves use of machine learning techniques such as: SVM, KNN, decision trees and Naives bayes.

### *A. Dataset*

The Wisconsin Diagnostic Breast Cancer dataset was obtained from the UCI machine learning depository (available at: http://archive.ics.uci.edu/ml). The dataset contains 357 cases of benign breast cancer and 212 cases of malignant breast cancer. The dataset contains 32 columns, with the first column being the ID number, the second column being the diagnosis result (benign or malignant), followed by the mean, standard deviation and the mean of the worst measurements of ten features. There were no missing values. The features are obtained from a digitized image of a fine needle aspiration biopsy of the tumour. These features describe the nuclei of the cell. The different features are as shown:

TABLE I. FEATURES USED

| | |
|---|---|
| Radius | Mean of distances from centre to points on the perimeter |
| Texture | Standard deviation of grey-scale values |
| Perimeter | The total distance between the snake points constitutes the nuclear perimeter. |
| Area | Number of pixel on the interior of the snake and adding one-half of the pixel in the perimeter |
| Smoothness | Local variation in radius length, quantified by measuring the difference between the length of a radial line and the mean length of lines surrounding it. |
| Compactness | Perimeter ^2 / area |
| Concavity | Severity of concave portions of the contour |
| Concave points | Number of concave portions of the contour |
| Symmetry | The length difference between lines perpendicular to the major axis to the cell boundary in both directions. |
| Fractal dimension | Coastline approximation. A higher value corresponds to a less regular contour and thus to a higher probability of malignancy. |

### *B. Methodology*

The dataset is divided into training set and testing set. 80% of the data is used to train the system and the remaining 20% is used for testing. From the dataset, we analyse and build a model to predict if a given set of symptoms lead to breast cancer. The machine learning algorithms are trained on the training data, and tested on the untrained data. If the model is excessively complex, such as having too many parameters, it is likely to lead to the problem of overfitting. Likewise, if the model is excessively simple that cannot capture the underlying trend of the data, underfitting occurs. Both overfitting and underfitting lead to poor predictive performance. There are several techniques to overcome overfitting, such as crossvalidation, regularization and drop out.. One of the most commonly used methods is k-fold crossvalidation, where the original data is randomly partitioned into k equal sized subsamples. Out of the k subsamples, one subsample is used to testing the model, and the remaining k-1 subsamples are used to train the model. The k results are then averaged to generate one single estimation. One advantage of k-fold cross validation is each testing subsample is used exactly once. Support vector machine (SVM), a binary classifier, searches the hyperplane leaving the largest possible fraction of points of the same class on the same side, while maximizing the distance of each class from the hyperplane. SVMs are a more recent approach of ML methods applied in the field of cancer prediction/prognosis. Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples.

## IV. FLOWCHART

The data obtained from the patients is used to form a dataset. The dataset is divided into training and testing data, and it is ensured that the dataset has no incomplete values. The four machine learning algorithms are applied and the accuracy of prediction of each algorithm is compared. The algorithm with best accuracy is used for prediction.

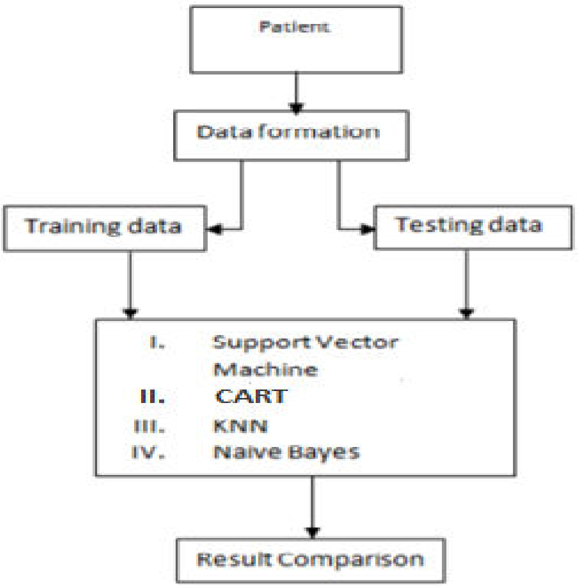

Fig 1. Flowchart

## V. RESULTS AND DISCUSSIONS

### *A. Data Exploration*

The distributions of the mean, standard error and worst average of the 10 features extracted from the fine needle aspiration slides show that compactness, concavity, fractal dimension, smoothness and symmetry each have relatively small values for the measurement. Perimeter, radius and texture each have relatively large values for the measurement, with areas that show the largest measurement value and amount of variation for all three measurements. From the distribution visualization, we can see overall the malignant diagnosis class has relatively higher mean for all the attributes.

### *B. Correlation*

Among the mean measurement of the 10 attributes, we can see several of them are highly correlated between each other. 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.

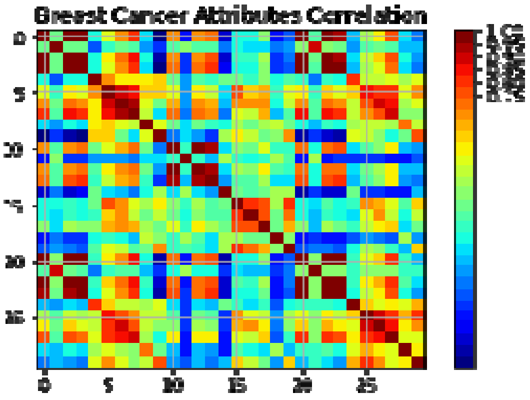

Fig 2. Correlation graph

### *C. Count of Benign(B) and Malignant(M):*

From the dataset we can observe that the number of BENIGN patients are more than the number of

MALIGNANT patients and this is proved by the bar graph below obtained in our jupyter notebook.
Malignant: 212
Benign:357

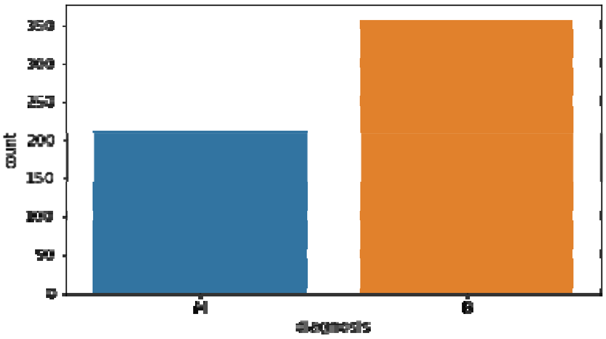

Fig 3. Count of patients diagnosed with benign and malignant cancer

### D. Performance Comparison

From the initial run, it looks like Gaussian NB, KNN and CART performed the best given the dataset (all above 92% mean accuracy). Support Vector Machine has a surprisingly bad performance here. However, if we standardise the input dataset, it's performance will improve.

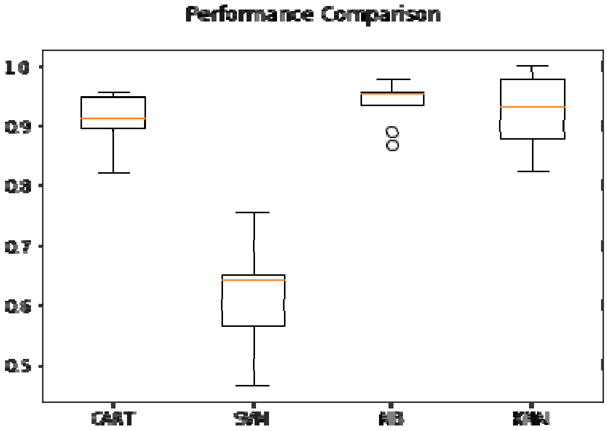

Fig 4. Box plot before standardising

After standardising the dataset, accuracy of SVM drastically improves as shown below

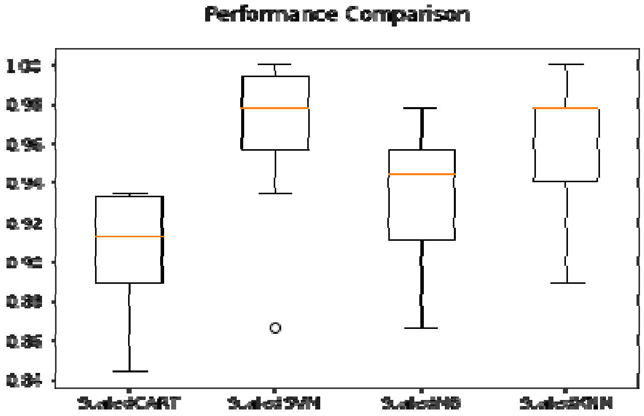

Fig 5. Box plot after standardising

### E. Calculation of Accuracy:

When we calculate accuracy we observe the output to be as shown below:
Accuracy score 0.991228

TABLE II. RESULTS

| Cancer type | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 1.00 | 0.99 | 0.99 | 75 |
| 1 | 0.97 | 1.00 | 0.99 | 39 |
| Average / Total | 0.99 | 0.99 | 0.99 | 114 |

### F. Confusion Matrix:

```
      M  B
 M [[74  1]
 B  [0 39]]
```

We can see that we achieve an accuracy of 99.1% 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.

## VI. CONCLUSION AND FUTURE SCOPE

Each algorithm performs in a different way depending on the dataset and the parameter selection. For overall methodology, KNN technique has given the best results. Naive Bayes and logistic regression have also performed well in diagnosis of breast cancer. SVM is a strong technique for predictive analysis and owing to the above finding, we conclude that SVM using Gaussian kernel is the most suited technique for recurrence/non-recurrence prediction of breast cancer.

The SVM that is used in the analysis in this paper is only applicable when the number of class variable is binary i.e. we can't have more than 2 classes. To solve this problem scientists have come up with multiclass SVM. Further research in this domain such as the creation of SVM classes like LIBSVM has taken place. Fine tuning of parameters used in algorithms can result in better accuracy. Furthermore, this can also be implemented on a cloud platform for ease of usage.

## REFERENCES

[1] UCI depository - http://archive.ics.uci.edu/ml/
[2] Maglogiannis, I., Zafiropoulos, E., & Anagnostopoulos - An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers, Applied intelligence journal, Volume 30,Issue 1, February 2009
[3] S.Kharya, D.Dubey, S.Soni -Predictive Machine Learning Techniques for Breast Cancer Detection, IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013, 1023-1028
[4] Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, Nikahat Kazi – Breast cancer Diagnosis and Recurrence prediction using Machine learning Techniques,IJRET- International Journal of Research in Engineering and Technology, April 2015
[5]National Breast Cancer Foundation Inc., http://www.nationalbreastcancer.org/about-breast-cancer.