# Using Machine Learning Algorithms For Breast Cancer Risk Detection And Diagnosis

**Sumitha BS[1], Kavalakuntla Chakradhar[2]**
[1, 2] Dept of Computer Science and Engineering
[1, 2] Atria Institute of Technology, Bengaluru, Karnataka, India

***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), Random forest 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: Random forest, KNN, Naives Bayes and CART and the accuracy of prediction for each algorithm is compared.*

***Keywords***- Breast Cancer, Random forest, 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. A mammogram is an X-ray of the breast. It is a medical technique used for the detection of breast cancer in women without any side effects deeming the procedure as safe. Women who get regular mammograms have a higher survival rate as compared to women who do not. It is recommended by the NBCF (National Breast Cancer Foundation) that women over the age of forty years of age should get a mammogram once a year. 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. 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.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).

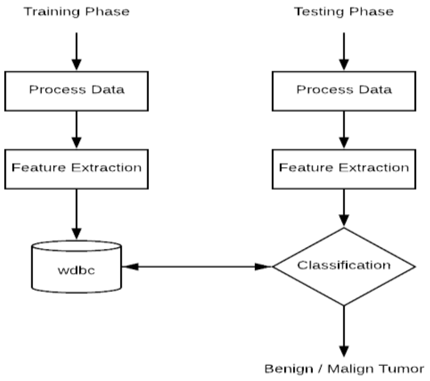


Figure 1: Proposed Breast Cancer Detection Model

## II. RELATED WORK IN BREAST CANCER

Breast cancer detection using Relevance Vector Machine [3], obtained an accuracy of 97% using Wisconsin original dataset which has 699 instances and 11 attributes, while [4] allots distinct weights to different attributes with regard to their capabilities of prediction and yielded an accuracy of 92% working with the weighted naïve bayes method. [5] built a hybrid classifier of Support Vector Machines and decision trees in WEKA and obtained an accuracy of 91%. [6] used Linear Discriminant Analysis for feature selection and trained the dataset by using one of the fuzzy inference method called Mamdani Fuzzy inference model and obtained an accuracy of 93%. Various differentiation between multiple techniques has been provided through this manuscript[7] like Bayes Network, Pruned Tree, kNN algorithm using WEKA on breast cancer dataset, it has a total of 6291 data and a dimension of 699 rows and 9 columns. The highest accuracy is 89.71% which belongs to bayes network.[11][12][13]. 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%).SVM is the most suited technique for recurrence/non-recurrence prediction of breast cancer.

*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.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. Given N training vectors in the Figure 3, kNN algorithm identifies the k nearest neighbors of regardless of labels.

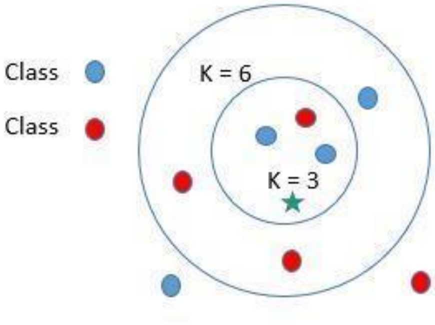
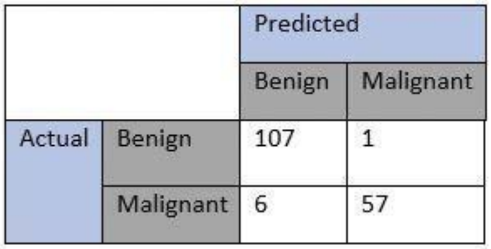


Figure 2: kNN Illustration

The accuracy of kNN is found to be 95.90% , there is only one observation that is misclassified as Benign and four observations are misclassified as Malignant as represented in Table 2. The results are comparatively better than Random Forest algorithm.



| | | Predicted | |
|---|---|---|---|
| | | Benign | Malignant |
| Actual | Benign | 107 | 1 |
| | Malignant | 6 | 57 |

Table 1: kNN Confusion Matrix

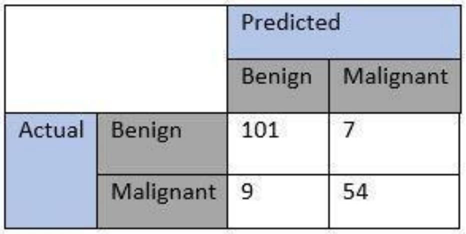
*B. Naives Bayes*

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.

There are sixteen misclassified observations, seven of them being benign and nine of them are malignant. The same 398 observations are used for training set and 171 observations for testing and the accuracy equals to 94.47%.



| | | Predicted | |
|---|---|---|---|
| | | Benign | Malignant |
| Actual | Benign | 101 | 7 |
| | Malignant | 9 | 54 |

Table 2: Naïve Bayes Confusion Matrix

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. Support 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.

*E. Random Forest*

It is a supervised learning algorithm. An ensemble of decision trees is created, the bagging method is used to train the system. The confusion matrix of random forest is quite promising. There are only five observations that are misclassified as Benign and four observations are misclassified as Malignant and the accuracy equals 94.74%. The ground methodology on which this technique is based is recursion. A random sample of size N is picked from the data set in each instance of an iteration.

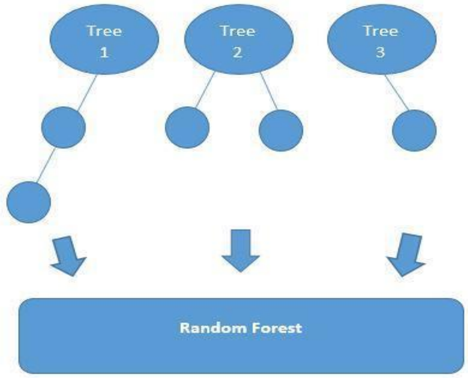
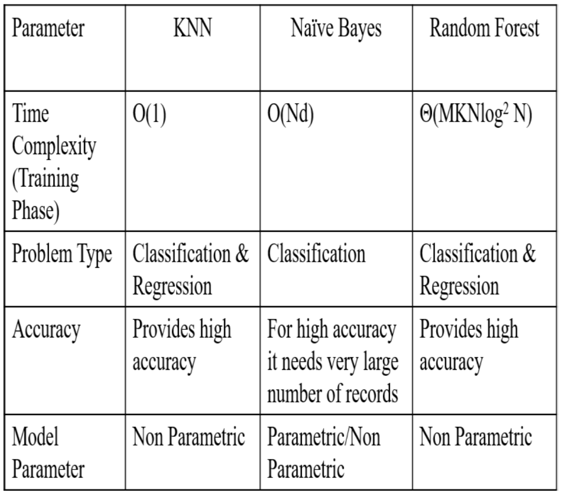


Figure 3: How Random Forest Works

The dataset has been divided into training and testing sets, there are 398 observations for training set and 171 observations for testing. The number of estimators are set to 72 thus it is ensured that every observation is predicted at least a few times. It is obvious that diagnosis, radius_mean, texture_mean, perimeter_mean are influential variables, the other variables are of moderate influence but none of them can be neglected to increase the model accuracy.

*F. Comparison Among Proposed Algorithms*

Each one of the three algorithm's – kNN, Naïve Bayes and Random Forest have their advantage and disadvantage over each other in terms of performance, the type of problem they handle etc. As shown in Table 4: kNN test time is O(1) without preprocessing of training set [8], in the case of Naïve Bayes: N is the number of training examples and d is the dimensionality of the features whereas for Random Forest [9]: N is the number of samples and K is the number of variables randomly drawn at each node. Naïve Bayes algorithm deal only with classification problems whereas both kNN and Random Forest can deal with classification as well as regression problems. In terms of accuracy both kNN and Random Forest can deliver high accuracy but Naïve Bayes algorithm need large number of records in order to yield a better accuracy. Algorithms that simplify the function to a known form are called parametric machine learning algorithms, Naïve Bayes algorithm can be expressed as parametric as well as non-parametric model.



| Parameter | KNN | Naïve Bayes | Random Forest |
|---|---|---|---|
| Time Complexity (Training Phase) | O(1) | O(Nd) | $\Theta(MKN\log^2 N)$ |
| Problem Type | Classification & Regression | Classification | Classification & Regression |
| Accuracy | Provides high accuracy | For high accuracy it needs very large number of records | Provides high accuracy |
| Model Parameter | Non Parametric | Parametric/Non Parametric | Non Parametric |

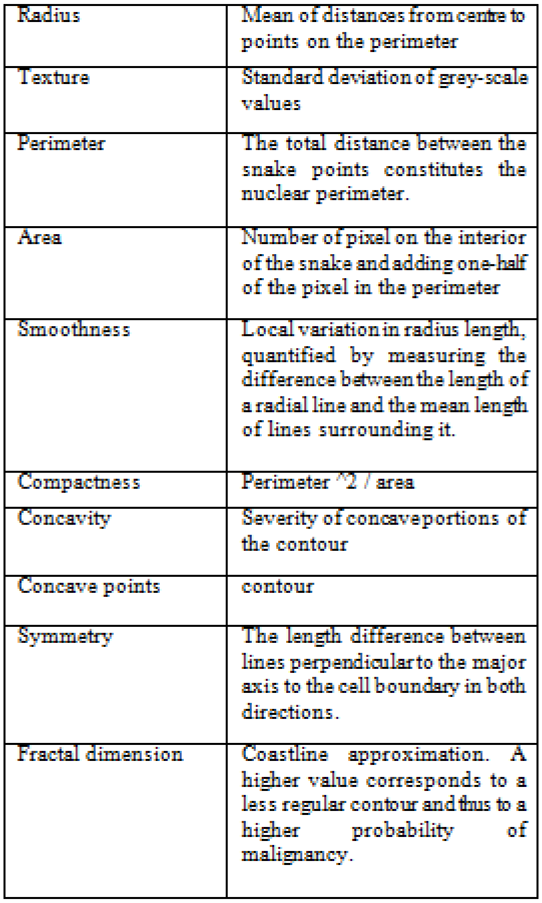
Table 3: Comparison among kNN, Naïve Bayes and Random Forest

## 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 ,Naives bayes and Random forest.

### 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:



| | |
|---|---|
| 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 concaveportions of the contour |
| Concave points | 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.

A confusion matrix for actual and predicted class is derived comprising of the standard five values namely TruePositive, FalsePositive, TrueNegative and FalseNegative to evaluate the performance.

*1. Accuracy*

Accuracy is a good predictor for the degree of correctness in the training of the model and how it may perform generally. It may be defined as the measure of the correct prediction in correspondence to the wrong ones. Thus the equation presented can be used to calculate the value of accuracy:

Accuracy = (TruePositive + TrueNegetive) /*(*TruePositive + FalsePositive + TrueNegative + False Negative*)*

*2. Recall*

Recall known as sensitivity in general terms, may be defined as the ratio of rightfully determined positive instances to the all observations. Recall may be seen as a measure for the effectiveness of the system in predicting positives and determining costs.

*Recall = TruePositive/ (TruePositive + FalseNegetive)*

*3. Precision*

The degree of correctness in determining the positive outcomes may be defined as precision. It is basically the ratio between true positives and the overall set of positives. This depicts the handling capacity of the system for positive values but does not provide insight into the negative values.

*Precision = TP /(TP + FP)*

*4. F1 Score*

It is the weighted average of Precision and Recall. This measure hence, considers both type of false values. F1 score is considered perfect when at 1 and is a total failure when at 0.

*F1 Score = 2*(Precision*Recall)/ (Precision + Recall)*

## 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.

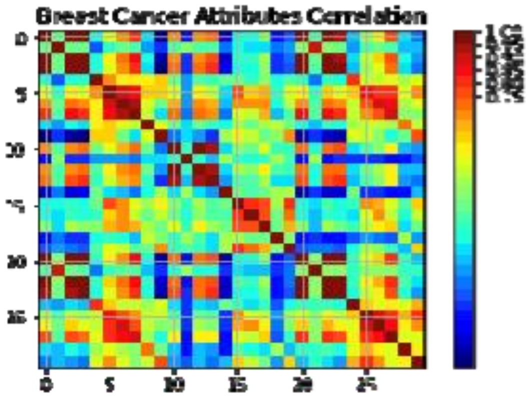
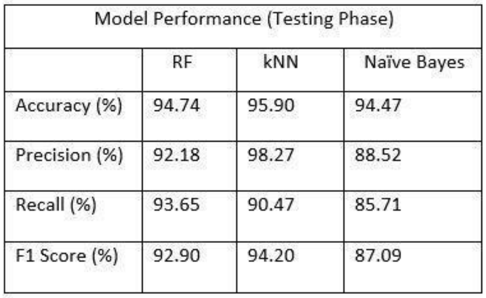


Fig 4. Correlation graph

*C. Performance Comparision*



| Model Performance (Testing Phase) | | | |
|---|---|---|---|
| | RF | kNN | Naïve Bayes |
| Accuracy (%) | 94.74 | 95.90 | 94.47 |
| Precision (%) | 92.18 | 98.27 | 88.52 |
| Recall (%) | 93.65 | 90.47 | 85.71 |
| F1 Score (%) | 92.90 | 94.20 | 87.09 |

Table 4: Performance Measure Indices

A comparative study using Random Forest, kNN (k-Nearest-Neighbor) and Naïve Bayes algorithm which are implemented in a computer having configuration as Intel Core i7 with 16GigaBits RAM has been proposed. We have used numpy, pandas and Scikit-learn which are open source machine learning libraries in Python. An open source web application named as Jupyter Notebook is used to run the program. The classifier was tested using the k − fold cross validation

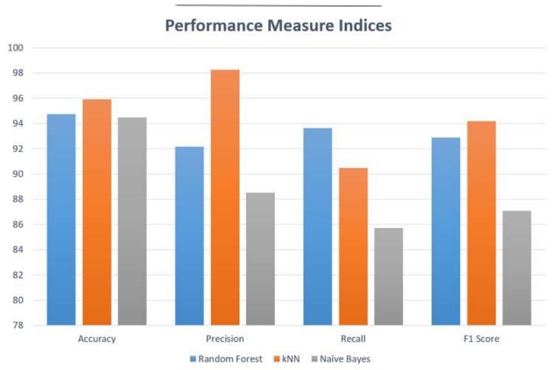


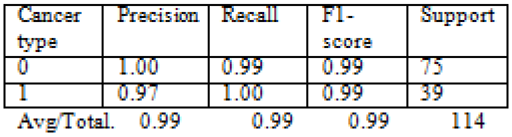
Figure 5: Graphical representation of Performance Measure Indices

*D. 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 |
| Avg/Total. | 0.99 | 0.99 | 0.99 | 114 |

*E. 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] National Institute of Cancer Prevention and Research, cancer statistics [Online], Available: http://cancerindia.org.in/statistics/

[2] WHO breast cancer statistics [Online]. Available : http://www.who.int/cancer/prevention/diagnosis-screening/breast- cancer/en/

[3] B.M. Gayathri, Dr. C.P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer" 2016

[4] S Kharya and S Soni,"Weighted Naïve Bayes classifier – Predictive model for breast cancer detection", January 2016

[5] Sivakami, "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model" 2015

[6] B.M.Gayathri and C.P.Sumathi,"Mamdani fuzzy inference system for breast cancer risk detection", 2015.

[7] Mohd,F.,Thomas,M, "Comparison of different classification techniques using WEKA for Breast cancer" 2007.

[8] Time complexity and optimality of kNN [Online] Available: https://nlp.stanford.edu/IR-book/html/htmledition/time-complexity-and-optimality-of-knn-1.html

[9] Gilles Louppe, “Understanding Random Forests from theory to practice” 2015.

[10] U.S. Breast Cancer Statistics [Online] Available:https://www.breastcancer.org/symptoms/understand_bc/statistics

[11] T Choudhury, V Kumar, D Nigam ,An Innovative Smart Soft Computing Methodology towards Disease (Cancer, Heart Disease, Arthritis) Detection in an Earlier Stage and in a Smarter Way- International Journal of Computer Science and Mobile Communication (IJCSMC) 2014.

[12] T Choudhury, V Kumar, D Nigam, B Mandal ,Intelligent classification of lung & oral cancer through diverse data mining algorithms, International Conference on Micro-Electronics and Telecommunication Engineering 2016

[13] T Choudhury, V Kumar, D Nigam,Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm - International Journal of Advanced Research in Computer Science and Software Engineering, 2015

[14] Shubham Sharma, Archit Agarwal, Tanupriya Choudary – Breast Cancer Detection Using Machine Learning Algorithms - Computational Techniques , Electronics and Mechanical Systems (CTEMS), 2018

[15] Anusha Bharat, Pooja N,R Anishka Reddy -Using Machine Learning algorithms for breast cancer risk prediction and diagnosis -IEEE, 2018