Breast cancer prediction techniques : A review

Anjaly Nelson[1], Bony M Sunny[2], Jibymol Joseph[3] and Ms. Shelly Shiju George[4]

[123] PG Scholars,Department of MCA, AmalJyothiCollege of Engineering
[4]Assistant Professor, Department of MCA, AmalJyothi College of Engineering

Abstract:Breast cancer is the second leading cause of cancer death among women. The mortality rate due to cancer can be reduced if it can be detected at a relatively early stage. Big data analytics is the process of collecting and organizing large sets of data and analyzing it to find useful information. This paper reviews different algorithms for predicting breast cancer. The review focuses on the various algorithms and datasets used in the prediction.

Keywords: **Big Data, Breast cancer, Prediction, K-NN algorithm, Eucledian distance.**

I.INTRODUCTION

Cancer is caused due to uncontrolled growth of the cells which can spread to other parts of the body. Cancer can be of different types, among them breast cancer is the most commonly diagnosed cancer in women. It is also the second leading cause of cancer deaths in women. Breast cancer occurs when cells in the breast begin to grow out of control. These cells usually form a tumor. Early detection and diagnosis can help to a greater extent in reducing the mortality rate [8]. The primary test for breast cancer is mammography. If anomalies are found, suggestions for the next level of diagnosis are made. Fine Needle Aspiration (FNA) is a type of biopsy that is suggested if abnormalities were found after the mammography. In this study, we are using the results of FNA for prediction purpose, which is obtained from various datasets[7].

Prediction is a form of data analysis that is used to predict future data trends. With the latest technologies the prediction and diagnosis of cancer arepossible.

II.LITERATUREREVIEW

Shailaja et al., [1] uses the K nearest neighbor (KNN) algorithm for classification of breast cancer tumors as benign and malignant. It is implemented using the R tool. The dataset used in the study is the Wisconsin Breast Cancer (original) Dataset (WBCD) from UCI machinelearning repository. The authors concluded that the precision, accuracy, recall and f-measure were increased compared to other models and it is found to be 97.65% accurate in prediction.

Sivakami et al., [2] have utilized Decision Tree and Support Vector Machine (DT-SVM), which are hybrid methods used for prediction. WEKA tool was used for performing the experiment and Wisconsin Breast Cancer Dataset (WBCD) was used in the study. The algorithm provided 91% accurate prediction.

Lavanya et al., [3] used a hybrid approach consisting of CART classifier with feature selection and bagging technique to evaluate the performance and accuracy for classification using different datasets. They used the WBDC (original), WBDC (diagnostic) and WDBC datasets for comparison. The experiment showed that WBDC (original) dataset showed the best result of 97% accuracy

Sankari et al., [4] performed prediction of breast cancer using a hybrid of Logistic Regression and Random Forest algorithms. The dataset used is the Wisconsin Breast Cancer Dataset (WBCD). Logistic Regression helped in building interpretable models for breast cancer prediction and Random Forest acted as a learning method for classification.

Sindhu et al., [5] compared the use the decision tree and Naïve Bayes algorithm in the prediction of breast cancer. It was found that Naïve Bayes is less accurate but take lesser time compared to decision tree. Thus the research concluded that decision tree is better for theprediction.

Nalini et al., [6] compared the performance of J48 and Naïve Bayes algorithm in the prediction of breast cancer. The study showed that Naïve Bayes provided better accuracy of 64% while J48 was 60% accurate. In terms of execution time Naïve Bayes took less time compared to J48.Thus it was concluded that J48 decision trees has a better accuracy than the Naïve Bayesmodel.

III.IMPLEMENTATION

The Nearest Neighbor (NN) algorithm is one of the decision procedures which is used for classification. It assigns a sample the class label of its closest neighbor. But this has limitations when it comes to large samples. An extension of this algorithm is the K-Nearest Neighbor (K-NN) algorithm. It takes the class labels of „k" nearest neighbors of the sample and assigns the label which occur the most to the sample. When k=1, it is considered as Nearest Neighbor algorithm.

K-NN is a supervised learning algorithm. When a new data point is given, K-NN calculates the distance from this point to all other points in the training dataset. A value, called the

„k value" is selected. The accuracy of prediction or classification depends on this value. Depending on the k value, the majority of class label in the points with k minimum distance is chosen as the final classifier for the new data point.

The research focuses on the prediction of breast cancer using KNN algorithm. The dataset used is Wisconsin Breast Cancer Dataset (WBCD) from the UCI machine learning repository and it is taken as the training set. The data inputted by the user is taken as the testset

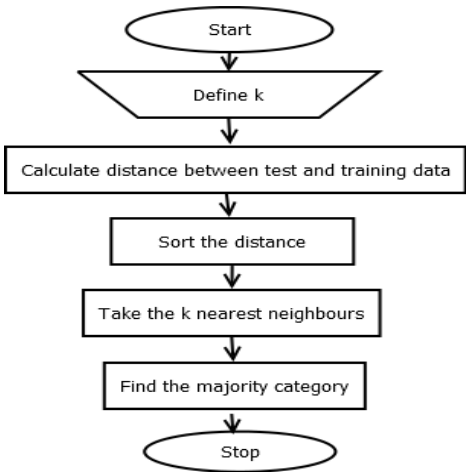

Figure 1: Flowchart of K-NN algorithm

id	diagnosis	radius_mean	texture_mean	perimeter_mean
842302	M	17.99	10.38	122.8
842517	M	20.57	17.77	132.9
84300903	M	19.69	21.25	130
84348301	M	11.42	20.38	77.58

Figure 2: Sample of dataset

IV.METHOD OFIMPLEMENTATION

The research is implemented using PHP. Euclidean distance is used for the distance calculation between test and training data. The „k" value selected is the square root of number of observations. The steps followed in the implementation are given below:

1. Determine the parameter k
2. Calculate the Euclidean distance between input data and trainingdata.

```
for ($i = 0; $i< $n; $i++)

{           $sum =0;
for($j =0; $j <8;$j++)
{
$sum += ($rows[$i][2] - $instance[$j]) *
($rows[$i][2] - $instance[$j]);
}

$distance[$i] = sqrt($sum);
}
```

3. Sort the distance and determine the nearest neighbors based on k minimumdistances.

```
asort($distance);
$x=1;
foreach($distance as $key => $d)
{ if(++$x <= $k)
{

$kclosest[$rows[$key][1]]++;
}
}
```

4. Use the majority of the class labels of nearest neighbours as the prediction value of the inputinstance.

```
arsort($kclosest)
$diagnosis = key($kclosest);
```

V.RESULT

The experiment used only the first 8 attributes and first 100 rows of the WDBC dataset. The dataset is taken from University of Wisconsin Hospitals, Madison from Dr. William. H. Wolberg. The data in the dataset is obtained from digitized image of a fine needle aspirate of breast. The study provided approximately 85% accurate prediction.

VI.CONCLUSION

The paper presented a review of different algorithms used for predicting breast cancer and an implementation of K-NN algorithm. Each algorithm used a different technique for classification purpose. The dataset used also varied in different studies. In terms of implementation, the experiment using KNN is comparatively easy. In terms of accuracy in prediction, KNN gave the most accurate prediction(97.65%).

VII.FUTUREWORK

In the future, predictions of other types of cancer using different algorithms are planned. Machine learning and artificial intelligence promise yet more efficient and accurate methods for prediction, which could also be subjects for research in the comingfuture.

REFERENCES

[1] K. Shailaja, B.Seetharamulu, M.A.Jabbar, "Prediction of Breast Cancer Using Big Data Analytics", International Journal of Engineering and Technology, Volume 7,2018

[2] K.Sivakami, "Mining Big Data: Breast ancer Prediction using DT-SVM Hybrid Model", International Journal of Scientific Engineering and AppliedScience, Volume 1, 2015

[3] D.Lavanya, Dr.K.Usha Rani, "Ensemble decision tree classifier for breast cancer data", International Journal of Information Technology Convergence and Services (IJITCS), Volume 2, No.1,2012.

[6]Ms. L.Sankari, Mr.R.Rajbharath, "Predicting Breast Cancer using Novel Approach in Data Analytics", International Journal of Engineering Research and Technology, volume 6,2017

[5]Sindhu A M, Saleema J S, "Prediction of BreastCancer",InternationalJournalofResearchinEngineering, IT and Social Sciences, Volume 7, 2017

[6] Dr. C Nalini, D.Meera, "Breast cancer prediction system using Data mining methods", International Journal of Pure and Applied Mathematics, Volume 119, No: 12, 2018

[7]Mulazim Hussain Bukhari, MadihaArshad,Shahid Jamal,ShahidaNiazi, ShahidBashir, Irfan M. Bakhshi,Shaharyar, "Use of Fine-Needle Aspiration in the Evaluation of Breast Lumps", Pathology Research International, Volume2017

[8]Peter Adebayo Idowu, KehindeOladipo Williams, Jeremiah AdemolaBalogun and AdeniranIsholaOluwaranti, "Breast Cancer Risk Prediction Using Data Mining Classification Techniques", Volume 3, Issue2