ISSN No:-2456-2165

Role of Machine Learning in Diagnosis of Breast Cancer

Alqua Anjum
Department of Computer Science and Engineering, SVIET,
Banur, Chandigarh

Abstract: - Breast cancer has become the second main cause of cancer deaths in the women, nowadays, after lung cancer. It has become very common and is spreading very quickly and undesirably. Though there are a number of medical facilities available to treat and diagnose breast cancer but still it is very long process and takes time over its diagnoses which just increase the probability of cancer spreading more quickly. For this reason scientists decided to develop a computer based system that could be used for diagnosis of breast cancer, so that the time taken would be reduced and eventually the rate of spreading can be reduced. This paper provides the overview of survey of breast cancer using various machine learning algorithms. This will help us to know a lot about the work done on this system and eventually will help the researchers to improve the system.

Keywords: *- Machine Learning Approach, Neural Nets, Breast Cancer*

I. INTRODUCTION

Breast cancer has become a serious problem in the world today, affecting the health of women all around the globe very badly. From last decades, it has been seen this cancer is spreading more quickly than the attempts made to control it. Death rates due to breast cancer has also increased and main cause for this is late or time-taking diagnosis process. According to a survey in US, on increasing rates of breast cancer, it has been found that each year about 252,710 new cases of breast cancer are being reported and about 40,610 women die[1]. According to some other report provided by WHO, each year around 1.5 women are affected by invasive breast cancer [2].

When the survey was carried out in Asia it was observed that in country Pakistan most of the cases of breast cancer are reported. About 90,000 cases are reported each year there [3].

Because of the fact that the breast cancer are not easily detectable in early stages, its diagnostic process becomes very long and takes a lot of time. For this reason also it has become so invasive. Medical diagnosis include mammography and Biopsies. If abnormalities are found in mammography of a patient has to go through Biopsy, which is very painful, time-taking and costly.

An automation system would be very beneficial here, in this scenario. The analysis using the automated system can help the radiologists or experts to enhance the diagnosis accuracy. An when an automated system is provided with a proper dataset, it can perform so well that a patient need not to go through the painful biopsies and the other tests. This paper gives an overview of some major works done machine learning over diagnosis of Breast cancer. This paper will provide the different works done in ANN and various classification algorithms, which helped in developing an automation system to predict the breast cancer. This paper will help the future researchers to find out the important information regarding to the any research in this field.

II. DATA MINING TECHNIQUES USED FOR BREAST CANCER ANALYSIS

Data mining & Knowledge Discovery of Data of (KDD) are extracting novel, understandable, and useful information, knowledge or patterns from huge amounts of available data[4]. As the data collected from the different sources is in very large volumes with the huge noise along with it , it becomes very difficult to analyze this data and get results with satisfying accuracy rates. Her the data mining techniques can be used which have the capabilities to analyze the large sets of data, finding hidden relationships between various features of these data and extracting satisfying results. For example in the health centers doctors observe the various symptoms of any disorder in a patient and then analyze that obtained information and try to find the relation , after that, according to which they predict the disease. The problem with this manual method is that it is slow, and expensive and also the patient has to go through a series of tests and biopsies.

Data mining involves the following steps;

- Selection of data
- Preprocessing: It involves the removing of columns or rows with the missing data
- Transformation: It includes normalization of data and selecting only the important features among all the features chosen, on basis of which prediction could be made.
- Predictive Tasks: involves using of various classification and association rules of datmining to predict results on the basis of features.
- Interpretation/Evaluation: includes statistical

validation , qualitative review and accuracy review.

III. SOME RESEARCH WORKS APPLYING THE DATA MINING TECHNIQUES FOR BREAST CANCER DIAGNOSIS

- "Vikas Chaurasia [5] et al" used the three popular data mining algorithms (Naïve Bayes, RBF network,j48) to develop the prediction models in order to predict the breast cancer. They used a large dataset with 683 instance which they obtained from UCI machine learning repository. In order to measure the accuracy free from any bias they used the 10-fold cross validation procedure. In obtained results they found that Naïve Bayes comes with largest accuracy of 97.36%. RBF came out as second best with 96.77% of classification accuracy and then j48 which has the accuracy of 93.41%.
- "D.Lavanya and Dr.K.Usha Rani [6]" Showed in their study that for the effective classification with better accuracy results, the selection of features is very important as it eliminates the useless and irrelevant attributes. In their paper they showed the performance and accuracy of CART algorithm with and without the feature selection and concluded with the fact that proper selection of features help in gaining the more accurate results.
- "Shweta Kharya [7]" presented a report over the survey of various classification algorithms made to classify breast cancers. In her paper she presented the performances and accuracy rates of different techniques including Naive Bayes, Decision Trees, logistic regression, support vector machines, association rule and ANN and she find out the technique with best accuracy for the classification is Decision Trees with the accuracy of 93.62%.
- "M. Tahmooresi[8] et al" used a hybrid model in their research which comprises of different techniques including SVM, ANN, K-nearest neighbor, and decision tree for effective detection of breast cancer. In this study it can be found that SVM(single or hybrid) has achieved the greatest accuracy among all with rates of accuracy as 99.8% which can be improved to 100% .In that study it was also observed that using an optional ANN on MRI can also increase the accuracy upto 100%.
- "Jabeen Sultana and Abdul Khader Jilani [9"] has proposed their work in which they compare the efficiency of
- 10 different classifiers including Simple Logistic regression, MLP, Multi-class classifiers ,DT trees, REP tree, K-star, IBK, decision table, PART, and random forest. Among these all Simple Logistic Regression was found with the best accuracy.
- "Peter Adebayo Idowu, Kehinde Oladipo Williams[10] et al" carried out their work to predict breast cancer risks using two data mining classification techniques, naïve bayes and j48 decision tree. Using their experimental model it can be observed j48 decision tree has more accuracy than the naïve bayes. It showed j48 decision tree is better model in predicting breast cancer risks than the naïve bayes algorithm in terms of not only accuracy but also has better values of recall, precision, and error rates among the two.
- "Roulan Xu and Qiongjia Xu [11]" applied five machine learning techniques Logistic Regression, Naive Bayes, Linear SVC, SVM with linear kernel and Random Forest and three feature selection techniques PCA, RFE and Heat map for breast cancer diagnosis. They concluded that the Random forest algorithm and SVM with the linear kernel yield the results which are more accurate than the other models included in the experiment. The accuracy of Random Forest was 98% and for SVM with linear kernel it was 97%.
- "Y.Ireaneus Anna Rejani and Dr.S. Thamarai Selvi[12]" presented a paper in which they had used SVM classification algorithm to classify breast cancers. First they used the mammogram images to detect the breast cancer. These images were filtered using Gaussian filter based on standard deviation and matrix dimensions such as rows an columns. Then filtered image is used for contrast stretching and the background of image is eliminated using top hat operation. The top hat output is decomposed into two scales and then JaDWT (Discrete Wavelet Transform) is used for reconstruction to create a reconstructed image on which then segmentation is performed using Thresholding method. Then from a segmented tumor area the features are extracted and classified. The classification is done using SVM.
- "Hiba Asri, Hajar Mousannif [13]" et al have written an article over their work, where they used where they showed the comparison among the four machine learning algorithms used to predict the breast cancer. The algorithms included SVM, decision tree(C4.5), naïve bayes, and k- nearest neighbor over Wisconsin breast cancer (original) dataset. Experimental results showed SVM gives the highest accuracy of 97.13% with lowest error rate.
- "Jaimini Majali[14]" presented the system for diagnosis and prognosis of breast cancers using FP growth algorithm. The researcher first compared the results of FP growth with the results of various other classifying techniques like Bayesian classifier, decision tree and neural networks. FP growth was used to recognize the type of breast cancer and DT was used to predict possibility of breast cancer.

ISSN No:-2456-2165

IV. NEURAL NETWORKS AND ITS ROLE IN ANALYSIS OF DATA

A neural network is a complex algorithm that is inspired from the structure of a biological human brain. It is developed in a way that it could mimic the decision-making ability of human brain. Though human brain is more complex as compared to this artificial neural network but still it is able to extract the patterns and meaningful information from huge complex data, which is very difficult to be recognized and detected in other ways. A neural network develops an ability within itself with help of which it can detect patterns in input data and predict an output data free of any noise. The structure of neural network algorithm consist of 3- layers which are given as;

- The Input Layer: This layer takes the data items from outside environment and feeds it to the next following layer, that's called hidden layer.
- The Hidden Layer: This layer consists of certain complex functions according to which input data is processed and predictions/results/decisions are made. These function are not visible to users but are just embedded in an algorithm used.
- The Output layer: The decisions made in the hidden layers are collected here and final decisions are made which eventually provide the final predictions.

Figure 1, shows more clearly how the artificial neural network work, is given as follows:

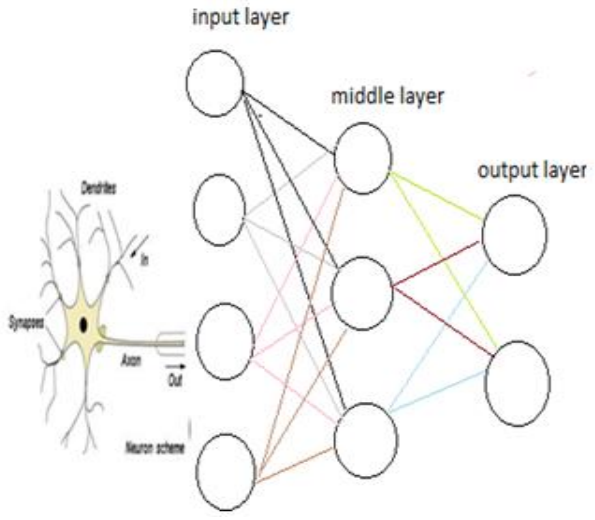

Fig 1:- Neural Net Representation

Neural networks provide the results with very high accuracy even if the data is very noisy. However it has a disadvantage that accuracy of predictions made is valid only for a specific period of time in which data was collected. A lot of the work over the analysis of breast cancer related data and prediction of different types of tumors, using ANN algorithms has been done, with a satisfying accuracy. Some of the major works will be given in another Section of this paper.

V. RELATED WORKS IN NEURAL NETWORKS IN PREDICTION OF BREAST CANCERS

- "Chandra Prasetyo Utomo, Aan Kardiana, Rika Yuliwulandari" [15] have presented a paper in which they have implemented ANN with the extreme learning for diagnosis of Breast Cancer. In their work they have used Wisconsin dataset. They later have compared the results of Extreme Learning techniques with the Back Propagation algorithms and find out that Extreme Learning Machine Neural Network(ELM NN) is better classifier than BP ANN. Although the specificity rates of BP ANN were found better than the later, but still from the results it could be observed that accuracy rates and sensitivity of ELM NN were much improved.
- "Mihir Borkar, Prof. Khushali Deulkar, and Abhinav Garg"[16] , in their study developed an artificial neural network that was able to predict weather a patient has the breast cancer or not that is weather the patient has malignant tumor or benign tumor. The attributes used in dataset were taken from the cell of patients and total 699 patients were the sources of data. 599 instances were created to train the network and the final network provided the accuracy of 96.49%.
- "Larrisa Westerdijk along with prof. Dr. Sandjai Bhulai" [17] in their research have used five machine learning models to predict the benign and malignant breast cancers. The 5 models included logistic regression, random forest, support vector machine, neural networks and ensemble models. The performances and accuracies of these all models were found. The accuracy results were 97.35%, 97.35%, 98.23%, 97.35% and 98.23%.
- "Xin Yao and Yong Liu" [18] have proposed their work where they used two neural network based approaches to diagnose the breast cancer. The two approaches used were Evolutionary based and Ensemble based approach. It was found out that Evolutionary approach can be used to design compact neural networks automatically by evolving network architectures and weights and the Ensemble approach could tackle with the large problems.
- "Harsh Vaizirani, Rahul Kala" [19] et al used the Modular Neural Network for diagnosis of Breast Cancer. They used the two different neural network models, over four modules to solve problem. They used BPNN and RBFN as two models to train and test the 4 different modules. Module 1 and 3 were trained and tested using BPNN and 2 and 4 using RBFN. Training and Testing accuracies for BPNN for module 1 and 3 were found as 89.50% &96.44% (for module1) and 91.50% & 94.67%(for module3)respectively. Training and testing accuracies for RBFN for module 2 and 4 were found as 94.75% &96.44% (for module2) and 97.50% & 97.63% (for module4)respectively.

- "D. Thuthi Sarabal and Dr. K. Arthi" [20] proposed their paper over the work of diagnosis of breast cancer, in which they used the data provided by the Adayar Cancer Institute Chennai. They used the improved Fuzzy Cognitive Mapping to classify trained features i.e; they used Cat Swarm Optimization based learning neural network for feature classification. Final classification of cancer (malignant or benign) is performed using fuzzy If and Then rule based system. Performances are found by calculating accuracy, sensitivity and specificity metrics.
- "S. Swathi, S.Rizwana" [21] et al in their research work have studied the performances of different neural network structures including Radial Bias Functions(RBF), General Regression Neural Network (GRNN), Probabilistic Neural Network(PNN), Multi-layer Perceptron Model and Back Propagation Neural network(BPNN) on Wisconsin Breast Cancer Data. After comparing the performances of classifying the different tumors, it has been found that BPNN have good diagnostic performance than the rest that is 99.28%.
- "Dr. A.Santhakumaran and F. Paulin" [22] have applied the feed forward neural network architecture for the classification of breast cancer and the Backpropagation algorithm has been used to train the network. Missing values in a dataset are replaced using median missing value replacement method and min-max normalization method is also used to normalize the datset. The diagnostic performance calculated was 99.28%.
- "Htet Thazin Tike Thein and Khin Mo Mo Tun" [23] in their research tried to propose an effective tool for building neural networks that could later help in proper classification of breast cancers. In their research they constructed a feed forward network and used Island Differential Evolution propagation algorithm to train the network. Through the result analysis it was found that computing time is reducing and solution quality improved significantly.
- "Bibhuprasad Sahu, Sachi Nandan Mohanty" [24] et al in their study have combined the artificial intelligence based technique with multivariate statistical method in order to develop a prediction model to predict the breast cancer. The features were selected here, also with a hybrid feature selection method that used Pricipal Component Analysis(PCA) along with ANN. Dataset used was Wisconsin Breast Cancer Dataset. The proposed algorithm was tested using 10-fold cross validation method during classification and its results were also compared to different classifier algorithms on the same database and it was seen that proposed algorithm gained much better accuracy.

VI. CONCLUSION

In this survey the performances of the various machine learning classifiers in diagnosis of Breast cancer and the performances of Neural Networks in the same subject has been summarized. Some of the interesting works with so good accuracy rates have been proposed in this paper. This paper may guide and help the young researchers to easily get an idea of how much work has been done in the field of diagnosis of tumors using machine learning concepts ,so that then they could simply proceed and try to develop a system with much satisfying results .

REFERENCES

[1]. Seigel RL,Miller KD,Jenal A Cancer
[2]. http://www.who.int/prevention /diagnosis-screening/breast cancer/
[3]. www.dawn.com/news/1344915
[4]. Vikas Chaurasia,Surabh Pal and BB Tiwari, "Prediction Of Benign And Malignant Breast Cancer using data mining techniques", journal of Algorithms and Computer technology
[5]. D. Lavanya and Dr. K. Usha Rani, "Ananlysis of Feature Selection with Classification:Breast Cancer Datasets",Indian Journal of Computer Science and engineering(IJCSE)
[6]. Kharya S. "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Diesease",International Journal Of Computer Science ,Engineering and Information Technology(IJCSEIT) 2012
[7]. M. Tahmooresi, A. Afshar, B. Bashari Rad, K. B. Nowshath and M. A. Bamiah, "Early detection of Breast Cancer using Machine Learning Techniques", Journal of Telecommunication, Electronic and Computer Engineering, e-ISSN: 2289-8131 Vol. 10 No. 3-2
[8]. Jabeen Sultana and Abdul Khadir Jilani, "Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers", International Journal of Engineering & Technology,2018,22-26
[9]. Peter Adebayo Idowu, Kehinde Oladipo Williams, Jeremiah Ademola Balogun and Adeniran Ishola Oluwaranti, "Breast Cancer Risk Prediction Using Data Mining Classification Techniques", Transactions On Networks and Communications, Society of Science and Education, UK, volume 3 No 2, April 2015
[10]. Ruolan Xu and Qiongjia Xu, "Applying Different Machine Learning Models To Predict Breast Cancer Risk"
[11]. Y Rejani- "Early detection of breast cancer using SVM", 2009
[12]. Hiba Asria, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", Procedia Computer Science 83 (2016) 1064 – 1069

[13]. Milani J, Niranjan R, Phatak V, Tadakhe O, International Journal of Computer Science and Information Technologies,2014
[14]. Chandra Prasetyo Utomo, Aan Kardiana, Rika Yuliwulandari, "Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Learning Techniques", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 3, No. 7, 2014
[15]. Mihir Borkar, Prof. Khushali Deulkar, and Abhinav Garg, "Prediction of Breast Cancer using Artificial Neural Networks", International Journal Of Engineering Research & Technology(IJERT),ISSN 2278-0181
[16]. Larrisa Westerdijk and Prof. Dr. Sandjai Bhulai, "Predicting malignant tumor cells in breasts", Vrije Universiteit Amsterdam, July 2018
[17]. Xin Yao and Yong Liu, "Neural Networks For Breast Cancer Diagnosis", 0-7803-5536-9/99, 1999IEEE
[18]. Harsh Vaizirani, Rahul Kala,Anupam Shukla, Ritu Tiwari, "Diagnosis of Breast Cancer by Modular Neural Network" , Third IEEE International Conference on Computer Science and Information Technology, Chengdu, China,pp 115-119
[19]. D. Thuthi Sarabal and Dr. K. Arthi, "Efficient Breast Cancer Classification Using Improved Fuzzy Congnitive Maps with Csonn", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 2,number4(2016)
[20]. S. Swathi, S.Rizwana, G. Anjan Babu, P. Santosh Kumar, P.V.G.K Sarma, "Classification of Neural Network Structures For Breast Cancer Diagnosis", International Journal of Computer Science and Teechnology, Vol 3,No.1, Jan-June 2012
[21]. Dr. A.Santhakumaran and F. Paulin, "Back Propagation Neural Networks By Comparing Hidden Neurons: Case Study on Breast Cancer Diagonosis", International Journal of Computer Applications (0975-8887) volume 2_ No. 4, June 2010
[22]. Htet Thazin Tike Thein and Khin Mo Mo Tun, "An Approach for Breast Cancer Diagonosis Classification Using Neural Network", Advanced Computing: An International Journal (ACI, Vol. 6, No. 1, January 2015
[23]. Bibhuprasad Sahu, Sachi Nandan Mohanty and Saroj Kumar Rout, "A Hybrid Approach for Breast Cancer Classification and Diagnosis", EAI Endorsed Transactions on Scalable Information Systems, 21 January 2019