MAT JOURNALS

Journal of Image Processing and Artificial Intelligence
e-ISSN: 2581-3803
Volume 5 Issue 3

Diagnosis of Breast Cancer using Machine Learning Algorithms

Divya D*[1], Amudhasurabhi A[1], Rekha P M[2]
[1]UG Student, [2]Assistant Profesor
[1,2]Department of Information Science, JSS Academy of Technical Education, Bangalore, Karnatala, India
***Email:** *divyadivuuu28@gmail.com*
DOI:

Abstract

Breast cancer is one on the most well-known kind of malignant growth influencing ladies from all around the globe. If not diagnosed at an early stage it can lead to many complications or even death. This disease can affect an individual due to genes, environmental affects and many other factors. In today's world with very advanced technologies and surgical procedures, the treatment of cancer has somewhat become easier. Machine learning and artificial intelligence techniques has played a vital role in detection, analysis and visualization of cancer cells in the human body. In this paper a review about the various machine learning algorithms used for the detection of cancerous cells have been presented. Some of the algorithms reviewed are K-nearest neighbors, random forest, and support vector machine etc., for detection of cancerous tumor.

Keywords: *Bayesian network, diagnosis, mammogram images, random forest, support vector method*

INTRODUCTION

Artificial intelligence is a growing technology in present society and Machine learning is a subset of it. Its application is spread over in different fields like business, medical, defense research, medicine etc. Machine learning is taking over the field of medicine and plays a crucial role in human life. Different machine learning algorithms are random forest (RF), support vector method (SVM), Bayesian network (BN), k nearest neighbor (KNN) etc.

Breast cancer is deadly heterogeneous disease which is wide spread among middle aged women [1, 2]. Around 25% of the females in the US have breast cancer. 43% of females in UAE are diagnosed with breast cancer.

Malignant tumor is the cancerous it spreads throughout the body by several mutations. It converts the healthy tissues to cancerous ones. On the other side, benign is non-cancerous tumor which is bounded to only the affected area and doesn't spread among other parts of body, it is considered as harmless. Women having untreated malignant tumor or delay in identifying will lead to death. Identifying or differentiating between these tumors is a difficult task. ML has been connected with diagnosis of breast cancer from long time but still effective and good accuracy has not been obtained.

There are nine different characteristics of cancer depending on which classification of cancer is possible:

- Thickness of clump
- Cell size Uniformity
- Marginal adhesion
- Cell shape Uniformity
- Bare nuclei
- Single epithelial cell
- Normal nuclei
- Bland chromatin
- Mitoses Ease of Use

In this section, the classifiers used in the detection of breast cancer are presented.

ALGORITHMS

Random Forest Method

Like in a court, many juris assembles and play key role in taking a decision, similar way many decision trees are assembled to make a forest of trees. It is seen that stability is increased by using many decision trees rather than a single tree. RF is best suited to handle data minorities. RF is a recursive method. Cases are classified depending upon the majority vote of decision trees as shown in Fig. 1 [3].

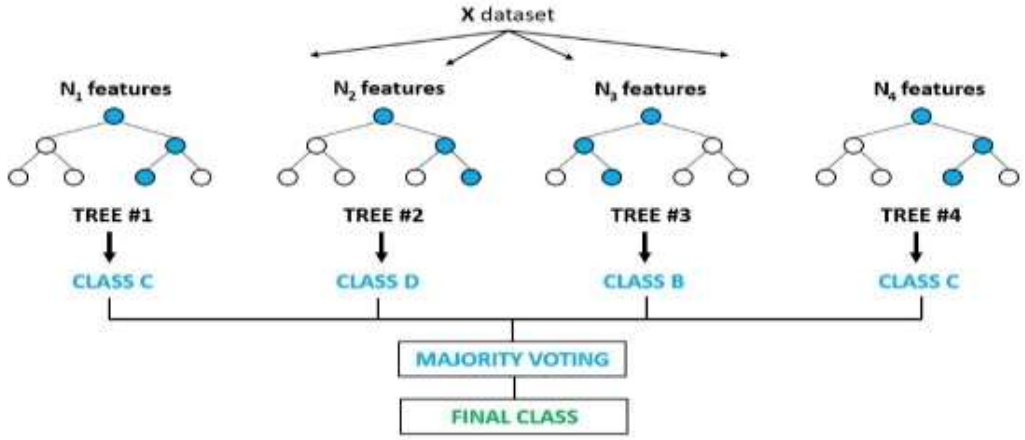

Figure 1: *Random forest.*

K Nearest Neighbours' Method

It is a type of supervised ml classifier technique used in diagnosis of breast cancer. Given a case its property are analysed and is placed in a suitable proper class. Classifying the given case depends on its k neighbours. Even majority vote is considered while placing any given case in a class [4]. Euclidean distance equation is used to find the distance between the given case and its neighbouring points as shown in Fig. 2.

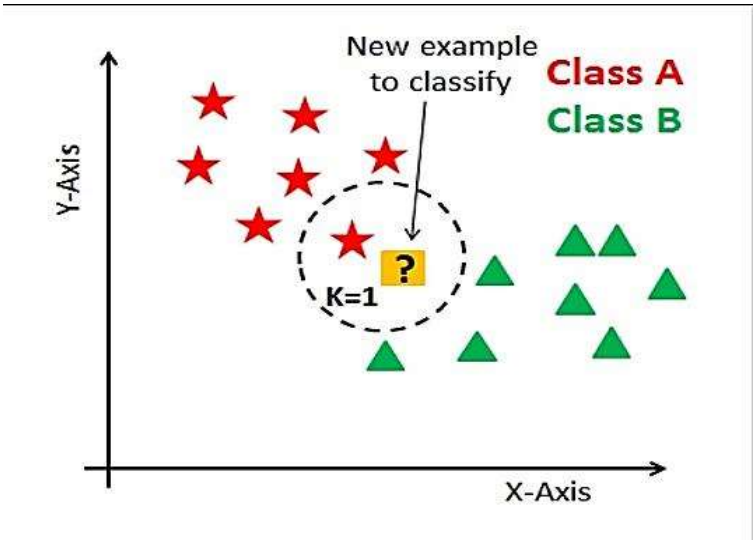

Figure 2: *K-nearest neighbors.*

Support Vector Method

SVM is one of the machine learning technique. It is a widely used classifier technique for cancer diagnosis. SVM

makes a mapping between a high dimensional space to an input vector. It finds a hyperplane such that it divides the given data into classes. SVM's aim is to find a hyperplane such that its distance to the nearest data point is maximum. It is called as marginal distance. Data set that are closer to the boundary (decision boundary) are called support vectors and SVM depends on these vectors as shown in Fig. 3.

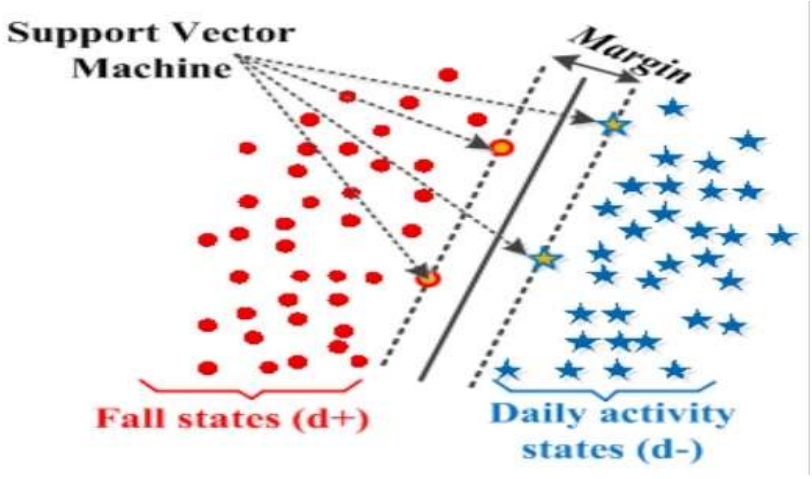

Figure 3: *Support vector method.*

Bayesian Network

BN is subset of PGM (probabilistic graphical model) used in prediction uncertain domains. The chart contains numerous hubs and edges between them. Every hub speaking to one arbitrary variable and each edge relating to reliance among hubs. To gauge these conditions, numerous measurable techniques are fused. All the variables may or may not depend on their ancestral nodes but conditionally independent of their non predecessors as shown in Fig. 4 [5, 6].

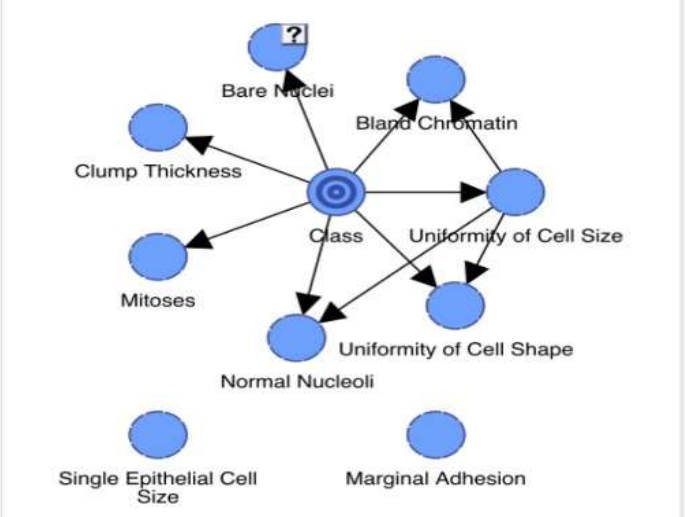

Figure 4: *Bayesian network.*

Image Processing

Mammography images are used in identifying the tumor type. Mammography images are processed and the data collected from it is used as input to ml classifiers. Many image filtering processes are done enhance the images. These images can be converted to binary form (black and white) to

get clear detection or can be kept in grey scale as shown in Fig. 5 and Fig. 6.

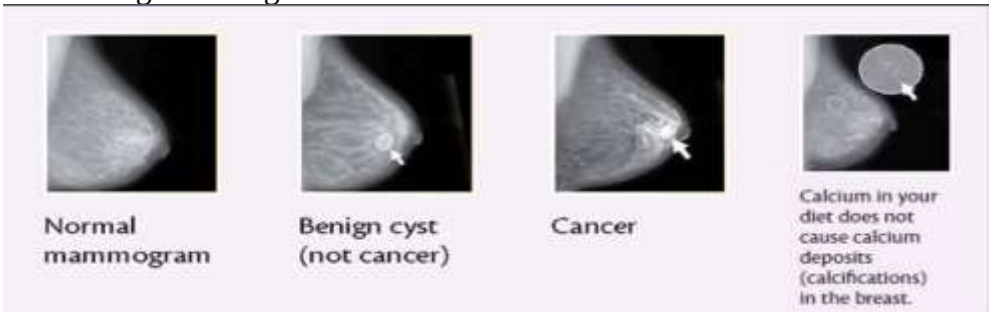

Figure 5: *Mammography image.*

BLOCK DIAGRAM: FLOW OF DIAGNOSIS OF BREAST CANCER

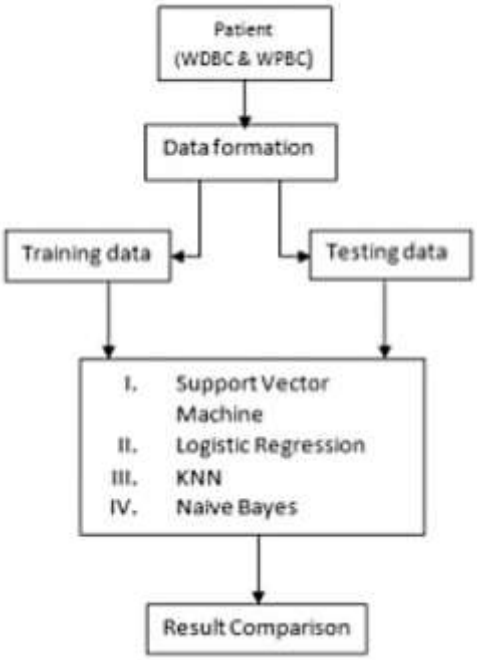

Figure 6: *Flow diagram of diagnosis of breast cancer.*

LITERATURE SURVEY

In the year 2013, Farzana Kabir Ahmad et al., have proposed a method in classification of breast cancer by using fine needle aspiration (FNA) biopsy data. Random forest ml algorithm was incorporated as it highly tolerates noisy data. Spilt with low impurity was selected by using Gini criterion. FNA dataset was obtained from UCI machine learning repository. 458 cases were of benign and 241 cases were of malignant out of 700 data. The performance of RF method was measured used ROC (receiver operating characteristic). Out of all the features that were used as input to RF, bland chromatin, mitoses, uniformity of cell size and single epithelial cell size were most relevant features. RF gave an accuracy of 72% and sensitivity of 75% [7].

In the year 2013, Xiufeng yang et al., have discussed a method in classification of breast cancer. In this paper isomap-svm and SVM with kernel RBE was used on the Wisconsin data set, which yielded a test accuracy of 96.4497 % and 97.633% respectively. Authors utilized the calculation bolster vector machine with numerous pieces which depend on

isometric component mapping in the characterization of bosom malignant growth. Isomap was utilized to extend high dimensional bosom malignant growth information to a lower dimensional space. Secondly, they used SVM algorithms with multi kernels to classify lower dimensional space [8]. There experiment concluded that the isomap-svm with combinational kernels has a higher test accuracy then traditional SVM.

In the year 2016, Dana Bazazeh et al., have proposed an approach in detection and diagnosis of breast cancer by using supervised ml classification techniques. They are SVM (support vector machine), random forest method (RF), Bayesian network (BN). The accuracy in the prediction of the cancer was increased by removing the support vector which was very near to the decision boundary. Probabilistic graphical model was used to find the joint probability value of network variable. To handle data minorities, RF was used. The original cancer data set of Wisconsin was used as data set. It consisted a total of 669 instances of which 458 was of benign cancer and 241 of malignant. K overlap cross approval technique was utilized as preparing set, with k as 10. He utilized WEKA (Waikato condition for learning examination) an open source device for programming recreation. Out of all classifiers, SVM have highest accuracy, specificity and precision but RF's probability of identifying tumour correctly was highest with ROC 99.9%.

In the year 2018, Meriem amrane et al., have proposed an approach in classification of breast cancer, using two supervised ml classifier technique NB, KNN. Bayes theorem was incorporated in determining to which class a particular case belongs. Euclidean distance equation was incorporated to find distance between a particular sample and other points and then to place it in appropriate class. Wisconsin breast database was taken as dataset. K fold cross validation technique is used as validation set. Out of two techniques adopted, KNN has highest accuracy of 97.51%.

In the year 2016, Moh'd Rasoul Al Hadidi et al., have proposed another strategy to recognize bosom Cancer. There are two sections in recognizing sort of bosom malignant growth, initial one is handling mammography pictures to gather valuable information and second part is utilizing this information in ml calculations. BPNN and LR supervised techniques were incorporated. 209 images from 50 patients were used as dataset. Wiener filter was used to remove blur effects and noise from images. Algorithms gave an output of 1's for tumor images and 0's for non-tumor images. BPNN technique gave an accuracy of 93%.

In the year 2018, Siyabend Turgut et al., proposed a different method used in classification of breast cancer tumor. The authors used world cancer research fund international breast cancer statistics data set. They used the SVM and decision tree, SVM had the highest accuracy and the later was the least accurate. The increase in the number of neurons and layers didn’t have any effect on the total accuracy. The first dataset had 1919 proteins belonging to 133 individuals out of which 122 had cancer. The second dataset had 24481 proteins belonging to 97 persons out of which 46 had a cancer relapse and the rest didn’t. Algorithms were applied to the dataset without applying feature selection and later two feature selection methods were applied.

In the year 2017, Aman Sharma et al., have discussed a method in classification of breast cancer. The authors used the Wisconsin dataset and sixteen important features were selected using the recursive

feature elimination algorithm using the chi2 method. Later NN and logistic algorithm was applied individually and the accuracy was calculated. At the end they used, proposed voting method to compute the accurate method for the diagnosis of breast cancer. They acquired an accuracy of 98.50% by applying voting algorithm to the top sixteen features of the dataset.

PERFORMNACE METRICS

Performance factors are used by researchers to evaluate the efficiency of machine learning algorithm. Few performance measures are as follows (Table 1).

True Positive

It represents number of people who has breast cancer and is predicted correct.

True Negative

It represents number of people who doesn't have cancer and is predicted correct.

False Positive

It represents number of people who does not have cancer but predicted wrong

False Negative

It represents number of people who has cancer but predicted wrong.

Accuracy

It represents number of correctly predicted values out of all other values.

Sensitivity

From patient database to categories patients who are having breast cancer, sensitivity is used.

Accuracy, sensitivity and specificity are given by equation 5.1, 5.2 and 5.3

Accuracy = (TP+TN) /(TP+TP+FP+FN) eqt 5.1
Sensitivity= TP /(TN+TP) eqt 5.2
Specificity=TN /(TN+FP) eqt 5.

Table 1: *Comparison of different machine learning algorithms.*

Author and Year	Type of Disease	Methods and Algorithm Used	Tools Used	Data Set Used	Accuracy
Farzana Kabir 2013	Breast cancer	Random forest	Gini criterion	Fine needle aspiration (FNA) biopsy data	75%
Dana Bazazeh 2016	Breast cancer	SVM (support vector machine) random forest method Bayesian network	WEKA (Waikato environment for knowledge analysis)	Wisconsin data set	99.9%
Xiufeng yang 2013	Breast cancer	Isomap-svm SVM with kernel RBE	K fold cross validation (k=10)	Wisconsin data set	97.633%
Meriem amrane 2018	Breast cancer	NB (naïve Bayesian), KNN (k nearest neighbors' method)	K fold cross validation (k=10)	Wisconsin breast database	97.51%
Moh'd rasoul 2016	Breast cancer	BPNN (back propagation neural network), LR (logistic regression)	Mat lab	209 images from 50 patients	93%
Siyabend Turgut 2018	Breast cancer	SVM and decision tree	K fold cross validation (k=5)	world cancer research fund international breast cancer statistics data set	___
Aman Sharma 2017	Breast cancer	NN (neural networks) logistic regression	Chi 2 method	Wisconsin dataset	98.50%

CHALLENGES AND ISSUES

The main challenge in using the machine learning algorithm for diagnosis of breast cancer is consideration of all possible risk factors as it is tremendously time consuming. Even with computer of high performance optimizing all the possible approaches in time will consume lots of time. Analysis of data gathered from patients is key factor for diagnosis which is not an easy task to do.

CONCLUSION

The survey on implementing automation in the field of classification of breast cancer cells is presented. The various machine learning algorithms implemented for diagnosis of breast cancer has been summarized. A few specialists have executed AI calculations on huge informational index by thinking about precision and affectability as parameters for guess of bosom malignant growth. The contribution of authors and the accuracy obtained is described in this paper. Still many challenges and issues need to be addressed in machine learning algorithm for real time application.

REFERENCES

1. Mohd Rasoul Al-headed, Abdulsalam Alarabeyyat (2017), "Breast Cancer Detection using K-nearest Neighbours Machine Learning Algorithm", *IEEE*, Liverpool, UK, pp. 75–85.
2. Farzana Kabir Ahmad, Nooraini Yusoff (2013), "Classifying Breast Cancer Types Based on Fine Needle Aspiration Biopsy Data Using Random Forest Classifier", *IEEE,* Bangi, Malaysia.
3. Meriem Amrane, Saliha Oukid, Ikram Gagaoua, Tolga Ensar (2018), "Breast Cancer Classification Using Machine Learning", *IEEE*, Istanbul, Turkey.
4. Dana Bazazeh, Raed Shubair (2016), "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis", *IEEE,* Ras Al Khaimah, United Arab Emirates.
5. N Marline Joys Kumari, Krishna Kishore KV, MIEEE (2018), "Prognosis of Diseases Using Machine Learning Algorithms: A Survey", *IEEE,* Coimbatore, India.
6. Xiufeng Yang， Hui Peng, Mingrui Shi (2013), "SVM with Multiple Kernels based on Manifold Learning for Breast Cancer Diagnosis", *IEEE,* Yinchuan, China.
7. Aman Sharma, Rinkle Rani (2017), "Classification of Cancerous Profiles using Machine Learning", *IEEE*, Noida, India.
8. Siyabend Turgut, Mustafa Da tekin, Tolga Ensari (2018), "Microarray Breast Cancer Data Classification Using Machine Learning Methods", *IEEE,* Istanbul, Turkey.

Cite this article as: