TEST Engineering & Management

March - April 2020
ISSN: 0193-4120 Page No. 11563 - 11568

Review on Machine Learning Technique to Detect Breast Cancer

[1]Priya Porwal, [2]Ajay Shanker Singh, [3]Thirunavukkarasu K

[1]PhD Scholar, [2, 3]Professor,
[1,2,3]School of Computer Science and Engineering, Galgotias University, Greater Noida, India
[1]porwal.priya4@gmail.com, [2]drajay.cse@gmail.com, [3]thiruk.me@gmail.com

Article Info
Volume 83
Page Number: 11563 - 11568
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September2019*
Accepted*: 15 February2020*
Publication*: 16 April 2020*

Abstract

The breast cancer is a malignant breast tumour, which is characterized by uncontrolled cell growth in the tissue of the breast. For the survival of patients, detection of breast cancer in the early stage is very helpful. To identify breast cancer using imaging techniques are mammography and ultrasound. Radiologists' technique is used to read the mammograms to detect the sign of breast cancer. Still, the radiologists may miss up to 30% of breast cancer depending upon the density of the breast. To improve the accuracy, different machine learning techniques are used. In this paper, we presented a review of the latest machine learning techniques on breast cancer. After review analysis, the higher accuracy is achieved by SVM (Support Vector Machine) and ANN (Artificial NeuralNetwork).

Keywords: *Machine Learning, SVM, ANN, Decision tree, Breast Cancer, NaiveBayes*

1. Introduction

Breast cancer became a problem nowadays in rural areas as well as developed countries in women. The discovery rate of breast cancer is continuously increasing, and from the review of previous research, the survival rate is below 88% after five-year from diagnosis and 80% after ten years from diagnosis. The percentage of the existence of breast cancer in a rural area is higher than the developed countries. To overcome this problem is to detect breast cancer in an early stage. Breast cancer becomes a serious threat to women's life and health. Breast cancer is the biggest problem of death for women in the world. The prediction of breast cancer in the early stage is one of the essential works. Basically, there are two types of breast tumour benign tumour and malignant tumour. Before providing appropriate treatment to the patients, symptoms related to tumour must be studied, which help for automatic prediction tool. Using data mining techniques extract useful signs from a large amount of data generated from hospital websites or social media.

Different type of machine learning techniques is used in the prediction of breast cancer, and the result of different machine learning techniques varies. We focus on the accuracy of the result by using different features applied in different countries dataset. CAD (Computer-Aided Design) system is a handy tool for medical radiology.

But sometimes the accuracy of the result of a CAD system is not satisfied. To improve the efficiency of breast cancer applies the different classification techniques. Detection of breast cancer in the early stage is beneficial to reduce the death rate due to breast cancer. The CAD system is essential tool for medical radiology. Breast cancer is the most common cancer which is observed mostly in women. Survivals of patient had breast cancer depending on the stage of disease [24] when it is diagnosed. So, it is very important to detect the cancer at its early stage. If detected in the early stage, it is helpful to reduce the rate of death due to breast cancer detection.

2. Data Collection and Preprocessing

A. Data Collection

For detect the breast cancer first collect the dataset from UCI machine learning repository. Some of the datasets are available

Table 1: Datasets used by different Authors

Author &Year	Dataset
D. A. Almuhaidib and F. M. Albusayyis(2018)	Breast Cancer Wisconsin DataSet[18]

March - April 2020
ISSN: 0193-4120 Page No. 11563 - 11568

Sarah B. Sakri, Nuraini B. ABDUL RASHID(2018)	Wisconsin Breast Cancer Prognostic Dataset[19]
R. Alyamil and J. Alyamil (2018)	Breast Cancer Wisconsin (Original) Data Set[20]
B. Dai & Rung C. Chen (2018)	Irvine Breast Cancer Wisconsin (Diagnostic) Dataset[21]
L. Husain, W. Aziz (2018)	The Digital Database for Screening Mammography (DDMS)database[22]

B. Data Preprocessing

After collecting the dataset, apply to preprocess for converting the raw data into clean data. Data preprocessing task are data cleaning, data integration, data transformation. In data cleaning remove the noisy data, the meaning of noisy data is the data which contains error or any incomplete data. After cleaning the data combine the data from multiple sources into coherent data store which is called data integration. In the data transformation transforming or consolidating data into appropriate form for mining.

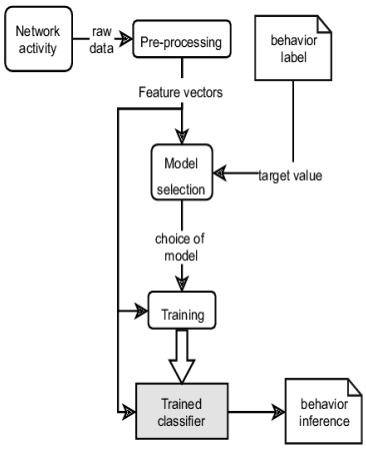

Figure 1: Flow of Work

3. Machine Learning Techniques

In this phase applying the different classification algorithm for prediction such as SVM, DT, ANN etc.

SVM (Support Vector Machine)

SVM is a supervised machine learning technique. In which the classification uses the hyperplane in the form of a linear function to separate the two class of data.SVM technique is used when the data contains precisely two levels. By using the hyperplane separate the data into two categories.

The objective of the SVM technique is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

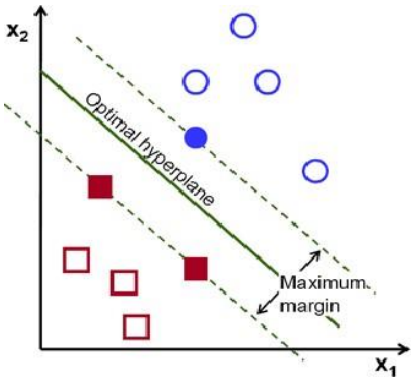

Figure 2: SVM Classification

The data points which is closest to the hyper plane is called support vector. SVM provides a unique solution to a problem. But the limitation of SVM is the deal with the data as two classes.

DT (Decision Tree)

DT is also a supervised classification technique which can solve the problem of both regression and classification. It uses the tree representation to solve the problem from the root node which denotes a test of an attribute, and each branch of this root node represent the outcome of the trial, each leaf node holds a classlevel.

ANN (Artificial Neural Network)

Artificial Neural Network (ANN) are multi-layer connected neural network. It consists of an input layer, hidden layer and output layer. We can increase the hidden layer, so the system becomes deeper. One layer is connected to every other node.

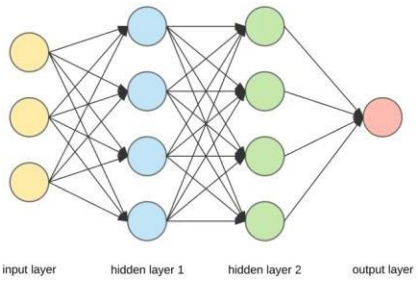

Figure 3: Artificial Neural Network

Take the weighted sum of its input and passes it through a non- linear activation function. This output becomes the input of another node to the next layer. The final output is calculated when all the nodes are performing this procedure for the nodes.

The output takes either 1 or 0, 1 for input above threshold and 0 for others. Bias is added with the output, bias is an input for all the input & always has the value 1.

It helps the model to train when all the input features are 0. The above equations look as follows with the bias included-

March - April 2020
ISSN: 0193-4120 Page No. 11563 - 11568

$$z = f(b + x \cdot w) = f\left(b + \sum_{i=1}^{n} x_i w_i\right)$$

$$x \in d_{1\times n},\ w \in d_{n\times 1},\ b \in d_{1\times 1},\ z \in d_{1\times 1}$$

K-Nearest Neighbor

K-Nearest Neighbor is a straightforward supervised algorithm which is used for both regression and classification. K is the small and positive integer, and K denotes the number of classes in the dataset. K- Nearest Neighbor is one of the first classification algorithms where there is no prior knowledge about the distribution data. An object is classified by a majority vote of its neighbour, with the objective being assigned to the class meet common among its K nearest neighbour.

Naive Bayes

It is based on Bayes theorem. Bayes theorem is a result in probability theory that relates conditional probability. If A and B denote the conditional probability of A occurring, given that B occurs. It finds the conditional probability of an event occurring given the likelihood of another game that has already happened.

Mathematically represented as the following equations.

P(A/B)= (P(A/B)*P(A))/P(B)

P(A/B) is a posterior probability

P(A) is the prior probability

4. Literature Review

In 2018, Daad Abdullah Almuhaidib and his co-operative [2] researcher proposed a system for detection of breast cancer by using a Decision tree, Naïve Bayes technology. The proposed method provides a website that is helpful for physicians to enter features related to breast cancer recurrence.

Experiments are done in R-studio; the result from the Random Forest and Decision Tree gives the high accuracy. Random Forest uses different decision trees and combines their findings to produce one prediction. Random forest with accuracy 0.6522, sensitivity 0.6250 and specificity 0.6593, the decision tree with accuracy 0.6521, sensitivity 0.63656 and specificity 0.62500, finally Naïve Bayes with accuracy 0.5913, sensitivity 0.4889 and specificity 0.6571. Aim of this tool is to help physicians decide on the treatment to provide to a patient. This tool accepts input from users. If users enter the irrelevant data, the online tool will discard that data and will only keep the relevant features.

In 2018, Sapiah B. SAKRI, NURAINI BINTI ABDUL RASHID AND ZUHAIRA MUHAMMAD ZAIN [3] proposed a system which is used to extract the features to detecting the breast Cancer by using machine learning techniques such as K nearest neighbour, DT, and Bayesian classifier. In this paper, try to improve the accuracy of classification algorithm by using feature selection techniques to reduce the number of features. They used dataset available publicly by the university of south Florida, and different elements are extracted such as texture, morphological entropy-based, scale-invariant features transform. These features passed as input to ML Classifiers. The result predicted in terms of specificity, sensitivity after experiment naïve Bayes produce a better result with PSO or without PSO(Particle swarm optimization).

In 2018, Bin Dai, Rung-Ching Chen2, Shun-Zhi Zhu1, Wei-Wei Zhang [5] proposed a system for Breast cancer detection by using Random Forest Classifier. In this paper, the random forest algorithm can combine the characteristics of multiple eigenvalues, and mixed results of numerous decision trees can improve the prediction accuracy. Random Forest algorithm is used to discuss the case of breast cancer case diagnosis and obtain high prediction accuracy. Here we randomly select the dimensions to build a decision tree, but the proportion can not be too large neither too small. The number of dimension selections can be log2N+1.

In 2018, Lal Hussain, Wajid Aziz & his co-operative researcher [1] proposed a system to distinguish the standard mammograms from that of breast cancer. The system work on mainly four stages is preprocessing, feature selection, classification, testing of data. In the preprocessing phase extracted different features texture, morphology, SIFT and EFDs. By using different classification techniques such as SVM(Support vector machine), DT(Decision tree), Bayesian classifier distinguish the classify images into normal and malignant images. The performance was measured on the basis of extracted different features.

In 2018, Reem Alyami1 & Jinan Alhajjaj [4] predicted breast cancer based on classification techniques ANN(Artificial Neural Network) and SVM (Support Vector Machine). To detect whether a patient has breast cancer or not is consider as a big challenge. In this paper, they try to improve the accuracy of detection by using SVM and ANN classification techniques. The SVM algorithm is used when the data contains two classes exactly where the classification is done by defining a hyperplane that separates the samples belong to one class A form those belong to the other class B. Neural networks, in general, contains at least two physical elements, Neurons which are the process element and a weighted link to connect these elements together. There are three types of neurons, input neurons, hidden neurons and output neurons. By the experiment, compare both classification techniques. SVM predicts the better result on classifying the sample with 97.1388% accuracy while ANN achieved 96.7096 % accuracy.

In 2017, Prannoy Giri and K Saravankumar [7] detected breast cancer by using image processing techniques. The system work on four stages Pre-processing, Segmentation, Feature extraction and apply the classification technique ANN(Artificial neural network). In the phase of preprocessing filter the image by using a noise removal algorithm. After Preprocessing segment the picture, the purpose of segmentation is to focus on the ROI(Region of interest), means removing the unwanted part of the image and only segment the part which has the higher pixel density. In the next phase, extract the features and the extracted features passed

through the neural network as input. The back propagation algorithm is used for self-learn and adjust the weight accordingly with error correction rule. The accuracy of the system is impossible to determine because the performance of algorithm sensitive to the self-learning rate. The performance of the back-propagation can be improved drastically by allowing the learning rate to adapt and change based on the complexity of the error and variance in the input received.

In 2017 Benzheng Wei and his co-operative researcher [6] proposed a novel breast cancer histopathological image classification method based on deep convolution neural networks, named as CNN model. This tool is very efficient, which provides higher accuracy(up to 97%) and good generalization and robustness. CNN model has strong ability to feature learning with multiple hidden layers and able to learn inherent features of the clinical image database. In this model, overcome the over-fitting problem by reducing the training samples.

Table 2: Literature Survey comparison table by different Authors

Author & Year	**Paper Title**	**Technique used**	**Accuracy**	**Advantages**	**Disadvantages**
D. A.	Ensemble learning	Decision Tree	Decision Tree	The system is	It does not
Almuhaidib	method for	Naïve Bayes	0.6261	capable of predict	work when the
and F. M.	prediction of breast	Random Forest	Naïve Bayes	breast cancer	data comes in
Albusayyis	cancer		0.5913	recurrence	different shape
(2018)			RF		and from
			0.6522		various sources
Sarah B.	Particle swarm	Naive Bayes	Naive Bayes 81.3%	It increases the	Number of
Sakri Nuraini	optimization	K-Nearest	KNN 80%	accuracy level of	features to
B. ABDUL	features selection	Neighbor	Fast Decision Tree	the prediction model	reduce
RASHID	for breast cancer	Fast Decision	80%		
(2018)	recurrence	Tree			
	prediction				
B. Dai &	Using a random	Random Forest		It has practical	Over-fitted, the
Rung C. Chen	forest algorithm for			significance for	model due to a
(2018)	breast cancer			auxiliary medical	large number of
				diagnosis	prediction
L. Husain, W.	Automated breast	SVM	SVM 96%	The system	It has an
Aziz	cancer detection	Decision Tree	Decision Tree 97%	distinguishes cancer	insufficient
(2018)	using machine	Bayesian	Bayesian Classifier	mammograms from	amount of
	learning techniques	Classifier	95%	healthy subject	dataset is
	by extracting				available
	different feature				
	extracting strategies				
R. Alyamil	Investigating the	ANN	--	It able to work with	Difficult to
and J.	effect of correlation-	SVM	--	linear and no- direct	understand an
Alyamil	based features			data	algorithm
(2018)	selection on breast				
	cancer diagnosis				
	using ANN & SVM				
Thirunavukka	Classification of	KNN	KNN-96%	It influence how	It is really only
rasu K., Ajay	IRIS Dataset using			you weight the	suitable when
S. Singh,	Classification			importance of	there are an
Prakhar Rai,	Based KNN			different	equal number
Sachin Gupta	Algorithm in			characteristics in the	of observations
	Supervised			results	in
	Learning				each class

P. Giriand, K. Saravanakumar (2017)	Breast cancer detection using image processing techniques	Digital image processing	---	It can extract the texture features from the ROI of mammograms. The features are selected based on PCA(Principal Component Analysis) for better identification.	It can not be work when the process data comes in the form of multiple arrays
B. Wei (2017)	Deep learning model based breast cancer histopathological image classification	Deep learning	Deep learning 97.02%	Provide the image classification method named as CNN model. It provides higher accuracy with excellent robustness and generalization.	Time-consuming

5. Conclusion

An overall review of literature gives the overview of current technology used for the detection of breast cancer and also shows the improvement in technology from the last few years. Many research works have been done in the field of breast cancer detection, but the need for improvement in accuracy of result still persists. In future improve the efficiency by using feature reduction technique. From the review, analyzed the number of false-positive and sensitivity. This research work can be further extended by providing the training on the different type and shape of the module to improve the detection of breastcancer.

References

[1] L. Hussain, W. Aziz, S. Saeed, S. Rathore, M. Rafique “Automated Breast Cancer Detection using Machine Learning Techniques by Extracting Different Feature Extracting Strategies” IEEE International Conference On Big Data Science and Engineering

[2] D. A. Almuhaidib, F. M. Albusayyis, H. A. Saiba, M. A. Alzaid, N. G. Alharbi, R. M. Almadhi, S. M. Alotaibi "Ensemble Learning Method for the Prediction of Breast Cancer Recurrence"IEEE(2018).

[3] S. B. Sakri, N. B. Abdul Rashid and Z. M. Zain “Particle Swarm Optimization Features Selection for Breast Cancer Recurrence Prediction” vol. 6, pp 2169-3536IEEE(2018).

[4] A.Adorama,R.Permatasari,P.W.Wiranwan, A. Wibowo, A. Suriwo "Support Vector Machine- Recurrence Feature Elimination (SVM-RFE) for Selection of Breast Cancer" pp 978-1-5386-7440-6 ICICOS(2018).

[5] B. Dai, R. Chen, S. Zhu, W. Zhang “Using Random Forest Algorithm for Breast Cancer Diagnosis” pp 978-1-5386-7036-1IS3C(2018).

[6] B. Wei, Z. Han, X. He and Y. Yin “Deep Learning Model Based Breast Cancer Histopathological Image Classification” pp. 978-1-5090-4499-3ICCCBD(2017).

[7] P. Giri and K. Saravanakumar “Breast Cancer Detection using Image Processing Techniques” vol. 10, ORIENTAL JOURNAL OF COMPUTER SCIENCE &TECHNOLOGY(2017)

[8] M. Gupta and B. Gupta "Survey of Breast Cancer Detection using Machine Learning Techniques in Big Data" vol. 21 Journal of Cases on Information Technology(2019)

[9] M. Gupta and B. Gupta “ A Comparative Study of Breast Cancer Diagnosis using Supervised Machine Learning Techniques” pp. 978-1-5386-3452-3 IEEE(2018)

[10] https://medium.com/@datamonsters/artificial-neural-networks-for-natural-language-processing-part-1-64ca9ebfa3b2

[11] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[12] https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

[13] https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6

[14] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[15] http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

[16] http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-

Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf
[17] https://www.researchgate.net/figure/Machine-Learning-process_fig3_326459139
[18] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%2.
[19] https://archive.ics.uci.edu/ml/datasets/Breast_Cancer_Wisconsin_(Prognostic)
[20] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29
[21] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
[22] http://marathon.csee.usf.edu/Mammography/Database.html
[23] Thirunavukkarasu K., Ajay S. Singh, Prakhar Rai, Sachin Gupta "Classification of IRIS Dataset using Classification Based KNN Algorithm in Supervised Learning "IEEE(2018)
[24] Thirunavukkarasu K., Ajay S. Singh, Md Irfan, Abhishek Chowdhury "Prediction of Liver Disease using Classification Algorithms" IEEE(2018)