Review Of ReseaRch
ISSN: 2249-894X
Impact Factor : 5. 7631(UIF)
UGC Approved JoUrnAl no. 48514

BREAST CANCER DETECTION USING WBCD

Kunal Prasad[1], Mahendra Kanojia[2], Brian Dsouza[3], Niketa Gandhi[4]
[1]Department of Computer Science, Sheth L.U.J. and Sir M.V. College, Mumbai, Maharashtra, India.
Email: kunalprasad91@gmail.com
[2] JJT University, Jhunjhunu, Rajasthan, India.
Email: kgkmahendra@gmail.com
[3] Department of Computer Science, Mithibai College, Mumbai, Maharashtra, India.
Email: dsouzabrian18@yahoo.in
[4] Machine Intelligence Research Labs (MIR Labs), Auburn, WA, USA.
Email: niketa@gmail.com

ABSTRACT

Cancer is a serious disease caused by the abnormal growth of cells. Early detection of cancer helps in diagnosis and complete cure of the disease. Research has been carried on a large scale to detect cancer using information technology. Machine Learning techniques helps to make a system able to learn. Machine learning technique is used in various sectors such as health, finance, education, defence etc. Researchers are in search of best analysis mechanism for bio-medical data with information technology. This paper discusses some machine learning techniques used to detect cancer tissues in the breast. Breast cancer tumours are of two types benign and malignant. Machine learning techniques are an efficient way to detect the early stage of cancer. This paper aims to provide aid to all researchers by giving them an insight into various machine-learning algorithms and the best parameters to detect the malignant in the breast. These algorithms can be implemented to detect cancer at the micro level. level. It gives the guidelines for the implementation of the algorithms on Wisconsin Breast Cancer Dataset (WBCD) for cancer detection. The study reports the comparative study of the algorithms to achieve high accuracy.

KEYWORDS — *artificial neural network, breast cancer, machine learning, support vector machine, Wisconsin breast cancer dataset.*

INTRODUCTION

Cancer is the most deadly disease, which has raised death rates in recent years. According to the report [1] new cases of cancer recorded are 18.1 million and the death rate due to cancer is 9.6 million in 2018. Ratio of developing cancer throughout their lifetime is 1:5 in men and 1:6 in women, whereas the death ratio is 1:5 and 1:11 in men and women respectively.

Cancer is of different types like lung cancer, breast cancer, oral cancer etc. Cancer can be detected through MRI imaging, CT-scans, Mammography etc. Treatment of cancer depends on the type of cancer. Cancer treatment may follow one line of treatment or may include multiple lines of treatment like surgery, chemotherapy radiation therapy etc. [2]. Late realization of cancer leads to a higher rate of death, so it is necessary to detect it as early as possible.

This paper focuses on work done recently for the detection of breast cancer. This paper discusses machine learning algorithms, which has proven to give better and accurate results for breast cancer detection. It will help other researchers to develop a system using machine learning algorithms to detect breast cancer at an early stage.

LITERATURE REVIEW

Research by [3] has conducted a systematic study to evaluate breast cancer detection using SVM classifier. For this study, two types of sequential filters were used. These filters were classified based on their properties. Data were analyzed and results were generated using algorithms of image processing. The result showed that final stage classification using SVM classifier. Thus, the study concluded that SVM Classifier gives the best classification rates.

Another research by [4] conducted a study using three machine learning algorithms on small subset feature set of breast cancer and reported high accuracy. For this study, all three techniques were classified to check the best accurate result. Data was analysed and result was generated using all three classifications. The result showed that SVM and Naive Bayes give almost the same accuracy whereas Naive Bayes in less time complexity gives the predicted result. Thus, the study concluded that Naive Bayes gives the best algorithm to gain accuracy.

This study [5] helps in the detection of cancer using three common machine-learning algorithms, namely SVM, Random Forest (RF) and Bayesian Network (BN). Precision equals to the ratio positive and negative values that were identified by true values, average precision values reported for detection of breast cancer for SVM is 97.0%, RF is 96.6% and BN is 97.2%.

Lothe Savita *et al*., 2017, has performed research on computer-aided detection used to check the experimentation of next elements. Five processes were involved to check the identification of breast cancer. Data were collected and results generated using SVM classifier. The result classified benign or malignant cells through the classifier. Thus, the study concluded that CAD system could be used to detect and classify breast mass.

A study by [7] implemented machine learning algorithm on feature set of Wisconsin Breast Cancer Dataset (WBCD). For this study, machine learning algorithms used were GRU-SVM, Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbor (NN) search, Softmax Regression and SVM. Data were collected from an online survey. Data was partitioned to 70% and 30% as training-set and test-set respectively. Seven parameters were considered to prove the result, which were test accuracy, epochs, number of data points, false positive rate (FPR), false negative rate (FNR), true positive rate (TPR) and true negative rate (TNR). Data were analysed and results were generated using six machine learning algorithms. The study concluded that all six machine learning algorithms could give us an accuracy of 99.04% whether breast cancer is benign or malignant.

Another study by [8] has completed work on automated detection and classification of breast cancer. For this study, mammography images were used. Data was collected from various hospitals in the form of images, and the result was generated using IDL technique. The result showed 82.50% of accuracy in detecting breast cancer.

Another research by [9] proposes a synthesized model with multiple machine-learning algorithms including SVM, Artificial Neural Network (ANN), K- Nearest Neighbor (KNN) and Decision Tree (DT). This model can be used on different types of data such as image, blood etc.

This study [10] used multiple machine learning algorithms and processed routine blood test results to understand the effectiveness of these algorithms in cancer detection. Algorithms used were ANN, Extreme Learning Machine (ELM), SVM, and KNN. Ten attributes of UCI dataset were used in this study. Accuracy rate achieved by ANN is 79.4304% in 0.4282 sec, ELM performed pretty well with an accuracy of 80% in 0.0075 sec, KNN with little less accuracy 77.5% in 0.15781 sec and SVM with 73.5% accuracy in 0.1866 sec.

WBCD DESCRIPTION

WBCD stands for Wisconsin Breast Cancer Dataset (Database) [11]. The dataset was originally recorded by Dr. William H. Wolberg (Physicist) from University of Wisconsin Hospitals. From attribute, 2-10 represents instances. Each instance has one of two class possibility i.e. benign or malignant.

The data samples were compiled periodically. The database is grouped by month and year sequentially. The information for the grouping is shown in Table I.

TABLE I

NUMBER OF INSTANCE UPDATE IN WBCD DATASET.

Group	Number of instances	Month	Year
1	367	January	1989
2	70	October	1989
3	31	February	1990
4	17	April	1990
5	48	August	1990
6	49	January	1991
7	31	June	1991
8	86	November	1991
Total of 699 instances as of donated database			

Breast tissues collected using fine needle aspiration technique were digitized. Features were extracted from these digitised image. The image's cell nuclei characteristics are described in the dataset. [12]

Attributes:

1. ID
2. Diagnosis (M – Malignant or B - Benign)

Feature extracted from each cell nucleus.

- Radius (Mean of distance from centre to the circumference)
- Texture (Standard deviation of grey scale values in the image)
- Perimeter
- Area
- Smoothness (variation in radius lengths)
- Compactness (perimeter^2 / area - 1.0)
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions)
- Symmetry
- Fractal dimension (“Coastline approximation” - 1)

RESEARCH METHODOLOGY

WBCD dataset was used as an input data. This WBCD dataset contains 699 instances and 10 attributes. The next step was to train the data. In addition, the data was trained and tested using various machine-learning algorithms. Through testing, results were generated. Result shows whether cancer is benign or malignant. Fig. 1 depicts the workflow of implementing machine learning algorithms on WBCD.

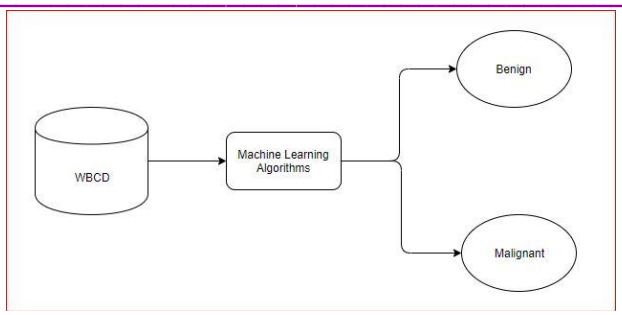

Fig. 1 General flow diagram for breast cancer detection.

Machine learning used are:

- Support Vector Machine (SVM)
- Naive Bayes classifier
- Linear Regression
- Multilayer Perceptron

A. Support Vector Machine (SVM)

Support vector machine is a supervised classification and regression algorithm. SVM incorporates various machine learning algorithms. SVM with machine learning algorithms follows the ANN model.

Fundamentally, SVM is a binary classifier. Its main objective is to find the optimal hyperplane between the SVM classes for regression and classification. Breast features under classification can be classified as benign or malignant. Therefore, SVM becomes a choice for classification of breast cancer feature set.

Algorithm:

Step 1: Pattern Classifier for separating two classes $\{a(k), b(k)\}_{k=1}^{p}$,

Where; k = input vector

S = input space

a and b = class index , a=1 or 0 and b = 1 or 0.

Step 2: Data set is trained using learning algorithm.

Step 3: Then training sets are paired $(a_{(k)},b_{(k)}),k=1,.....1$ where $a_{(k)} \epsilon Q^{un}$ and $b_{(k)} \epsilon (1.-1)^1$ [13] from step1

Step 4: mapping function Ø maps x(i) trained vectors to higher dimensional space.

Step 5: Higher dimensional space function Ø is iteratively implemented until a linear separating hyper-plane is found.

Step 6: Implementing the kernel method in SVM model is the final step. These include linear polynomial, RBF and sigmoid. RBF is the best method to implement in SVM in feature classification for breast cancer detection.

$K(x, y)) = {}_e-||x - y||/2\sigma 2$ is a combination of Gaussian Function

Where,

e is the expression

K is the Kernel Function

x and y are the classes.

Step 7: Radial Basis Kernel Function is implemented using $K(x, y)) = {}_e-||x - y||/2\sigma 2$

In this study, SVM with pattern recognition helps to predict malignancy in the dataset.

B. Naive Bayes

Naive Bayes is a collection of algorithm working on same principle of Bayes Theorem. The principle states that all the feature pairs under classification are independent of each other. Naive Bayes is a simple algorithm but shares a very good prediction. It shares amazing features, which any other algorithm does not have to implement. Hence, the Naïve Bayes Algorithm is the best algorithm to predict breast cancer.

Algorithm:

Step 1: Assuming all variables are independent.

Step 2: Each instance in the set be are trained and let T be the training set.

Step 3: Each tuples be calculate with A attributes such that each tuple will have A values.

Step 4: Suppose there are n classes having labels N1, N2,....Nk for any new tuple X, then the classifier will predict that class has the highest probability of X.

Step 5: Probability of having highest accuracy is

P(NiX)>PNjX where 1<j<n [14]

C. Linear Regression (LR)

Linear Regression (LR) is a type of regression technique based on supervised learning. The goal is to find out the relationship between variables and forecasting. LR is used to find out a dependent variable based on given information of the independent variable. Hypotheses function is involved in LR. Training of data can also be done through LR. Therefore, LR is the best way to find out Breast Cancer. WBCD dataset is classified using LR and threshold output is applied [7]. To check that there is no loss in predicting the output the following equation is applied.

L(y, θ, x) = $\frac{1}{N}\sum_{i=0}^{N}(y_i - (\theta_i . x_i + b))2$ [7] (Mean Squared Error (MSE))

Where y represents the actual class.

(θ, x + b) represents the predicted class.

D. Multilayer Perceptron

Multilayer Perceptron (MLP) is the class of neural network structures used to find up to two hidden layers. MLP is a function made up of two predictors also called as input and independent variables. The main goal of MLP is to produce an output, which is prevented from producing in neural networks. Hence, it works in two different ways to produce the desired output. Thus, MLP is the best way to predict breast cancer. In WBCD, MLP is used for considering the hidden layers through activation function such as tanh or σ [7]. Dataset is extracted and classified using MLP. The function is as follows:

f (hθ(p)) = hθ(p)+ =max(0,hθ(p)) [7]

h= is the hidden value

x= is the predictor.

Max= is to find the maximum prediction value.

RESULTS

TABLE II

MODEL / ALGORITHM COMPARISON

Ref No.	Model / Algorithms Used	Accuracy Achieved	Time complexity
[5]	Support Vector Machines	97.0%	-
[5]	Random Forest	96.6%	-
[5]	Bayesian Network	97.2%	-
[6]	Support Vector Machines	89.28%	-
[6]	Linear Regression	96.09375%	-

[6]	Multilayer Perceptron	99.038449585420729%	-
[4]	Naive Bayes Classifier	97.3978%	0.102023 ms
[8]	IDL, Interactive data language	82.50%	-
[10]	Artificial Neural Network	79.4304%	0.4282sec
[10]	Extreme Learning Machine	80.0%	0.0075 sec
[10]	K-Nearest Neighbor	77.5%	0.15781 sec
[10]	Support Vector Machines	73.5%	0.1866 sec

Research conducted by [5] shows the comparison of SVM, RF and BN algorithms. It shows the average precision values of all three algorithms.

A research by [6] used support vector machine algorithm on WBCD to compare it with other five machine learning algorithms to get the comparison result and support vector machine came out to be highest accuracy of 89.28%. Further linear regression algorithm was applied on the same dataset and it showed an accuracy of 96.09375% and finished its training on data in 35 seconds with an accuracy of 92.8906257%. Additionally multilayer perceptron algorithm was also used. This machine-learning algorithm completed its training of data in 28 seconds with an accuracy of 96.9286785% and showed classification accuracy of 99.038449585420729% on WBCD dataset.

Another research by [4] had applied Naive Bayes Classifier algorithm on the dataset consisting 203 malignant data and 66 benign data, in which algorithm predicted 201 correct malignant data and 2 incorrect and 61 correct benign data and 5 incorrect showing classification accuracy of 97.3978% with time complexity 0.102023 ms.

Another study by [8] aimed to collect a group of breast cancer images from various hospitals as the first-ever special data set in Sudan. 1170 images were collected from different hospitals. Images were classified by radiology specialists using BI-RADS and named it Samah Mammography Dataset (SMDS). Training and testing IDL method on SMDS dataset, result of 82.50% successful and accurate was obtained.

A study conducted by [10] shows the accuracy level of four different models. Various parameters were considered while applying all the models on the dataset.

CONCLUSION

In this paper after comparing all the machine learning techniques or algorithms reviewed all algorithm had a good performance on respective data. SVM is an algorithm widely used to classify or separate the data according to given parameters and showed accuracy of 89.28%, whereas linear regression has the accuracy of 96.09% and took 35 seconds to finish the training data. Comparing all the algorithms multilayer perceptron has the highest accuracy of 99.0384% on WBCD. Naive Bayes classifier classified both benign and malignant dataset with an accuracy of 97.3978%. IDL method showed the accuracy of 82.50% with the image data collected from different hospitals. Different models were applied on blood routines and ELM model showed the highest accuracy with minimum time complexity.

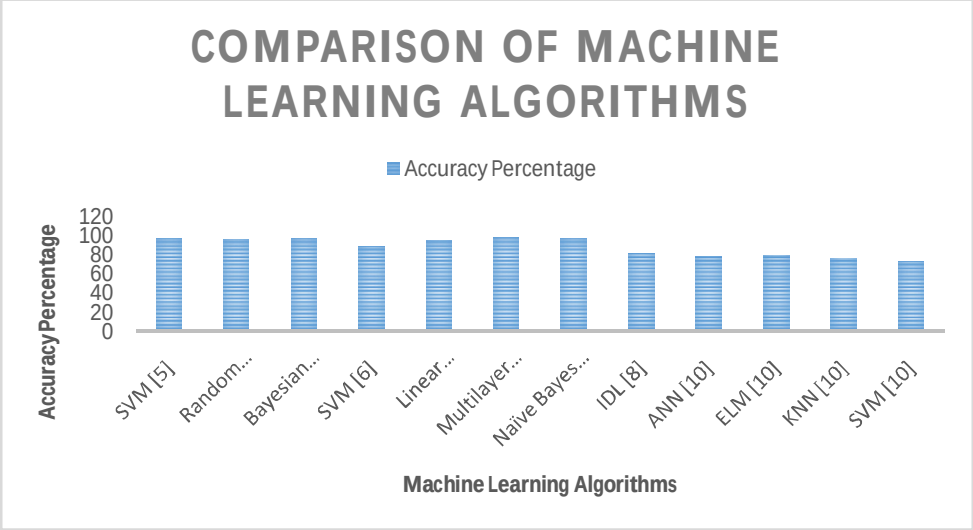

Fig. 2 Comparison of machine learning algorithms reviewed

REFERENCES

1. *Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: a cancer journal for clinicians, 68*(6), 394-424.*
2. *National Cancer Institute https://www.cancer.gov/about-cancer/treatment accessed on February 28th 2019.*
3. *Hiremath, B., & Prasannakumar, S. (2015). Automated Evaluation Of Breast Cancer Detection Using Svm Classifier. International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), 5(1), 11-20.*
4. *Hazra, A., Mandal, S. K., & Gupta, A. (2016). Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms. International Journal of Computer Applications, 145(2), 0975-8887.*
5. *Bazazeh, D., & Shubair, R. (2016, December). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)(pp. 1-4). IEEE.*
6. *Savita, L., Rupali, T., Almas, S., & Prapti, D.D. (2017). Detection and Classification of Breast Mass Using Support Vector Machine.*
7. *Agarap, A. F. M. (2018, February). On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (pp. 5-9). ACM*
8. *Mohamed, S.E., Wahbi, T.M., & Sayed, M.H. (April 2018). Automated Detection and Classification of Breast Cancer Using Mammography Images. International Journal of Science, Engineering and Technology Research (IJSETR), 7(4), 2278 -7798.*
9. *Tahmooresi, M., Afshar, A., Rad, B. B., Nowshath, K. B., & Bamiah, M. A. (2018). Early Detection of Breast Cancer Using Machine Learning Techniques. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(3-2), 21-27.*
10. *Aslan, M. F., Celik, Y., Sabanci, K., & Durdu, A. (2018). Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data. International Journal of Intelligent Systems and Applications in Engineering, 6(4), 289-293.*

11. *UCI Machine Learning repository Center for Machine Learning and Intelligent Systems. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original) accessed on 4th March 2019.*

12. *UCI Machine Learning repository Center for Machine Learning and Intelligent Systems. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) accessed on 4th March 2019.*

13. *Rejani, Y., & Selvi, S. T. (2009). Early detection of breast cancer using SVM classifier technique. arXiv preprint arXiv:0912.2314.*

14. *Nisha, S., & Kathija, A. (2016). Breast cancer data classification using SVM and Naive Bayes techniques. International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), 4, 21-167.*