Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 4s (2025)
https://www.jneonatalsurg.com

Detecting Breast Cancer Using Visual ML

Alok Chakravarty[1], Dr. Shweta Tewari[2]

[1]Dayananda Sagar University, India,
Email ID: alok_chakravarty@dsu.edu.in
[2]Dayananda Sagar University, India,
Email ID:shweta.tewari@dsu.edu.in

Cite this paper as: Alok Chakravarty, Dr. Shweta Tewari, (2025) Detecting Breast Cancer Using Visual ML. *Journal of Neonatal Surgery*, 14 (4s), 1211-1216.

ABSTRACT

Approximately 60% of Breast cancer patients are diagnosed in advanced stages. This paper examines the automation of identification of cancerous cells using visual machine learning approach. Results are obtained using two different datasets: Wisconsin and Coimbra. In Wisconsin dataset, predictors are extracted from the digitised image of a fine needle aspirate (FNA) of a breast mass. In Coimbra dataset, predictors are extracted from the blood analysis. Ten machine learning models are compared using a visual ML tool called Orange. Particular emphasis is placed on the metric "recall". Recall is defined as the ability to catch malignant cases out of total malignant cases present in the dataset. Highest recall of 0.982 in Wisconsin dataset is achieved using Stochastic Gradient Descent (SGD). Highest recall of 0.793 in Coimbra dataset is achieved using Gradient Boost algorithm. This tool can be deployed in hospitals where initial detection may be done using blood analysis and then for confirmation with digitised image of breast mass. In countries like India, where there is a scarcity of cancer specialists, this tool would fasten the detection process. Patients would spend more time in treatment than in diagnosis.

Keywords: *Breast Cancer, Machine Learning, Orange, Visual ML, Wisconsin, Coimbra*

1. INTRODUCTION

Among the various diseases that affect women, breast cancer is among the deadliest. WHO reports that in 2020 more than 2.3 million women were diagnosed with breast cancer and was responsible for 685000 deaths globally. In the period 2015-2020, 7.8 million women were diagnosed with breast cancer and survived. WHO has called breast cancer the "most prevalent cancer" and that 24% of all female cancer patients suffer from breast cancer.

Early detection of the cancer is critical. The chances of recovery are much higher if the cancer is detected early. (Wang Lu 2017). Machine learning and data mining offers medical professionals new tools which enhances early detection rates and reduces chances of diagnostic errors. (Ghassemi Marzyeh et.al 2018)

In case of breast cancer, a false negative can result in delayed diagnosis which can greatly increase the risk. National cancer Institute defines a false positive as 'A test result that indicates that a person does not have a disease or condition, when the person does have the disease or condition.

Studies have shown a false negative rate could be as high as 9.1%. (Shah VI et.al 2003). Our paper discusses how we can use ML and data mining to reduce the rate of false negatives and thereby reduce the risk considerably.

n this paper we would like to do following things differently:

- Focus on False Negative reduction and therefore optimise recall as a metric.
- Have two options made available to healthcare practioners for early detection. One based on blood analysis report parameters as predictors and second digitized image of a fine needle aspirate (FNA) of a breast mass
- Use visual ML tool like Orange for efficient deployment and easier training.
- Hyper parameter tuning to accomplish better performance metrics

2. LITERATURE REVIEW

Histopathology image samples of 683 patients has been analysed using Deep Neural Network with Support Value (DNNS) and accuracy of 0.929 has been achieved. (Anji Reddy et al 2020).

In the paper titled "Analysis of Breast Cancer Detection Using Different Machine Learning Techniques, authors have proposed resampling of data to address class imbalance to enhance the performance of Decision Tree (J48), Naïve Bayes and Sequential Minimal Optimization. Performance has been evaluated with True Positive, False Positive, ROC Curve, Standard Deviation and Accuracy. Highest accuracy of 99.56% has been achieved on WBC dataset. (Siham Mohammed et al 2020).

In the paper titled "Breast Cancer Detection using Machine Learning Way", authors have compared KNN, SVM and Naïve Bayes (NB) on Wisconsin dataset. Precision of 98.5% was obtained. (Sri Hari et al, 2019)

In the paper titled "Automated Breast Cancer Detection using Machine Learning Techniques by Extracting Different Feature Extracting Strategies" authors have employed Machine learning classification techniques such as Support vector machine (SVM) kernels and Decision Tree to distinguish cancer mammograms from normal subjects. Different features are proposed such as texture, morphological entropy based, scale invariant feature transform (SIFT), and elliptic Fourier descriptors (EFDs). Dataset was taken from publicly available database from University of South Florida. Performance: Naïve Bayes 0.9989 Sensitivity, 0.9991 Specificity; SVM Sensitivity 0.9066, Specificity 0.8981; Decision Tree Sensitivity 0.9588 Specificity 0.9606 (Lal Hussain et al, 2018)

In the paper titled "Breast Cancer Detection Using Machine Learning Algorithms", authors have compared Random Forest (RF), kNN and Naïve Bayes (NB) on Wisconsin dataset. Performance metrics reported RF: Accuracy 94.74, Precision 92.18, Recall 93.65;

KNN: Accuracy 95.9, Precision 98.27, Recall 90.47; Naïve Bayes: Accuracy 94.47, Precision 88.52, Recall 85.71 (Shubham Sharma et al 2018).

In the paper titled "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", authors have compared SVM, Decision Tree, Naïve Bayes and KNN on Wisconsin Dataset. Accuracy obtained were: SVM 97.13%, NB 95.99%, kNN 95.27%, C4.5 95.13%. (Hiba Asri et al 2016).

3. DATA ANALYSIS AND RESEARCH METHODOLOGY

In this paper, we use Wisconsin and Coimbra dataset and run 10 machine learning algorithms using a visual machine learning tool known as Orange. Model evaluation results are compared and particular emphasis is placed on the metric called 'recall'. Recall is defined as:

TP / (TP + FN)

Where, TP = True Positive, no of malignant cases correctly classified

FN = False Negative, no of malignant cases incorrectly classified

TP+FN = Total No of Malignant Cases present in the dataset

3.1 Machine Learning Algorithms

Machine learning is a subset of artificial intelligence which provides the machines the ability to learn automatically and learn from data without being explicitly programmed.

Machine learning algorithms are categorised as:

- Supervised,
- Unsupervised and
- Reinforcement

Supervised learning predicts continuous ranged values or discrete labels/classes based on the training it receives from examples with provided labels or values.

Unsupervised learning (for e.g. clustering) tries to club together samples based on their similarity and determine discrete clusters.

Reinforcement learning on the other hand, which is a subset of unsupervised learning, performs learning very differently. It takes up the method of "cause and effect".

Supervised Machine Learning: Here, we have some independent variables (aka predictors) which are used for prediction. Mathematically, it is expressed as:

y = f(X1,X2,X3…). Where y is the dependent variable, and X1,X2…are independent variables

If y is continuous (example Sales, Weight etc.), then this type of supervised algorithms are called Regression Algorithms

If y is categorical (example Male/Female, Approved/Not Approved etc.), then we call them Classification Algorithms. In this paper, our target variable is categorical (benign/malignant), hence our discussion would focus on classification algorithms used in this paper.

kNN Algorithm: k Nearest Neighbour algorithm doesn't do any modelling. It simply does classification by using the majority vote amongst 'k' nearest neighbours. Its advantage is that it is fast since no training is involved. kNN can slow down considerably as it has to calculate distances every time it does classification. Outliers and noisy data can also impact kNN's accuracy.

Naïve Bayes: Naïve Bayes algorithm is based on Bayes theorem and it does classification based on probability. It is mainly used in text classification, spam filtering and sentiment analysis. It's a simple, fast and effective algorithm. It assumes that the predictors are all independent which may not be so in reality.

Logistic Regression: Logistic regression is similar to linear regression, the difference is that instead of predicting numeric dependent variable, it predicts categorical variable and does so in terms of probabilities. Instead of fitting a regression line, it fits an 'S' shaped logistic function that predicts two maximum values (0 or 1).

Support Vector Machine (SVM): SVM algorithm creates best decision boundary (called hyper plane) that can create classes in an n dimensional space. SVM chooses the extreme points (also called support vectors) for creating the hyper plane. SVM can work for both linearly and non-linearly separable data.

Decision Tree: Decision Tree is a tree structured classifier that uses the predictors (independent variables) as decision nodes. Branches become the decision rules and leaves represent the outcomes. One of the advantage of Decision Tree is its explainability and is preferred in regulatory environments. Decision trees get impacted by outliers and have a tendency to do overfitting.

Random Forest: Random Forest belongs to ensemble category of machine learning algorithms. It uses multiple decision trees on multiple subsets within the dataset and takes the majority vote from these decision trees to perform classification. Greater number of trees results in higher accuracy and prevents overfitting.

Boosting is an ensemble modelling technique in which prediction power is improved by converting a number of weak learners to strong learners. There is a sequential approach behind boosting algorithms in the sense that every subsequent model learns from the errors made by previous model. Popular Boosting algorithms are: AdaBoost, Gradient Boosting and Xtreme gradient descent algorithm.

AdaBoost: Adaptive Boosting or AdaBoost uses one level decision trees called decision stumps. AdaBoost starts with a model that gives equal weightages to all data points. The subsequent models give higher weightages to misclassified data points.

Gradient Boosting: A variety of loss functions are used in the boosting technique. AdaBoost, also known as adaptive boosting, minimises the exponential loss function, which can make the algorithm more susceptible to outliers. Any differentiable loss function can be used with gradient boosting. AdaBoost is less resistant to outliers than the gradient boosting algorithm.

Stochastic Gradient Descent (SGD): Gradient descent algorithms are generic in nature and are capable of finding optimal solutions to wide variety of problems. Basic idea is to tune hyper parameters iteratively to minimise loss function. SGD does each iteration using a single sample, or a batch size of one. The sample is chosen and randomly shuffled in order to carry out the iteration.

Neural Network: In neural network, we can define relationships between the input signals received by the dendrites (x variables) and the output signal (y variable).

Each dendrite's signal is weighted (w values) according to its importance.

The input signals are summed by the cell body and the signal is passed on according to an activation function denoted by f. There are several variants of activation functions such as Relu, Sigmoid, Tanh, Linear, Unit Step etc. There are also hidden layers between input and output layers. A neural network's output and character can be changed by customizing no of hidden layers and the type of activation functions.

3.2 Dataset Description

Wisconsin Dataset

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

1) ID number

2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:
a) radius (mean of distances from centre to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

Coimbra Dataset

There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The predictors are anthropometric data and parameters which can be gathered in routine blood analysis. Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.

Quantitative Attributes:
Age (years)
BMI (kg/m2)
Glucose (mg/dL)
Insulin (µU/mL)
HOMA
Leptin (ng/mL)
Adiponectin (µg/mL)
Resistin (ng/mL)
MCP-1(pg/dL)

Labels:
1=Healthy controls
2=Patients

HOMA-IR (Homeostatic Model Assessment of Insulin Resistance) is a test used to determine a person's chances of developing diabetes. Optimal Range: 1.0 (0.5–1.4) Less than 1.0 means you are insulin-sensitive which is optimal. Above 1.9 indicates early insulin resistance. Above 2.9 indicates significant insulin resistance

MCP-1 (Monocyte chemoattractant protein-1), also known as Chemokine (CC-motif) ligand 2 (CCL2), is from family of CC chemokines. It has a vital role in the process of inflammation, where it attracts or enhances the expression of other inflammatory factors/cells.

Adiponectin is a hormone your adipose (fat) tissue releases that helps with insulin sensitivity and inflammation. Low levels of adiponectin are associated with several conditions, including obesity, Type 2 diabetes and atherosclerosis.

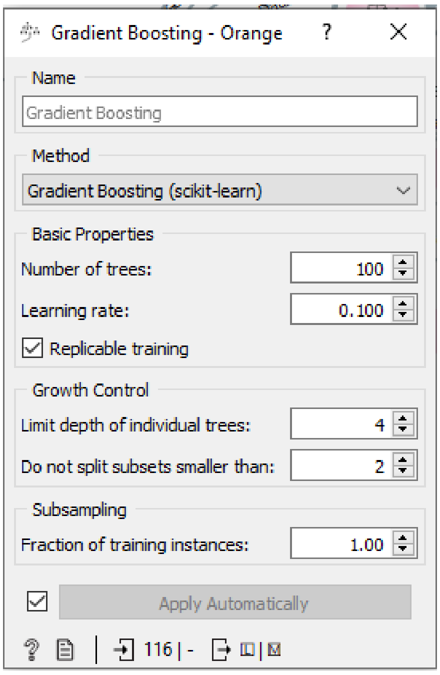
Leptin is a hormone your adipose tissue (body fat) releases that helps your body maintain your normal weight on a long-term basis. It does this by regulating hunger by providing the sensation of satiety (feeling full)

Resistin is a cysteine-rich hormone secreted from white adipocytes. Resistin is involved in insulin resistance and links obesity to diabetes in mice; it is involved in inflammation in humans. Resistin is exclusively expressed in white adipose tissue in rodents.

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

3.3 Hyper parameters tuning

Hyper parameters were tuned to obtain optimum results for the algorithms. For Coimbra dataset, best results were obtained with Gradient Boosting algorithm as shown below.



Name: Gradient Boosting

Method: Gradient Boosting (scikit-learn)

Basic Properties

Number of trees: 100

Learning rate: 0.100

☑ Replicable training

Growth Control

Limit depth of individual trees: 4

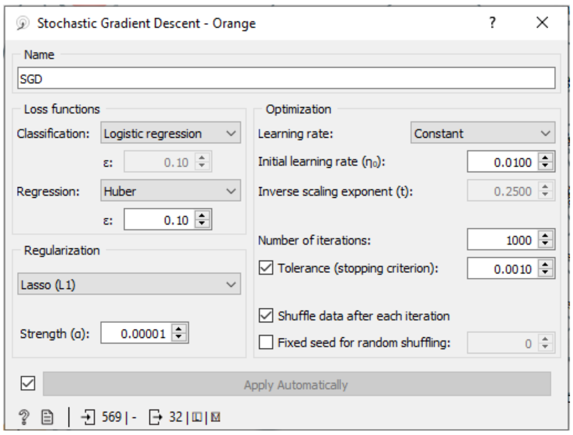
Do not split subsets smaller than: 2

Subsampling

Fraction of training instances: 1.00

☑ Apply Automatically

For Wisconsin dataset, best results were obtained with SGD:



Name: SGD

Loss functions

Classification: Logistic regression

ε: 0.10

Regression: Huber

ε: 0.10

Regularization

Lasso (L1)

Strength (α): 0.00001

Optimization

Learning rate: Constant

Initial learning rate (η_0): 0.0100

Inverse scaling exponent (t): 0.2500

Number of iterations: 1000

☑ Tolerance (stopping criterion): 0.0010

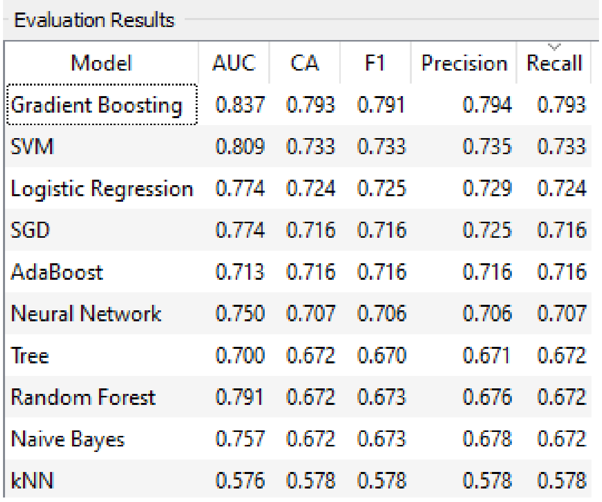
☑ Shuffle data after each iteration

☐ Fixed seed for random shuffling: 0

☑ Apply Automatically

3.4 Results

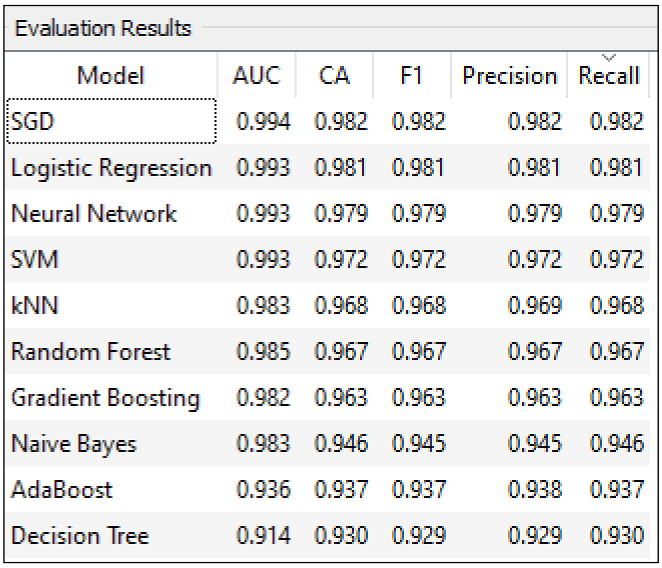
Table 1: Model Evaluation Results Using Coimbra Dataset



Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Gradient Boosting	0.837	0.793	0.791	0.794	0.793
SVM	0.809	0.733	0.733	0.735	0.733
Logistic Regression	0.774	0.724	0.725	0.729	0.724
SGD	0.774	0.716	0.716	0.725	0.716
AdaBoost	0.713	0.716	0.716	0.716	0.716
Neural Network	0.750	0.707	0.706	0.706	0.707
Tree	0.700	0.672	0.670	0.671	0.672
Random Forest	0.791	0.672	0.673	0.676	0.672
Naive Bayes	0.757	0.672	0.673	0.678	0.672
kNN	0.576	0.578	0.578	0.578	0.578

Table 2: Model Evaluation Results Using Wisconsin Dataset



Evaluation Results

Model	AUC	CA	F1	Precision	Recall
SGD	0.994	0.982	0.982	0.982	0.982
Logistic Regression	0.993	0.981	0.981	0.981	0.981
Neural Network	0.993	0.979	0.979	0.979	0.979
SVM	0.993	0.972	0.972	0.972	0.972
kNN	0.983	0.968	0.968	0.969	0.968
Random Forest	0.985	0.967	0.967	0.967	0.967
Gradient Boosting	0.982	0.963	0.963	0.963	0.963
Naive Bayes	0.983	0.946	0.945	0.945	0.946
AdaBoost	0.936	0.937	0.937	0.938	0.937
Decision Tree	0.914	0.930	0.929	0.929	0.930

4. CONCLUSIONS

In this paper, we compared results of 10 machine learning algorithms on **Wisconsin and Coimbra** dataset using a visual machine learning tool called Orange. **Particular focus was on minimising false negatives (FN), hence the algorithms were ranked on metric recall.**

In Coimbra dataset, highest recall of 0.793 was achieved with Gradient Boosting algorithm with AUC of 0.837.

In Wisconsin dataset, highest recall of 0.982 was achieved with Stochastic Gradient Descent algorithm with an AUC of 0.994.

Practical takeaway from this paper is that we can deploy this tool even in a primary healthcare facility. Technicians can be easily trained owing to its visual nature. As a first pass, patients can be screened using blood analysis report (similar to Coimbra dataset). For confirmation fine needle aspirate (FNA) of breast mass can be taken (similar to Wisconsin dataset). As 60% of breast cancers are detected in advanced stages, early detection using this approach can help lot of women fight this disease effectively.

As a next step, dataset similar to Wisconsin and Coimbra needs to be developed pan India. With more data, greater accuracy can be expected.

ACKNOWLEDGEMENTS

Authors would like to express their gratitude to the SCMS-PG, Dayananda Sagar University for their support in carrying out this research. Authors also declare they have no conflict of interest in carrying out this research.

REFERENCES

[1] Wang Lulu. Early diagnosis of breast cancer. Sensors (Basel, Switzerland). 2017; 17. https://doi.org/10.3390/s17071572 PMID: 28678153

[2] Ghassemi Marzyeh, Naumann Tristan, Schulam Peter, Andrew L. Beam, Irene Y. Chen, and Rajesh Ranganath. Opportunities in machine learning for healthcare. arxiv.org. 2018.

[3] Shah VI, Raju U, Chitale D, Deshpande V, Gregory N, Strand V. False-negative core needle biopsies of the breast: an analysis of clinical, radiologic, and pathologic findings in 27 consecutive cases of missed breast cancer. Cancer. 2003 Apr 15;97(8):1824-31. doi: 10.1002/cncr.11278. PMID: 12673707