Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 15s (2025)
https://www.jneonatalsurg.com

A Comparative Performance Analysis of Various Machine Learning Techniques in Breast Cancer Detection

Navita Rawat^{*1}, Kapil Sethi^{2}

[1]Department of computer Science, Bahra University, Shimla, India.

[2]Department of computer Science, Bahra University, Shimla, India.

.Cite this paper as: Navita Rawat, Kapil Sethi (2025) A Comparative Performance Analysis of Various Machine Learning Techniques in Breast Cancer Detection. *Journal of Neonatal Surgery*, 14 (15s), 756-764.

ABSTRACT

Cancer, a nemesis of humankind since eternity is a disease against which a relentless war has been waged by medical researchers, all guns blazing but with comparatively little success. There are more than 200 types of cancer which are known to occur in humans and breast cancer is the second most common form of cancer. This disease poses a daunting challenge to medical practitioners and every year copious numbers of fatalities occur due to this lumpy killer. Its detection at an early stage raises the chances of survival of the patients' manifolds. Use of artificial intelligence and machine algorithms has come in handy for detecting various kinds of diseases including breast cancer. Though in a nascent stage, these algorithms are turning out to be pivotal and nifty tools for medical practitioners to unravel and identify this fiendish enigma at an early stage. In this study, six classifiers i.e. Logistic Regression, Decision Tree, Random Forest, Support Vector, KNN and Naïve Bayes have been utilized and trained on Wisconsin data set to predict the occurrence of breast cancer. A comparative analysis of the performance of these algorithms have been done on the anvil of accuracy, precision, Recall and F1 score. On the basis of the empirical values obtained, it is evident that Random Forest is the most appropriate ML technique to detect the breast cancer.

Keywords: *Breast Cancer, Logistic Regression, Support Vector, Decision Tree, KNN, Naïve Bayes, Random Forest, Accuracy, Precision, Recall and F1 Score.*

1. INTRODUCTION

"There is none!" These are the laconic yet prophetic words of Egyptian physician Imhotep (circa 2600 BC) while prescribing therapy for bulging lumps in women's breasts which uncontrollably spread under their skin. The eerie and chilling echoes of these words still haunt and reverberate in modern world even after passage of several millennia. Cancer, despite phenomenal advances in medical research, still remains a formidable foe for the medical fraternity in particular and society in general. There have been many moments of exuberance and hope in recent past when it felt like the cure of this 'hideous adversary' was just round the corner, but the initial euphoria soon fizzled out like a damp squib and Cancer, like the proverbial Phoenix, remains almost as invincible as ever.

It is a disease in which physiology of certain body cells goes haywire, due to certain environmental, pathogenic or genetic factors, and they forget to 'switch off' their multiplication process and just go berserk. These recalcitrant cells soon form an ominous lump of cells which signals the onset of the dreaded malady. When the cells start invading other tissues and become extremely fulminant and aggressive, it is a stage which in medical parlance is known as 'metastasis'. Early detection of such rouge cells in the human body is a silver lining in otherwise bleak scenario which helps in adding more years to the life of patients and in case of some lucky ones, eradicating the malignancy all together. Many procedures like medical imaging, biopsy, physical examination, blood tests etc. have been utilized to detect the cancerous lumps. The most recent phenomenon in this direction is use of sophisticated Artificial Intelligence algorithms to study patterns and deciphering the onset of cancer at early stages.

2. LITERATURE REVIEW

Though, currently the usage of ML algorithms in detection of various diseases including Breast Cancer is only at a very primordial stage, the results obtained are so encouraging that research in this field in expanding exponentially by leaps and bounds with each passing day. The ultimate goal of researchers in this field is to zero-in on a ML Technique which predicts the occurrence of Breast Cancer with perfect accuracy and precision. Some of the seminal works done by various scholars using different algorithms and results obtained thereof are tabulated as under:

TABLE I. Related Earlier Studies: A Comparative Analysis

Sl. No	Author Name	Classifier Used	Accuracy %	Precision %
1.	Subham Sharma, Archit Aggarwal	K RF Naïve Bayes	95.90 94.74 94.47	98.27 92.18 88.52
2.	Madhuri Gupta	DT MLP SVM	96.9 90.9 93.9	100 99.2 97.21
3.	P. Sathiyanarayana	DT KNN	99.0 97.0	- -
4.	Tsehay Admassu,	DT SVM	87.12 91.92	87.1 88.2
5.	Dana Bazazeh	RF SM Bayesian	96.6 97.0 97.1	96.6 97.0 97.2
6.	M. Tahmooresi	NB ANN SVM	96.21 89.88 97.17	- - -
7.	Meriem Amrane	KNN Naïve Bayes	97.0 96.0	- -
8.	Sharmin Ara, Annesha Das, Ashim Day	SVM, RF	96.5	-
9.	Ganjar Alfian, Muhammad Syafrudin	SVM	80.23	-
10.	Shler Farhad Khorshid, Adnan Mohsin Abdulazeez	SVM LR NB	96.6 96.1 96.1	- - -

3. MACHINE LEARNING ALGORITHM

These are specialized programs which help the machine to arrive at specific results by studying the patterns which are perceptible in any given set of data. These patterns are learnt by machines by analyzing the trends in past data and results obtained thereof. There are many Machine Learning algorithms which are in vogue currently. Some of them are as under:

3.1 Logistic Regression

This supervised learning technique is used for binary classification, where independent inputs are classified in discrete classes of 0 and 1.

3.2 Decision Tree

A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It's a tree-like model of decisions and their possible consequences.

3.3 Random Forest

A Random Forest is an ensemble learning method used for both classification and regression tasks in machine learning. It builds upon the concept of decision trees to improve performance and robustness. Random Forest works good for large datasets. It has less variance, more flexible and possesses high accuracy.

3.4 Support Vector Classifier

A Support Vector Classifier is used for classification tasks in machine learning. The main goal of an SVC is to find the optimal boundary (or hyperplane) that separates data points of different classes with the maximum margin.

3.5 K Nearest Neighbor

KNN finds resemblance of new data with available classes and categorizes the new data into the available classes accordingly.

3.6 Naïve Bayes

Naive Bayes is a family of probabilistic algorithms used for classification tasks in machine learning. It means one feature does not depend on other and all the features are independent. Bayes means that it depends on the principle of Bayes theorem. This classifier is good for multiclass problems.

4. PROPOSED METHODOLOGY

4.1 Dataset Description

In this research paper, the data set utilized for training and testing the various ML algorithms is Wisconsin Diagnosis Breast Cancer (WDBC) data. The performance of the different algorithms is rated on the basis of four criteria: Accuracy, Precision, Recall and F1 score.

4.2 Feature Scaling

Feature scaling is a preprocessing technique used in machine learning to standardize the range of features or attributes in the dataset. This technique is also known as data normalization.

4.3 Data Visualization of empirical result

4.3.1 Histogram

Histograms are diagrammatic representations utilized to visualize the available data in the form of bars. In these diagrams, a bar represents a particular class or items belonging to a particular group.

In Figure 1, patients are classified in terms of Malignant and benign tumors. One bar represents patients with Benign tumor consisting of 357 patients (62%) while the other bar signifies patients with Malignant tumors which sum up to 212 patients (38%).

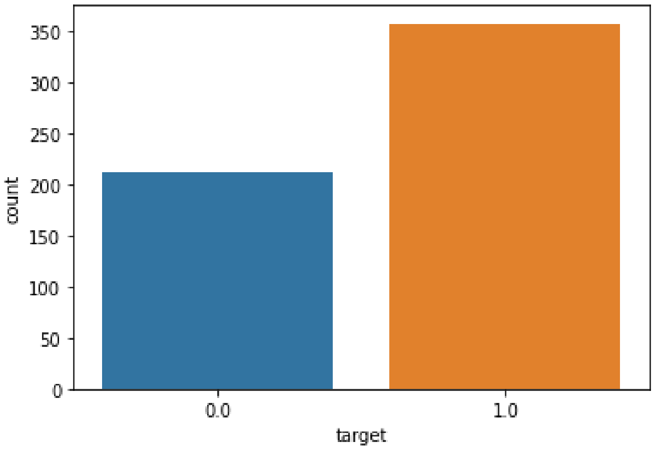


Figure 1 Class Distribution

4.3.2 Pairplot

The seaborn pairplot depicts the relationship between two continuous variables with the help of data visualization technique. The pairplot forms a grid of axes and plot values on the two axes of the graph. Nucleus features are shown against the target in Figure 2. A higher value obtained for the features is directly proportional to high propensity for a malignant tumor while lower values signify the probability of a benign tumor.

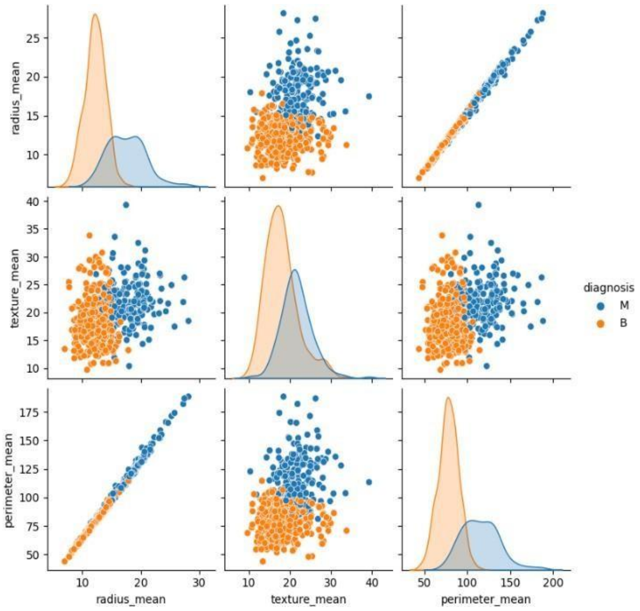
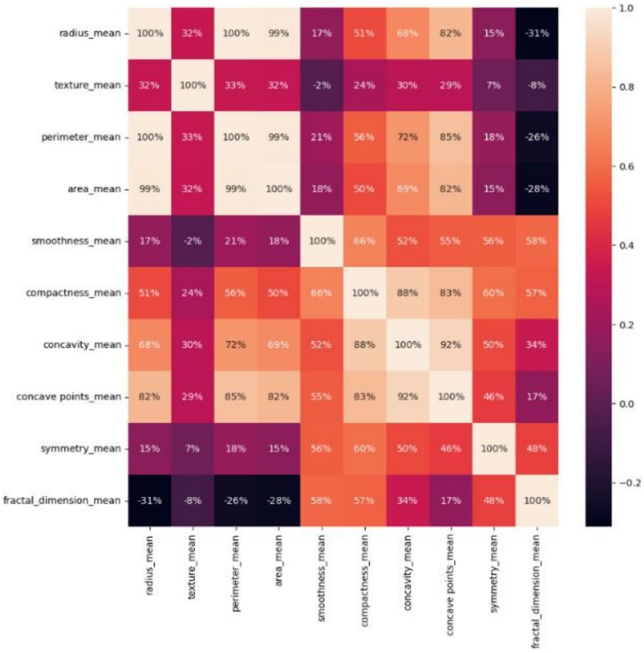


Figure 2 Nucleus Features vs Target

4.3.3 Heatmap

A Heatmap is pictorial depiction of the values of the dataset in form of a visualized matrix. It has a monochromatic scale, depicting a resultant correlation between the selected 30 features.



	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
radius_mean	100%	32%	100%	99%	17%	51%	68%	82%	15%	-31%
texture_mean	32%	100%	33%	32%	-2%	24%	30%	29%	7%	-8%
perimeter_mean	100%	33%	100%	99%	21%	56%	72%	85%	18%	-26%
area_mean	99%	32%	99%	100%	18%	50%	69%	82%	15%	-28%
smoothness_mean	17%	-2%	21%	18%	100%	66%	52%	55%	56%	58%
compactness_mean	51%	24%	56%	50%	66%	100%	88%	83%	60%	57%
concavity_mean	68%	30%	72%	69%	52%	88%	100%	92%	50%	34%
concave points_mean	82%	29%	85%	82%	55%	83%	92%	100%	46%	17%
symmetry_mean	15%	7%	18%	15%	56%	60%	50%	46%	100%	48%
fractal_dimension_mean	-31%	-8%	-26%	-28%	58%	57%	34%	17%	48%	100%

Figure 3 Correlation Matrix

5. MODEL PERFORMANCE AND RESULT ANALYSIS

When various ML algorithms are trained on the given data set, the values which are obtained can be either 0 or 1. If the value is 1, it depicts malignancy while 0 denotes that the person is benign. All the obtained values are plotted in a powerful tool called Confusion Matrix. In this matrix, the true and false instances obtained from a particular ML algorithm are displayed in a tabular form.

In this research paper, all the true and false instances obtained from the six ML algorithms have been plotted in their respective confusion matrices. The values so obtained are then analyzed on the basis of following four parameters:

Accuracy – It is the metric that shows how often a machine learning model correctly predicts the outcomes. It is the ratio of correct prediction to the all prediction made by the model.

Precision – It is a measure that indicates how well a model performs by measuring the quality of its positive predictions.

Recall –Recall, also known as sensitivity or true positive rate, is a performance metric used to evaluate classification models in machine learning. It measures how well a model identifies all the relevant instances of a class.

F1 Score – It is a metric that evaluates a model overall performance by combining factors of precision and recall.

On the basis of the aforesaid four parameters, the most suitable ML Algorithm for detecting the Breast Cancer is arrived at.

5.1 Logistic Regression

The logistic regression confusion matrix is enumerated below in table II.

TABLE II. Logistic Regression Confusion Matrix

	POSITIVE	NEGATIVE
POSITIVE	86(TP)	4(FP)
NEGATIVE	3(FN)	50(TN)

Accuracy= $\frac{\mathbf{TN+TP}}{\mathbf{TN+FN+TP+FP}}$ **(1)**

Testing Accuracy = 0.9510

Precision= $\frac{\mathbf{TP}}{\mathbf{FP+TP}}$ **(2)**

Testing Precision = 0.9555

Recall= $\frac{\mathbf{TP}}{\mathbf{TP+FN}}$ **(3)**

= 0.9662

F1 Score= $\frac{2X(\text{Precision x Recall})}{\text{Precision + Recall}}$ (4)

= 0.9608

5.2 Decision Tree

The Decision tree confusion matrix is represented below in Table III.

TABLE III. Decision Tree Confusion Matrix

	POSITIVE	NEGATIVE
POSITIVE	83(TP)	7(FP)
NEGATIVE	2(FN)	51(TN)

Accuracy=$\frac{\textbf{TN+TP}}{\textbf{TN+FN+TP+FP}}$ **(5)**

Testing Accuracy = 0.9370

Precision=$\frac{\textbf{TP}}{\textbf{FP+TP}}$ **(6)**

Testing Precision = 0.9222

Recall=$\frac{\textbf{TP}}{\textbf{TP+FN}}$ **(7)**

= 0.9764

F1 Score=$\frac{\text{2X(Precision x Recall)}}{\text{Precision+Recall}}$ (8)

= 0.9485

5.3 Random Forest

The Random Forest confusion matrix is depicted below in Table IV.

TABLE IV. Random Forest Confusion Matrix

	POSITIVE	NEGATIVE
POSITIVE	67(TP)	0(FP)
NEGATIVE	3(FN)	44(TN)

Accuracy=$\frac{\textbf{TN+TP}}{\textbf{TN+FN+TP+FP}}$ **(9)**

Testing Accuracy = 0.9736

Precision=$\frac{\textbf{TP}}{\textbf{FP+TP}}$ **(10)**

Testing Precision = 0.1

Recall=$\frac{\textbf{TP}}{\textbf{TP+FN}}$ **(11)**

= 0.9571

F1 Score=$\frac{\text{2X(Precision x Recall)}}{\text{Precision+Recall}}$ (12)

= 0.9620

5.4 Support Vector

The Support Vector confusion matrix is listed below in Table V.

TABLE V. Support Vector Confusion Matrix

	POSITIVE(1)	NEGATIVE(0)
POSITIVE(1)	88(TP)	2(FP)
NEGATIVE(0)	3(FN)	50(TN)

Accuracy=$\frac{\textbf{TN+TP}}{\textbf{TN+FN+TP+FP}}$ **(13)**

Testing Accuracy = 0.9650

Precision=$\frac{TP}{FP+TP}$ **(14)**

Testing Precision = 0.9777

Recall=$\frac{TP}{TP+FN}$ **(15)**

= 0.9670

F1 Score= $\frac{2X(Precision\ x\ Recall)}{Precision+Recall}$ (16)

= 0.9723

5.5 KNN

The KNN confusion matrix is exhibited below in Table VI.

TABLE VI. KNN Confusion Matrix

	POSITIVE (1)	NEGATIVE (0)
POSITIVE (1)	89(TP)	1(FP)
NEGATIVE (0)	6(FN)	47(TN)

Accuracy=$\frac{TN+TP}{TN+FN+TP+FP}$ **(17)**

Testing Accuracy = 0.9510

Precision=$\frac{TP}{FP+TP}$ **(18)**

Testing Precision = 0.9888

Recall=$\frac{TP}{TP+FN}$ **(19)**

= 0.9368

F1 Score= $\frac{2X(Precision\ x\ Recall)}{Precision+Recall}$ (20)

= 0.9620

5.6 Naïve Bayes

The Naïve Bayes confusion matrix is illustrated below in Table VII.

TABLE VII. Naïve Bayes Confusion Matrix

	POSITIVE (1)	NEGATIVE (0)
POSITIVE (1)	87(TP)	3(FP)
NEGATIVE (0)	5(FN)	48(TN)

Accuracy=$\frac{TN+TP}{TN+FN+TP+FP}$ **(21)**

Testing Accuracy = 0.9440

Precision=$\frac{TP}{FP+TP}$ **(22)**

Testing Precision = 0.9666

Recall=$\frac{TP}{TP+FN}$ **(23)**

= 0.9456

$$\text{F1 Score} = \frac{2X(\text{Precision x Recall})}{\text{Precision + Recall}} \quad (24)$$

= 0.9559

Table VIII Result Comparison Table

Proposed Objective	Classifier Applied	Accuracy %	Precision %	Recall %	F1 Score
Detection of breast Cancer	Logistic Regression	95.10	95.55	96.62	96.08
	Decision Tree	93.70	92.22	97.64	94.85
	Random Forest	97.36	100	95.71	96.20
	Support Vector	96.50	97.77	96.70	97.23
	KNN	95.10	98.88	93.68	96.20
	Naïve Bayes	94.40	96.66	94.56	95.59

6. CONCLUSION

In this research study, various algorithms have been utilized which are trained on a robust and trusted data set to obtain meaningful, verifiable and accurate results to diagnose and predict occurrence of breast cancer in patients. The results obtained from this model infer that Random Forest is the most appropriate ML algorithm to predict the breast cancer on the basis of aforementioned four parameters. This study is just a baby step in the long and arduous journey to find a permanent and lasting cure of the hydra-headed monster called Cancer. As it is rightly said:

All great enterprises begin with one small step.

Who knows. With the help of serendipitous advances in genetic engineering, potent clinical therapies and humungous utilization of artificial intelligence techniques, cure for cancer may soon become a tangible reality. Then, the soul of legendry physician Imhotep which might be squirming in anguish due to his helplessness of not having any cure for Cancer, may finally rest in peace. Amen.

REFERENCES

[1] Rathi M, Pareek V. Hybrid approach to predict breast cancer using machine learning techniques. International Journal of Computer Science Engineering.2016;5(3): 125-136.

[2] Tahmooresi M, Afshar, Bashari Rad B, Nowshath K B, Bamiah M A. Early detection of breast cancer using machine learning techniques. Journal of Telecommunication, Electronic and Computer Engineering.2018;10(3):21-27.

[3] Muhammet Fatih Aslam, Yunus Celik, Kadir Sabanci, Akif Durdu. Breast cancer diagnosis by different machine learning method using blood analysis data. International Journal of Intelligent System and Applications in Engineering.2018;6(4): 289-293.

[4] Bharat A, Pooja N, R Anishka Reddy. Using machine learning algorithms for breast cancer risk prediction and diagnosis. IEEE 3rd International Conference on Circuits, Control, Communication and Computing. GEU.2018 :1-4.

[5] Ebru Aydindag Bayrak, Pinar Kirci, Tolga Ensari. Comparison of machine learning methods for breast cancer diagnosis. Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT).2019;1-3.

[6] Shwetha K, Spoorthi M, Sindhu S S, Chaithra D. Breast cancer detection using deep learning technique. International Journal of Engineering Research & Technology.2018; 6(13): pp. 1-4, 2018.

[7] Ch. Shravya, K. Pravalika, Shaik Subhani. Prediction of breast cancer using supervised machine learning techniques. International Journal of Innovative Technology and Exploring Engineering.2019;8(6): 1106-1110.

[8] Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S. Breast cancer prediction using machine learning. International Journal of Recent Technology and Engineering.2019; 8(4):4879-4881.

[9] Kalyani Wadkar, Prashant Pathak, Nikhil Wagh. Breast cancerdetection using ANN network and performance analysis with SVM. International Journal of Computer Engineering and Technology2019;10(3):75-86.

[10] Vishal Deshwal, Mukta Sharma. Breast cancer detection using SVM classifier with grid search techniques. International Journal of Computer Application.2019;178(31):18-23.

[11] S. Shamy, J. Dheeba. A Research on detection and classification of breast cancer using k means GMM & CNN algorithms. International Journal of Engineering and Advanced Technology2020;8(6): 501-505.

[12] Panwar N, Sethi K, Breast Cancer: “Early Diagnosis by using Artificial Intelligence” IEEE 2nd International Conference on Innovative Sustainable Computation Technology (CISCT) 2022.

[13] Rawat N, Sethi K. Early Detection of Breast Cancer by using Machine Learning Algorithm. IEEE 3nd International Conference on Innovative Sustainable Computation Technology (CISCT) 2023.

[14] V Sansya Vijayan, Lekshmy P L. Deep Learning based prediction of breast cancer in histopathological images. International Journal of Engineering Research & Technology.2019;8(7):148-152.

[15] Puspanjali Mohapatra, Baldev Panda, Samikshya Swain. Enhancing histopathological breast cancer image classification using deep learning. International Journal of Innovative technology and Exploring Engineering. 2019;8(7):2024-2032.

[16] Chandra Churh Chatterjee, Gopal Krishan. A Noval methodfor IDC prediction in breast cancer histopathology images using deep residual neural networks. 2nd International Conference on Intelligent Communication and Computational techniques (ICCT). 2019; 95-100.

[17] Siddhartha Mukherjee. The Emperor of All Maladies – A Biography of Cancer. Scribner; 2010.