Early Breast Cancer Detection using Various Machine Learning Techniques

Chhaya Gupta, Kirti Sharma
Vivekananda School of Information Technology,
Vivekananda Institute of Professional Studies
Delhi, India

***Abstract*: Breast cancer is a deadly disease that is responsible for the death of women all over the world. Breast Cancer is detected as the most important concern for women demises. According to World Health Organisation (WHO), 2.3 million women are diagnosed with breast cancer and 685000 deaths globally, the data was released at the end of the year 2020. An early diagnosis is the need of the hour. If cancer can be detected at an early stage, the survival rate can be increased. As a solution to the problem, machine learning methods are an efficient way to classify data and diagnose the disease at an early stage. In this paper, different machine learning algorithms have been used to classify breast cancer on the Wisconsin Breast Cancer Dataset. The main objective is to calculate the performance comparison of various methods according to performance metrics. In this paper, eleven classifiers have been used which are Logistic Regression, Support Vector Machine, Extra Trees Classifier, Ada Boost Classifier, Light Gradient Boosting Machine, K-Nearest Neighbour classifier, Ridge Classifier, Random Forest Classifier, Naïve Bayes, Gradient Boosting Classifier and Decision Tree Classifier. Experimental results have shown that Logistic Regression performs immensely well and achieves an accuracy of 97.89%, F1-score of 97.35%, specificity of 90.69%, and sensitivity of 100%.**

Keywords: Wisconsin, Breast Cancer, Machine Learning, Classification, Naïve Bayes, Random Forest, Logistic Regression

1. INTRODUCTION

Breast Cancer is not a transmissible or infectious disease, and mostly half of the breast cancer develops in women who have no identifiable breast cancer risk factor other than age and gender [1]. Breast cancer is one of the most common cancer in women in India and approximately 14% of all cancers in women in India as per the cancer statistics released by the Indian Council of Medical Research (ICMR) [2]. Indian women are diagnosed with breast cancer every four minutes. The disease is completely unpredictable. The time when a woman comes to know that she is suffering from Breast Cancer, it goes beyond its original state. Breast Cancer is a very common and deadly disease in women that is created by abnormal cell mutation and spreads throughout the body.

Breast Cancer is categorized as Malignant or Benign, Benign is non-cancerous but malignant is cancerous as well as harms other organs too. This disease infects women's milk ducts and it may spread to other organs of the body with the bloodstream. Various techniques are used to diagnose this disease like ultrasound, mammography, Biopsy, etc.

Machine learning and Deep Learning techniques are very effective ways to analyse and classify the data at a faster rate too. As there is a lack of skilled people in the field, a Computer-Aided Diagnosis (CAD) system is proposed for better classification and accurate results. Machine learning algorithms are one of the options for detecting and classifying breast cancer at an early stage with high accuracy.

In this paper, Wisconsin Breast Cancer Dataset has been used with different Machine learning algorithms to predict whether a person is Benign or Malignant. Eleven machine learning methods have been analysed and compared based on Accuracy, F1-score, Recall, Precision, AUC, Sensitivity, and Specificity.

2. LITERATURE SURVEY

In this section, a brief literature survey of the various methods used in the state-of-art is discussed.

Chhaya Gupta et al. [3] have compared the performance analysis of various machine learning models like SVM, KNN, Random Forest, Decision Tree, and Extreme Learning Machine. The results showed that Extreme Learning Machine surpassed all other models and provides the highest accuracy. Manav Mangukiya et al. [4] have compared performance metrics for different machine learning models on Wisconsin Breast Cancer datasets and the results showed that XGBoost has the better accuracy over all other models.

Essam H. Houssein et al. [5] provided a review on machine learning and deep learning techniques for medical-imaging-based breast cancer. The review majorly reflects the classification of breast cancer with various methods. Jiande Wu et al. [6] have evaluated four classification models like SVM, KNN, Naïve Bayes, and Decision Tree, and deduced that the Support Vector Machine(SVM) was able to classify breast cancer more accurately.

Gunjan Chugh et al. [7] presented a survey on machine learning and deep learning techniques for classifying breast cancer at an early stage. The authors of this study also discussed the limitations and research gaps of various methods. Abdur Rasool et al. [8] proposed a data exploratory technique (DET) which is an ensembled technique for predicting breast cancer at an early stage. The dataset used in this study are Wisconsin Breast Cancer Dataset and Breast Cancer Coimbra Dataset and acquired better results in terms of accuracy.

Mohammad H. Alshayeji et al. [9] utilized a Shallow Artificial Neural Network model with just one hidden layer to diagnose breast cancer at an early stage. Wisconsin breast cancer dataset and Wisconsin Diagnostic Breast Cancer dataset are used for this study. The results showed that ANN acquired a remarkable accuracy. Pronab Ghosh et al. [10] have done a comparative analysis of seven deep learning techniques which were applied to the Wisconsin Breast Cancer Dataset and out of those techniques, LSTM (Long Short Term Memory) and Gated Recurrent Unit (GRU) were among the most effective methods to diagnose breast cancer at an early stage.

Tsehay Admassu [11] employed an optimized K-Nearest Neighbour model for the detection of breast cancer. The dataset used was Wisconsin Breast Cancer Dataset and when the results of optimized KNN were compared with the results of the classic KNN model, the former was declared better over the latter one. The optimized KNN achieved an accuracy of 94.35%. M. Divyavani et al. [12] utilized the Wisconsin Breast Cancer Dataset to provide a comparison between SVM and ANN to detect breast cancer at an early stage. Meerja Akhil Jabbar [13] used Wisconsin Breast Cancer Dataset and proposed an ensemble model built with a Bayesian network and Radial Basis Function and the results show that the proposed model achieved an accuracy of 97%.

Anuradha Reddy [14] evaluated the Support Vector Machine classifier on the Wisconsin Breast Cancer dataset and achieved an accuracy of 96.09%. Wathiq Laftah Al-Yaseen et al. [15] proposed a modified K-means algorithm that handles noise and irregularity in data. Two datasets were used namely, the Wisconsin Breast Cancer dataset and Wisconsin Diagnostic Breast Cancer.

3. METHODOLOGY

a. Dataset Used

In this study, Wisconsin Breast Cancer (Diagnostic) Dataset has been used from the Kaggle repository that consists of 569 records and 32 attributes with features and diagnoses. Each record has an instance of cancerous and non-cancerous cells. The dataset is divided into training and testing sets and the split ratio is 0.2. The dataset describes whether each record entered in the dataset is of a benign patient or a malignant patient. If a person is benign then it means that he/she is not having cancer and if a person is malignant then it means that he/she is having cancer.

b. Data Visualization

This section graphically represents the Wisconsin Breast Cancer (Diagnostic) Dataset in terms of the features provided in the dataset like mean_radius, mean_texture, mean_perimeter, mean_area, mean_smoothness, mean_compactness, mean_concavity, mean_concavepoints, mean_symmetry, and mean_dimension_fractal. Fig. 1 represents the Dataset features in two categories: Benign and Malignant.

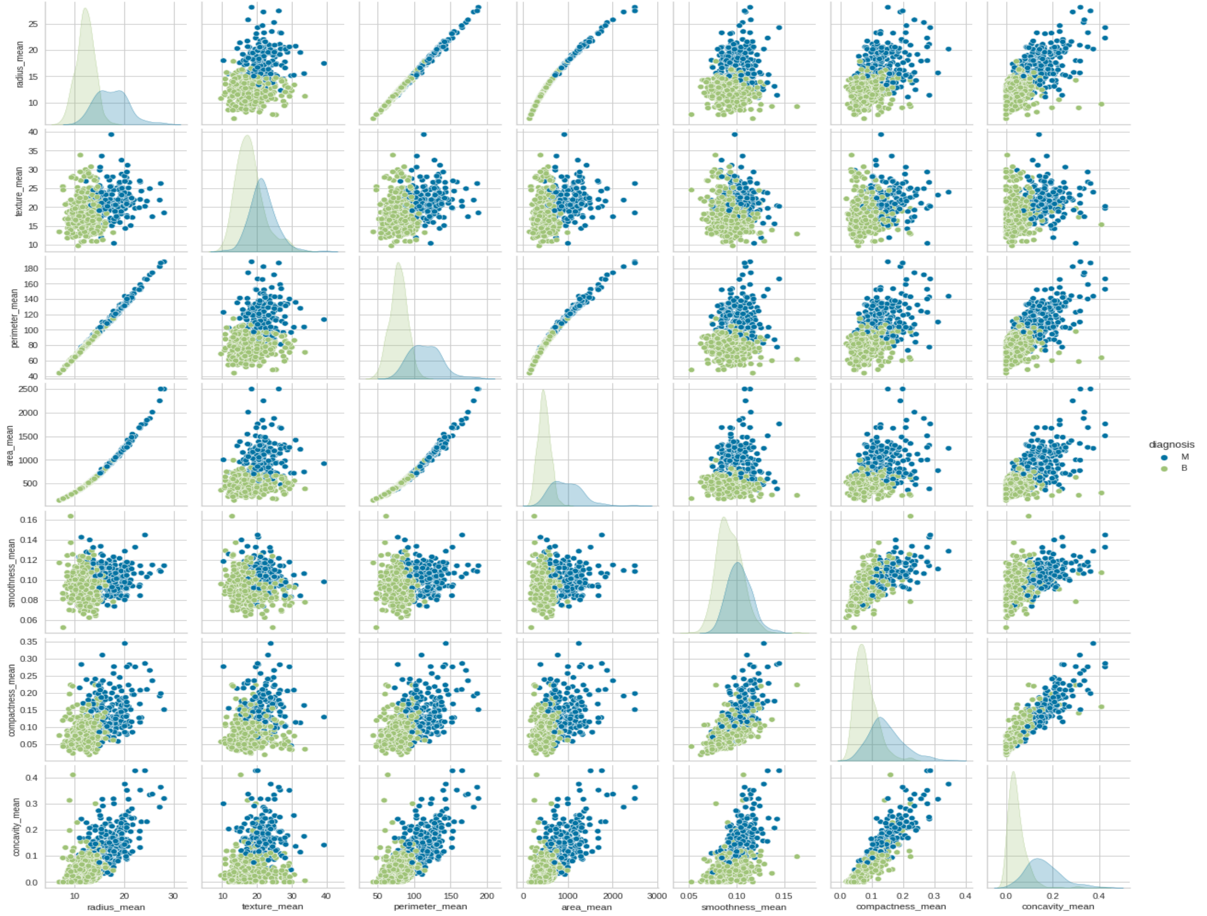

Fig.1 Pairplot of various Features (mean_radius, mean_texture, mean_perimeter, mean_area, mean_smoothness, mean_compactness, mean_concavity, mean_concavepoints, mean_symmetry and mean_dimension_fractal)

c. Methods Used

In this research, Anaconda Spyder version 3.8 was used. In this study, the authors used so many classification techniques like Logistic Regression [16], Support Vector Machine [17], Extra Trees Classifier [18], Ada Boost Classifier [19], Light Gradient Boosting Machine [20], K-Nearest Neighbour classifier [21], Ridge Classifier [22], Random Forest Classifier [23], Naïve Bayes [24], Gradient Boosting Classifier [25] and Decision Tree Classifier [26] for early detection of Breast Cancer. In addition, with all the methods, Outlier Removal and Normalisation techniques are also used. Fig. 2 provides an outline of system architecture.

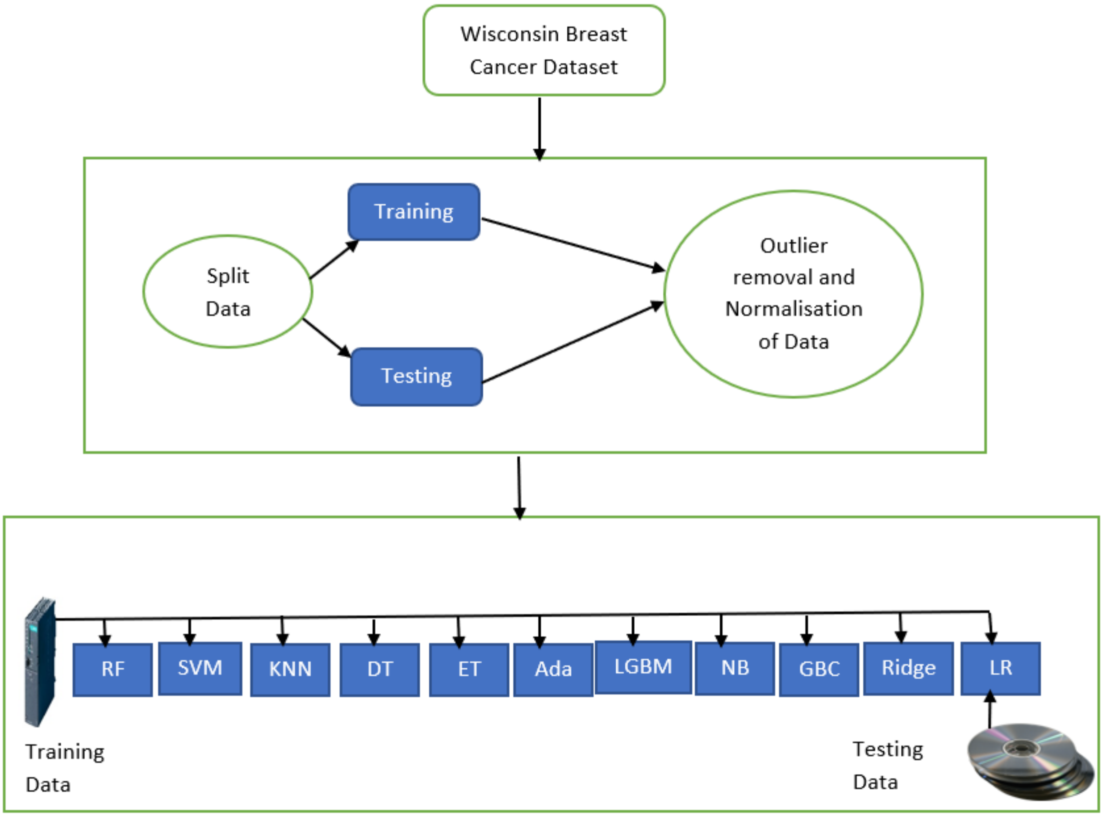

Fig. 2 System Architecture

An outlier is an observation point that is distant from other observations, which means that it is a mistake that happens during data collection. Normalization is a technique that is used to help reduce data duplication when designing datasets, also resulting in an improvement in data integrity. Table 1 provides an overview of all the methods used with their accuracy, F1-Score, AUC, Recall, and Precision. The results clearly show that Logistic Regression achieved the highest accuracy of 97.89%, precision of 98.79%, and F1-score of 97.35%.

Table 1. Performance metrics calculated for different methods used in the study

Model Name	Accuracy	AUC	Recall	Precision	F1-score
Logistic regression	97.89	99.53	96.17	98.79	97.35
Extra Trees Classifier	96.29	99.09	94.21	96.98	95.40
Ada Boost Classifier	96.04	98.97	93.67	96.91	95.01
SVM	96.02	0.001	97.46	93.55	95.38
Light Gradient Boosting Machine	95.50	99.16	92.96	96.24	94.42
KNN	95.25	98.23	92.33	96.23	94.07
Ridge Classifier	94.99	0.002	89.13	98.66	93.50
Random Forest	94.97	99.04	93.62	94.44	93.92
Naïve Bayes	93.91	98.88	93.00	92.60	92.62
Gradient Boosting Classifier	93.38	98.79	92.33	91.95	92.04
Decision Tree	91.54	91.08	88.50	91.19	89.52

4. RESULTS AND DISCUSSIONS

The study utilizes Wisconsin Breast Cancer (Diagnostic) dataset which is freely available on the Kaggle repository [27]. All the methods used in this experiment have been developed on Anaconda Spyder version 3.8. Out of eleven methods used, the results show that Logistic Regression has achieved the highest accuracy of 97.89%, F1-score of 97.35%, Recall of 96.17%, and Area Under Curve (AUC) accuracy of 99.53%. In this study, the performance of different methods used is also analysed using different performance metrics and confusion matrices have been prepared.

Different Performance metrices are described below:

$$\text{Accuracy} = (TP+TN)/(TN+FP+TP+FN) \quad (1)$$

$$\text{Sensitivity} = TP/(TP+FN) \quad (2)$$

$$\text{F1-score} = (2*TP)/(2*TP+FP+FN) \quad (3)$$

$$\text{Specificity} = TN/(TN+FP) \quad (4)$$

Where TP = True Positive = Correctly predicted malignant cases
FP = False Positive = malignant cases that are incorrectly classified
TN = True Negative = benign cases that are correctly classified
FN = False Negative = benign cases that are incorrectly classified

A confusion matrix is used to accommodate and describe the performance of a model in terms of TP, FP, TN, and FN. The basic structure of a confusion matrix is shown in Fig. 3 [28], [29].

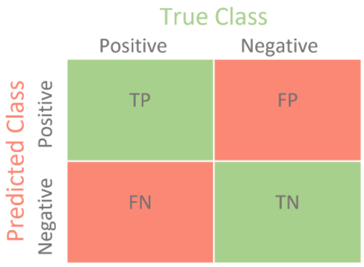

Fig. 3. The basic Confusion Matrix.

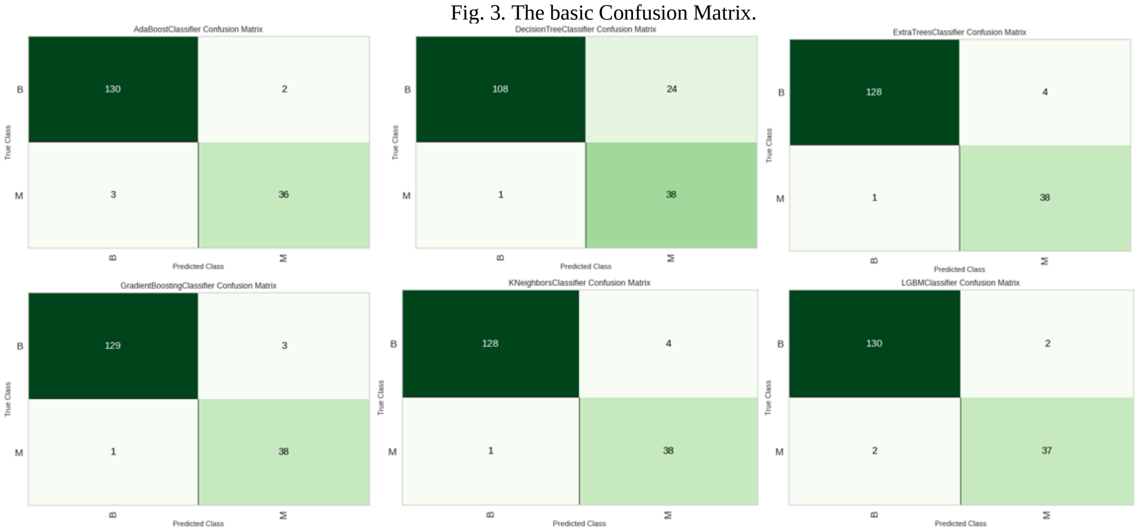

Fig. 4 represents the confusion matrices of all the different models used in the study.

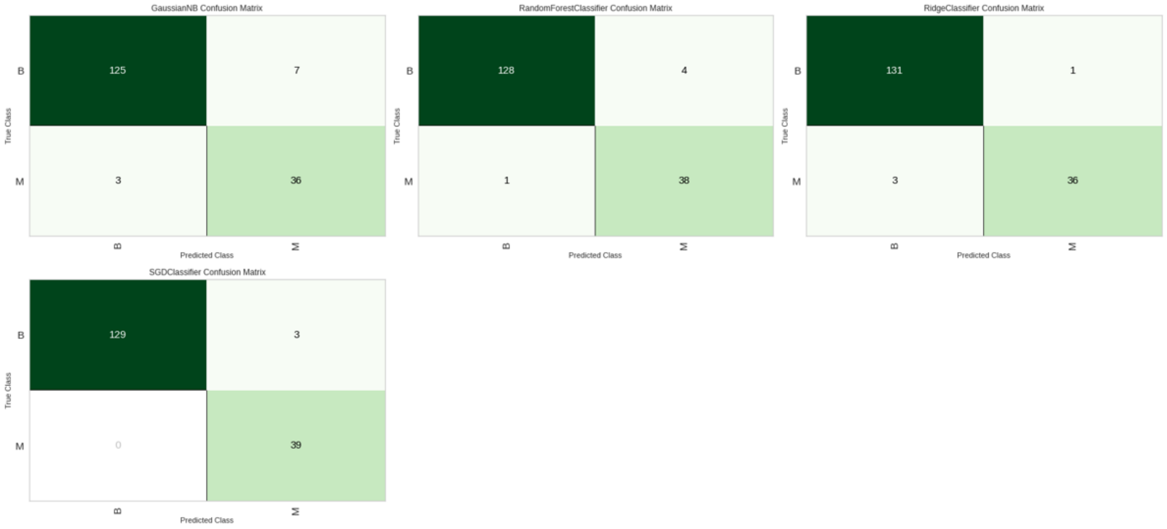

Fig. 4 Confusion matrices for various methods used.

Fig. 5 represents the confusion matrix for Logistic Regression and Table 2 presents a comparison between the above-stated methods in terms of F1-score, Accuracy, sensitivity, and specificity.

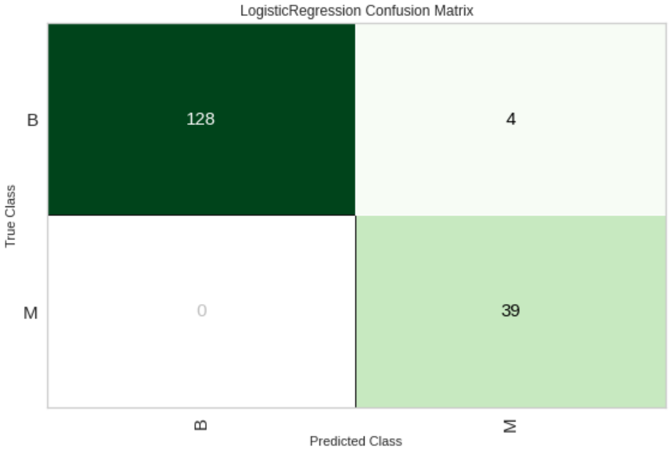

Fig. 5 Confusion matrix for Logistic Regression

Table 2. Performance evaluation of different models

Model Name	Accuracy	F1-score	Sensitivity	Specificity
Logistic regression	97.89	97.35	1.00	90.69
Extra Trees Classifier	96.29	95.40	99.22	90.47
Ada Boost Classifier	96.04	95.01	97.74	94.73
SVM	96.02	95.38	99.34	92.85
Light Gradient Boosting Machine	95.50	94.42	98.48	94.87
KNN	95.25	94.07	99.23	90.48
Ridge Classifier	94.99	93.50	97.76	97.29
Random Forest	94.97	93.92	99.44	90.49
Naïve Bayes	93.91	92.62	97.65	83.72
Gradient Boosting Classifier	93.38	92.04	99.23	92.68
Decision Tree	91.54	89.52	99.08	61.29

Table 2 clearly shows that Logistic regression has surpassed all the models with an accuracy of 97.89%, F1-score of 97.35%, specificity of 90.69%, and sensitivity of 100%. The graphical representation of the results is shown in Fig. 6.

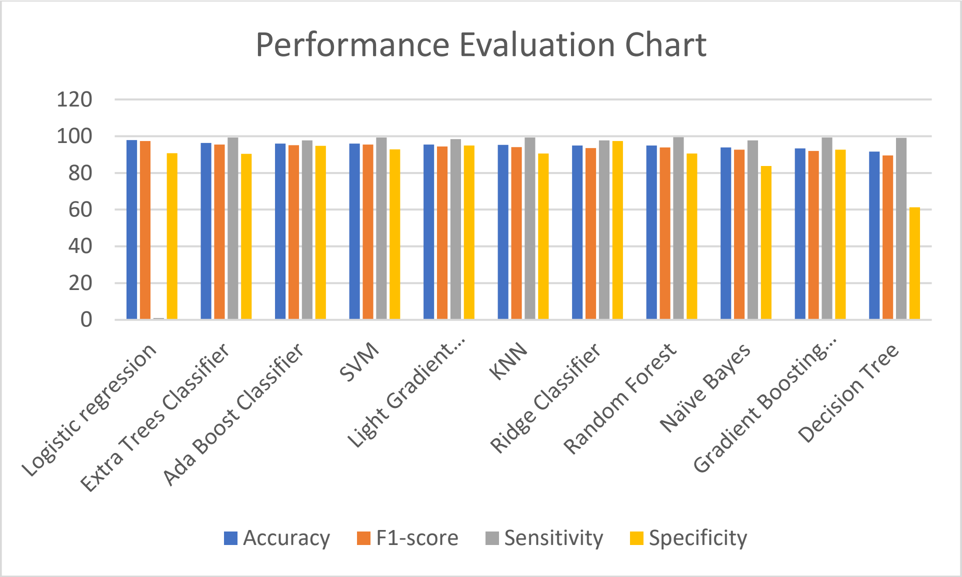

Fig. 6 Performance Evaluation Chart of various methods used

5. CONCLUSION

The main objective of this paper is to analyse different machine learning methods for the prediction of Breast Cancer at an early stage. For this, Wisconsin Breast Cancer (Diagnostic) dataset has been used which consists of 569 records and 32 features. The dataset was analysed with eleven different machine learning methods namely, Logistic Regression, Support Vector Machine, Extra Trees Classifier, Ada Boost Classifier, Light Gradient Boosting Machine, K-Nearest Neighbour classifier, Ridge Classifier, Random Forest Classifier, Naïve Bayes, Gradient Boosting Classifier and Decision Tree Classifier. The experimental results clearly show that logistic Regression achieves the highest accuracy of 97.89%, F1-score of 97.35%, specificity of 90.69%, and sensitivity of 100%. In the future, a larger dataset will be worked on to increase the efficiency and scalability of the algorithm.

REFERENCES:

[1] "Breast cancer." https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed May 27, 2022).

[2] "Cancer Statistics - India Against Cancer." http://cancerindia.org.in/cancer-statistics/ (accessed May 27, 2022).

[3] C. Gupta and N. S. Gill, "Machine Learning Techniques and Extreme Learning Machine for Early Breast Cancer Prediction," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 4, pp. 163–167, 2020, doi: 10.35940/ijitee.d1411.029420.

[4] M. Mangukiya, A. Vaghani, and M. Savani, "Breast Cancer Detection with Machine Learning February 2022," no. February, 2022, doi: 10.22214/ijraset.2022.40204.

[5] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," *Expert Syst. Appl.*, vol. 167, p. 114161, 2021, doi: 10.1016/j.eswa.2020.114161.

[6] J. Wu and C. Hicks, "Breast cancer type classification using machine learning," *J. Pers. Med.*, vol. 11, no. 2, pp. 1–12, 2021, doi: 10.3390/jpm11020061.

[7] G. Chugh, S. Kumar, and N. Singh, "Survey on Machine Learning and Deep Learning Applications in Breast Cancer Diagnosis," *Cognit. Comput.*, vol. 13, no. 6, pp. 1451–1470, 2021, doi: 10.1007/s12559-020-09813-6.

[8] A. Rasool, C. Bunterngchit, L. Tiejian, M. R. Islam, Q. Qu, and Q. Jiang, "Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis," *Int. J. Environ. Res. Public Health*, vol. 19, no. 6, 2022, doi: 10.3390/ijerph19063211.

[9] M. H. Alshayeji, H. Ellethy, S. Abed, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomed. Signal Process. Control*, vol. 71, p. 103141, Jan. 2022, doi: 10.1016/J.BSPC.2021.103141.

[10] P. Ghosh, S. Azam, K. M. Hasib, A. Karim, M. Jonkman, and A. Anwar, "A Performance Based Study on Deep Learning Algorithms in the Effective Prediction of Breast Cancer," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2021-July, Jul. 2021, doi: 10.1109/IJCNN52387.2021.9534293.

[11] T. A. Assegie, "An optimized K-Nearest neighbor based breast cancer detection," *J. Robot. Control*, vol. 2, no. 3, pp. 115–118, 2021, doi: 10.18196/jrc.2363.

[12] M. Divyavani, G. Kalpana, and P. D. Research Scholar, "an Analysis on Svm & Ann Using Breast Cancer Dataset," no. January, 2021, [Online]. Available: https://www.researchgate.net/publication/348869189.

[13] M. A. Jabbar, "Breast cancer data classification using ensemble machine learning," *Eng. Appl. Sci. Res.*, vol. 48, no. 1, pp. 65–72, 2021, doi: 10.14456/easr.2021.8.

[14] A. reddy, "Support Vector Machine Classifier For Prediction Of Breast Malignancy Using Wisconsin Breast Cancer Dataset," *J. Artif. Intell. Mach. Learn. Neural Netw.*, vol. VII, no. 21, pp. 1–8, 2022, doi: 10.55529/jaimlnn.21.1.8.

[15] W. L. Al-Yaseen, A. Jehad, Q. A. Abed, and A. K. Idrees, "The Use of Modified K-Means Algorithm to Enhance the Performance of Support Vector Machine in Classifying Breast Cancer," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 2, p. 190, 2021, doi: 10.22266/ijies2021.0430.17.

[16] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," *Proc. - 2018 Int. Conf. Robot. Intell. Syst. ICRIS 2018*, pp. 157–160, 2018, doi: 10.1109/ICRIS.2018.00049.

[17] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, "SVM and SVM ensembles in breast cancer prediction," *PLoS One*, vol. 12, no. 1, pp. 1–14, 2017, doi: 10.1371/journal.pone.0161501.

[18] V. Chaurasia and S. Pal, "Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer," *SN Comput. Sci.*, vol. 1, no. 5, 2020, doi: 10.1007/s42979-020-00296-8.

[19] M. Sharifmoghadam and H. Jazayeriy, "Breast Cancer Classification Using AdaBoost- Extreme Learning Machine," *5th Iran. Conf. Signal Process. Intell. Syst. ICSPIS 2019*, no. December, pp. 1–5, 2019, doi: 10.1109/ICSPIS48872.2019.9066088.

[20] D. Wang, Y. Zhang, and Y. Zhao, "LightGBM: An effective miRNA classification method in breast cancer patients," *ACM Int. Conf. Proceeding Ser.*, pp. 7–11, 2017, doi: 10.1145/3155077.3155079.

[21] H. Rajaguru and S. R. Sannasi Chakravarthy, "Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer," *Asian Pacific J. Cancer Prev.*, vol. 20, no. 12, pp. 3777–3781, 2019, doi: 10.31557/APJCP.2019.20.12.3777.

[22] Y. Mate and N. Somai, "Hybrid Feature Selection and Bayesian Optimization with Machine Learning for Breast Cancer Prediction," *2021 7th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2021*, pp. 612–619, 2021, doi: 10.1109/ICACCS51430.2021.9441914.

[23] B. Dai, R. C. Chen, S. Z. Zhu, and W. W. Zhang, "Using random forest algorithm for breast cancer diagnosis," *Proc. - 2018 Int. Symp. Comput. Consum. Control. IS3C 2018*, pp. 449–452, 2019, doi: 10.1109/IS3C.2018.00119.

[24] F. Paquin, J. Rivnay, A. Salleo, N. Stingelin, and C. Silva, "Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic semiconductors," *J. Mater. Chem. C*, vol. 3, pp. 10715–10722, 2015, doi: 10.1039/b000000x.

[25] A. Derangula, S. R. Edara, and P. K. Karri, "Feature selection of breast cancer data using gradient boosting techniques of machine learning," *Eur. J. Mol. Clin. Med.*, vol. 7, no. 2, pp. 3488–3504, 2020, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096914275&partnerID=40&md5=476b9182339725809c258a0b63a14a48.

[26] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Comput. Biol. Med.*, vol. 128, p. 104089, 2021, doi: 10.1016/j.compbiomed.2020.104089.

[27] "Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle." https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data (accessed May 27, 2022).

[28] "Simple guide to confusion matrix terminology." https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/ (accessed Feb. 11, 2022).

[29] "Confusion Matrix for Your Multi-Class Machine Learning Model | by Joydwip Mohajon | Towards Data Science." https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826 (accessed Feb. 11, 2022).