Applying Machine Learning Algorithms to Diagnose Breast Cancer: A Comparative Study

Harkesh Kumar[1], M. Tech (CSE) Scholar, Sunder Deep Engineering College, Ghaziabad, UP, India, hkmaurya22@gmail.com

Kavish Tomar[2], Assistant Professor, Sunder Deep Engineering College Ghaziabad, UP, India, ktomarcs17@gmail.com

Prof. (Dr.) Sandeep Gupta[3], Director, Sunder Deep Engineering College, Ghaziabad, UP, India, Guptasandeep1093@gmail.com

Abstract: Breast cancer poses a substantial global health challenge for women, emphasizing the critical importance of early detection for effective treatment. The adoption of machine learning algorithms in the detection of breast cancer has shown promise in recent times. These algorithms enhance patient outcomes. In these research nine distinct classification techniques for breast cancer detection are employed encompassing Logistic Regression (LR), Classification and Regression Trees (CART), Stochastic Gradient Descent Classifier (SGDC), Support Vector Machine (SVM), Gaussian Naive Bayes (Gaussian NB), Random Forest Classifier (RFC), Linear Discriminant Analysis (LDA), Gradient Boosting Classifier (GBC), and K-Nearest Neighbors (KNN). Two distinct data splits are used: one with 70% training and 30% testing split, and another with an 80% training and 20% testing split. The performance of each algorithm was assessed using five metrics: the area under the receiver operating characteristic curve (AUC) for the positive class, F1 score, accuracy, recall, and precision. SVM emerged as the top performer and CART was the worst among all.

Keywords — Benign, Breast Cancer Detection, Classification Techniques, Machine Learning Algorithms, Malignant, SVM

I. INTRODUCTION

Breast cancer is a condition marked by the unregulated growth of malformed cells within the breast, resulting in tumor development. These tumors can be classified into two distinct types: benign, which are non-cancerous and typically left untreated upon diagnosis, and malignant, which are cancerous and have the potential to be aggressive by invading and harming surrounding tissue [1]. Breast cancer stands as a significant contributor to mortality among women globally. As per the 2022 statistics provided by the World Health Organization, 2.3 million women universally are affected by breast cancer, leading to 670,000 fatalities. The goal of the WHO GBCI (Global Breast Cancer Initiative) is to annually decrease worldwide mortality of breast cancer by 2.5%, aiming to prevent 2.5 million deaths caused by breast cancer globally from 2020 to 2040 and one of the essential pillars for attaining this objective is the early detection of breast cancer [2]. Numerous early detection methods, including screening, are available for identifying breast cancer at an early stage. Furthermore, the progression of artificial intelligence has resulted in the development of various machine learning techniques, which can support experts' decisions across various domains. The utilization of machine learning methods is experiencing a rapid rise, aiding medical professionals in disease diagnosis. In breast cancer research, machine learning algorithms are vital in detecting and forecasting cancer [3]. In this study, nine machine learning methods including LR, CART, SVM, Gaussian NB, SGDC, LDA, RFC, GBC, and KNN are applied to the Wisconsin Breast Cancer Dataset (WBCD) taken from the UCI repository to detect breast cancer. The later sections are organized in this manner: Section II summarizes the relevant research conducted by other scholars. Section III outlines the dataset employed and elucidates the methodology employed in the study. Section IV delineates the results. Finally, Section V offers conclusions on the study's outcomes and discusses potential future directions.

II. LITERATURE SURVEY

Numerous researchers conducted studies aiming to predict breast cancer using various machine learning algorithms. This section provides the research conducted by scholars in the existing literature, as depicted in TABLE 1. In Table 1, the first column comprises references to the studies, followed by the corresponding year in the second column. The third column specifies the datasets utilized by the authors and the fourth column enumerates the machine learning algorithms or models employed. The fifth column highlights the evaluation metrics applied, and the sixth

 DOI : 10.35291/2454-9150.2024.0299

column presents the results of the research.

Table 1: Summary of Related Work Done in This Area

Studies	Year	Data Set	Models Used	Evaluation Metrics	Results
[4]	2024	Wisconsin Malignant Breast Diagnostic Dataset	KNN, SVM, NB, Decision Tree, GBC, CN2 rule inducer, Neural Network (NN), SGDC, Multilayer Perceptron (MP), Neural Decision Forests.	Matthews Correlation Coefficient (MCC), Accuracy, AUC, F1-score, Precision, and Recall.	CN2 and GB classifiers outperformed traditional models, while MLP excelled overall in breast cancer detection.
[3]	2023	Wisconsin Diagnostic Dataset	KNN, DT, SVM, NB, RFC, LR, MP.	Accuracy, Recall, Precision, and F1-score.	KNN had the best while DT had the least effective performance.
[5]	2023	WBCD and Mammographic Breast Cancer Dataset (MBCD)	SVM, KNN, DT, NB, and Ensemble Learning (EL).	Accuracy, Recall, Precision, and F1-score.	SVM demonstrated the highest accuracy in both data sets.
[6]	2023	Wisconsin Diagnostic Dataset	RF, DT, KNN, LR, Support Vector Classifier (SVC), Linear SVC	Accuracy, Recall, Precision, F1-score.	RFC shows the highest accuracy value.
[7]	2023	Basavatarakam Indo-American Cancer Hospital and Research Institute	LR, KNN, DT, SVM, Linear SVM, Radiant SVM, GBC, and XGBoost.	Accuracy, Recall, Precision, F1 score, and AUC-ROC curve.	DT performed best.
[8]	2022	WBCD	KNN, SVM, DT	Accuracy, Recall, Precision	SVM provides the best results.
[9]	2022	WBCD	SVM, KNN, LR, NB, RFC, DT, Artificial Neural Network (ANN)	Accuracy, Precision, Recall, and F-measure.	RFC performed best.
[10]	2021	Wisconsin Diagnosis of Breast Cancer (WDBC), Wisconsin (Original) Breast Cancer (WBC)	SVM, PCA, Auto encoder	Accuracy, Precision, Specificity, MCC, Sensitivity, F1-score, and AUC.	Relief SVM is better suited for breast cancer detection.
[11]	2021	WDBC	SVM, LR, KNN, DT, NB, and RFC.	Accuracy	RFC and SVM achieve higher accuracy.
[12]	2021	WDBC	SVM, KNN, DT, RFC, Ada-Boost Classifier, XGBoost Classifier.	Accuracy	XG Boost Classifier gave higher accuracy.
[13]	2021	WBCD	KNN, SV, RFC, NB, LR, GBC, ANN.	Accuracy, Cross Validation, Sensitivity, and Specificity Gained.	RFC provides maximum accuracy.
[14]	2020	Wisconsin Diagnostic Dataset	SVM, KNN, NB, DT, ANN.	Precision, Recall, and Accuracy.	ANN provides better prediction
[15]	2020	WBC and Breast Cancer dataset	DT (J48), NB, Sequential Minimal Optimization (SMO)	ROC, Standard Deviation, Accuracy	SMO's superior performance on WBC and J48 outperforms others on the Breast Cancer dataset.
[16]	2020	WBC	LR, RFC, GBC, DT, SVM.	Accuracy, Specificity, Sensitivity.	RF demonstrates superior performance.
[17]	2020	WDBC	SVM, DT, LR, RFC, KNN	Accuracy, Specificity, MCC, Precision, Recall, F1-score, False-Positive Rate (FPR) and False-Negative Rate (FNR).	RF outperformed all.
[18]	2019	WBC	SVM, ANN	Accuracy, Precision,	SVM showed

 DOI : 10.35291/2454-9150.2024.0299

				Recall, ROC Curve	better accuracy.
[19]	2018	WBC, WDBC	Genetic Programming with ML	Accuracy, Specificity, and Sensitivity.	WBC – 100% Accuracy, WBCD-98.24% Accuracy
[20]	2018	Breast Cancer Coimbra Dataset, WBCD	DT, SVM, LR, RFC, and NN	Accuracy, Precision, Recall, and F1 score.	RFC performed best.

III. METHODOLOGY

Fig (1) shows the methodology followed in this study. The first step is data preparation to collect and clean it, followed by data exploration to understand its characteristics. After that, the dataset is divided into training and test sets, and then feature scaling is applied to normalize the features eventually different classifiers are trained and evaluated.

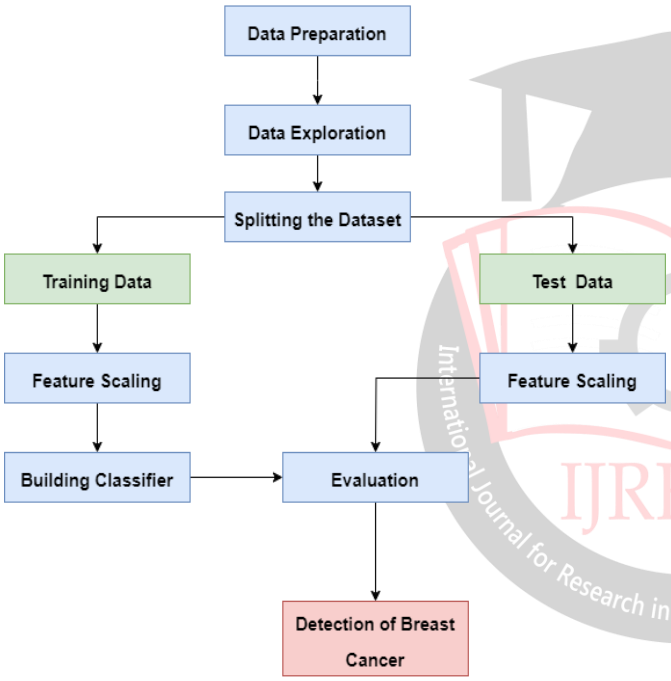

Fig. 1: Operational Procedure

A. *Data Preparation*

The breast cancer dataset used in this research is publicly available on the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic) and was created by Dr. William H. Wolberg. It contains 30 features, including 10 real-valued features computed for each cell nucleus (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension) and their mean, standard error, and "worst" or largest values. The dataset includes 569 instances, with 357 benign and 212 malignant cases. The goal is to classify the diagnosis based on these features.

B. *Data Exploration*

In this study, we utilized the Wisconsin Breast Cancer dataset to predict the malignancy or benignancy of tumors using 30 features. We conducted our analysis using Google Colab. After examining, we found that the dataset contains 569 rows and 32 columns. The 'diagnosis' column serves as the target variable, distinguishing between malignant (M) and benign (B) cancer types. Out of these, 357 instances are labeled as B (benign) and 212 as M (malignant). To prepare the data for analysis, we employed Label Encoder from the SciKit Learn library in Python. Label Encoder converts categorical or text data into numerical values, making it easier for our predictive models to interpret. Specifically, we mapped the labels B and M to 0 and 1, where 0 represents benign and 1 represents malignant.

C. *Splitting the Dataset*

The data is typically divided into training and test sets. The training set, comprising known outputs, is used for the model to learn and generalize to new data. The test data is essential for evaluating the model's performance after it has been trained on the training set. It contains unseen data that the model has not been exposed to during training. Two separate splits are employed: one with a 70% training and 30% testing ratio, and another with an 80% training and 20% testing ratio.

D. *Feature Scaling*

Feature scaling is essential in machine learning as it ensures that each feature receives fair consideration in the learning algorithm, particularly when features exhibit varying scales or ranges. This is crucial for algorithms employing distance measures like k-nearest neighbors, SVMs, and neural networks, preventing bias towards features with larger scales or ranges. In this study, the Standard Scaler method from the SciKit-Learn library is applied for feature scaling.

E. *Models*

The classification models utilized for predicting breast cancer are:

a) LR: LR is a type of supervised learning to address the problems of classification (multi or binary class). It works on the principle of probability.

b) CART: Classification and Regression Trees (CART) is a decision tree algorithm that can be used for classification and regression tasks. It works by recursively partitioning the feature space into smaller regions, and creating a decision tree that maps each region to a specific class label or continuous output value. The algorithm uses a criterion such as Gini impurity or cross-entropy to determine the best split at each node and continues splitting until a stopping criterion is met.

 DOI : 10.35291/2454-9150.2024.0299

c) **RFC:** Random Forest Classifier is a powerful ensemble method based on decision trees, initially developed by Tin Kam Ho in 1995 [18]. This method constructs multiple decision trees and classifies data samples independently. The final classification is determined by the majority vote of the individual trees, which helps in reducing overall errors. This combination of decision trees is called 'bagging'.

d) **SVM:** Support Vector Machines (SVMs) are a type of supervised machine learning algorithm that can be used for both classification and regression tasks, but are most commonly applied to classification problems. The goal of SVMs is to find a hyper plane that can best separate different classes, to minimize misclassifications. This is achieved through an iterative process that adjusts the hyperplane until the optimal separation between classes is found [19].

e) **Gaussian NB:** Gaussian NB is a machine-learning classification method that employs a probabilistic approach using the Gaussian distribution. It assumes that each feature independently contributes to predicting the target variable. The combined predictions of all features determine the probability of the dependent variable belonging to each class. The class with the highest probability is chosen as the final prediction [20].

f) **SGDC**: SGDC is a variant of the traditional gradient descent algorithm and, it works by iteratively updating model parameters so that the loss functions can be minimized, thereby enhancing the efficiency of training and convergence.

g) **GBC**: This classifier minimizes errors by focusing on instances where the previous models faltered, thereby enhancing overall prediction accuracy. It achieves this by progressively constructing a collection of weak learners, typically decision trees, in sequential manner [4].

h) **KNN**: KNN is a supervised learning algorithm utilized for both classification and regression tasks. The underlying principle of KNN is that similar kinds of data points tend to be located in close proximity to each other. This idea is used to classify any new data point; the algorithm identifies the 'K' nearest neighbors to the new data point and classifies it based on the category with the maximum number of neighbors within that group. In other words, 'K' neighbors are encircled, and the new data sample is classified according to the majority class of those neighbors [14].

i) **LDA**: LDA is a versatile machine learning technique that can be applied to both dimensionality reduction and classification problems. In classification tasks, it identifies a linear combination of features that maximizes the separation between various categories or classes. The objective of LDA is to lower the dimensionality of the feature space while retaining the essential information about the separability of the classes.

F. Evaluation Metrics

We assessed the performance of machine learning models using evaluation criteria including accuracy, precision, recall, F1 score, and AUC for positive class. Fig. (2), Fig. (3), Fig. (4), and Fig. (5) show the mathematical representation of these evaluation metrics where true positive refers to the total number of correctly predicted positive classes, true negative means that the model correctly predicted the negative class, false positive would occur when the actual value is negative and the model predicted it positive and false negative shows that the model indicates an actual positive class as negative.

i) Accuracy: Accuracy refers to the proportion of accurate predictions relative to the total number of predictions made.

$$\text{Accuracy} = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + True\ Negative + False\ Positive + False\ Negative)}$$

Fig. 2: Mathematical representation of accuracy

ii) Precision: Precision is determined by dividing the total number of positive samples predicted correctly by the total number of positive samples.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Fig. 3: Mathematical representation of precision

iii) Recall: The recall is determined by dividing the total number of positive samples accurately categorized as positive by the entire count of false negative and true positive samples.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Fig. 4: Mathematical representation of accuracy

iv) F1 Score: The F1 score serves as a classification performance metric used alongside other evaluation metrics to gauge algorithm performance. It enables the assessment of a machine learning model's effectiveness specifically in binary classification tasks.

$$\text{F1 score} = \frac{2*(Precision*Recall)}{(Precision+Recall)}$$

Fig. 5: Mathematical representation of accuracy

v) AUC: The AUC is a measure utilized to assess the

 DOI : 10.35291/2454-9150.2024.0299

performance of binary classification models. It indicates the model's ability to differentiate between negative and positive classes. In the context of breast cancer detection, the positive class refers to malignant cases. Therefore, the AUC for the positive class signifies the model's capability to detect malignant cases correctly.

IV. RESULTS

The outcomes of both splits of the machine learning classifier models trained to detect breast cancer are provided below in Table 2 & Table 3. According to both tables, SVM emerged as the top-performing classifier across both scenarios with CART consistently demonstrating inferior performance. Fig. (6) and Fig. (7) shows the comparison of evaluation metrics in the form of chart of all models used in this research with data split (70-30) and (80-20) respectively.

Table 2: **Results of Machine Learning Models in 70-30 Data Split**

Train-Test Data Split – 70% and 30 %

	Algorithm	Accuracy	Precision	Recall	F1 score	AUC
0	LR Method	0.976608	0.983607	0.952381	0.967742	0.993974
0	CART	0.906433	0.821918	0.952381	0.882353	0.916005
0	RFC	0.964912	0.938462	0.968254	0.953125	0.995885
0	SVM	0.976608	0.983607	0.952381	0.967742	0.997354
0	Gaussian NB	0.912281	0.863636	0.904762	0.883721	0.982510
0	SGDC	0.929825	0.869565	0.952381	0.909091	0.966343
0	GBC	0.976608	0.983607	0.952381	0.967742	0.997795
0	KNN	0.959064	0.982759	0.904762	0.942149	0.983613
0	LDA	0.970760	1.000000	0.920635	0.958678	0.995738

Table 3: Results of Machine Learning Models in 80-20 Data Split

Train-Test Data Split- 80% and 20%

	Algorithm	Accuracy	Precision	Recall	F1 score	AUC
0	LR Method	0.964912	0.957447	0.957447	0.957447	0.993331
0	CART	0.903509	0.846154	0.936170	0.888889	0.908384
0	RFC	0.964912	0.957447	0.957447	0.957447	0.996666
0	SVM	0.982456	1.000000	0.957447	0.978261	0.999047
0	Gaussian NB	0.903509	0.875000	0.893617	0.884211	0.984122
0	SGDC	0.973684	0.978261	0.957447	0.967742	0.989838
0	GBC	0.973684	0.958333	0.978723	0.968421	0.998095
0	KNN	0.956140	1.000000	0.893617	0.943820	0.982375
0	LDA	0.964912	1.000000	0.914894	0.955556	0.997142

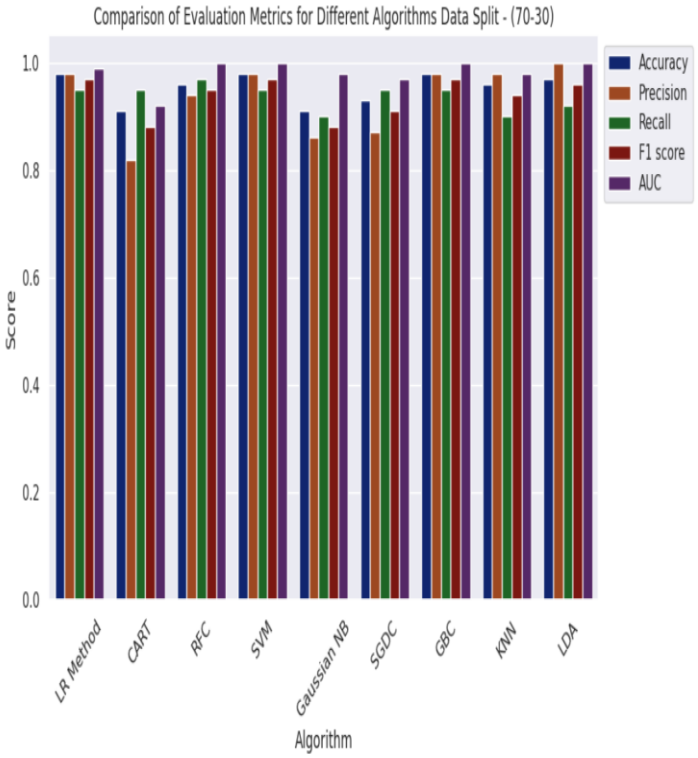

Fig. 6: Data Split (70-30) Comparison 1

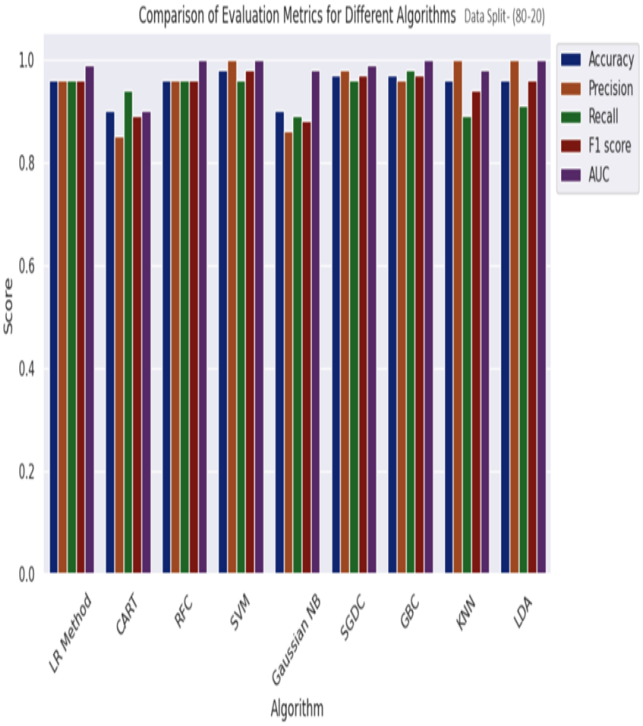

Fig. 7: Data Split (80-20) Comparison 2

V. CONCLUSION

Breast Cancer is a serious and life-threatening condition affecting women worldwide. Early detection is imperative for the successful treatment of breast cancer, as it can be fatal if not diagnosed early. Incorporating insights from the literature, it is evident that machine learning is increasingly being utilized as a decision-support system in diagnosing diseases, particularly in the field of cancer. This study

 DOI : 10.35291/2454-9150.2024.0299

evaluated multiple machine learning classifiers as LR, CART, RFC, SVM, Gaussian NB, SGDC, GBC, KNN, and LDA across two distinct training and test split scenarios: a 70% training and 20% test split, and an 80% training and 20% test split and compared the results using various metrics. In our comparison, SVM proved to be the most reliable classifier with 97.66% and 98.24% accuracy, 98.36% and 100% precision, 95.23% and 95.74% recall, 96.77% and 97.83% F1 score with 99.73% and 99.90% AUC value for positive class in both splits respectively. Additionally, CART is the worst in almost all metrics with 90.64% and 90.35% accuracy, 82.19% and 84.61% precision, 95.24% and 93.62% recall, 88.23% and 88.89% F1 score with 91.60% and 90.84% AUC value. In future research, we plan to perform multiple classifications on a more extensive and diverse dataset to fully realize the potential of this research in clinical settings.

REFERENCES

[1] L. D. Shockney, "Breast Tumors," *National Breast Cancer Foundation, Inc*, 2023. https://www.nationalbreastcancer.org/breast-tumors/ (accessed May 02, 2024).

[2] "Breast cancer," *World Health Organization*, 2024. https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed May 02, 2024).

[3] M. Karakoyun, "2 nd International Conference on Scientific and Academic Research," no. March, 2023.

[4] S. Devi, R. Kaul Ghanekar, J. A. Pande, D. Dumbre, R. Chavan, and H. Gupta, "Prediction and Diagnosis of Breast Cancer Using Machine and Modern Deep Learning Models," *Asian Pac. J. Cancer Prev.*, vol. 25, no. 3, pp. 1077–1085, 2024, doi: 10.31557/APJCP.2024.25.3.1077.

[5] E. Akkur, F. TURK, and O. Erogul, "Breast Cancer Diagnosis Using Feature Selection Approaches and Bayesian Optimization," *Comput. Syst. Sci. Eng.*, vol. 45, no. 2, pp. 1017–1031, 2023, doi: 10.32604/csse.2023.033003.

[6] A. Khalid *et al.*, "Breast Cancer Detection and Prevention Using Machine Learning," *Diagnostics*, vol. 13, no. 19, pp. 1–21, 2023, doi: 10.3390/diagnostics13193113.

[7] M. Botlagunta *et al.*, "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," *Sci. Rep.*, vol. 13, no. 1, p. 485, 2023.

[8] B. Raj, K. Raj, L. Vetrivendan, and M. Arvindhan, "Breast Cancer Detection using Machine Learning," *Proc. - IEEE 2023 5th Int. Conf. Adv. Comput. Commun. Control Networking, ICAC3N 2023*, vol. 4, no. 3, pp. 275–280, 2023, doi: 10.1109/ICAC3N60023.2023.10541841.

[9] H. El Massari, N. Gherabi, S. Mhammedi, H. Ghandi, F. Qanouni, and M. Bahaj, "An Ontological Model based on Machine Learning for Predicting Breast Cancer," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 7, pp. 108–115, 2022, doi: 10.14569/IJACSA.2022.0130715.

[10] A. U. Haq *et al.*, "Detection of Breast Cancer through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques," *IEEE Access*, vol. 9, pp. 22090–22105, 2021, doi: 10.1109/ACCESS.2021.3055806.

[11] S. Ara, A. Das, and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," *2021 Int. Conf. Artif. Intell. ICAI 2021*, pp. 97–101, 2021, doi: 10.1109/ICAI52203.2021.9445249.

[12] S. Pawar, P. Bagal, P. Shukla, and A. Dawkhar, "Detection of Breast Cancer using Machine Learning Classifier," *2021 Asian Conf. Innov. Technol. ASIANCON 2021*, pp. 1–5, 2021, doi: 10.1109/ASIANCON51346.2021.9544767.

[13] K. Mridha, "Early Prediction of Breast Cancer by using Artificial Neural Network and Machine Learning Techniques," *Proc. - 2021 IEEE 10th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2021*, pp. 582–587, 2021, doi: 10.1109/CSNT51715.2021.9509658.

[14] T. Thomas, N. Pradhan, and V. S. Dhaka, "Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey," *Proc. 5th Int. Conf. Inven. Comput. Technol. ICICT 2020*, no. February, pp. 192–196, 2020, doi: 10.1109/ICICT48043.2020.9112464.

[15] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of Breast Cancer Detection Using Different Machine Learning Techniques.," *Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings*, vol. 1234. pp. 108–117, 2020. doi: 10.1007/978-981-15-7205-0_10.

[16] M. J. Rasool, A. S. Brar, and H. S. Kang, "Risk Prediction of Breast Cancer From Real Time Streaming Health Data Using Machine Learning," no. 11, pp. 409–418, 2020, doi: 10.5281/zenodo.4284315.

[17] S. Sakib, N. Yasmin, A. K. Tanzeem, F. Shorna, K. Md. Hasib, and S. B. Alam, "Breast Cancer Detection and Classification: A Comparative Analysis Using Machine Learning Algorithms," *Lect. Notes Electr. Eng.*, vol. 844, no. March, pp. 703–717, 2022, doi: 10.1007/978-981-16-8862-1_46.

[18] E. A. Bayrak, P. Kırcı, and T. Ensari, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019, pp. 1–3. doi: 10.1109/EBBT.2019.8741990.

[19] H. Bhardwaj, A. Sakalle, A. Tiwari, M. Verma, and A. Bhardwaj, "Breast cancer diagnosis using simultaneous feature selection and classification: a genetic programming approach," in *2018 IEEE symposium series on computational intelligence (SSCI)*, 2018, pp. 2186–2192.

[20] Y. Li and Z. Chen, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl Comput Math*, vol. 7, no. 4, pp. 212–216, 2018.

[21] T. Kam, "Random decision forests," *(No Title)*, vol. 1, p. 278, 1995.

[22] S. H. Shetty, S. Shetty, C. Singh, and A. Rao, "Supervised machine learning: algorithms and applications," *Fundam. methods Mach. Deep Learn. algorithms, tools Appl.*, pp. 1–16, 2022.

[23] M. Carla, "Gaussian Naive Bayes Explained With Scikit-Learn," 2023. https://builtin.com/artificial-intelligence/gaussian-naive-bayes

 DOI : 10.35291/2454-9150.2024.0299