Breast Cancer Prediction by Ensembling Machine Learning Algorithms and Explainable AI

M Sobhana
Faculty at the Department of CSE
V R Siddhartha Engineering College
Vijayawada, Andhra Pradesh, India
sobhana@vrsiddhartha.ac.in

Anil kumar Palaketi
Department of CSE
V R Siddhartha Engineering College
Vijayawada, Andhra Pradesh, India
anilkumarpalaketi@gmail.com

Ramya Nalabothu
Department of CSE
V R Siddhartha Engineering College
Vijayawada, Andhra Pradesh, India
ramyamohanraonalabothu@gmail.com

***Abstract*—A primary cause of death is cancer, which is a result of abnormal cell growth. Globally, breast cancer is a significant contributor to female fatalities, and its prevention is challenging due to unidentified causes. However, early detection is pivotal for reducing risk and improving survival rates. Advanced imaging techniques like mammography and ultrasound are instrumental in diagnosing breast cancer. This model integrates machine learning and Explainable AI to predict breast cancer. Trained on a dataset with diverse features from fine needle aspiration of breast masses, the model not only determines whether a patient is positive or negative but also sheds light on the importance of specific features of the cancerous cell. In cases of a positive diagnosis, early detection empowers patients to promptly seek essential treatment, significantly enhancing their chances of survival.**

Keywords- Breast Cancer, Fine needle aspiration, ensemble, Machine Learning, Explainable AI.

I. INTRODUCTION

Breast cancer, affecting one or both breasts, can result from genetic factors or lifestyle choices. It is classified into different types, including ductal and lobular carcinoma, in situ and invasive cancers, and less common variants. Staging from I to IV determines the extent of the disease. Breast cancer in India is a significant concern, comprising 14% of women's cancers. In 2018, there are 162,468 cases of newly diagnosed breast cancer were reported, resulting in 87,090 documented fatalities, with over 50% diagnosed in advanced stages. The post-cancer survival rate for Indian women is 60%, lower than the U.S. at 80%, with Kerala having the highest cancer rate. Raising awareness and early diagnosis are essential for improving outcomes [1]. The American Cancer Society predicts that around 297,790 new cases of invasive breast cancer will be diagnosed in the US in the year 2023, affecting women. Furthermore, it is anticipated that around 55,720 additional instances of invasive and in situ breast cancers will be diagnosed. Unfortunately, it is projected that over 43,700 women die due to breast cancer in the same time frame [2].

Clinical methods for breast cancer detection include Breast Self-Examination for assessing lumps or texture changes, Clinical Breast Examination, and Mammography, which may yield false results. Breast Ultrasound aids in distinguishing solid from fluid-filled cysts. Magnetic Resonance Imaging assesses disease extent but is costly. Breast Biopsy is definitive but may cause discomfort, while Genetic Testing raises ethical concerns. Machine learning models, having undergone training on extensive datasets, possess the capability to recognize and interpret subtle patterns and nuances in data. Breast tissue abnormalities, reduce human error and interpretation discrepancies found in clinical methods like mammography. They aid in early breast cancer detection by incorporating diverse data sources, allowing for personalized risk assessments and recommendations based on individual risk factors. This precision can reduce unnecessary invasive procedures, minimizing patient discomfort and anxiety. These models efficiently process large data volumes, continuously improving their accuracy with new information.

Ultimately, machine learning has the potential to lower costs by reducing late-stage diagnoses and extensive treatments. This framework utilized machine learning, training a model on a dataset: with attributes of patient ID, diagnosis, and specific characteristics of tumor cells. Employing ensemble learning for enhanced accuracy, we scrutinized the performance of multiple algorithms in predicting breast cancer. Integrating the model into a user-friendly website, users(doctors) input health and tumor attributes, and the backend processes this information to deliver a clear output on whether the patient has breast cancer or not. Incorporating Explainable AI (shape tool) serves the purpose of providing nuanced insights into the specific features of tumor cells that predominantly contribute to breast cancer. This integration enhances the interpretability of the machine learning model's outcomes, instilling a heightened level of confidence in users. The transparent nature facilitated by Explainable AI increases trust in the model prediction.

II. LITERATURE REVIEW

Reshma et al. [3] introduce an automatic segmentation approach using Fourier Transform, ensuring spatial information incorporation, independence of magnification, and automatic determination of morphological operation inputs. The proposed method enhances image clarity for pathologists efficiently and exhibits speed advantages. Arooj et al. [4] study addresses the critical issue of breast cancer, emphasizing early detection and classification through a proposed model employing deep learning and transfer learning techniques. The customized

CNN-AlexNet model showcases the effectiveness of transfer learning in achieving superior accuracy.

Shafique et al. [5] Involves breast cancer prediction through fine-needle aspiration features. It employs three feature selection techniques: PCA, Chi2, and SVD, optimizing classifier performance. Addressing imbalanced data using SMOTE enhances classifier accuracy, with KNN outperforming others. Botlagunta et al. [6] study aims to create a non-invasive breast cancer classification system using ML. Utilizing text mining from Electronic Medical Records facilitates blood profile data separation for metastatic breast cancer identification.

Islam et al. [7] research presents a comprehensive prediction of breast cancer by using five ML algorithms, underscoring the notable efficacy of artificial neural networks (ANNs) in terms of precision, F1 score, and achieving the highest accuracy. Liza et al. [8] explored eight machine learning techniques, including AdaBoost and Randomforest, which demonstrate high accuracy (99.20%) and effectiveness in predicting breast cancer.

Humayun et al. [9] Leverages advancements in gene expression analysis and deep learning techniques, particularly transfer learning with InceptionResNetV2. Jakhar et al. [10] the proposed SELF framework employs a stacked-based ensemble learning approach, utilizing AdaBoost, Extra tree, Gradient Boosting, Random Forest, and KNN9 classifiers on wbcd datasets and Breakhis.

Ayoola et al. [11] employs various feature selection algorithms, including Tree-based techniques, Genetic Algorithm, Fisher_Score, Chi-square test, Mutual Information Gain, Correlation, Variance, Coefficient, Lasso, Ridge Regressors with L2 regularization, and Linear Regressors with L1 regularization. Samieinasab et al. [12] introduces the Meta-Health Stack, an ensemble-based framework utilizing the Extra Trees classifier for efficient breast cancer prediction. The framework integrates attributes from Pearson's Correlation, Variance Inflation factor, and Information Gain for feature selection and employs Boosting, Bagging, and Voting approaches combined through Stacking.

Ebrahim et al. [13] proposed an analysis that compares classical and deep learning techniques, making use of a dataset including 1.7 million data records from the National Cancer Institute (NIH), USA. Deep learning techniques like probabilistic neural network (PNN), deep neural network (DNN), and recurrent neural network (RNN) are compared with classical methods like SVM, DT, logistic regression (LR), and linear discriminants (LD). Ak et al. [14] employs data visualization and various machine learning techniques, including KNN, SVM, naïve Bayes, DT, random forest, rotation forest, and logistic regression. The application of these techniques is carried out using R, Minitab, and Python.

Vaka et al. [15] highlights the urgency of early detection, given the alarming statistics that one woman is diagnosed every two minutes, and one woman dies every nine minutes due to breast cancer. Amoroso et al. [16] used hierarchical clustering and adaptive dimensional reduction to create an explainable AI framework for tailored breast cancer treatment. They discovered important clinical characteristics, emphasizing the role of molecular subtypes. Benefits include the potential for data-driven therapies and strong alignment with established guidelines. The tiny dataset and the requirement for additional validation in larger populations are drawbacks, though. All in all, the method provides information for customized interventions, but careful evaluation of its existing limitations is necessary.

Yue et al. [17] underscores the importance of continued algorithm development and exploration of novel ML approaches to enhance BC diagnosis and prognosis. Alsabry et al. [18] synthesizes existing literature on breast cancer risk factors, data sources, and ML algorithms for prediction, providing insights into the complex nature of the disease and evaluating the performance of diverse ML models using clinical, genomic, and lifestyle data sources.

Amethiya et al. [19] study evaluates the utilization of diverse ML algorithms and biosensors for the early detection of breast cancer, comparing their performance metrics, including accuracy, precision, and recall, across diverse databases and analytes. It emphasizes the capability of combining biosensors and machine learning to enable rapid and effective breast cancer detection. Jansen et al. [20] study used LIME and SHAP for model interpretation in order to apply machine learning to the prediction of 10-year overall survival in patients with breast cancer. LIME showed some discrepancies despite its overall consistency, underscoring evaluation difficulties. Transitions between survival outcomes were explained by the feature ranges' identified turning points. Benefits include increased model transparency, although in certain situations, reliability raises questions. To validate and generalize these approaches of interpretability for application in clinical practice, more research is required.

III. METHODOLOGY

This model made use of the Wisconsin dataset which encompasses a diverse set of information related to tumor cells derived from patients. The dataset comprises various metrics of the cell undergone fine needle aspiration. These features not only describe the physical characteristics of the cells but also provide diagnostic information. Each measurement is assessed at different stages of the tumor, including mean, standard error, and worst-case scenarios. The data for these features were obtained by analyzing the cell cytology images taken from the fine needle aspiration technique. The dataset consists of 569 unique patient records, with 357 containing information on benign tumor cells and 212 on malignant tumor cells.

TABLE 3.1
Attribute Description

Attribute	Description
ID number	A unique number is allocated to every specimen.
Diagnosis	'M' indicates malignant and 'B' indicates Benign
Radius	mean distances between the center and outermost locations.
Texture	The grayscale values' standard deviation.
Perimeter	The cell nucleus's periphery.
Area	section of the nucleus of a cell.
Smoothness	Variation of radius lengths locally.
Compactness	perimeter^2 / area - 1.0
Concavity	The extent of concavity in specific contour features.
Concave points	represent the quantity of concave areas along the contour.
Symmetry	nuclear symmetry in a cell.
Fractal dimension	"Coastline approximation" - one

Table 3.1 displays the cell nuclei are included in the collection; each nucleus is given a unique ID and is classified as benign or malignant. Ten real-valued attributes are included in it: Area, Compactness, Concave spots, Concavity, Fractal dimension, Radius, Smoothness, Symmetry, and Texture. Thirty characteristics (mean of the three biggest values) indicating the "worst" circumstances, standard error, and mean are computed for each nucleus. The fields that the Mean Radius, Radius SE, and Worst Radius refer to are, respectively, fields 3, 13, and 23. The dataset has been carefully selected, with all feature values up to four significant digits accurately recorded and no missing attribute values. Regarding the distribution of classes, 357 cases are categorized as benign and 212 as malignant. Ten features are determined for every cell nucleus.

This research employs a systematic approach to breast cancer prediction, initially analyzing diverse machine learning algorithms trained on a dataset featuring cell nucleus attributes associated with breast cancer. Subsequently, an accurate ensemble model is constructed by integrating the most effective individual models. A user-friendly interface is developed to allow input of cell features, which are then processed by the ensemble model to predict the likelihood of breast cancer presence. Enhancing interpretability, the SHAP (Shapley Additive exPlanations) library, a tool for Explainable AI, is utilized to evaluate feature importance. This comprehensive methodology aims for both accurate predictions and a nuanced understanding of the specific contributions of individual features in the context of breast cancer diagnosis. The methodology diagram shown in Figure 3.1 describes the proposed methodology.

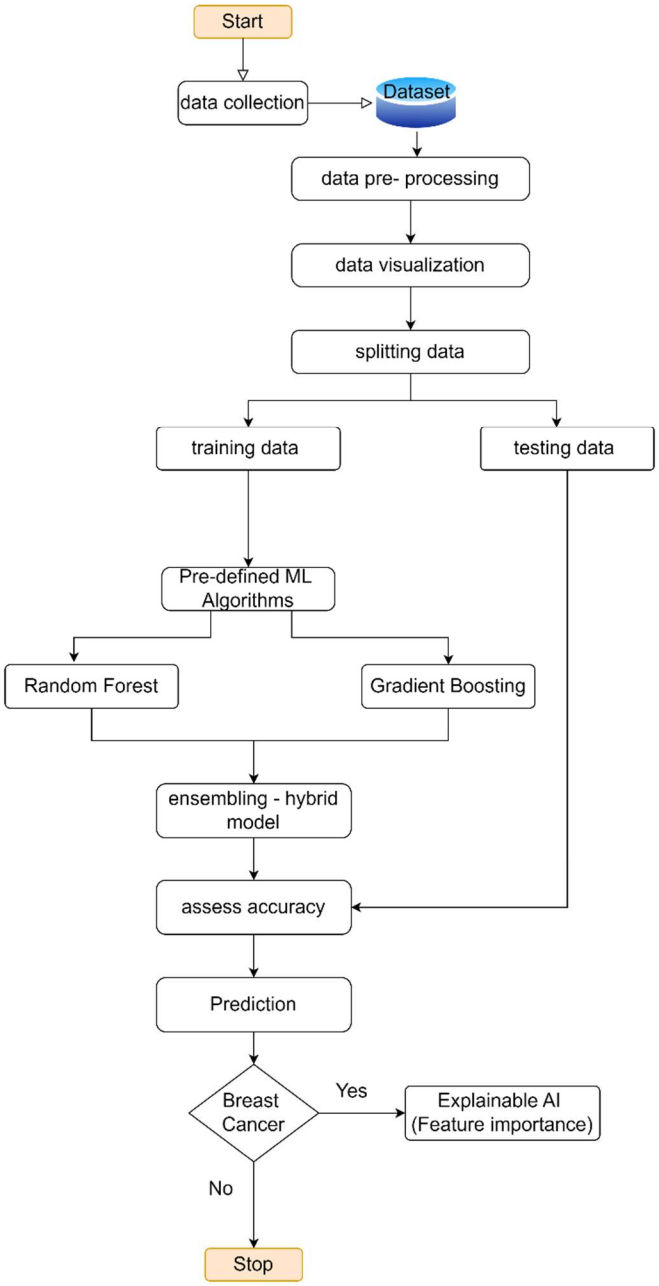

Figure 3.1 Methodology Diagram

A. Dataset collection

This model made use of the Wisconsin dataset which encompasses a diverse set of information related to tumor cells derived from patients. The dataset comprises various metrics of the cell undergone fine needle aspiration. These features not only describe the physical characteristics of the cells but also provide diagnostic information. Each measurement is assessed at different stages of the tumor, including mean, standard error, and worst-case scenarios. The data for these features were obtained by analyzing the cell cytology images taken from the fine needle aspiration technique. The dataset consists of 569

unique patient records, with 357 containing information on benign tumor cells and 212 on malignant tumor cells.

B. Data Preprocessing

In data preprocessing, it is important to handle null values in the raw dataset to avoid false positives. Attributes with null values and unwanted are removed to ensure the accuracy of the model. Additionally, character values in the dataset are encoded into integers to ensure algorithm compatibility.

C. Encoding the Target variable

Transform the target variable "diagnosis" into two distinct classes: 1 for malignant and 0 for benign, ensuring a clear representation of each class.

D. Data Visualization

Data visualization is a powerful tool that can help identify the correlation between attributes and determine highly correlated features. By extracting these highly correlated features, data visualization can enable better and more accurate decision-making for the model.

E. Removing highly correlated features

Removing highly correlated features in machine learning is common to eliminate redundancy and reduce model complexity. This practice improves model interpretability, prevents overfitting, and enhances generalization to new data. Additionally, it streamlines computational efficiency by reducing the number of dimensions and mitigates collinearity issues, particularly in linear models, promoting the development of more robust and interpretable machine learning models.

F. Feature Selection

To solve the data fitting problems, eliminate the linked features using a user-defined function in the code. The features that remain after removing the highly correlated features:'Smoothness_se', 'compactness_mean', 'symmetry_se', 'radius_mean', 'texture_se', 'fractal_dimension_mean', 'texture_mean', 'smoothness_mean', 'symmetry_mean', 'symmetry_worst'.

G. Splitting the data

The dataset is divided into testing and training datasets. The testing dataset is used to test the models, and the training dataset is used to train them. There are 75-25 (75%training,25%testing) divisions in the dataset.

H. Machine Learning Algorithms

A diverse set of machine learning algorithms, including RandomForest, DecisionTree, Gradient Boosting, and AdaBoost, was selected for this analysis. Each algorithm was individually trained on the dataset, subsequently tested, and evaluated to measure its unique accuracy.

I. Ensemble model or Hybrid model

Identifying the top two models with the highest accuracies and creating an ensemble using a Voting Classifier for improved breast cancer prediction. By leveraging the diverse strengths of these models, the ensemble approach aims to surpass individual predictive capabilities. This strategic combination enhances accuracy and provides a more robust prediction of breast cancer presence.

J. Feature Importance – ExplainableAI

Within the Explainable AI (XAI) framework, the Python package SHAP (Shapley Additive exPlanations) facilitates the interpretation of machine learning model output. Using Shapley values, which have their roots in cooperative game theory, determines how much each feature contributes to a model's prediction. SHAP improves the interpretability of complex models by offering a consistent and comprehensible measure of feature importance. It supports several machine-learning models, is model-agnostic, and provides both local and global interpretability. All things considered, SHAP is a useful addition to the Explainable AI environment, giving users access to information about the variables that affect model predictions. Following the breast cancer prediction in this process, a table showing the outcomes communicates the significance of each input feature.

•If the first value is positive and the second is negative, it indicates that an increasing value of the feature may cause the cell to fall into the malignant class, while a decreasing value may indicate the benign class.

•If the first value is negative and the second is positive, it suggests that a decreasing value of the feature may cause the cell to fall into the malignant class, and an increasing value may indicate the benign class.

IV. Result and Analysis

The selected machine learning algorithms include RandomForest, GradientBoosting, AdaBoost, and Decision Tree. These classifiers were trained on a dataset, with a partition of 75% for training and 25% for testing. Subsequently, the models were tested, and the individual accuracies in predicting breast cancer are presented in Table 4.1.

TABLE 4.1

Classifiers and their accuracies

Classifier	**Accuracy**
RandomForest	94.41%
GradientBoosting	92.31%
AdaBoost	91.61%
DecisionTree	81.82%
Proposed model-RandomForest, GradientBoosting (Ensemble)	95.10%

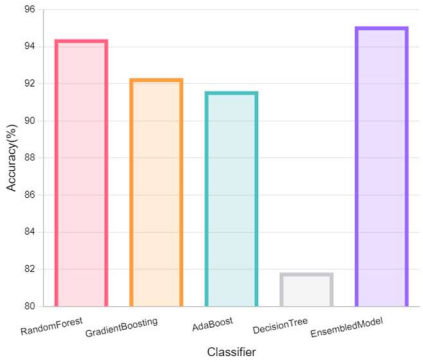

Figure – 4.1 Classifiers and their accuracies

Figure 4.1 reveals that RandomForest and GradientBoosting exhibit the highest accuracies at 94.41% and 92.31%, respectively. when the dataset is divided into 75-25 RandomForest and GradientBoosting consistently achieve the highest accuracies. Consequently, an ensemble model is crafted by combining these two classifiers, resulting in accuracies of 95.10% for the 75-25 partition of the dataset.

The table which is used to evaluate the performance of the model is called a confusion matrix. It offers an overview of the model's actual and anticipated classifications for a given batch of data. When dealing with binary classification problems—problems in which the results are classified into two classes—the matrix is especially helpful (e.g., positive and negative).

TABLE 4.2
Confusion matrix

	Predicted Positive	Predicted Negative
True Positive	TP	FN
True Negative	FP	TN

In the Table 4.2 indicates the description of confusion matrix. A True Positive (TP) is recorded when the model predicts the positive class correctly. A True Negative (TN) is recorded when the negative class is correct. False Positive (FP) describes situations where the model predicts the positive class incorrectly; these are also referred to as Type I errors. Type II mistakes, or false negatives (FN), happen when the model makes an incorrect prediction about the negative class. The ensemble model, GradientBoosting, and RandomForest models' confusion matrices are as follows

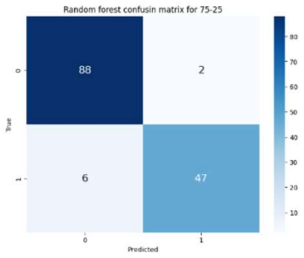

Figure – 4.2 Random Forest Confusion matrix

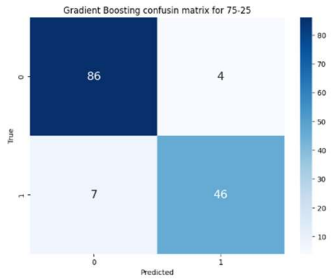

Figure – 4.3 GradientBoosting Confusion matrix

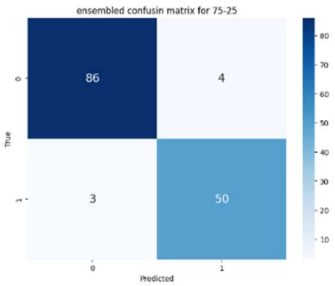

Figure – 4.4 Ensembled model Confusion matrix

Figure 4.2, 4.3, 4.4 shows the confusion matrices of RandomForest model, GradientBoosting model, and Ensembled model respectively.

Figure – 4.5 Input page

Figure 4.5 shows the Input page where the user gives features of cell nuclei as input.

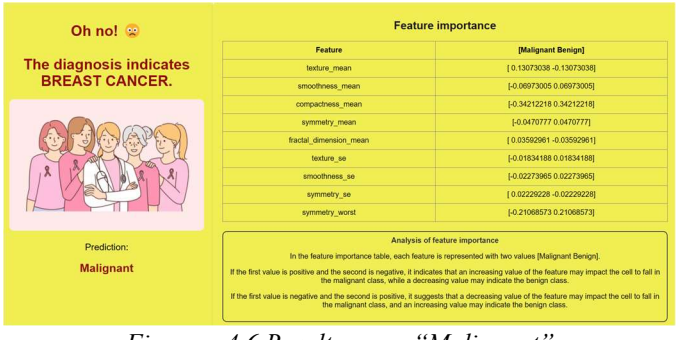

Figure – 4.6 Result page - "Malignant"

Figure 4.6 shows the Result page where the cell falls under Malignant class.

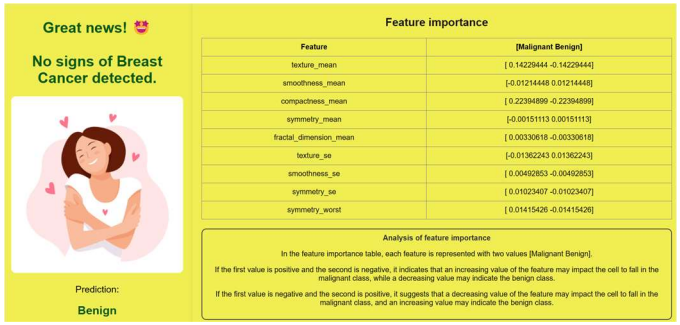

Figure – 4.7 Result page Cancer – "Benign"

Figure 4.7 shows the Result page where the cell falls under Benign class.

V. Conclusion

This paper revolves around the development of a hybrid model that seamlessly integrates various ML algorithms, to elevate the accuracy of predicting breast cancer. Exclusive to cell cytology data derived from Fine Needle Aspirations (FNA) of breast masses, our ensembled model ensures accuracy in discerning between benign and malignant cells. A key focus lies in strategically minimizing false positive diagnoses, and prioritizing accuracy, an imperative consideration in healthcare applications. Integral to our ensembled model is the incorporation of Explainable AI, particularly SHAP (Shapley Additive explanations) values. This inclusion allows for a comprehensive exploration of the model's inner workings, shedding light on the significance of each input feature. This transparency enriches the interpretability of the model, offering a lucid understanding of the factors influencing its predictions. Crucially, our hybrid model operates exclusively as a predictive tool, abstaining from providing treatment recommendations to patients.

VI. References

[1] "Statistics of Breast Cancer in India," cytecare.com. https://cytecare.com/blog/breast-cancer/statistics-of-breast-cancer/ (accessed Dec. 16,2023).

[2] American Cancer Society, "Key Statistics for Breast Cancer ," cancer.org https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html (accessed Dec. 16,2023).

[3] Reshma, V. K., et al. "Detection of breast cancer using histopathological image classification dataset with deep learning techniques." BioMed Research International 2022 (2022).

[4] Arooj, Sahar, et al. "Breast cancer detection and classification empowered with transfer learning." Frontiers in Public Health 10 (2022): 924432.

[5] Shafique, Rahman, et al. "Breast cancer prediction using fine needle aspiration features and upsampling with supervised machine learning." Cancers 15.3 (2023): 681.

[6] Botlagunta, Mahendran, et al. "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms." Scientific Reports 13.1 (2023): 485.

[7] Islam, Md Milon, et al. "Breast cancer prediction: a comparative study using machine learning techniques." SN Computer Science 1 (2020): 1-14.

[8] Liza, Fatema Tabassum, et al. "Machine Learning-Based Relative Performance Analysis for Breast Cancer Prediction." 2023 IEEE World AI IoT Congress (AIIoT). IEEE, 2023.

[9] Humayun, Mamoona, et al. "Framework for detecting breast cancer risk presence using deep learning." Electronics 12.2 (2023): 403.

[10] Jakhar, Amit Kumar, Aman Gupta, and Mrityunjay Singh. "SELF: a stacked-based ensemble learning framework for breast cancer classification." Evolutionary Intelligence (2023): 1-16.

[11] Ayoola, Joyce A., and Tokunbo Ogunfunmi. "A Comparative Analysis of Hybridized Genetic Algorithm in Predictive Models of Breast Cancer Tumors." IEEE Access (2023).

[12] Samieinasab, Mina, et al. "Meta-Health Stack: A new approach for breast cancer prediction." Healthcare Analytics 2 (2022): 100010.

[13] Ebrahim, Mohamed, Ahmed Ahmed Hesham Sedky, and Saleh Mesbah. "Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer." Data 8.2 (2023): 35.

[14] Ak, Muhammet Fatih. "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications." Healthcare. Vol. 8. No. 2. MDPI, 2020.

[15] Vaka, Anji Reddy, Badal Soni, and Sudheer Reddy. "Breast cancer detection by leveraging Machine Learning." ICT Express 6.4 (2020): 320-324.

[16] Amoroso, Nicola, et al. "A roadmap towards breast cancer therapies supported by explainable artificial intelligence." Applied Sciences 11.11 (2021): 4881.

[17] Yue, Wenbin, et al. "Machine learning with applications in breast cancer diagnosis and prognosis." Designs 2.2 (2018): 13.

[18] Alsabry, Ayman, Malek Algabri, and Amin Mohamed Ahsan. "Breast Cancer-Risk Factors and Prediction Using Machine-Learning Algorithms and Data Source: A Review of Literature." JAST 1.2 (2023).

[19] Amethiya, Yash, et al. "Comparative analysis of breast cancer detection using machine learning and biosensors." Intelligent Medicine 2.2 (2022): 69-81.

[20] Jansen, Tom, et al. "Machine learning explainability in breast cancer survival." Digital Personalized Health and Medicine. IOS Press, 2020. 307-311.