Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

An Extensive Analysis of Machine Learning Models to Predict the Breast Cancer Recurrence

D P Singh
Amity University Uttar Pradesh Greater Noida Campus

Abstract: This paper presents an in-depth exploration of various machine-learning models for predicting breast cancer recurrence. Leveraging a comprehensive dataset, we systematically evaluate the performance of multiple algorithms, considering factors such as accuracy, sensitivity, specificity, and computational efficiency.

The forecasting system was developed using fourteen varied machine learning (ML) methods to predict the likelihood of breast cancer recurrence. The prognostic model's performance was evaluated using multiple measures including the area under the curve (AUC), accuracy, sensitivity, specificity, and F1 score. Following this assessment, the most appropriate machine learning (ML) algorithm was selected, and the significance of features was identified.

Through rigorous experimentation and analysis, we identify key insights into the strengths and limitations of different approaches, offering valuable guidance for selecting the most suitable model for breast cancer recurrence prediction. The findings contribute to enhancing the accuracy and interpretability of breast cancer recurrence prediction systems, thereby facilitating informed clinical decision-making.

Keywords: Machine Learning, Breast Cancer Recurrence, Prediction Models, Algorithm Evaluation, Accuracy, Sensitivity, Specificity, Features and Predictive Models.

1. Introduction

Breast cancer recurrence remains a significant concern in oncology, prompting the exploration of advanced predictive tools such as machine learning models. With the wealth of available data, including patient demographics, tumor characteristics, and treatment history, machine learning presents a promising avenue for improving recurrence prediction accuracy.

The Centers for Disease Control and Prevention report highlights that breast cancer ranks among the most prevalent cancers in women globally[31]. Approximately 10% to 15% of women worldwide will experience this condition at some point in their lives[10].

In, urban women facing a 1 in 22 chance and rural women a 1 in 60 chance of developing breast cancer in their lifetime [15]. A 2020 study by the World Health Organization revealed over 2,000,000 new cases and more than 600,000 deaths globally due to breast cancer within a year[30].

Breast cancer led to 670,000 fatalities worldwide in 2022. About half of all cases affect women without any identifiable risk factors besides their gender and age. It was the predominant cancer among women in 157 out of 185 countries in 2022, and it's prevalent across all nations.

Detecting and forecasting breast cancer at an early stage can significantly enhance survival rates by facilitating prompt medical interventions. Moreover, being informed about breast cancer beforehand could impact a woman's choice regarding a particular medication for breast cancer prevention, which may elevate the risk of another form of cancer. Early anticipation of breast cancer or recurrence timing can incentivize high-risk women to undergo screening and adhere to screening protocols. Besides, it can also recommend chemoprevention and other suitable actions to reduce one’s risk[4,25].Machine learning encompasses an amalgamation of techniques and instruments

aimed at devising algorithms capable of aiding in forecasting, categorization, and detecting patterns. One of the advantages of using machine learning models over statistical models is the amount of flexibility in capturing high-order interactions between the data, which might result in better predictions[7].
Triple-negative breast cancer carries a significant likelihood of distant recurrence within the initial 3 to 5 years post-diagnosis[6]. Therefore, it's crucial to create predictive models for breast cancer recurrence to support diagnosis and monitoring. Breast cancer is identified through histological examination using established pathological criteria. The majority of cases are categorized as invasive ductal carcinoma (occurring in 60-75% of patients) or invasive lobular carcinoma (present in 5-15% of patients), with a smaller proportion classified as special type carcinomas[41].

In the context of breast cancer (BC), essential indicators such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) are pivotal in determining the course of treatment. Given the multifaceted nature of BC, effective management, early detection, and tailored treatment are pivotal for enhancing patient outcomes[12]. Targeting the endoplasmic reticulum signaling through hormonal medications stands as the primary systemic therapy for ER-positive or PR-positive breast cancers. Meanwhile, HER2, which is overexpressed in about 20% of breast cancer cases, correlates with a poorer prognosis if not addressed with systemic therapy[2].

Individuals with breast cancer characterized by an overexpression of HER2 experience advantages from treatments aimed at HER2, which involve anti-HER2 antibodies like trastuzumab and pertuzumab, as well as small-molecule tyrosine kinase inhibitors such as lapatinib and neratinib[48]. The primary focus of breast cancer therapy lies in diagnosing and monitoring the condition[1]. Essential aspects of this process include gathering information from patient and primary tumor characteristics, such as tumor size, nodal status, tumor grade, and the therapeutic approaches employed. These details are instrumental in constructing prognostic models like predict [8].

However, despite significant efforts towards early detecting recurring disease, research indicates that only a small proportion of recurrent cases are identified during the asymptomatic phase [13, 19]. Collaborative research or data analysis is crucial to assist physicians in forecasting breast cancer recurrence. With the rapid advancements in artificial intelligence (AI) and its integration into clinical cancer research, the accuracy of cancer prediction has soared [16, 28]. AI techniques, particularly machine learning (ML) and deep learning (DL), can sift through vast datasets to extract clinical insights, aiding in sound clinical decision-making [9, 14]. These AI methods offer non-invasive means of diagnosing diseases without posing risks to patients. ML, being an objective and replicable approach, amalgamates various quantitative factors to enhance diagnostic precision [32]. In population studies, ML proves effective in characterizing breast cancer risk, foreseeing outcomes, and identifying biomarkers without preconceived causal assumptions [33,37,38]. While most breast cancer recurrence models hinge on imaging and pathological data [29,36,42,45], incorporating ML into these models enhances their predictive capabilities.

Previous research suggests that various cancers share hormonal and epidemiological risk factors, complicating treatment[50]. For example, while tamoxifen can prevent breast cancer, it may also increase the risk of endometrial cancer, as reported by Visvanathan et al[5] and Vogel et al[3].

This paper undertakes an extensive analysis of various machine learning models applied to predict breast cancer recurrence. By synthesizing existing research and exploring the performance of different algorithms, we aim to shed light on the efficacy and limitations of these models in clinical practice. The objective was to leverage readily available clinical information to develop a decision support system for clinicians. This system aims to identify patients prone to cancer recurrence, facilitating early intervention strategies.

2. Related Work

Previous research has applied machine learning methods to forecast cancer recurrence and survival rates. For example, one study used three predictive models based on digitized images of fine needle aspirations from breast masses to accurately predict breast cancer recurrence within a year[39]. In another instance, Tahmassebi et al. leveraged machine learning with multiparametric magnetic resonance imaging to predict pathological complete

response and survival in patients receiving neoadjuvant chemotherapy[26]. Additionally, another study created a framework to evaluate the risk of breast cancer recurrence and metastasis from histopathological images using image features and machine learning technologies[36].

At present, there's a growing utilization and advancement of AI methods and statistical approaches in clinical oncology. These are employed to diagnose cancers, forecast patient prognoses, and guide treatment strategies. Specifically, the availability of extensive imaging and molecular data has spurred the adoption of machine learning (ML) and/or deep learning (DL) techniques. A recent study by Manoj Sharma et al.[35] compared traditional methods of extracting features manually with DL frameworks for categorizing colon and lung cancers.

Deep convolutional neural networks (CNNs) have shown a notable enhancement in classifier performance when utilized for feature extraction. Employing DenseNet-121 extracted features, the random forest (RF) classifier demonstrates exceptional ability in distinguishing colon and lung cancer tissue. Additionally, the authors suggest a combined method for predicting survival in hepatocellular carcinoma cases, achieving superior accuracy and sensitivity compared to existing models[43].

Yala et al.[27] introduced a deep learning (DL) model aimed at prioritizing mammograms. They established a high-sensitivity threshold to ensure that most cases predicted as negative were indeed negative. Various advanced studies have explored breast cancer prediction. Sharma and his team employed an ensemble model[44], incorporating three pre-trained convolutional neural networks (CNNs), to predict grades for the Databiox dataset, containing histopathological images of patients diagnosed with invasive ductal carcinoma breast cancer. They attained an impressive accuracy of 94% in classifying grades.

Dhahri et al.[24] introduced an approach that merges Machine Learning and Genetic Programming to distinguish between benign and malignant breast tumors. They employed electronic health records from 569 individuals extracted from the Wisconsin Breast Cancer dataset. Among seven classifiers tested, the AdaBoost classifier demonstrated the highest performance, achieving a commendable accuracy of 98.23%. This indicates its potential suitability for early breast cancer detection within controlled parameter settings.

Whitney et al. [22] utilized machine learning (ML) and deep learning (DL) techniques to examine standard H&E-stained images of early-stage ER Positive breast cancer patients. Their aim was to forecast the Oncotype DX recurrence risk. Meanwhile, Bremer et al. [23] formulated a biological marker called DCISionRT. This marker utilizes a nonlinear model to calculate a personalized decision score (DS) by combining molecular markers and clinical factors associated with the recurrence or progression of ductal carcinoma in situ (DCIS) in patients who have had breast-conserving surgery.

Some research has primarily relied on traditional statistical methods with restricted clinical data, while others have focused on machine learning techniques using image and pathology data[47]. However, there's been a scarcity of studies predicting breast cancer recurrence solely using readily available clinical information and routine lab data alongside machine learning. This study aims to fill this gap by utilizing clinical and lab data from electronic medical records and testing 14 different machine learning models to develop a predictive model for breast cancer recurrence. The goal is to offer a practical tool for clinicians and decision-makers.

3. Machine Learning Models:

Several algorithms were utilized in constructing the prediction system, categorized as:

3.1 Logistic Regression: Logistic regression, a form of supervised machine learning, is used to categorize data by estimating the probability of an observation belonging to a specific group. Based on statistical principles, this method analyzes the relationship between pairs of variables in the dataset. It's employed to predict the outcome of a categorical dependent variable, where the outcome is discrete, such as Yes or No, 0 or 1, true or false, etc[50].

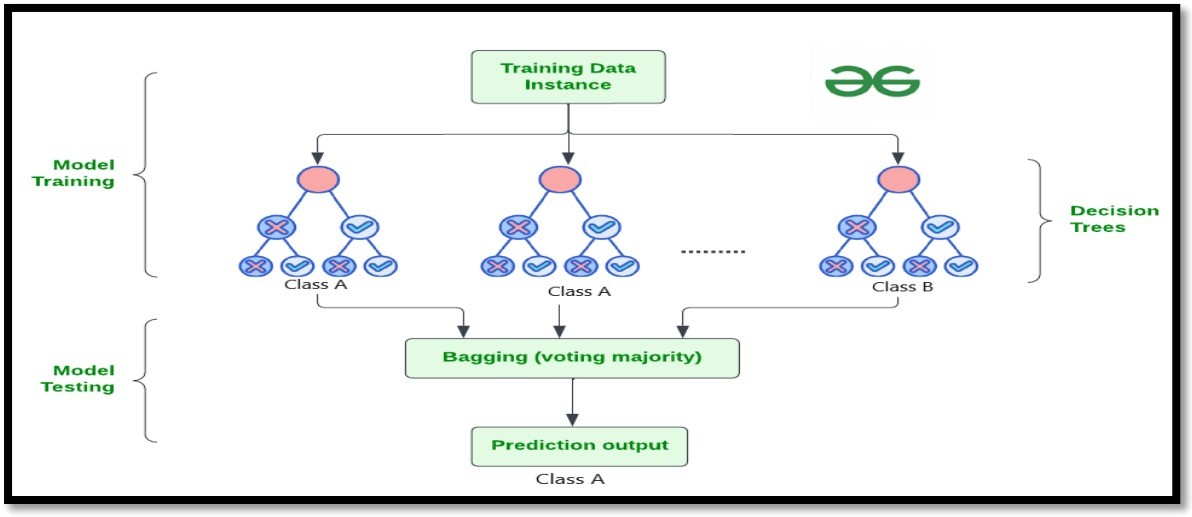
The logistic regression model uses the sigmoid function to forecast outcomes for both positive and negative categories. The sigmoid function's equation is $(z) = \frac{1}{1-e^{-z}}$, and thus, the logistic regression formula becomes: $P(X, b, w) = \frac{1}{1+e^{-wX+b}}$.

3.2 Random Forest:

The Random Forest algorithm is a robust method for tree-based learning in Machine Learning. It operates by generating multiple Decision Trees during training.

Random Forest improves upon the performance of Decision Trees by reducing variance. This is accomplished by introducing more randomness into the model through increasing the number of trees. Rather than solely focusing on the most influential feature during node splitting, the algorithm selects the optimal feature from a random subset of features. This strategy contributes to a more efficient model.

Random forests are extensively employed for both classification and regression tasks, renowned for their capacity to manage intricate data sets, mitigate overfitting, and deliver dependable predictions across various scenarios.



Each tree is built using a random sample of the dataset to evaluate a random subset of features within each division. This randomness adds diversity among the trees, lowering the chance of overfitting and enhancing predictive accuracy[20].

3.3 Support Vector Machines (SVM): The Support Vector Machine (SVM) stands out as a potent tool in machine learning, adept at tackling tasks ranging from linear or nonlinear classification and regression to outlier detection. Its adaptability spans various domains such as text and image sorting, spam filtering, handwriting and facial recognition, gene expression analysis, and anomaly detection. SVMs shine particularly in handling complex, high-dimensional data and capturing nonlinear relationships, making them versatile and effective across diverse tasks. They excel at pinpointing the best hyperplane for distinguishing between various classes in the target feature. This is accomplished by maximizing the margin, which refers to the distance between the nearest points of each class and the decision boundary. While SVM typically aims for the hyperplane with the widest margin, there are scenarios where it prioritizes accurate class prediction over margin maximization. Consequently, careful selection of hyper parameters is pivotal in SVM implementation as they significantly influence accuracy[21].

3.4 Extreme Gradient Boosting: XGBoost stands as a meticulously crafted distributed gradient boosting solution, tailored for efficient and scalable machine learning model training. It leverages ensemble learning, amalgamating predictions from diverse weak models to bolster forecasting accuracy. Abbreviated for "Extreme Gradient Boosting," XGBoost has gained immense popularity and adoption within the machine learning

community. Its strengths lie in its capacity to manage sizable datasets and deliver top-notch performance across various tasks like classification and regression[46].

XGBoost stands out for its adept management of missing data, streamlining the handling of real-world datasets by minimizing the need for extensive pre-processing. Moreover, it boasts inherent parallel processing capabilities, facilitating efficient model training even with vast datasets.

This versatile tool finds application across a spectrum of tasks like Kaggle contests, recommendation systems, and predicting click-through rates. Its adaptability shines through customization options, enabling users to fine-tune model parameters for optimal performance.

3.5 Gradient Boosted Decision Trees: Gradient Boosted Decision Trees is a method utilized in conjunction with another machine-learning algorithm. In essence, gradient boosting consists of two types of models:

A "feeble" machine-learning model, often exemplified by a decision tree.

A "robust" machine-learning model composed of numerous feeble models.

In gradient boosting, in every step, a fresh weak model is trained to forecast the "error" of the current robust model, which is called the pseudo response. We'll delve into the concept of "error" later; for now, think of it as the disparity between the prediction and a regression label. The weak model (or "error") is then incorporated into the strong model with a negative sign to diminish the error of the strong model.

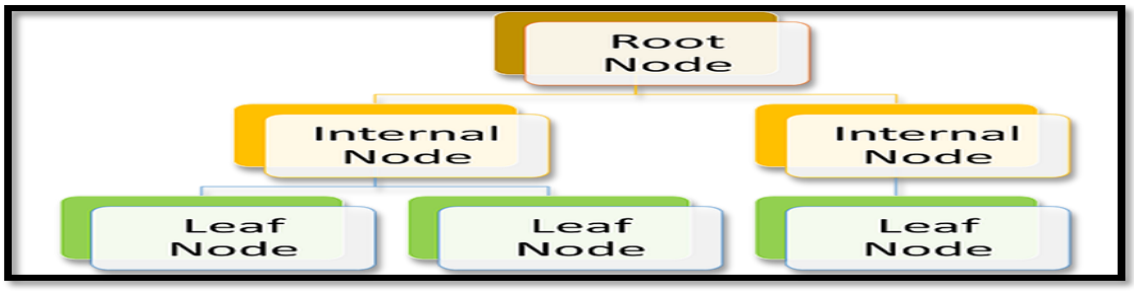
Gradient boosting operates through iteration. At each step, it combines a strong model from the previous iteration with a weak model specific to that iteration. This procedure persists until it meets a stopping criterion, such as hitting a predefined limit on iterations or identifying overfitting in a validation dataset[50].

Each iteration invokes the following formula: $F_{(i+1)} = F_i - f_i$

In this context, F_i represents the robust model at the ith step, while f_i denotes the less powerful model at the same step.

To illustrate, let's consider a basic regression scenario where the goal is to predict y based on x. We start with a simple constant strong model initialized to zero.i.e $F_0(x) = 0$.

3.6 Decision Tree: A decision tree serves as a supervised learning method suitable for both categorical and continuous input and output data. It accommodates the prediction of categorical outcomes (classification tree) as well as continuous values (regression tree). Its visual depiction simplifies comprehension for humans and aids in decision-making processes. Essentially, a decision tree resembles a flowchart characterized by if-else conditions. Beginning at the root node, questions are posed and traversal through the tree occurs via branches corresponding to specific groupings until a leaf node, where a final decision or outcome is reached.



During training, the Decision Tree algorithm selects the best attribute for dividing the data based on criteria such as entropy or Gini impurity, which assess the level of disorder or randomness in the subsets. The goal is to find the attribute that maximizes information gain or minimizes impurity most efficiently after the division[34].

3.7 Multi-layer Perceptron: In simpler terms, a multi-layer perceptron is a brain-inspired computer program that learns from data. It's built up of layers of interconnected processing units, and these units can find complex patterns thanks to special functions they use. This makes them useful for various machine-learning tasks, like sorting things into categories, predicting values, and recognizing patterns[50].

3.8 Linear Discriminant Analysis: Linear Discriminant Analysis (LDA), alternatively referred to as Normal Discriminant Analysis or Discriminant Function Analysis, is a method employed in supervised classification scenarios. Its primary function involves reducing the dimensionality of data, aiding in the differentiation between multiple groups or classes. LDA achieves this by projecting features from a higher-dimensional space into a lower-dimensional one. Within the realm of machine learning, LDA functions as a supervised algorithm tailored for classification purposes. Its objective is to pinpoint a linear combination of features that effectively separates different classes within a dataset[40].

For instance, consider a scenario where we have two categories that need to be effectively distinguished. Each category may possess various attributes. Relying solely on one attribute for classification could lead to overlap, as illustrated in the figure below. Therefore, to ensure accurate classification, we gradually incorporate more features.

3.9 AdaBoost: AdaBoost, or Adaptive Boosting, is a machine learning technique employed in both classification and regression assignments. It falls under ensemble learning approaches, where several weak learners are amalgamated to form a robust learner. Within AdaBoost, these weak learners are usually decision trees characterized by just one split, often referred to as decision stumps[11].

The algorithm works by iteratively training weak learners on the dataset, each time adjusting the weights of incorrectly classified instances. In subsequent iterations, more emphasis is placed on the misclassified instances, effectively "boosting" the performance of the model.

In the prediction phase, AdaBoost aggregates the forecasts made by all the weak learners using a weighted sum. These weights are assigned based on the accuracy of each weak learner. This ensemble approach often leads to improved generalization and robustness compared to individual weak learners[18].

3.10 Gaussian Naive Bayes: Gaussian Naive Bayes, a variant of the Naive Bayes approach, deals with continuous attributes by assuming that the data's features adhere to a Gaussian distribution across the dataset. In the Sklearn library, Gaussian Naive Bayes is categorized as an algorithm for classification, tailored specifically for continuous features following a normal distribution, and it operates based on the principles of the Naive Bayes algorithm. Understanding the basics of Gaussian Naive Bayes is crucial before diving deeper into this topic[17].

3.11 Light Gradient Boosting Machine: LightGBM, a gradient boosting framework created by Microsoft, is an open-source tool known for its speed, scalability, and accuracy. It utilizes decision trees to enhance model efficiency and minimize memory consumption. Innovative methods like Gradient-based One-Side Sampling (GOSS) are employed to prioritize instances with significant gradients during training, optimizing both memory and time. Moreover, LightGBM implements histogram-based algorithms for swift tree construction. These strategies, combined with features such as leaf-wise tree expansion and streamlined data storage formats, bolster LightGBM's efficiency, making it a standout choice among gradient boosting frameworks[50].

3.12 K-Nearest Neighbors (KNN): K-nearest neighbors (KNN) serves as a fundamental yet crucial method in machine learning, particularly within supervised learning. It holds significant importance across various domains like pattern recognition, data mining, and intrusion detection.

The KNN algorithm, known for its simplicity and flexibility, is widely employed in machine learning tasks. Unlike many other approaches, it doesn't make assumptions about data distribution, making it suitable for diverse datasets. Its flexibility extends to handling both numerical and categorical data, making it a practical choice for classification and regression tasks.

Being a non-parametric technique, KNN predicts outcomes by measuring the similarity between data points, making it robust against outliers. It functions by pinpointing the K closest neighbors to a specific data point utilizing a selected distance measure like Euclidean distance. Afterwards, the data point's category or value is established by either majority consensus or averaging among its nearest neighbors. By leveraging local data structures, this method enables the algorithm to effectively adapt to varying patterns[50].

3.13 Neural Networks: Neural networks, devoid of predefined understanding, extract unique features from data. Their elements consist of neurons, connections, weights, biases, propagation functions, and a mechanism for learning. Neurons manage inputs using thresholds and activation functions, while connections oversee the flow of information through weights and biases. The learning process involves input processing, output generation, and iterative refinement stages, improving performance across various tasks[50].

At first, the system takes in diverse attributes like the content of emails, details about the sender, and subject headings. These attributes are then enhanced with adjusted weights and processed through concealed layers. Over time, as the system undergoes training, it becomes adept at recognizing patterns that signify either spam or genuine emails.The output layer, using a binary activation function, predicts the email's spam status as either 1 (spam) or 0 (not spam). Backpropagation adjusts weights, improving spam detection accuracy over time. Neural networks emulate human brain functions through layers including input, hidden, and output layers.

3.14 Naive Bayes: Naïve Bayes methods are classification algorithms rooted in Bayes' Theorem, assuming that predictors are independent. This theorem assesses the likelihood of an event-taking place given the occurrence of another event. Bayes' Theorem is formulated as (A/B)=P(B/A)P(A)/P(B), with A and B representing events and P(A) & P(B) indicating the probabilities of A and B happening separately(A|B) indicates the likelihood of A happening given that B is true, while P(B|A) denotes the probability of B happening given that A is true[50].

4. Confusion Matrix in Machine Learning:

A confusion matrix summarizes a machine learning model's performance on a test dataset, visually displaying both accurate and inaccurate predictions. It is commonly used to evaluate classification models, which assign categorical labels to input data. This matrix is essential for assessing a classification model's performance, providing detailed counts of true positives, true negatives, false positives, and false negatives. It enables a deeper understanding of the model's recall, accuracy, precision, and overall ability to distinguish between classes by showing the frequency of predicted outcomes on the test dataset[50].

4.1 Accuracy: Accuracy measures a model's effectiveness by calculating the ratio of correctly classified instances to the total number of instances.

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, where TP= True positives, TN= True negatives, FP= False positives and FN= False negatives.

4.2 Precision: Precision refers to the accuracy of a model's positive predictions. It is measured by the ratio of true positive predictions to the total number of positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP}$$

4.3 Recall: Recall measures how well a classification model can identify all the relevant instances within a dataset. It is calculated by dividing the number of true positive (TP) cases by the total number of true positives and false negatives (FN).

$$Recall = \frac{TP}{TP + FN}$$

4.4 Specificity: Specificity, an essential metric for evaluating classification models, particularly in binary cases, measures how accurately a model identifies negative instances, also known as the True Negative Rate.

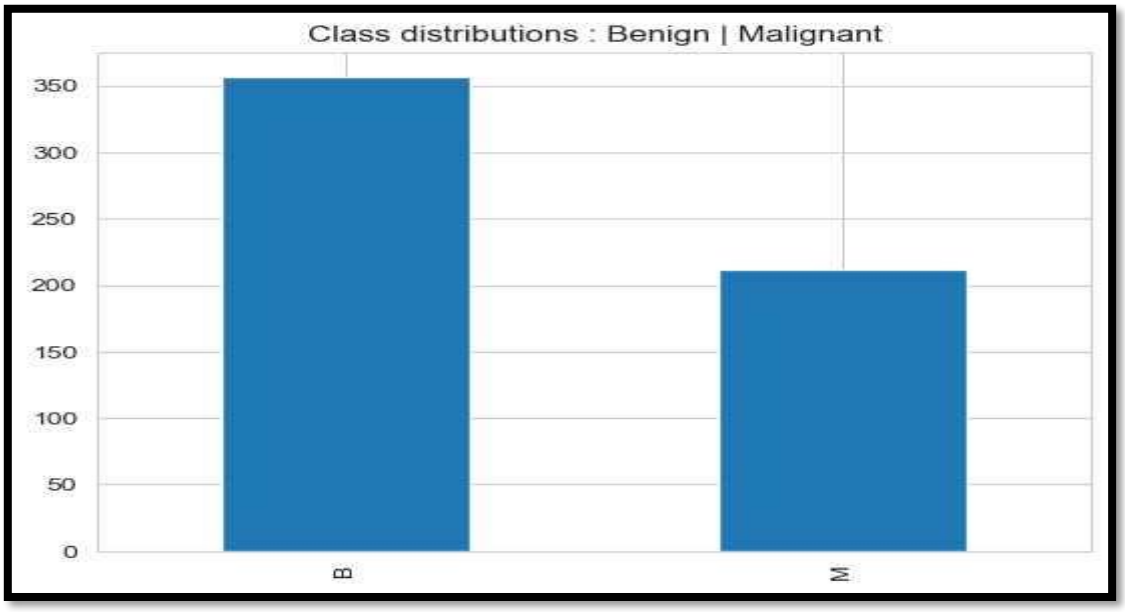
$$Specificity = \frac{TN}{TP + FP}$$

5. Data Cleaning and Feature Engineering: To test breast cancer recurrence, we collected data from(https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset), including 31 clinical features, and employed them to construct models for forecasting breast cancer recurrence. Table 1 is showing that there is no missing value in the data.

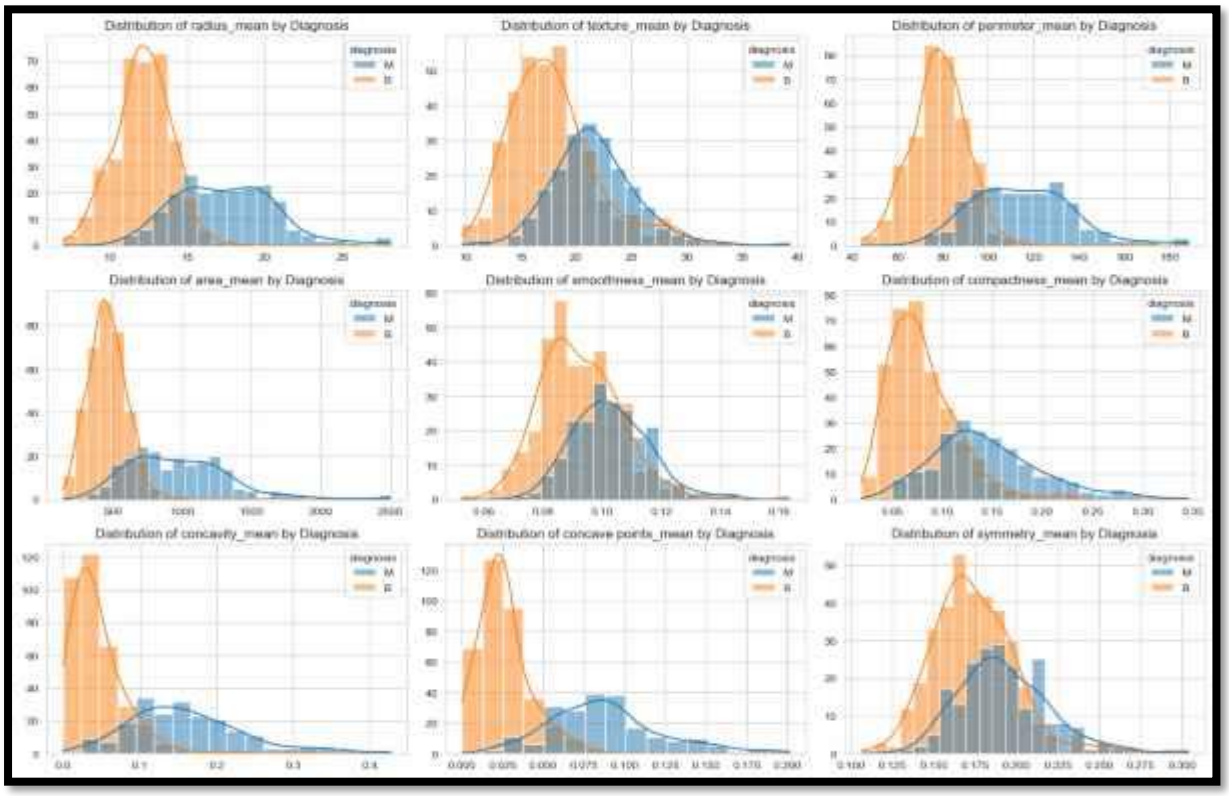
Table1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   diagnosis                569 non-null    object
 1   radius_mean              569 non-null    float64
 2   texture_mean             569 non-null    float64
 3   perimeter_mean           569 non-null    float64
 4   area_mean                569 non-null    float64
 5   smoothness_mean          569 non-null    float64
 6   compactness_mean         569 non-null    float64
 7   concavity_mean           569 non-null    float64
 8   concave points_mean      569 non-null    float64
 9   symmetry_mean            569 non-null    float64
 10  fractal_dimension_mean   569 non-null    float64
 11  radius_se                569 non-null    float64
 12  texture_se               569 non-null    float64
 13  perimeter_se             569 non-null    float64
 14  area_se                  569 non-null    float64
 15  smoothness_se            569 non-null    float64
 16  compactness_se           569 non-null    float64
 17  concavity_se             569 non-null    float64
 18  concave points_se        569 non-null    float64
 19  symmetry_se              569 non-null    float64
 20  fractal_dimension_se     569 non-null    float64
 21  radius_worst             569 non-null    float64
 22  texture_worst            569 non-null    float64
 23  perimeter_worst          569 non-null    float64
 24  area_worst               569 non-null    float64
 25  smoothness_worst         569 non-null    float64
 26  compactness_worst        569 non-null    float64
 27  concavity_worst          569 non-null    float64
 28  concave points_worst     569 non-null    float64
 29  symmetry_worst           569 non-null    float64
 30  fractal_dimension_worst  569 non-null    float64
dtypes: float64(30), object(1)
memory usage: 137.9+ KB
```

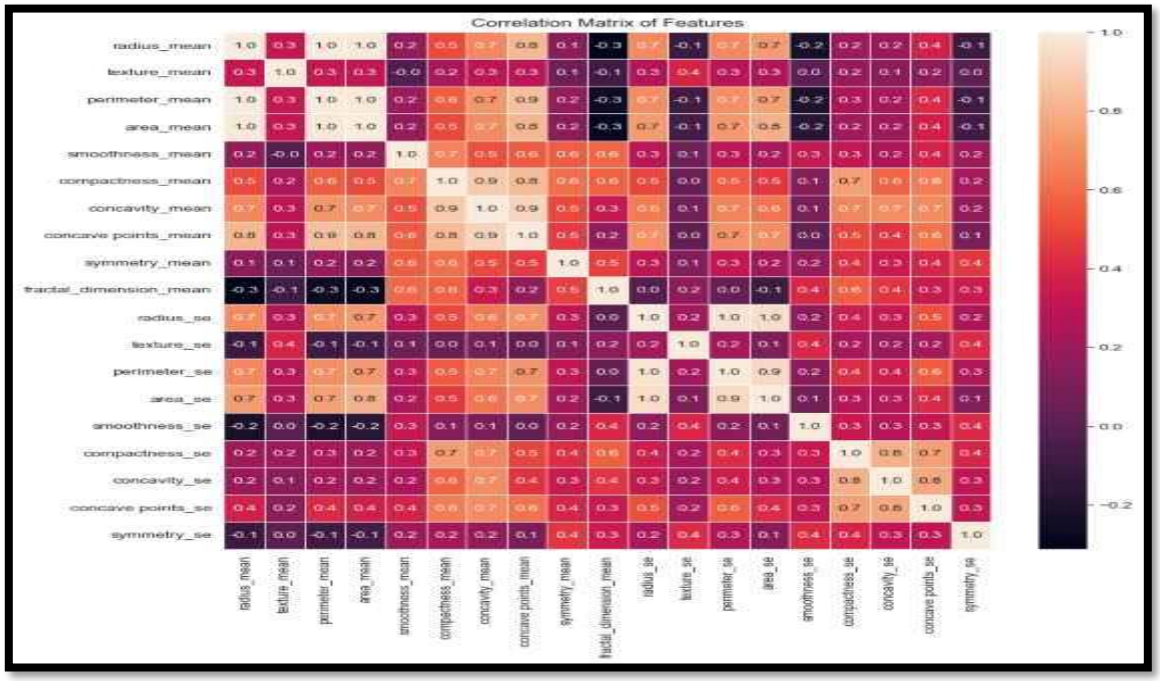
Data has distributed in class benign and malignant which is shown in below graph:



In the histogram shown below, B represents benign and M stands for malignant. These attributes are considered abnormal and unreliable because they fall outside the expected range.



The heat map has been utilized to establish the correlations among all clinical features:



6. Prediction of Breast Cancer Recurrence:

We examined the effectiveness of fourteen machine-learning algorithms. Afterwards, we evaluated how well these fourteen machine-learning models could serve as clinical decision support systems in forecasting breast cancer recurrence.

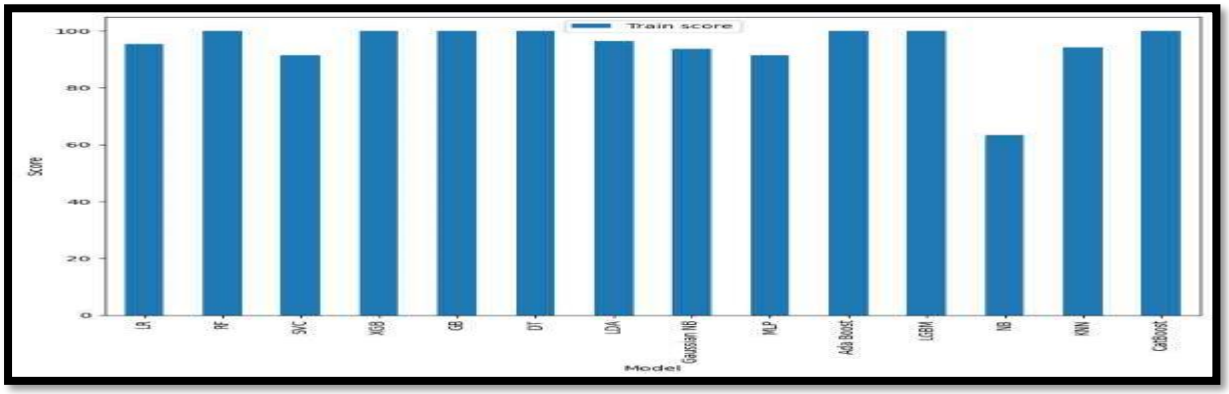
6.1. Examine the effectiveness of Machine Learning models:

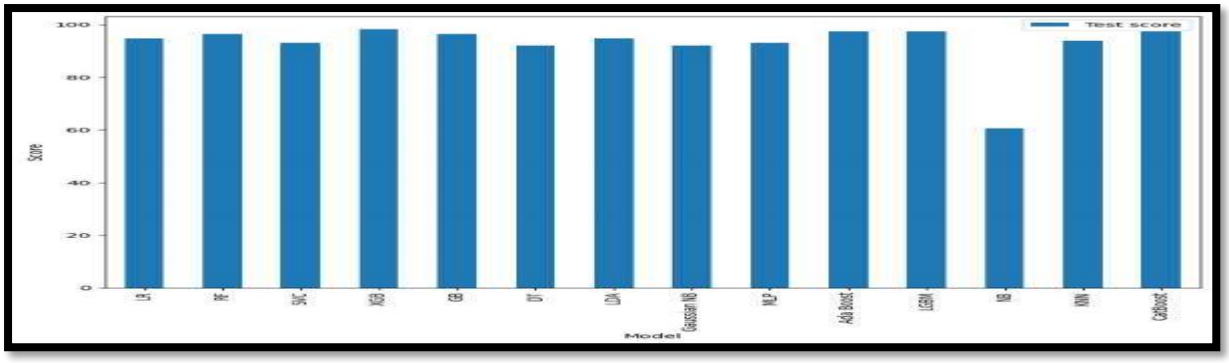
Divide the data into training, validation, and test sets to evaluate the model's performance. Standardize the data to ensure uniformity, which is essential for many ML algorithms. In the process of developing the model, we randomly selected clinical features from 80% of the patients for training purposes. We then performed a 3-fold internal cross-validation to assess the model's ability to predict outcomes using this training data. Furthermore, we validated the model's predictive accuracy externally by testing it on an independent sample size, as detailed in Table 2.

Table2

	Model	Train score	Test score
0	LR	95.384615	94.736842
1	RF	100.000000	96.491228
2	SVC	91.428571	92.982456
3	XGB	100.000000	98.245614
4	GB	100.000000	96.491228
5	DT	100.000000	92.105263
6	LDA	96.483516	94.736842
7	Gaussian NB	93.626374	92.105263
8	MLP	91.428571	92.982456
9	Ada Boost	100.000000	97.368421
10	LGBM	100.000000	97.368421
11	NB	63.296703	60.526316
12	KNN	94.285714	93.859649
13	CatBoost	100.000000	97.368421

Furthermore, we represented above data by bar graphs of train & test score to explain the performance in a better way:





6.2. Selection of Models for Predicting Breast Cancer Recurrence:

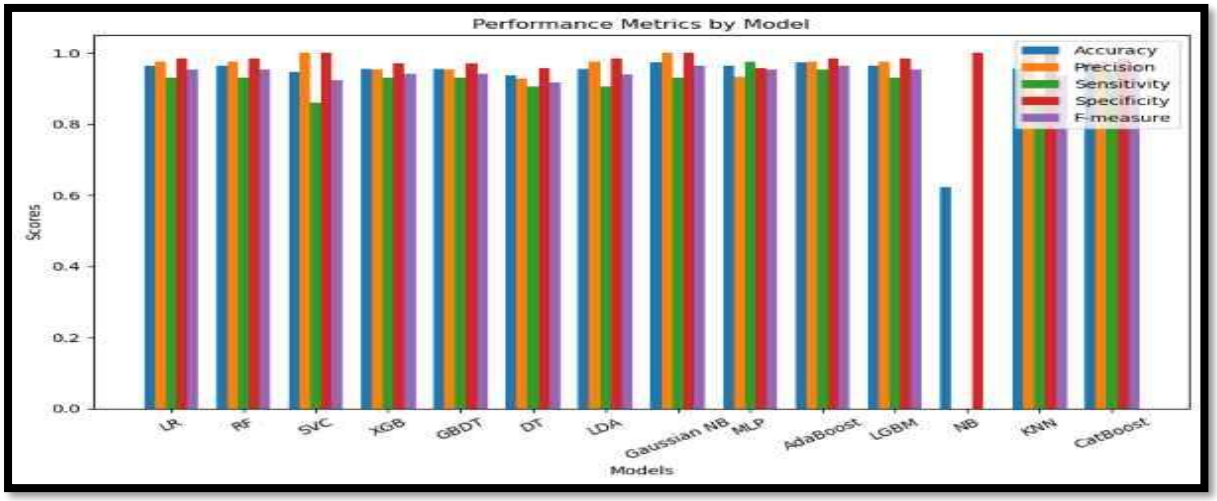
Fourteen ML models were compared in terms of their predictive capabilities, and the most accurate one was chosen. Among these models, Ada Boost and Cat Boost are best models for breast cancer recurrence. In predicting breast cancer recurrence, both models attained a accuracy of 97.36% , precision of 97.61%, sensitivity of 95.34%, specificity of 98.59% and F1 score of 96.47% .

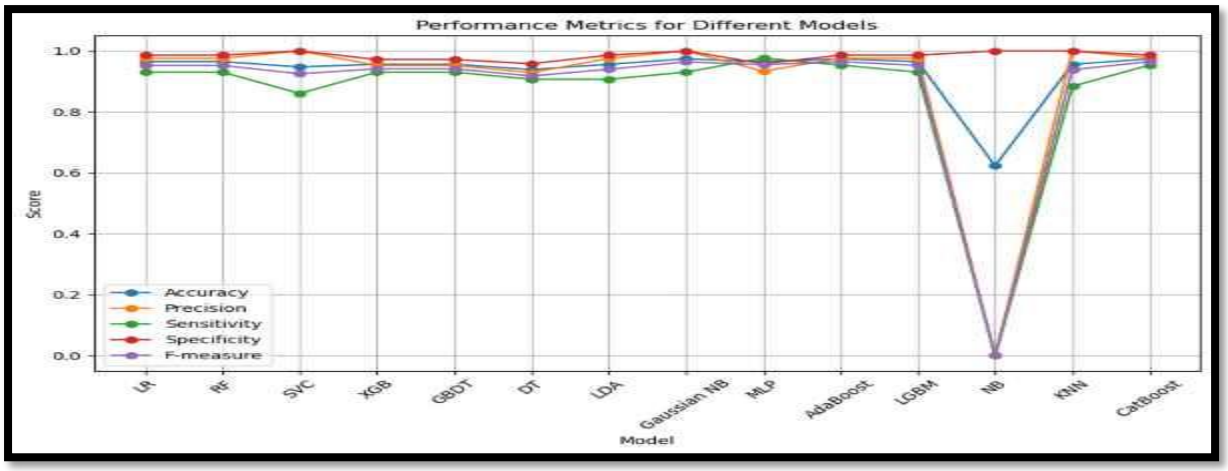
Accuracy , Precision , Sensitivity , Specificity and F1 score of all the models are shown in table3:

Table3

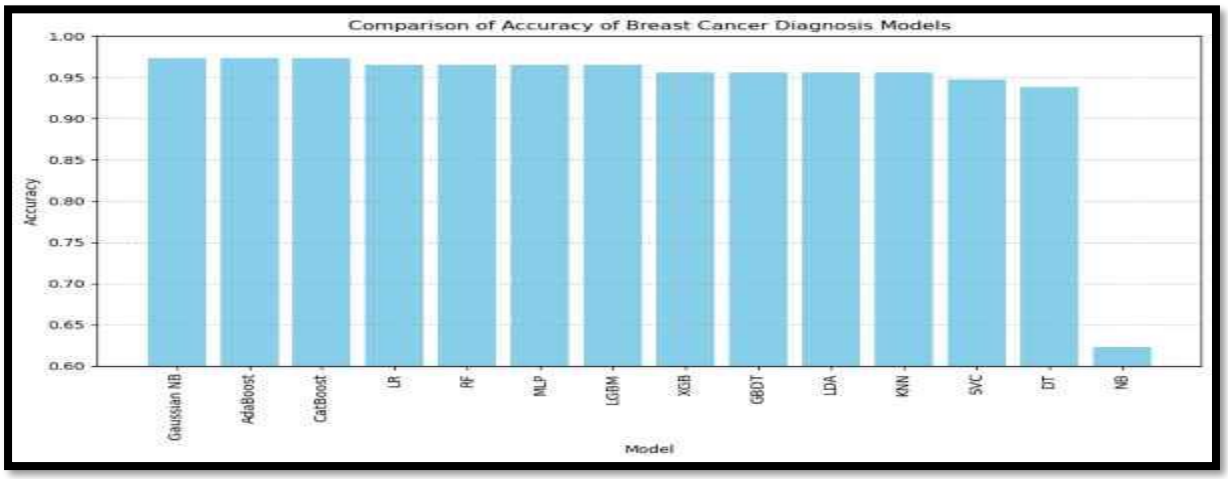
	Accuracy	Precision	Sensitivity	Specificity	F-measure
LR	0.964912	0.975610	0.930233	0.985915	0.952381
RF	0.964912	0.975610	0.930233	0.985915	0.952381
SVC	0.947368	1.000000	0.860465	1.000000	0.925000
XGB	0.956140	0.952381	0.930233	0.971831	0.941176
GBDT	0.956140	0.952381	0.930233	0.971831	0.941176
DT	0.938596	0.909091	0.930233	0.943662	0.919540
LDA	0.956140	0.975000	0.906977	0.985915	0.939759
Gaussian NB	0.973684	1.000000	0.930233	1.000000	0.963855
MLP	0.964912	0.975610	0.930233	0.985915	0.952381
AdaBoost	0.973684	0.976190	0.953488	0.985915	0.964706
LGBM	0.964912	0.975610	0.930233	0.985915	0.952381
NB	0.622807	0.000000	0.000000	1.000000	0.000000
KNN	0.956140	1.000000	0.883721	1.000000	0.938272
CatBoost	0.973684	0.976190	0.953488	0.985915	0.964706

Furthermore, we represented above data by the graphs to understand the performance of the models in a better way:





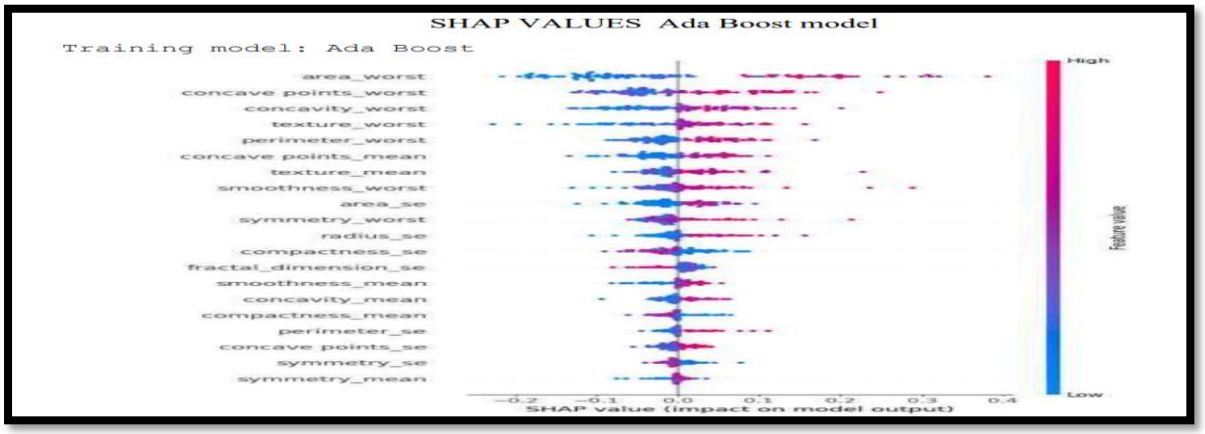
A comparison of accuracy is shown by bar graph:

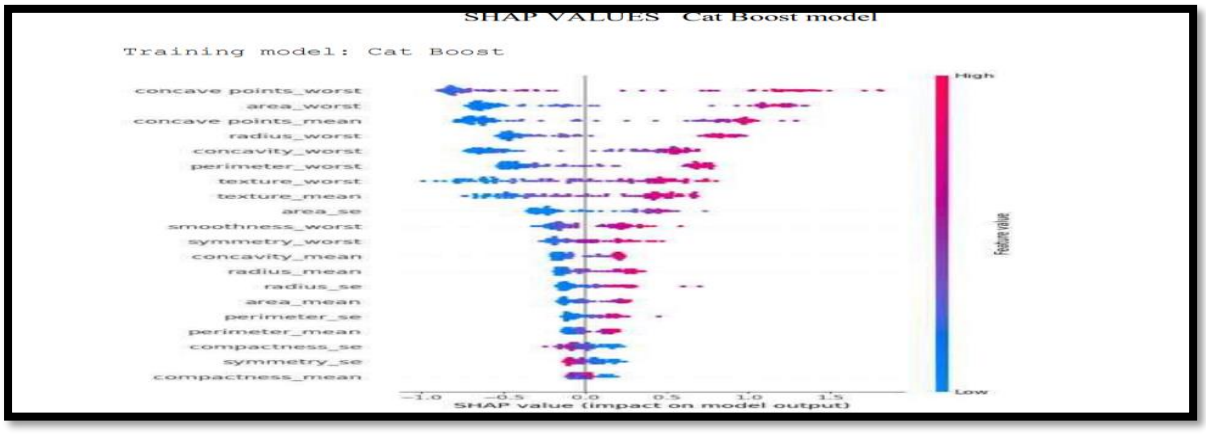


The prediction models for breast cancer recurrence can be utilized the AdaBoost and Cat Boost algorithms. This innovation introduces novel perspectives into learning algorithms design. Despite employing numerous base classifier instances, AdaBoost and Cat Boost models demonstrates rare occurrences of overfitting and effectively minimizes the exponential loss function by constructing a stepwise additive model.

6.3. SHAP Method to Improve Prediction of BC Recurrence:

To delve deeper into how the features in this model influence prediction outcomes, we assessed the SHAP values of each feature of both models. We chose the top 20 features based on their importance ranking, determined by the average absolute SHAP value derived from the AdaBoost and Cat Boost algorithm models are given below:





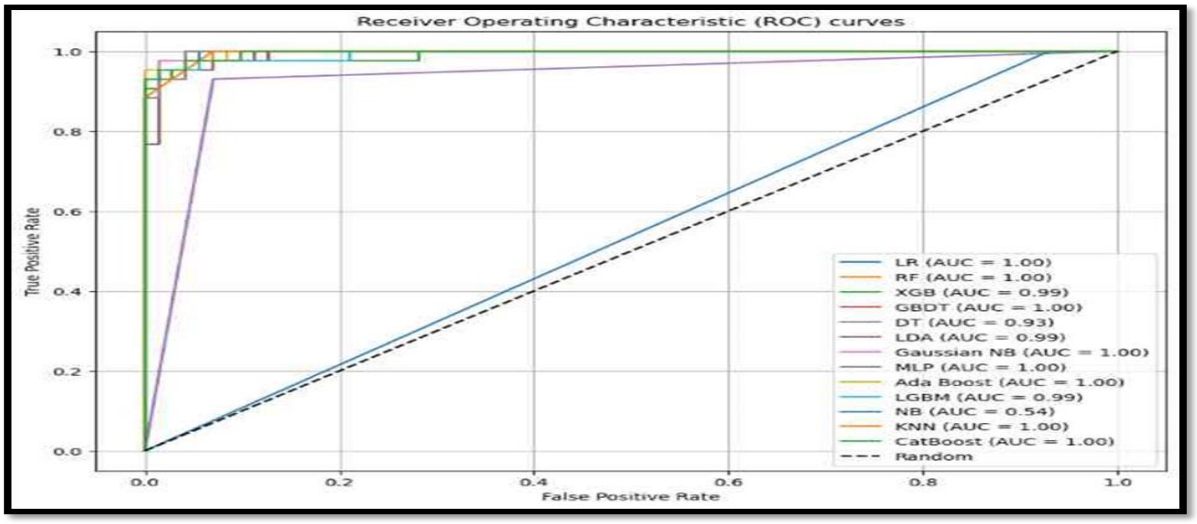
SHAP is a framework designed to interpret the outputs of machine learning models, inspired by cooperative game theory and Shapley values. Unlike other methods, SHAP offers a detailed understanding of how each feature

contributes to predictions, fostering fairness and improving clarity. By providing Shapley values, it helps in comprehending complex models and the influence of input features on their predictions.

7. Result

In this research, we evaluated 14 machine learning algorithms using 31 readily available clinical features from electronic medical records. Comparing the predictive performance of individual clinical features with machine learning-based algorithms using a combination of these features, we found that AdaBoost and Cat Boost demonstrated significantly better performance.

Our research focused on testing multiple models based on various algorithms to predict breast cancer recurrence. We discovered that AdaBoost and Cat Boost achieved a high accuracy of 97.36% , precision of 97.61%, sensitivity of 95.34%, specificity of 98.59% and F1 score of 96.47% in predicting recurrence and non-recurrence.



The AdaBoost algorithm (97.36% accuracy, ROC–AUC score of 1.00) and the CatBoost algorithm (97.36% accuracy, ROC–AUC score of 1.00) stand out as the leading models, outperforming the others.

Therefore, either the AdaBoost or CatBoost algorithm could be regarded as the top model for predicting breast cancer recurrence, based on whether your priority is accuracy or overall performance across different classes.

In the realm of predicting breast cancer recurrence, AdaBoost and CatBoost models can be proven valuable tools for physicians. Additionally, using SHAP (SHapley Additive explanations) helps interpret the predictions made by these models by examining SHAP values and feature interaction scores.

We believe our study is the initial one to combine traditional laboratory indicators with easily accessible clinical data from electronic medical records into AdaBoost and Cat Boost models for predicting breast cancer recurrence.

8. Conclusion:

The analysis of machine learning models for predicting breast cancer recurrence demonstrates both advancements and challenges in leveraging computational methods for clinical decision-making. While certain algorithms exhibit high accuracy and robustness, others may struggle with generalization or interpretability issues. The selection of an appropriate model must consider not only predictive performance but also practical considerations such as computational efficiency and ease of implementation in real-world settings.

In this study, models using AdaBoost and Cat Boost are developed and assessed to support decision-making, aiming to predict the likelihood of breast cancer recurrence.

Moving forward, further research is warranted to address these challenges and enhance the utility of machine learning in guiding personalized treatment strategies for breast cancer patients, ultimately improving outcomes and quality of life.

References:

[1] Saphner T, Tormey DC, Gray R. Annual hazard rates of recurrence for Breast cancer after primary therapy. J Clin Oncol. 1996;14(10):2738–46.
[2] Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive Breast cancer. N Engl J Med. 2005;353(16):1659–72.
[3] Vogel VG, Joseph Costantino MP, Lawrence Wickerham DD, et al. Effects of Tamoxifen vs Raloxifene on the Risk of Developing Invasive Breast Cancer and Other Disease Outcomes The NSABP Study of Tamoxifen and Raloxifene (STAR) P2 Trial. 2006 www.jama.com.
[4] DGR Evans, A. Howell, Breast cancer risk-assessment models, Breast Cancer Res 9 (2007) 213.
[5] K Visvanathan, RT Chlebowski, P Hurley, et al., American Society of Clinical Oncology Clinical Practice Guideline Update on the Use of Pharmacologic Interventions Including Tamoxifen, Raloxifene, and Aromatase Inhibition for Breast Cancer Risk Reduction, J Clin Oncol 27 (2009) 3235–3258.
[6] Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative Breast cancer. N Engl J Med. 2010;363(20):1938–48.
[7] JM Jerez, I Molina, PJ García-Laencina, et al., Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artif Intell Med 50 (2010) 105–115.
[8] Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, et al. PREDICT: a new UK prognostic model that predicts survival following Surgery for invasive Breast cancer. Breast Cancer Res. 2010;12(1):R1.
[9] Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309(13):1351–2.
[10] A LG, E AT, Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence, J Heal Med Informatics 04 (2013) 1–4.
[11] Cao Y, Miao Q-G, Liu J-C, Gao L. Advance and prospects of AdaBoost Algorithm. Acta Automatica Sinica. 2013;39(6):745–58.
[12] Rautenberg T, Siebert U, Arnold D, Bennouna J, Kubicka S, Walzer S, et al. Economic outcomes of sequences which include monoclonal antibodies against vascular endothelial growth factor and/or epidermal growth factor receptor for the treatment of unresectable metastatic Colorectal cancer. J Med Econ. 2014;17(2):99–110.
[13] Pourzand A, Tajaddini A, Pirouzpanah S, Asghari-Jafarabadi M, Samadi N, Ostadrahimi AR, et al. Associations between Dietary Allium vegetables and risk of Breast Cancer: a hospital-based Matched Case-Control Study. J Breast Cancer. 2016;19(3):292–300.
[14] Kolker E, Özdemir V, Kolker E. How Healthcare can refocus on its supercustomers (Patients, n = 1) and customers (doctors and nurses) by leveraging lessons from Amazon, Uber, and Watson. Omics. 2016;20(6):329–33.
[15] Chaurasia V, Pal S. A Novel Approach for Breast Cancer Detection Using Data Mining Techniques. undefined2017.
[16] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43.
[17] Ontivero-Ortega M, Lage-Castellanos A, Valente G, Goebel R, Valdes-Sosa M. Fast Gaussian Naïve Bayes for searchlight classification analysis. NeuroImage. 2017;163:471–9.
[18] Baig MM, Awais MM, El-Alfy E-SM. AdaBoost-based artificial neural network learning. Neurocomputing. 2017;248:120–6. Publisher's Note Springer Nature remains neutral with regard to jurisdictional cla

[19] Zapater-Moros A, Gámez-Pozo A, Prado-Vázquez G, Trilla-Fuertes L, Arevalillo JM, Díaz-Almirón M, et al. Probabilistic graphical models relate immune status with response to neoadjuvant chemotherapy in Breast cancer. Oncotarget. 2018;9(45):27586–94.
[20] Chen Z, He N, Huang Y, Qin WT, Liu X, Li L. Integration of a deep learning classifier with a Random Forest Approach for Predicting Malonylation sites. Genomics Proteom Bioinf. 2018;16(6):451–9.
[21] Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 2018;15(1):41–51.
[22] Whitney J, Corredor G, Janowczyk A, Ganesan S, Doyle S, Tomaszewski J, et al. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER + Breast cancer. BMC Cancer. 2018;18(1):610.
[23] Bremer T, Whitworth PW, Patel R, Savala J, Barry T, Lyle S, et al. A Biological signature for breast ductal carcinoma in situ to predict Radiotherapy Benefit and assess recurrence risk. Clin Cancer Res. 2018;24(23):5895–901.
[24] Dhahri H, Al Maghayreh E, Mahmood A, Elkilani W, Faisal Nagi M. Automated Breast Cancer diagnosis based on machine learning algorithms. J Healthc Eng. 2019;2019:4253641.
[25] GF Stark, GR Hart, BJ Nartowt, J. Deng, Predicting breast cancer risk using personal health data and machine learning models, PLoS One 14 (2019), e0226765.
[26] Tahmassebi A, Wengert GJ, Helbich TH, Bago-Horvath Z, Alaei S, Bartsch R, et al. Impact of machine learning with Multipara metric magnetic resonance imaging of the breast for early prediction of response to Neoadjuvant Chemotherapy and survival outcomes in Breast Cancer patients. Invest Radiol. 2019;54(2):110–7.
[27] Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to Triage Screening mammograms: a Simulation Study. Radiology. 2019;293(1):38–46.
[28] Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. Cancer Lett. 2020;471:61–71.
[29] Yu Y, Tan Y, Xie C, Hu Q, Ouyang J, Chen Y, et al. Development and validation of a Preoperative Magnetic Resonance Imaging Radiomics-Based Signature to Predict Axillary Lymph Node Metastasis and Disease-Free Survival in patients with early-stage Breast Cancer. JAMA Netw Open. 2020;3(12):e2028086.
[30] Age standardized (World) incidence rates, breast, all ages. 2020 https://gco.iarc.fr /today/data/factsheets/cancers/20-Breast-fact-sheet.pdf (accessed April 6, 2021).
[31] Breast Cancer Statistics | CDC. https://www.cdc.gov/cancer/breast/statistics/inde x.htm (accessed April 5, 2021).
[32] Daimiel Naranjo I, Gibbs P, Reiner JS, Lo Gullo R, Sooknanan C, Thakur SB et al. Radiomics and Machine learning with multiparametric breast MRI for Improved Diagnostic accuracy in Breast Cancer diagnosis. Diagnostics (Basel). 2021;11(6).
[33] Chen Z, Wang M, De Wilde RL, Feng R, Su M, Torres-de la Roche LA, et al. A machine learning model to predict the Triple negative Breast Cancer Immune Subtype. Front Immunol. 2021;12:749459.
[34] Wang L, Zhu L, Jiang J, Wang L, Ni W. Decision tree analysis for evaluating Disease activity in patients with rheumatoid arthritis. J Int Med Res. 2021;49(10):3000605211053232.
[35] Kumar N, Sharma M, Singh VP, Madan C, Mehandia S. An empirical study of handcrafted and dense feature extraction techniques for lung and colon Cancer classification from histopathological images. Biomed Signal Process Control. 2022;75:103596.
[36] Liu X, Yuan P, Li R, Zhang D, An J, Ju J, et al. Predicting Breast cancer recurrence and Metastasis risk by integrating color and texture features of histopathological images and machine learning technologies. Comput Biol Med. 2022;146:105569.
[37] Ma M, Liu R, Wen C, Xu W, Xu Z, Wang S, et al. Predicting the molecular subtype of Breast cancer and identifying interpretable imaging features using machine learning algorithms. Eur Radiol. 2022;32(3):1652–62.
[38] Rasool A, Bunterngchit C, Tiejian L, Islam MR, Qu Q, Jiang Q. Improved Machine Learning-based predictive models for Breast Cancer diagnosis. Int J Environ Res Public Health. 2022;19(6).

[39] Gupta SR. Prediction time of Breast cancer Tumor recurrence using machine learning. Cancer Treat Res Commun. 2022;32:100602.
[40] Xu L, Raitoharju J, Iosifidis A, Gabbouj M. Saliency-based Multilabel Linear Discriminant Analysis. IEEE Trans Cybern. 2022;52(10):10200–13.
[41] Rakha EA, Tse GM, Quinn CM. An update on the pathological classification of Breast cancer. Histopathology. 2023;82(1):5–16.
[42] Romeo V, Cuocolo R, Sanduzzi L, Carpentiero V, Caruso M, Lama B et al. MRI Radiomics and Machine Learning for the prediction of Oncotype Dx Recurrence score in invasive Breast Cancer. Cancers (Basel). 2023;15(6).
[43] Sharma M, Kumar N. Improved hepatocellular carcinoma fatality prognosis using ensemble-learning approach. J Ambient Intell Humaniz Comput. 2022;13(12):5763–77.
[44] Kumaraswamy E, Kumar S, Sharma M. An invasive ductal carcinomas Breast Cancer Grade classification using an ensemble of convolutional neural networks. Diagnostics [Internet]. 2023; 13(11).
[45] Lee J, Yoo SK, Kim K, Lee BM, Park VY, Kim JS, et al. Machine learning–based radiomics models for prediction of locoregional recurrence in patients with Breast cancer. Oncol Lett. 2023;26(4):422
[46] Guan X, Du Y, Ma R, Teng N, Ou S, Zhao H, et al. Construction of the XGBoost model for early Lung cancer prediction based on metabolic indices. BMC Med Inform Decis Mak. 2023;23(1):107.
[47] D P Singh, J S Jassi, Sunaina, (2023), Exploring the Significance of Statistics in the Research: A Comprehensive Overview, Eur. Chem. Bull., 12(Special Issue 2),2089-2102
[48] Waks AG, Winer EP. Breast Cancer Treatment: a review. JAMA. 2019;321(3):288–300. Zuo et al. BMC Medical Informatics and Decision Making (2023) 23:276 Page 13 of 14
[49] Pfeiffer RM, Park Y, Kreimer AR, et al. Risk Prediction for Breast, Endometrial, and Ovarian Cancer in White Women Aged 50 y or Older: Derivation and Validation from Population-Based Cohort Studies. DOI:10.1371/journal.pmed.1001492.
[50] https://www.geeksforgeeks.org/machine-learning/Dr D P Singh drdps97@gmail.com ID: https://orcid.org/0000-0001-9494-4296