Prediction of Breast Cancer using Traditional and Ensemble Technique: A Machine Learning Approach

Tamanna Islam[1], Amatul Bushra Akhi[2], Farzana Akter[3], Md. Najmul Hasan[4], Munira Akter Lata[5]
Dept. of Computer Science Engineering, Daffodil International University, Dhaka, Bangladesh[1, 2, 4]
Dept. of IoT and Robotics Engineering, Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh[3]
Dept. of Educational Technology, Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh[5]

Abstract*—Breast cancer is a prevalent and potentially life-threatening disease that affects millions of individuals worldwide. Early detection plays a crucial role in improving patient outcomes and increasing the chances of survival. In recent years, machine learning (ML) techniques have gained significant attention in the field of breast cancer detection and diagnosis due to their ability to analyze large and complex datasets, extract meaningful patterns, and facilitate accurate classification. This research focuses on leveraging ML algorithms and models to enhance breast cancer detection and provide more reliable diagnostic results in the real world. Two datasets from Kaggle have been used in this study and Decision tree (DT), Random Forest (RF), Logistic Regression (LR), K-Nearest Classifier (KNN) etc. are applied to identify potential breast cancer cases. On the first dataset, A, the test's accuracy using Logistic Regression, SVM, and Grid SearchCV was 95.614%, however in dataset B, the accuracy of Logistic Regression and Decision Tree increased to 99.270%. The accuracy of Boosting Decision Tree was 99.270% when compared to other algorithms. To defend the performances, various ensemble models are used. To assign the optimal parameters to each classifier, a hyper-parameter tweaking method is used. The experimental study examined the findings of recent studies and discovered that LRBO performed best, with the highest level of accuracy for predicting breast cancer being 95.614%.

Keywords—Breast cancer; prediction; machine learning algorithms; ensemble models; voting; stacking

I. Introduction

Multiple tissues are being harmed or developing out of control, which is known as cancer, since sickness is the worst aspect of our daily lives. Breast cancer is a type of cancer that develops when unregulated tissue or damaged tissue does so [1]. This patient's prevalence is significantly rising. However, finding or recognizing the injured region at the time of diagnosis is the key issue. Machine learning may be the most effective component of a crucial factor in predicting the presence of breast cancer from responsive health datasets by examining various variables and patient diagnosis records. We looked at the patient's diagnosis papers for our work and discovered certain key factors to pinpoint the condition. The dataset dealt with the size and structure of a woman's bodily tissues as well as determining whether or not she had breast cancer [7]. In order to employ machine learning algorithms to recognize the cancer tissue in the body, several different researchers have worked together. However, their method and accuracy were not appropriate nor smooth for predicting breast cancer [12]. We suggest our method to increase the accuracy rate of breast cancer prediction in a woman's body. There are two different kinds of machine learning techniques. One of them is under supervision, while the other is not. Working with labeled data, supervised learning creates outputs from inputs based on examples of input-output pairings. The dataset's training data is used as the working data. Unsupervised learning works with the unlabeled data and creates the model to work with its patterns and information which was not detected previously. Unsupervised learning uses unlabeled data to build models that can make use of previously undetected patterns and information.

II. Background Literature

Breast cancer is the most common and rapidly developing illness in the world. Breast cancer is more commonly detected in women. Breast cancer can be controlled if it is detected early. A hybrid approach-based methodology that uses machine learning has been presented. This method was put into practice utilizing MRMR feature selection using four classifiers to determine the optimal outcomes. The four classifiers SVM, Naive Bayes, Function tree, and End Meta were utilized by the author, and they were all compared. SVM was discovered to be an effective classifier. to ascertain better outcomes [1]. To achieve the most accurate result machine learning is the most reliable technique. We have used a few machine learning classifiers to categorize breast cancer, and they are appropriate for the job we are proposing. To execute decision models, machine learning algorithms that are based on decision tree models are known as "tree structures" [2]. similarly proposed a comparison between Random Forest, Naïve Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (K-NN) and they found the SVM is the best classifier with an accuracy of 97.9% compared with K-NN, RF, and NB, they are based on Multilayer perceptron with 5 layers and 10 times cross-validation using MLP. The author F. M. Javed Mehedi Shamrat et al. [3] focused on the enhancement of the accuracy value using the Wisconsin Breast Cancer Diagnostic Dataset (WBCD) by applying an ML-based system for the early prediction of breast cancer disease. Six supervised classification techniques are used which are: SVM, NB, KNN, RF, DT, and LR. According to the analysis of breast cancer prediction performance, SVM had the highest performance and the highest classification accuracy (97.07%). While NB and RF have attained the second-highest prediction accuracy. Mumine Kaya Keles [4] predicted and detected breast cancer early where Random Forest outperformed all the other algorithms giving an average accuracy of 92.2 percent. K.Anastraj et al. [5] depicts that the support vector machine had given better results (94%). In the experimental results [6], BayesNet was the best classification method with an accuracy rate of 97.13%. Ch. Shravya et al. [7] provided relative study on the implementation of models using Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbor (KNN) on the dataset taken from the UCI Repository. With respect to the results of accuracy, precision, sensitivity, specificity and False Positive Rate the efficiency of each algorithm is measured and compared and focused in the advancement of predictive models to achieve good accuracy in predicting valid disease outcomes using supervised machine learning methods. The results analysis shows that the combination of multidimensional data with various feature selection, classification, and dimensionality reduction techniques can offer advantageous tools for inference in this field. This study has shown that SVM is the best accuracy of 92.7%. The authors Ertel Merouane et al. in [8] provide a cloud-based Extreme Learning Machine (ELM) architecture for the classification of breast cancer. Cloud computing increases categorization accuracy and provides access around the clock. When compared to standalone systems, the ELM model executed faster and with higher accuracy when it was put on the Amazon EC2 cloud platform. Future additions may improve the framework's functionality in image-processing applications like medical imaging and character recognition [9]. Probability is constantly between 0 (never happens) and 1 (occurs). In the case of binary classification, using our COVID-19 example, the likelihood of testing positive and not testing positive will total up to the logistic function or sigmoid function to compute probability in logistic regression. The logistic function is a straightforward S-shaped curve that converts input into a value between 0 and 1 [17].

III. Research Methodology

The Dataset sourced from Kaggle [9] [10].The size of the dataset A is 32x569 and B is 10x683. The frequency of breast cancer is categorized in the diagnostic and Class column. Malignant (M) and Beginning (B) conditions are used to categorize patients. There, 0 represents "B" and 1 represents "M." 212 individuals were at the malignant stage, leaving 357 patients in the initial stage in dataset A. Another dataset B had 239 patients in the malignant stage and 444 individuals in the initial stage. Fig. 1 for dataset A and Fig. 2 for dataset B below display the ratio. The dataset was split into a testing and training set. We've chosen 20% for the exam portion and another 80% for the learning portion. The dataset contains nominal values and there were no missing or incorrect values. A comprehensive explanation of the dataset with its range is displayed in Tables I and II.

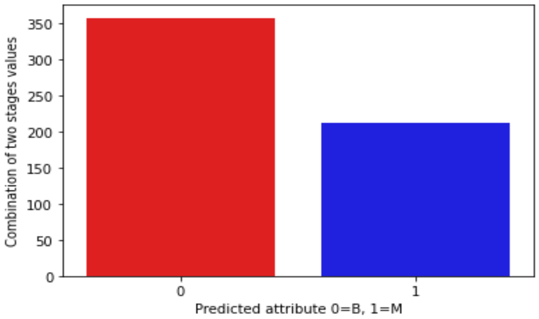


Fig. 1. Number of target values dataset A.

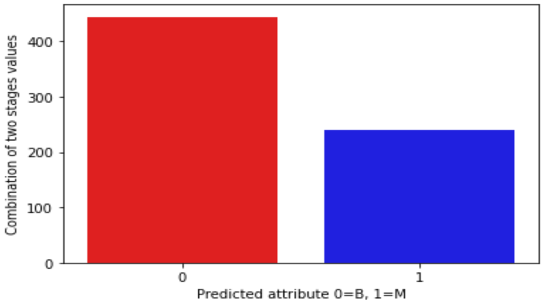


Fig. 2. Number of target values dataset B.

TABLE I. Details of the Dataset A

Attributes	**Description**	**Value Range**	**Types of values**
Diagnosis	Malignant or Begin	0 and 1	Integer
Radius_mean	Radius of Lobes	6.98 to 28.1	Float
Texture_mean	Mean of Surface Texture	9.71 to 39.28	Float
Perimeter_mean	Outer Perimeter of Lobes	43.8 to 188.5	Float
Area_mean	Mean Area of Lobes	143.5 to 2501	Float
Smoothness_ mean	Mean of Smoothness Levels	0.05 to 0.163	Float
Compactness_ mean	Mean of Compactness	0.02 to 0.345	Float
Concavity_ mean	Mean of Concavity	0 to 0.426	Float
Concave points_mean	Mean of Concave Points	0 to 0.201	Float
Symmetry_ mean	Mean of Symmetry	0.11 to 0.304	Float
Fractal_ dimension_ mean	Mean of Fractal Dimension	0.05 to 0.1	Float
Radius_se	SE of Radius	0.11 to 2.87	Float
Texture_mean	SE of Texture	0.36 to 4.88	Float
Perimeter_se	Perimeter of SE	0.76 to 22	Float
Area_se	Area of SE	6.8 to 542	Float
Smoothness_se	SE of Smoothness	0 to 0.03	Float
Compactness_se	SE of Compactness	0 to 0.14	Float
Concavity_se	SE of Concavity	0 to 0.4	Float

TABLE II. DETAILS OF THE DATASET B

Attributes	Description	Value Range	Types of values
Clump Thickness	Thickness of Clump	1 to 10	Integer
Uniformity of Cell Size	Cell size	1 to 10	Integer
Uniformity of Cell Shape	Cell shape	1 to 10	Integer
Marginal Adhesion	Adhesion Marginal value	1 to 10	Integer
Single Epithelial Cell Size	Cell size	1 to 10	Integer
Bare Nuclei	Number of Nuclei	1 to 10	Integer
Bland Chromatin	Number of Bland Chromatin	1 to 10	Integer
Normal Nucleoli	Number of Normal Nucleoli	1 to 10	Integer
Mitoses	Number of Mitoses	1 to 10	Integer
Class	Malignant or Begin	0 and 1	Integer

A. *Statistical Analysis*

Statistical analysis is one of the most crucial segments of any research. In this work, we employed four distinct kinds of algorithms in this work, including Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Classifier (KNN), Adaboost Classifier (ABC), Decision Tree (DT), GridSearch CV (GS), XGBoost Classifier (XGB), Gaussian Naïve Bayes (GNB), and Support Vector Classifier (SVC). Logistic Regression, Support Vector Classifier and Grid SearchCV was 95.614% accuracy for dataset A. Logistic Regression and Decision Tree was 99.270% accuracy for dataset B. Following the use of bagging, boosting, stacking and voting algorithms. Hyperparameter tweaking and 10-fold cross-validation have both been employed.

B. *Proposed Methodology*

The dataset for the system's training and testing was initially introduced. Next, we used data preparation techniques such as the Standard Scaler Transform. Categorical data conversion to numeric data. We utilized 80% for the training portion and 20% for the testing portion. After that we implemented algorithms and assessed the outcomes. Then, to get the highest forecast accuracy, we employed ensemble methods. Bagging, Boosting, Stacking and Voting are ensemble algorithms. The outcomes of the ensemble algorithms that were used were then assessed. Then we used Hyper Parameter Tuning to verify the outcome. Then, using outcome analysis, we assessed the models that had been put into practice. Fig. 3 displays the recommended model technique.

C. *Implementation Requirements*

A number of filtering techniques is used to clean the dataset. Then, data preprocessing techniques like Standard Scaler Transform were used. We utilized 80% for the training portion and 20% for the testing portion. After that, we implemented algorithms and assessed the outcomes. Then, in order to get the highest forecast accuracy, we employed ensemble methods. Bagging, Boosting, Stacking and Voting are the ensemble algorithms. The outcomes of the ensemble algorithms that were used were then assessed. Then we used Hyper Parameter Tuning to verify the outcome. Then, using outcome analysis, we assessed the models that had been put into practice. After that we need to execute the data analysis part to start the learning process. Later, to execute model learning and fit the method of predictions. Finally, we need to bagging, boosting, stacking and voting the models to get the best accuracy. Then we can decide the best model to implement considering the best accuracy, precision, recall, and F-1 score. The learning process must then be initiated by carrying out the data analysis step. Next, we must put model learning into practice and fit the predictions approach. To acquire the best accuracy, we must then vote, boost, and bag the models. The best model may then be chosen for implementation based on accuracy, precision, recall, and F-1 score.

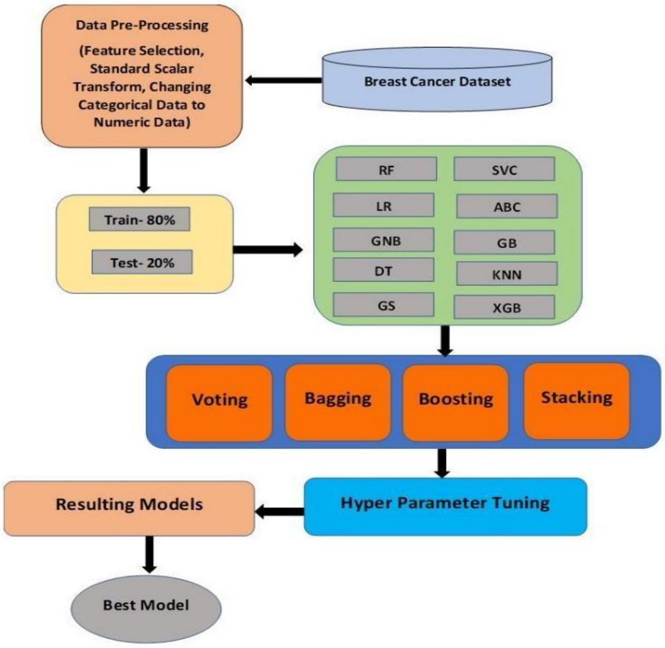


Fig. 3. Methodology of breast cancer prediction.

A correlation subplot has been used to underlying the relationships between two variables or how one variable changes as a result of a change in another. The greater the dependency between variables, the more likely it is that one variable may be successfully predicted from another. It suggests a greater understanding of the dataset and facilitates our capacity to pinpoint the important variables [11].

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. *Classifier Algorithms*

We have implemented some different classifiers named Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Classifier (KNN), Adaboost Classifier (ABC), Decision Tree (DT), GridSearch CV (GS), XGBoost Classifier (XGB), Gaussian Naïve Bayes (GNB), Support Vector Classifier (SVC) algorithms [13].

1) Adaboost classifier: AdaBoost is a boosting classifier that joins a number of ineffective classifiers to create a powerful classifier. 1000 samples are used by ABC to forecast

TA. Weights that differ for classifiers and samples are fixed by ABC [22]. This makes it challenging for classifiers to concentrate on the end outcome. The final formula to achieve TA is,

$$H_k(P) = l - (\sum_{k=1}^{k} \quad a_k h_k(P)).........(1)$$

Here, N=frequency of training data, k = total number of weak classifiers combined to use, hk = output of weak classifier (lower limit 1 to upper limit k), ak = weight of classifier. The notion is depicted in Fig. 4.

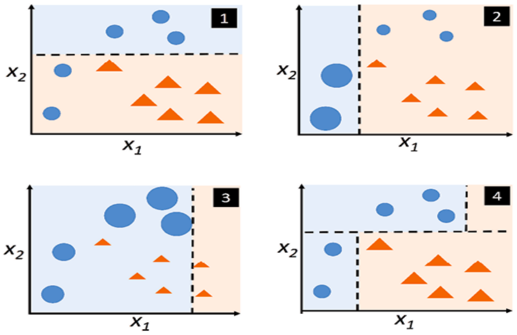


Fig. 4. Adaboost classifier.

2) *Gaussian NB Classifier (GNB):* Gaussian NB Classifier calculates the likelihood of an event occurring given the chance of another event occurring as expressed. Here, every pair of features being categorized is independent of each other (equation 3). The concept is shown below Fig. 5.

$$P(B) = \frac{P(A)P(A)}{P(B)}........................(2)$$

For each feature in Gaussian NB, the continuous value is assumed to have a Gaussian distribution. The resulting histogram looks like bell curve, with all points being equal distance from the curve's center. The conditional probability is provided by equation (4) if the feature likelihood is Gaussian.

$$P(y) = \frac{1}{\sqrt{2\pi\sigma y^2}} exp(-\frac{(x_i-\mu_y)^2}{2\pi\sigma y^2})......(3)$$

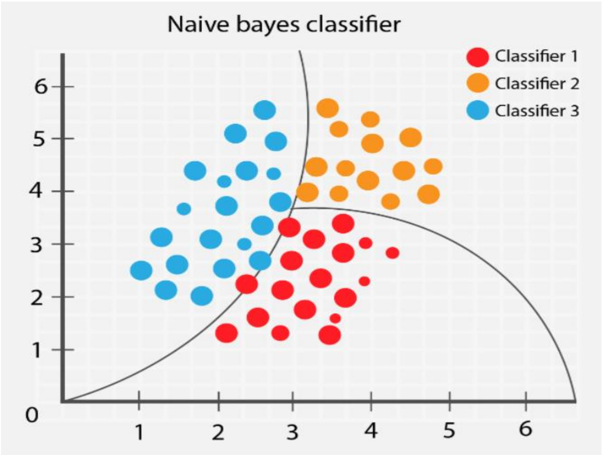


Fig. 5. Gaussian NB classifier.

3) *K-Nearest Classifier (KNN):* K-Nearest Neighbors (KNN) calculates the Euclidean distance between new (x1, x2) and existing (y1, y2) data.

$$Euclidean\ Distance = \sqrt{}(x2 - x1)2 + (y2 - y1)2 \quad (4)$$

The concept is shown in below Fig. 6.

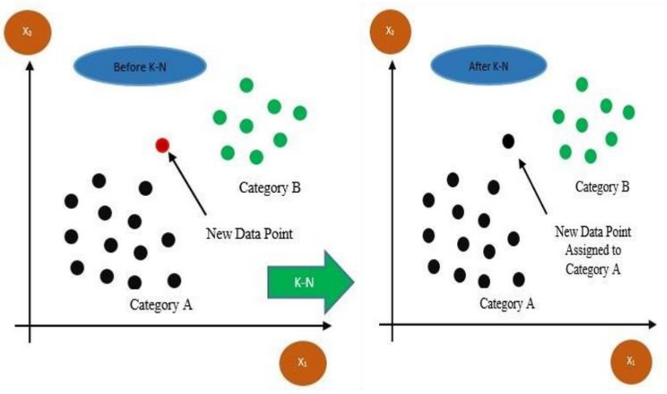


Fig. 6. K-Nearest classifier.

4) *Grid Search CV (GS):* Grid Search CV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. The performance of a model significantly depends on the value of hyperparameters. The concept is shown below in Fig. 7.

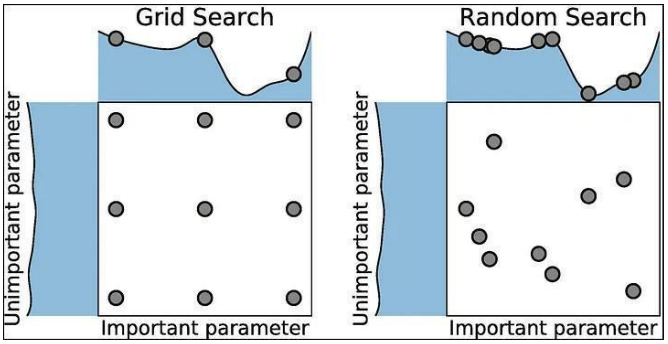


Fig. 7. Grid search CV classifier.

5) *Decision Tree (DT):* Decision trees categorize occurrences by branching them out from a central node to a collection of "leaf" nodes that offer the categorization. To assign a category to an instance, we look at the attribute pointed out by the root node of the tree and then follow the branch of the tree that corresponds to the attribute's value. It is usual practice to calculate two additional metrics to identify the "Best Attribute," "Entropy," as shown in (2), and "Information Gain," as shown in (3) [14]. The "best characteristic" is the trait that provides the most valuable data. Entropy measures dataset homogeneity, whereas information gain measures the rate of change in entropy of attributes. The concept is shown below Fig. 8.

$$E(D) = -P\ (positive)log2P\ (positive) - P(negative)\ log\ log\ 2P\ (negative) \quad (5)$$

$$Gain\ (Attribute\ X) = Entropy\ (decision\ Attribute\ Y) - Entropy\ (X, Y) \quad (6)$$

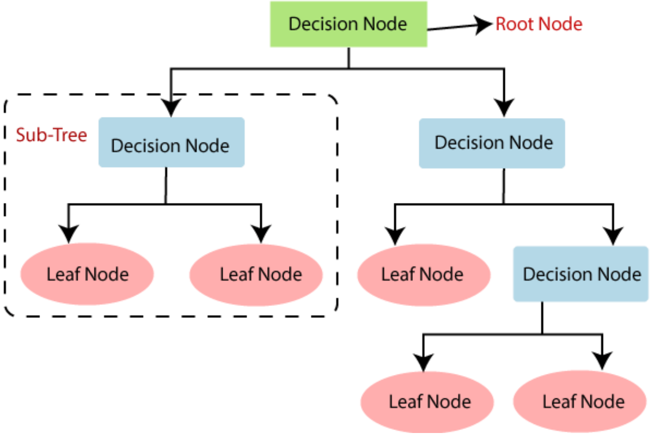


Fig. 8. Decision tree.

6) Logistic Regression (LR): A classifier approach based on machine learning called logistic regression (LR) contains two categories for the class label: yes or no, like a binary (0/1) scale. Although it permits the combined value of continuous variables and discrete predictors, logistic regression is appropriate for discrete variables [16]. The idea is depicted in Fig. 9 below. Logistic regression adopts the supervised machine learning approach. The fundamental equation is shown below [17].

$$h(x) = \frac{1}{1} + e - (\beta o + \beta 1X) \quad (7)$$

'hΘ(x)' is the output of the logistic function, where 0 ≤ hΘ(x) ≥1.

'β1' is the slope.

'βo' is the y-intercept.

'X' is the independent variable.

(βo + β1X) – derived from the equation of a line Y (predicted) = (βo + β1X) + Error.

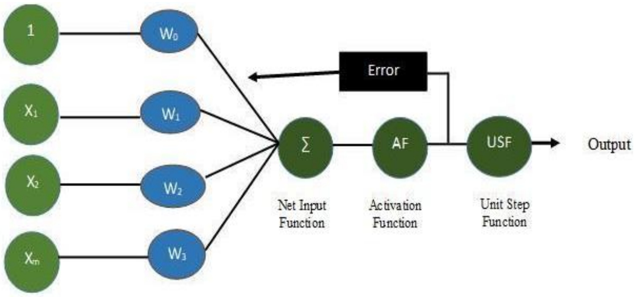


Fig. 9. Logistic regression classifier.

7) Random Forest (RF): Different Decision Tree algorithms make up the Machine Learning (ML) based classifier ensemble approach known as Random Forest (RF) [18]. In order to provide an ideal decision model with more accuracy than the single decision tree model, RF builds several decision trees while the algorithm is being trained. The notion is depicted in Fig. 10 below. All decision tree methods are calculated using the Random Forest algorithm [20].

$$j = \frac{1}{B} + \sum_{b=1}^{B} \quad fb(X') \quad (8)$$

Concerning X = {x1,x2,x3,.................. xn} with respect to Y = {y1,y2,y3,.................. yn} with the lower to upper limit is 1 to B.

Sample x′ = mean of the sum of the prediction $\sum^{B}{}_{b=1} (X^{'})$ for every summation.

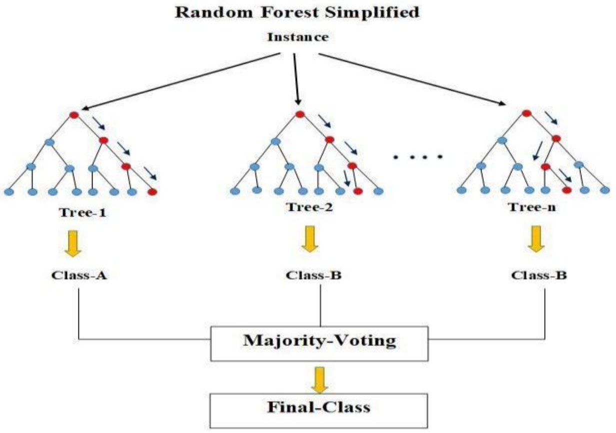


Fig. 10. Random forest classifier.

8) Gradient Boosting (GB): The loss function is the main component of the boosting method known as Gradient Boosting (GB). The notion is depicted in Fig. 11 below. It works by combining and optimizing weak learners to reduce a model's loss function. To improve an algorithm's performance, over fitting is eliminated [7]. Here fi(x) = loss function with correlated negative gradients (−ρi x gm(X)), m = number of iterations.

Feature increment (i) = 1, 2,3, ... , m. Therefore, the optimal function F (X) after m−th iteration is shown below.

$$F\ (X) = \sum_{i=0}^{m} \quad fi(x) \quad (9)$$

Here, gm = the path of loss function's fast decreasing F(X) = Fn − 1(X) the decision tree's target is to solve the mistakes by previous learners [21].

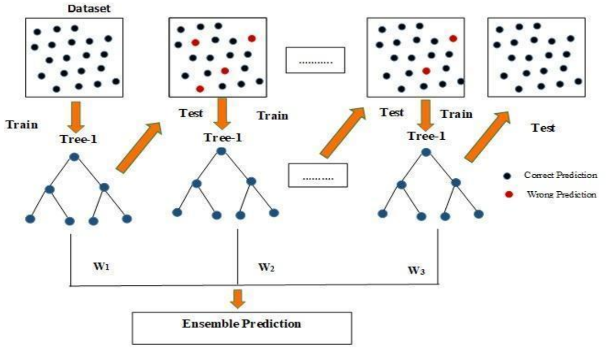


Fig. 11. Gradient boosting classifier.

9) Support Vector Classifier(SVC): The Support Vector Classifier (SVC) approach aims to discover a line, or decision boundary, that divides the space into classes in the most optimum way possible across all n dimensions in order to efficiently categorize new data points [15]. Fig. 12 is showing the working process of Support Vector Classifier (SVC).

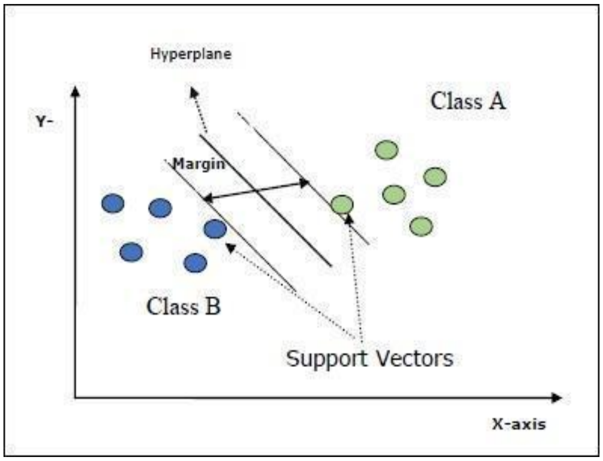


Fig. 12. SVC classifier.

B. Ensemble Methods of Machine Learning

The term "ensemble approach" refers to the use of several classifiers to turn weak classifiers into strong classifiers by producing the greatest accuracy and effectiveness. It was used in our investigation due to variable handling, bias, and uncertainty since it lowers variances, merges predictions from several models, and narrows the prediction spread [23]. In our investigation, four ensemble approaches were employed. Bagging, Boosting, Stacking and voting ensemble modeling was employed.

1) Bagging: Bagging describes the decrease of variance, diminishing handling, and missing variables. The Bagging model's classification formula is displayed below [24].

Here$f'(x)$ is the average of $fi(x)$ for i = 1,2,3,….T.

$$\mathrm{f'(x)} = \mathrm{sign}(\textstyle\sum_{i=1}^{T} fi(x)) \tag{10}$$

2) Boosting: The term "boosting" indicates a method that uses an weighted average to operate with several algorithms and create the loss functions [25]. In our study, the training and testing phase of the hybrid model construction uses the boosting method. The formula is displayed below.

Here, Υ_t = ½-ϵ_t (how much f_t is on the weighted sample).

$$\frac{1}{n}\sum_{i=1}^{n} \quad I(y_j g(x_i) < 0) \leq \prod_{t=1}^{T} \quad \sqrt{1 - 4\Upsilon_t^{\,2}} \tag{11}$$

3) Stacking: Stacking is used to explore many models for the same problem. The idea is that we may approach a learning issue with many models that can learn a piece of the problem but not altogether. Each learnt model can have its own intermediate prediction, allowing for the creation of several distinct learners. As a result, the intermediate prediction may be used to train a second model that will eventually learn the same goal [19]. Stacking outperforms any single model in terms of efficiency. It can offer a representation that uses Logistic regression as a joiner method to blend all conventional classifiers into a final prediction using a joiner technique.

4) Voting: Voting classifiers are a group of classifiers that are used to forecast the class with the best majority of votes. It implies that the model trains using many models to anticipate outcomes by aggregating the results of voting. The notion is depicted in Fig 16 below. The formula we employed is shown below [26] [27].

Here, w_j = weight that can be assigned to the j^{th} classifier.

$$\mathrm{y'} = \mathrm{argmax} \sum_{j=1}^{m} \quad w_j p_{ij} \tag{12}$$

C. Experimental Results and Analysis

First of all, we will clarify the judicial system of our proposed model. We have considered the accuracy, precision, recall and F-1 score shown in Fig. 13.

1) Accuracy: It speaks about the proportion of testing data predictions that were correct. Whereas accessibility of the measures with actual measurements is performed by accuracy. It is founded on a solitary variable. Accuracy only addresses deliberate mistakes. It is one of the most straightforward measurement methods for any model. We must strive to make our models as accurate as possible.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

2) Precision: It speaks about the percentage of positively expected observations that really occurred. The genuine true portion of all the cases where they correctly predicted true are identified by precision. For any type of model, a high recall might also be highly deceptive.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

3) Recall: It speaks about the percentage of positively anticipated observations from a model. High accuracy, though, might occasionally be deceptive. Normally Recall determines the proportion of expected positives to all positive labels.

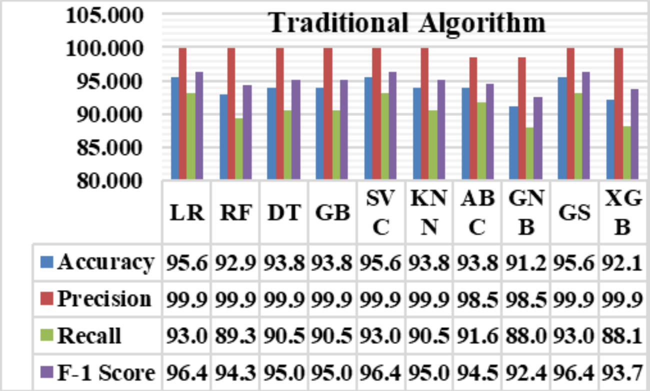
$$Recall = \frac{TruePositive}{TruePositive + \ FalseNegative}$$

4) F-1 Score: It speaks of precision and recall harmonic means. Both the recall and precision ratios are relevant. If the harmonic mean is smaller, the model is probably not very good.

$$F - 1\,Score = 2 * \frac{Recall * \ Precision}{Recall + Precision}$$

At first, we calculated the missing or incorrect values and filtered these from our dataset A and B[28]. The Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Classifier (KNN), Adaboost Classifier

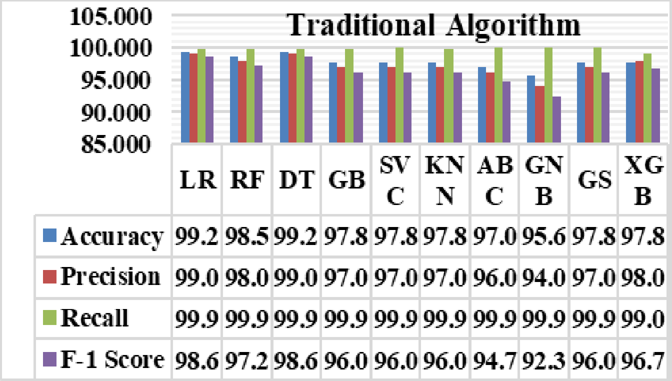
(ABC), Decision Tree (DT), GridSearch CV (GS), XGBoost Classifier (XGB), Gaussian Naïve Bayes (GNB), Support Vector Classifier (SVC) algorithms applied and their performance are measured. We have measured Confusion matrices Accuracy, Precision, Recall and F-1 Score for our proposed algorithms. We have evaluated Bagging, Boosting, Stacking and voting ensemble techniques for dataset A and B.



	LR	RF	DT	GB	SVC	KNN	ABC	GNB	GS	XGB
🟦 Accuracy	95.6	92.9	93.8	93.8	95.6	93.8	93.8	91.2	95.6	92.1
🟥 Precision	99.9	99.9	99.9	99.9	99.9	99.9	98.5	98.5	99.9	99.9
🟩 Recall	93.0	89.3	90.5	90.5	93.0	90.5	91.6	88.0	93.0	88.1
🟪 F-1 Score	96.4	94.3	95.0	95.0	96.4	95.0	94.5	92.4	96.4	93.7

Fig. 13. Experimental results of classifiers for dataset A.

Firstly, we considered the performances of algorithmic classifiers, the best accuracy obtained at 95.614% using Logistic Regression (LR), Support Vector Classifier (SVC), and Grid SearchCV (GS). The best precision score was obtained from Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Classifier (KNN), Decision Tree (DT), GridSearch CV (GS) and XGBoost Classifier (XGB) about 99.99%. The best recall score was obtained from Logistic Regression (LR) and Support Vector Classifier (SVC) with 93.055%. The best F-1 score was obtained at 96.402% from GridSearch CV (GS), Logistic Regression (LR) and Support Vector Classifier (SVC). The visualization is shown in Fig. 14.

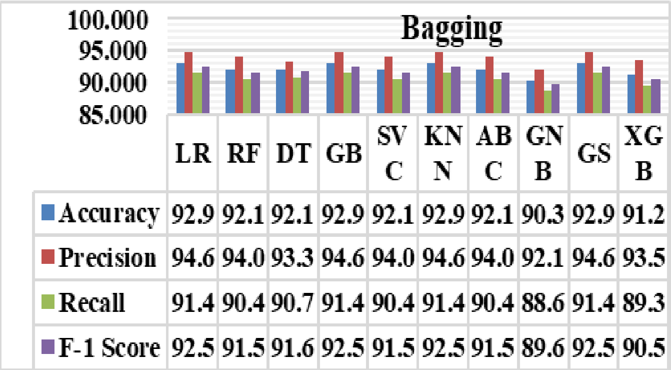


	LR	RF	DT	GB	SVC	KNN	ABC	GNB	GS	XGB
🟦 Accuracy	99.2	98.5	99.2	97.8	97.8	97.8	97.0	95.6	97.8	97.8
🟥 Precision	99.0	98.0	99.0	97.0	97.0	97.0	96.0	94.0	97.0	98.0
🟩 Recall	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.0
🟪 F-1 Score	98.6	97.2	98.6	96.0	96.0	96.0	94.7	92.3	96.0	96.7

Fig. 14. Experimental results of classifiers for dataset B.

We considered the performances of algorithmic classifiers, the best accuracy obtained at 99.270% using Logistic Regression (LR) and Decision Tree (DT). The best precision score was obtained from Logistic Regression (LR) and Decision Tree (DT) about 99.00%. The best recall score was obtained from Adaboost Classifier (ABC), GridSearch CV (GS), Gaussian Naïve Bayes (GNB) and Support Vector Classifier (SVC) with 99.99%. The best F-1 score was obtained at 96.402% from Logistic Regression (LR) and Decision Tree (DT).

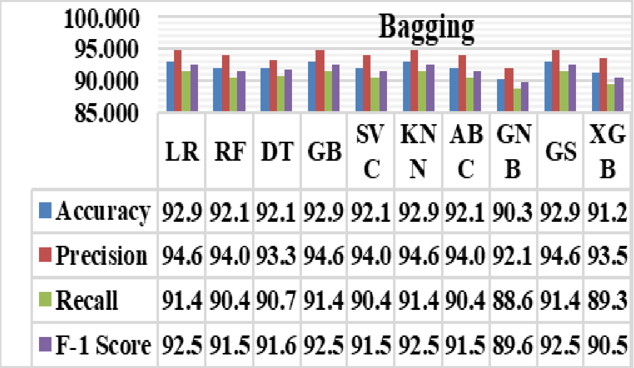
Fig. 15 showed that Decision Tree (DT) had acquired the best score of 99.99%. But GridSearch CV (GS), XGBoost Classifier (XGB) and Support Vector Classifier (SVC) had acquired 99.89%. Hence according to the above analysis as well as the detailed results with graphical representation, Decision Tree (DT) can be stamped as the best algorithmic classifier.



	LR	RF	DT	GB	SVC	KNN	ABC	GNB	GS	XGB
🟦 Accuracy	92.9	92.1	92.1	92.9	92.1	92.9	92.1	90.3	92.9	91.2
🟥 Precision	94.6	94.0	93.3	94.6	94.0	94.6	94.0	92.1	94.6	93.5
🟩 Recall	91.4	90.4	90.7	91.4	90.4	91.4	90.4	88.6	91.4	89.3
🟪 F-1 Score	92.5	91.5	91.6	92.5	91.5	92.5	91.5	89.6	92.5	90.5

Fig. 15. Experimental results of bagging for dataset A.

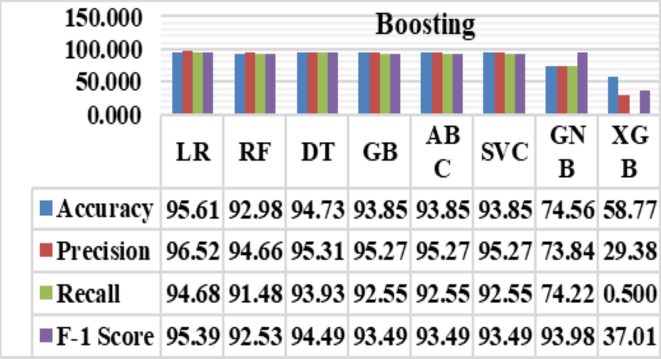
Secondly, we considered the performances of Ensemble Classifier Bagging, the best accuracy had obtained at 92.982% using Logistic Regression (LR) and Gradient Boosting Classifier (GB), KNeareast Classifier (KNN) and Grid Search CV (GS). The best precision score was obtained from Logistic Regression (LR) and Gradient Boosting Classifier (GB), KNeareast Classifier (KNN) and Grid Search CV (GS) about 94.666%. The best recall score was obtained from Logistic Regression (LR) and Gradient Boosting Classifier (GB), KNeareast Classifier (KNN) and Grid Search CV (GS) with 91.489%. The best F-1 score was obtained at 92.531% from Logistic Regression (LR) and Gradient Boosting Classifier (GB), KNeareast Classifier (KNN) and Grid Search CV (GS). The visualization is shown in Fig. 16.



	LR	RF	DT	GB	SVC	KNN	ABC	GNB	GS	XGB
🟦 Accuracy	92.9	92.1	92.1	92.9	92.1	92.9	92.1	90.3	92.9	91.2
🟥 Precision	94.6	94.0	93.3	94.6	94.0	94.6	94.0	92.1	94.6	93.5
🟩 Recall	91.4	90.4	90.7	91.4	90.4	91.4	90.4	88.6	91.4	89.3
🟪 F-1 Score	92.5	91.5	91.6	92.5	91.5	92.5	91.5	89.6	92.5	90.5

Fig. 16. Experimental results of bagging for dataset B.

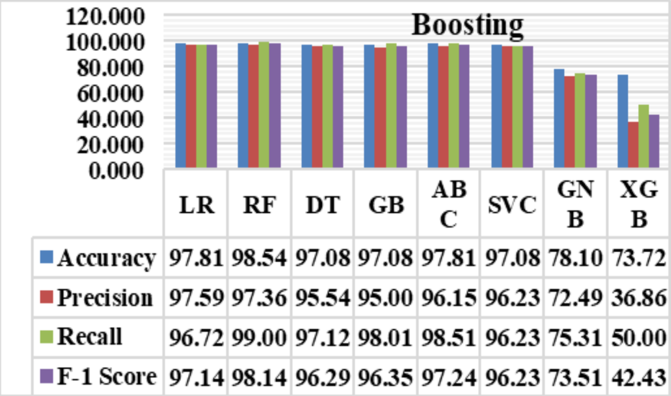
We considered the performances of Ensemble Classifier Bagging, the best accuracy obtained at 99.270% using Logistic Regression (LR) and Gradient Boosting Classifier (GB), XGBoost Classifier (XGB) and Grid Search CV (GS). The best precision score was obtained from Logistic Regression (LR) and Gradient Boosting Classifier (GB), XGBoost Classifier (XGB) and Grid Search CV (GS) about 98.648%. The best recall score was obtained from Logistic Regression (LR) and Gradient Boosting Classifier (GB), XGBoost Classifier (XGB) and Grid Search CV (GS) with 99.504%. The best F-1 score was obtained at 99.066% from Logistic Regression (LR) and Gradient Boosting Classifier (GB), XGBoost Classifier (XGB) and Grid Search CV (GS).



	LR	RF	DT	GB	ABC	SVC	GNB	XGB
🟦 Accuracy	95.61	92.98	94.73	93.85	93.85	93.85	74.56	58.77
🟥 Precision	96.52	94.66	95.31	95.27	95.27	95.27	73.84	29.38
🟩 Recall	94.68	91.48	93.93	92.55	92.55	92.55	74.22	0.500
🟪 F-1 Score	95.39	92.53	94.49	93.49	93.49	93.49	93.98	37.01

Fig. 17. Experimental results of boosting for dataset A.

The final consideration should be the performance obtained using boosting algorithms. After applying boosting algorithms, the best accuracy was obtained at 95.614% using Logistic Regression (LR). The best precision score was obtained from Logistic Regression (LR) about 92.527%. The best recall score was obtained from Logistic Regression (LR) with 94.680%. The best F-1 score was obtained at 95.392% from Logistic Regression (LR). The visualization is shown in Fig. 17.



	LR	RF	DT	GB	ABC	SVC	GNB	XGB
🟦 Accuracy	97.81	98.54	97.08	97.08	97.81	97.08	78.10	73.72
🟥 Precision	97.59	97.36	95.54	95.00	96.15	96.23	72.49	36.86
🟩 Recall	96.72	99.00	97.12	98.01	98.51	96.23	75.31	50.00
🟪 F-1 Score	97.14	98.14	96.29	96.35	97.24	96.23	73.51	42.43

Fig. 18. Experimental results of boosting for dataset B.

The final consideration should be the performance obtained using boosting algorithms. After applying boosting algorithms, the best accuracy had obtained at 98.540% using Random Forest Classifier (RF). The best precision score was obtained from Logistic Regression (LR) about 97.591%. The best recall score was obtained from Random Forest Classifier (RF) with 99.009%. The best F-1 score was obtained at 98.148% from Random Forest Classifier (RF). The visualization is shown in Fig. 18.

V. Conclusion

In this article, the researchers assess the influence rate of individuals employing algorithms. The prediction system may benefit from the diagnosing technology. People can gain from understanding if they will have an impact or not. They should presumably be aware about breast cancer. If individuals use this approach, they can quickly identify the various stages of breast cancer. Assuming the suggested model can also be beneficial to diagnosis authority. The time and difficulty involved in diagnosing breast cancer sickness have decreased because to new technology. The study have made an effort to provide the folks something fresh. A variety of widely used algorithms have been employed that are quick to construct, simple to use, and accurate. Two sets of data has been used and the size of the first dataset, A is 32x569 and the second dataset, B is 10x683. The frequency of breast cancer is categorized in the diagnostic and Class column. That provides the accuracy of 99.270%. Which made this study very relatable to the real life environment and the model can learn by itself which will make this a platform oriented and advance method to predict breast cancer. And early prediction is a cure of this kind of disease. This study has tried to simplify the process of predicting breast cancer in humans. Innovative models can assist people. It’s important to make sure the concept is workable and try to add a lot more features and work on more well-liked topics in the future.

VI. Limitations of a Study

The most crucial limitation of this study is the insufficiency of sample data sets and test dataset. Cell anatomy is evolving day by day and this limitation is a curse of disease prediction methods. Every human body is a box of mystery and it’s tough to fight against anything with a limited amount of data. To achieve this awareness needs to be raised. In this research, the methodology is used to forecast breast cancer in humans. But it is also observed that there is a shortage of knowledge and diagnostic tools. Cancer detection and symptom analysis are expensive in developing nations. This research work is attempting to use machine learning to address the issue.

VII. Recommendation for Future Research

Expanding the datasets used for breast cancer prediction can provide a more comprehensive understanding of the disease. Including data from different demographics, geographic regions, and medical institutions can help capture a broader spectrum of breast cancer cases and improve the generalizability of the models. Future work can focus on developing real-time prediction models that can assist healthcare professionals in making timely and informed decisions. Integration with electronic health records (EHRs) and clinical decision support systems can enable seamless integration of the prediction models into the clinical workflow. Applying emerging technologies, such as deep learning, reinforcement learning, and federated learning, can further enhance the performance and scalability of breast cancer prediction models.

References

[1] “World Cancer Research Fund International” Last Accessed: March 22, 2022. Available: https://www.wcrf.org/cancer-trends/breast-cancer-statistics/.

[2] L. Yang and A. Shami, “On hyperparameter optimization of machine learning algorithms: theory and practice,” Neurocomputing, vol. 415, pp. 295–316, 2020.

[3] Shamrat, F.J.M., Raihan, M.A., Rahman, A.S., Mahmud, I. and Akter, R., 2020. An analysis on breast disease prediction using machine learning approaches. International Journal of Scientific & Technology Research, 9(02), pp.2450-2455.

[4] Keleş, M.K., 2019. Breast cancer prediction and detection using data mining classification algorithms: a comparative study. Tehnički vjesnik, 26(1), pp.149-155.

[5] Anastraj, K., Chakravarthy, T., Sriram, K., Collge, A.S.P. and Poondi, T., 2019. Breast cancer detection either benign or malignant tumor using deep convolutional neural network with machine learning techniques. Adalya Journal, 8, pp.77-83.

[6] Erkal, B. and Ayyıldız, T.E., 2021, November. Using Machine Learning Methods in Early Diagnosis of Breast Cancer. In 2021 Medical Technologies Congress (TIPTEKNO) (pp. 1-3). IEEE.

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

[8] Merouane, E. and Said, A., 2022. Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers. International Journal of Advanced Computer Science and Applications, 13(2).

[9] “Breast Cancer Dataset”, Accessed: December 29, 2021, Available: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset .

[10] “Wisconsin Breast Cancer Database”, Accessed: December 29, 2021, Available: https://www.kaggle.com/datasets/roustekbio/breast-cancer-csv.

[11] “What is Correlation in Machine Learning?”, Accessed: August 6, 2020, Available: https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47.

[12] “What is Correlation in Machine Learning?”, Accessed: November 8, 2021, Available: https://medium.com/analytics /what-is-correlation.

[13] V. Lahoura, H. Singh, A. Aggarwal et al., “Cloud computing-based framework for breast cancer diagnosis using extreme learning machine,” Diagnostics, vol. 11, no. 2, p. 241, 2021.

[14] Nahar, Nazmun, and Ferdous Ara. "Liver disease prediction by using different decision tree techniques." International Journal of DataMining & Knowledge Management Process 8, no. 2 (2018): 01-09.

[15] Aljahdali, Sultan, and Syed Naimatullah Hussain. "Comparative prediction performance with support vector machine and random forest classification techniques." International journal of computer applications 69, no. 11 (2013).

[16] L. Mary Gladence, M. Karthi, V. Maria Anu. “A statistical Comparison of Logistic Regression and Different Bayes Classification Methods for Machine Learning” ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, Vol -10, No-14, August 2015.

[17] “Logistic Regression for Machine Learning”, Accessed: August 6, 2021, Available: https://www.capitalone.com /tech/machine-learning/what-is-logistic-regression/.

[18] Ghosh, Pronab, Asif Karim, Syeda Tanjila Atik, Saima Afrin, and Mohd Saifuzzaman. "Expert cancer model using supervised algorithms with a LASSO selection approach." International Journal of Electrical and Computer Engineering (IJECE) 11, no. 3 (2021): 2631.

[19] Shorove Tajmen, Asif Karim, Aunik Hasan Mridul, Sami Azam, Pronab Ghosh, Alamin Dhaly, Md Nour Hossain. “A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease”. In April 2022 The 10th International Conference on Computer and Communications Management in Japan.

[20] Aunik Hasan Mridul, Md. Jahidul Islam, Mushfiqur Rahman, Mohammad Jahangir Alam, Asifuzzaman Asif. “A Machine Learning-Based Traditional and Ensemble Technique for Predicting Breast Cancer”, In December, 2022. Conference: 22th International Conference on Hybrid Intelligent Systems (HIS 2022) online, 2022At: Auburn, Washington, USA.

[21] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms." ArtificialIntelligence Review 54, no. 3 (2021): 1937-1967.

[22] Hou, Zhi-Hua. Ensemble methods: foundations and algorithms. CRC Press, 2012.

[23] Emmens, Aurélie, and Christophe Croux. "Bagging and boosting classification trees to predict churn." Journal of Marketing Research43, no. 2 (2006): 276-286.

[24] Wang, Yizhen, Somesh Jha, and Kamalika Chaudhuri. "Analyzing the robustness of nearest neighbors to adversarial examples." In International Conference on Machine Learning, pp. 51335142. PMLR, 2018.

[25] Drucker, Harris, Corinna Cortes, Lawrence D. Jackel, Yann LeCun, and Vladimir Vapnik. "Boosting and other ensemble methods." Neural Computation 6, no. 6 (1994): 1289-1301.

[26] Sharma, Ajay, and Anil Suryawanshi. "A novel method for detecting spam email using KNN classification with spearman correlation as distance measure." International Journal of Computer Applications 136, no. 6 (2016): 28-35.

[27] Pasha, Maruf, and Meherwar Fatima. "Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection." J. Softw. 12, no.12 (2017): 923-933.

[28] Islam, Rakibul, Abhijit Reddy Beeravolu, Md Al Habib Islam, Asif Karim, Sami Azam, and Sanzida Akter Mukti. "A Performance Based Study on Deep Learning Algorithms in the Efficient Prediction of Heart Disease." In 2021 2nd International Informatics and Software Engineering Conference (IISEC), pp. 1-6. IEEE, 2021.