European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 02, 2020

Feature Selection of Breast Cancer Data Using Gradient Boosting Techniques of Machine Learning

Anusha Derangula[1]*, Prof. SrinivasaReddy Edara[2], Praveen Kumar Karri[3]

[1,2,3]Department of Computer Science and Engineering, AcharyaNagarjuna University, Andhra Pradesh, India
* Corresponding author's Email: d.anusha21@gmail.com

Abstract: Cancer is described as a very alarming disease among humankind. The second main reason for death among modern women is Breast cancer. It affects the physical, mental, social lifestyles of the people. It is possible to treat cancer in the early stages. The importance of cancer cells classification into begnin and malignant has led to many research areas in the medical field. Medical practitioners were adopting machine learning techniques to detect, classify, and predict the malignant tumour effectively. The machine learning algorithms yield better results in the diagnosis of malignant tissue. The learning algorithm performs well with optimal features. The objective of this paper is to identify optimal features in Wisconsin breast cancer Diagnostic data. The techniques used for feature selection here are Light Gradient Boosting Model (LGBM), Catboost and Extreme gradient boosting (XGB). The optimized features were given to the Naive Bayes classifier and got an accuracy of 96.49%.

Keywords: machine learning algorithms, Wisconsin Breast Cancer Diagnostic data, Light Gradient Boosting Model, Catboost, Extreme gradient boosting, Naive Bayes classifier**.**

1. Introduction

Cancer was considered one of the fatal diseases that have no cure until today. In this generation, many dangerous diseases have grown exponentially due to the high consumption of fast food and cancer is one among them. As per the World Health Organization (WHO) records, the number of women affected with breast cancer is 1.2 billion [6]. This dreadful disease has more than 100 subtypes, such as throat cancer, blood cancer, lung cancer, breast cancer, etc. [2]. Breast cancer can also show up among men and the chances are 1%. One in 1000 men can be examined with breast cancer [1]. Approximately 12.5 % of women had breast cancer all through the world [2]. Sometimes cancer makes massive tissue inside the body parts, which are called tumours. The tumours on growth can affect many organs of the body.

Scientists identified two types of cancerous tumours. They are Begnin and Malignant [2]. Begnin tumours are not very dangerous and they don't cause death. The growth of this type of tumour is limited to a particular body part and is very slow. Malignant type of tumours has uncontrollable growth with invasion lymph system destroys other healthy tissues of the body. New blood vessels were made by the tumour to feed itself which causes anaemia. Early detection of cancer has a 100 % rate of survivability. According to WHO, it is roughly calculated that the number of new cases recorded were 2 million among which 626679 deaths were estimated [10]. Many younger women were subjected to breast cancer in developing countries than in developed countries [2]. Though scientists did not identify the reason for cancer, some risk factors may help detect cancer early.

1.1 Factors of breast cancer

The factors leading to breast cancer were classified into tractable and non-tractable factors [1].Non-Tractable factors are:

- Age
- Gender
- Family members with breast cancer history
- Medical history of radiation therapy.

Tractable factors are:

- High Body Mass Index (BMI).
- Age during the firstchildbirth.
- Food habits.
- Alcohol.
- Number of children.
- Number of abortions.

1.2 Types of breast cancers

Breast cancer was of two types. In Non-invasive Breast Cancer, the cancerous cells were limited within the ducts and did not spread to further tissues. In Invasive Breast Cancer, the malignant cells have spread around into the remaining tissues [5]. By detecting the tumour in the beginning stages there is a higher chance of treating the patient effectively.

1.3Stages of breast cancer

There are five stages of Breast Cancer [1]. The details about the breast cancer stages and their respective survival rate (SR) were given below:

Stage 0: In this stage, the cancerous cells will be on the surface of the ducts. The tumour did not attack the surrounding tissues. The undergoing patient has an SR of 100%.

Stage 1: In this stage the size of the tumour is 2cm. The effect of cancerous cells on the lymph nodes aresignificantlyless whereas the patient has 98 % of SR.

Stage 2: Here, the size of the tumour is between 2cm-5cm. It may or may not spread to the lymph nodes. The rate of SR is %.

Stage 3: In this stage, the size of the tumour crosses 5cm or above. It may spread to a few or many lymph nodes. This stage has 52% of SR.

Stage 4: In this stage, the tumour has extended to different body organs, whereas SR is 16%.

1.4 Prognosis of breast cancer

For examining the various stages of breast cancer, Chest X-ray scan, CT scan, Bone scan, and PET scans are widely used by medical practitioners [4]. Initially, biologists have used a microscope to understand the tumour behaviour for breast cancer patients [4].

1.5 Indications of Breast Cancer

The signs of breast cancer were used for quick detection of the disease, which increases the chance of survivability for the undergoing patient. The symptoms are:

1. Mass in the breast area
2. Armpit or collarbone swelling
3. Redness
4. Inverting nipple
5. discomfort in the breast
6. Swelling of the breast
7. Skin dimpling

1.6Effectiveness of Machine Learning

Machine learning algorithms effectively classify between Begnin and Malignant, which helps the medical practitioner diagnose it. Identifying the subset of features is an essential task for machine learning classifiers. This paper presents an effective feature selection method by using gradient boosting techniques such as LGBM, CATBOOST and XGB techniques. The gradient boosting techniques uses a gradient descent approach and minimizes the loss to yield very accurate results. Hence this method is very effective than the existing feature selection methods.

2. Literature Survey

In recent times researchers have done a significant amount of research work to apply machine learning to clinical data. Many researchers have used breast cancer data for classification, prediction, and detection of the presence of malignant tissue.

2.1 Existing system

Some of the existing methodologies for classification and prediction of breast cancer data was given.

BiplobDey et al. [1] portrayed a detailed report on breast cancer. They also gave complete information about its causes, types, stages, factors that lead to breast cancer, treatments, and the detailed history of breast cancer.

RozillaJamiliOskouei et al. [2] have given a complete study of implementing data mining techniques, including the basic concepts of data mining and detailed information about breast cancer. They provided information regarding the frequently used datasets used by researchers and showed how data mining techniques were implemented to diagnose the malignant tissue.

Ajay Kumar et al. [3] implemented machine learning models like K-Nearest Neighbour, Support Vector Machine, Decision Tree, Bayesian Network and Naïve Bayes on datasets taken from the UCI repository. They concluded that Bayesian Network yields high accuracy with fewer features, whereas Support Vector Machine shows high accuracy with more features.

AnimeshHazra et al. [4] implemented two feature selection methods. They are Pearson's correlation methods and principal component analysis. The feature selection methods were implemented with Naive Bayes, Support Vector Machine and Ensemble classifier. According to the study, they concluded that Naive Bayes classifier was the best technique.

Ganesh N. Sharma et al. [5] provided a complete review of breast cancer, which includes the various types of techniques used in breast cancer diagnosis, various surgeries that can be done for the disease. Finally, they provided some of the ongoing researches in the field of breast cancer.

PuneetYadav et al. [6] implemented the decision tree and support vector machine (SVM) on breast cancer data, which has an accuracy of 90 % to 94 % and 94.5 % to 97 %, respectively.

PouriaKavianiet al. [7] performed a survey of Naive Bayes Classifier, its advantages and disadvantages and its application in various areas.

MedisettyHari Krishna et al. [8] tested various machine learning approaches on data retrieved from the UCI repository and concluded that Support Vector Machine shows high accuracy among all models.

BazilaBanu A et al. [9] portrayed a report on the performance of different Bayes classifiers like Tree Augmented Naïve Bayes, Boosted Augmented Naïve Bayes and Bayes Belief Network on Wisconsin Breast Cancer data and showed Tree Augmented Naive Bayes has the best performance.

Ch. Shravya et al. [10] performed the classification of breast cancer data. They have taken data from the UCI repository and used Logistic Regression, Support Vector Machine and K Nearest Neighbor. They showed that the Support Vector Machine got the highest accuracy among the remaining classifiers.

DestaMulatu et al. [11] surveyed different data mining techniques like Naive Bayes classification and prediction algorithm, Rotation forest model, Decision tree, Support Vector Machine, Artificial Neural Networks, etc. for breast cancer data. They stated that the Decision Tree, Naive Bayes, Support Vector Machine are giving more accurate results.

Ms.Sindhuja et al. [12] provided a review of some of the data mining models like k-nearest, Bayes, fuzzy-c-means, neural network, thresholding, etc. for breast cancer diagnosis.

Subrata Kumar Mandal et al. [13] performed the Pearson correlation Coefficient on Wisconsin Breast Cancer data and checks the accuracy with Decision Tree, Logistic Regression and Naïve Bayes models. They compared the three models' results and shown that the highest accuracy was achieved with Logistic Regression, whereas Naïve Bayes was the least.

Akhilesh Kumar Shrivas et al. [14] applied different data mining techniques like CART, C4.5, Multi-layer perceptron, Bayesian net, Support Vector Machine, and also Radial Basis Function on breast cancer data and compared their performance. They proved that the Bayesian net had got the highest accuracy. They also applied a feature selection technique called the infogain. They implemented with Bayes net as well as Support Vector Machine, among which they got the highest accuracy with Bayes net.

Ajay Kumar et al. [15] applied and compared many classification algorithms, namely Naïve Bayes, Decision Tree, K Nearest Neighbour, Support Vector Machine and Bayesian Network. They collected data from the UCI Machine Learning Repository. They implemented all the algorithms and proved that Bayesian Network for less featured dataset yields high accuracy, whereas Support Vector Machine gives the best accuracy for the more featured dataset.

NehaKumari at al. [16] gave detailed information about several Machine Learning algorithms used to diagnose breast cancer.

Ms.ShwetaSrivastava et al. [17] gave complete information about the filter, wrapper, and embedded feature selection approaches. They also gave information about various areas where feature selection can be applied.

K. Sutha et al. [18] provide a complete survey of distinct feature selection algorithms and mentioned their pros and cons.

B. Senthil Kumar et al. [19] uses novel techniques like an improved firefly and random forest algorithms for selecting the features on Pima dataset. The accuracy is tested for various classification algorithms like Support Vector Machine, Naïve Bayes, Artificial Neural Networks Random Forest, K Nearest Neighbours, Hybrid Random Forest.

Ms.ManjiriMahadev et al. [20] has reviewed several machine learning and provided brief information about various classification algorithms.

2.2 Proposed System

We perform feature selection for Wisconsin breast cancer diagnostic data in the proposed system by implementing gradient-based machine learning techniques such as LGBM, CATBOOST, and XGB. The accuracy of the selected features was tested by using the Naive Bayes classifier.

3.Description of the Dataset

The dataset used in this paper is Wisconsin Breast Cancer (Diagnostic) data (WDBC) from the standard UCI Machine learning repository. The dataset has 569 instances and 32 attributes.

Table 1. Wisconsin breastcancer (diagnostic) data

Dataset	No. of attributes	No. of Instances	No of classes
Wisconsin breast cancer (diagnostic) Data	32	569	2

The dataset consists of two classes, namely Begnin and Malignant, 357 and 212 instances. The following figure plots the number of Begnin and Malignant classes in the dataset.
Each attribute in WDBC dataset has three columns with three values calculated for each attribute. They are Mean, Standard Error and worst mean.
They are defined as:

$$Mean = \frac{\sum f(x)}{f} \quad (1)$$

$$Standard\ Error = \frac{Standard\ deviation}{Square\ root\ (N)} \quad (2)$$

and*Worst* is the worst or largest mean.

The attribute values of each attribute and their corresponding mean, standard error and worst mean were given in 10 columns per each.

In the dataset, all the mean values were given from column 3 to column 12 and standard error values are from column 13 to 22, whereas column 23 to 32 consists of worst mean values. Id and diagnosis were given in column 1 and column 2, respectively.

The following table shows the attributes of WDBC along with their description..

Table 2. Attributes of WDBC

s no	Name of the attribute	Description of the attribute
1	Radius	Mean of the distances from the center to the perimeter.
2	Texture	Standard deviation of grey scale values.
3	Perimeter	Total distance between continuous snake points.
4	Area	Measure of the number of each pixels inside the snake points and add half of the pixels to the parameter
5	Smoothness	variation in radius lengths
6	Compactness	perimeter^2/area-1.0
7	Concavity	Severity of the ---portions of the contour
8	Concave_points	Number of concave points of the contour
9	Symmetry	The longest chord through center should be found and the lines which are perpendicular to the major axis confines to the nuclear boundary in every direction is measured.
10	Fractal_dimension	“costline approximation”-1

The distribution of the classes Begnin and Malignant in the dataset was shown in the following plot diagram.

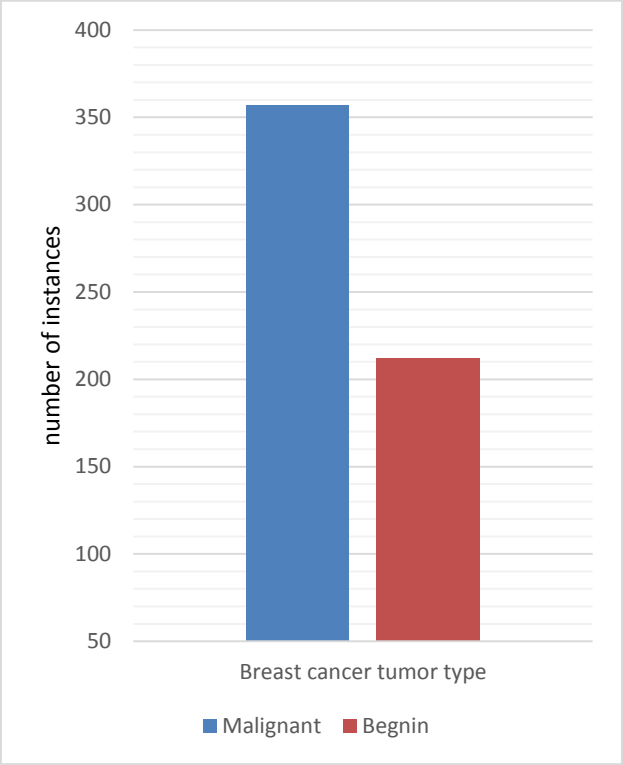

Figure 1: Plot diagram of number of malignant and begnin

4.Data Preprocessing and selection

Data preprocessing is the foremost step before implementing any machine learning technique. It includes finding the missing values, replacing the missing values, finding the outliers, encoding categorical variables, feature scaling, etc. The column id was dropped from the dataset as it does not influence the class label. The data preprocessing step improves the quality of data and makes it useful for modeling. During this stage, data was partitioned into training and testing data. Training data is used to train the model and testing data was used at the classification stage. In this paper 399 instances of 31 attributes were taken for training and 170 instances of 31 attributes were taken for testing data.

Data selection is used to reduce the number of features of the dataset by selecting important features. Here gradient boosting techniques were used to determine the essential elements from the dataset.

5. Methodology

In this paper we perform feature selection of WDBC by three gradient boosting techniuques that are LGBM, CATBOOST and XGB.

Feature selection means the selection of a subset of optimal features from the given dataset [10]. It is considered an important task in classification algorithms [18]. The overall accuracy of the learning model can be increased by feeding the model with optimized features.

Initially, the classifiers were run independently and the top 10 important features from each algorithm were noted. Among all three feature subsets, we select 7 feature that are common in the three feature subsets. In this way the hyperparameter tuning was done in this paper. We tested the hyperparameters by using the Naive Bayes classifier. The learning algorithm shows better performance of precision, recall and accuracy of 97%, 95%, 96% respectively.

The architecture of the proposed system is shown in the following figure.

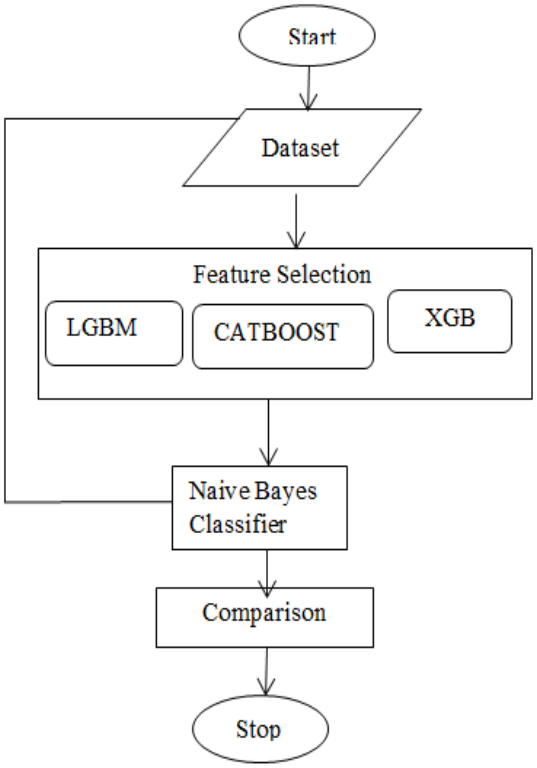

Figure 2: Workflow of the proposed methodology

5.1 Gradient boosting

Ensemble learning combines the predictions of many algorithms to produce a better prediction value. The main reason ensembling was used is that the number of models trying to predict the same target variable than a single predictor gives a better prediction. Bagging and Boosting are two ensemble techniques. In Bagging, we create independent learners and combines them using model average techniques (example: voting), whereas in boosting predictors were constructed sequentially shown in the following figure.

Gradient Boosting algorithm is a type of ensembling technique in which classification and prediction are made by combining weaker models. This technique has three elements: 1) The loss function, which needs to be optimized. The loss function is a metric for calculating how good the model coefficients are at fitting the data. It has no definition and depends on what the programmer wants to optimize. 2) A Weak learner, for performing classification or prediction (decision trees were used as a weak learner here), 3) additive model, which adds one tree at a time.

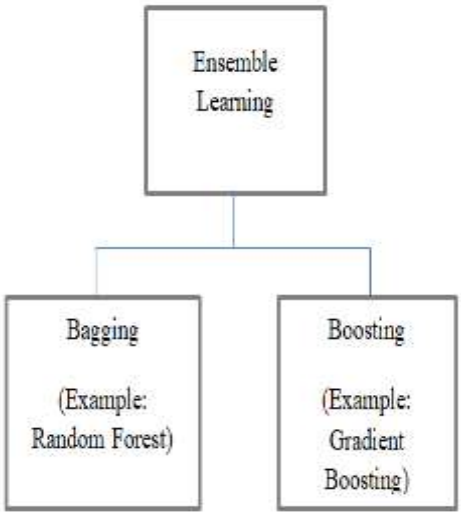

Figure 3: Types of ensemble based techniques

After the calculation of loss function, the new tree gets added to the existing model to perform gradient descent procedure in order to minimize the loss function. Asgradient boosting models work highly on optimizing the loss, maximum accuracy will be achieved.

5.1.1 Gradient Boosting Algorithm Procedure

Step i:

Calculate the loss function and minimize it. Here loss function is the mean squared error of target and predicted values.

$$\text{Loss} = \sum (x_i - x_i^p)^2 \quad (3)$$

Whereas, $x_i$is the ith target value and x_ip is the ithprediction value.

Step ii:

By applying gradient descent and Update the predictions

$$x_i^p = x_i^p + \alpha$$

$$\times \delta \sum \frac{\left(x_i - x_i^p\right)^2}{\delta x_i^p} \quad (4)$$

whereasα is the learning rate, $\sum (x_i - x_i^p)^2$ is the loss function.

Hence by updating the predictions, the loss function is minimized and predicted values were approximately close to near values.

5.2 LGBM

This is a very fast model and hence it is named 'Light Gradient Boosting Model'. LGBM is one of the gradient boosting frameworks that use tree-based learning algorithms. It is used for classification, feature ranking and can perform many machine learning tasks. This model is based on a decision tree algorithm and splits the tree using a leaf wise approach.

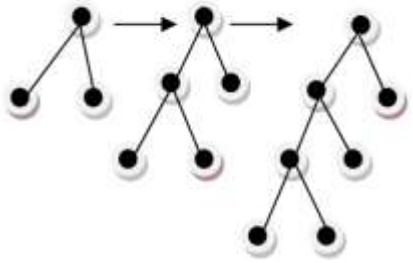

Figure 4: Leaf-wise growth of the tree

LGBM is highly efficient, provides better accuracy, can handle data on a very large scale and uses lower memory. The classifiers LGBM, CATBOOST and XGB, are available as packages in sklearn in python. LGBM classifier was applied to the given dataset to generate the important features of the dataset. As a result, we considered the top 10 features that are listed as follows:

1. texture_mean
2. concavity_worst
3. texture_worst
4. area_worst
5. smoothness_mean
6. concavepoints_worst
7. symmetry_se
8. perimeter_worst
9. smoothness_worst
10. concavepoints_mean

5.3 Catboost

The word CATBOOST is derived from two words "category" and "boosting". This machine-learning algorithm can handle multiple categories of data like text, numerical, audio, video, which is its main feature. Visualization tools were also included in this model.

Catboost model is trained with the dataset to generate the important features. The top 10 important features given by the model are:

1. concave points_worst
2. area_worst
3. concave points_mean
4. texture_worst
5. radius_worst
6. perimeter_worst
7. area_se
8. texture_mean
9. concavity_worst

10. smoothness_worst

5.4XGB:

Extreme Gradient Boosting (XGB) is an implementation of gradient boosting machine learning techniques. It is a software library that can be installed and accessed from various interfaces. The execution time of XGB is very less when compared to the other gradient boosting techniques. The model performance is very high when performed classification and regression on structured or tabular datasets. In this paper, the XGB classifier is used to find the optimized features for the given dataset. The ten important features that were obtained from the learning model are:

1. perimeter_worst
2. radius_worst
3. concave points_worst
4. area_worst
5. concave points_mean
6. area_mean
7. texture_mean
8. smoothness_worst
9. texture_worst
10. concave points_se

We have implemented LGBM, CATBOOST and XGB classifiers and got three subsets of optimal features. By comparing the important features generated by the three algorithms, we selected a subset of features that are common in all three classifiers. The subset of features was identified as hyperparameters of the Wisconsin Breast Cancer (Diagnostic) dataset are the following:

1. texture_mean
2. texture_worst
3. area_worst
4. concave points_worst
5. perimeter_worst
6. area_se
7. concave points_mean

The identified feature subset has a higher probability of breast cancer detection. The following figures show the probability of each hyperparameter's contribution to a particular instance for identifying malignant or begnin.

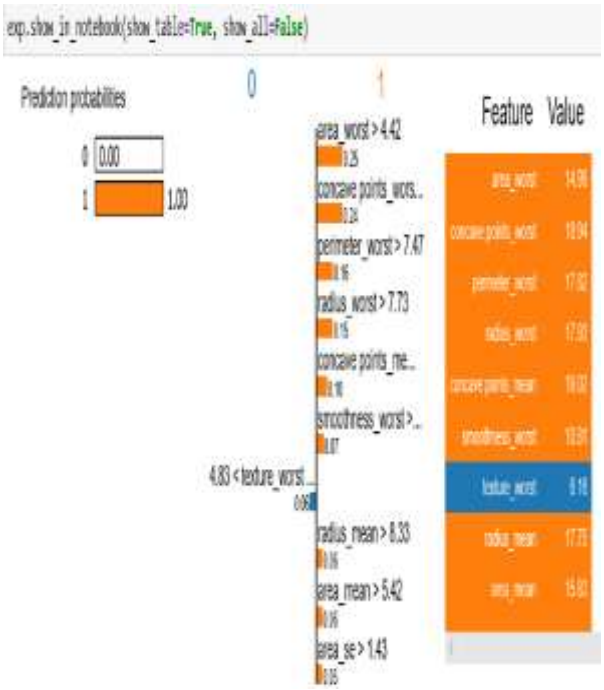

Figure 5: The contribution of each optimal parameter in identifying the malignant tissue

In figures 8 and 9, the prediction probability 0,1 indicates classes 0 and 1, respectively whereas, class 0 is begnin and class 1 is malignant.

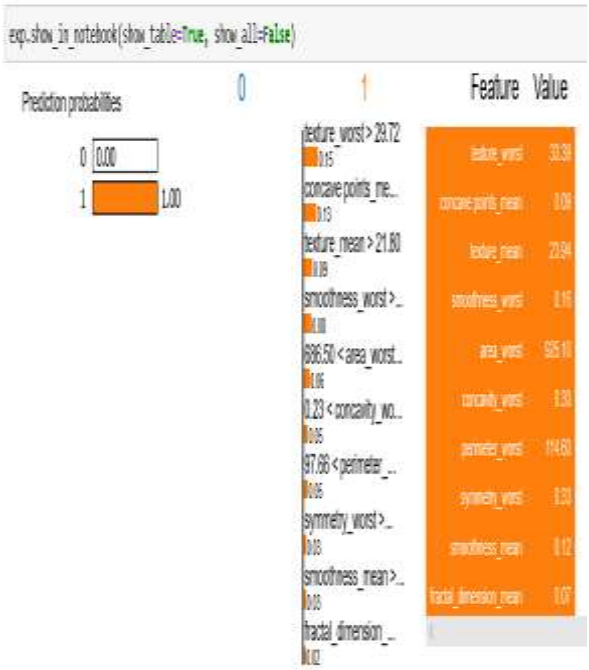

Figure 6: The contribution of each optimal parameter in identifying the malignant tissue

5.5Naive Bayes classification algorithm

Naive Bayes classification algorithm is based on Bayes theorem. This technique is commonly in supervised learning [10]. It is a subset of Bayesian decision theory. As the formulation is based on naive assumptions, hence it is called Naive [7]. Naive Bayes is one of the statistical classifier [10]. It is fast, easy to implement, effective and simple. The algorithm assumes that each feature given to the model was independent and equal to the target. Naive Bayes algorithm can be implemented effectively for very large datasets [8]. By giving the probability of an event that has already occurred, we can find an event's probability by using Bayes theorem.

$$P\left(\frac{p}{q}\right) = \frac{P\left(\frac{q}{p}\right) P(p)}{P(q)} \quad (5)$$

$P\left(\frac{p}{q}\right)$= posterior probability of class
P on predictorq

$P(p)$= prior probability of the vector.

6. RESULT ANALYSIS

The Confusion matrix presents the results of the classification. It shows the relationship between the actual classes and predicted classes. It is used to depict the outcomes of both. It also shows how many features that are actually true were predicted as true as well as false. Similarly, how many numbers of actually false were predicted as true as well as false.

Table 3: Confusion matrix

	Actual Values	
Predicted Values	TN	FP
	FN	TP

i) True Negative (TN): The predicted values were correctly predicted as actual negative.
ii) True Positive (TP): The predicted values were correctly predicted as actual positive.
iii) False Positive (FP): The predicted values were incorrectly predicted as actual positive. i.e., the values which are negative were predicted as positive.
iv) False Negative (FN): The predicted values were incorrectly predicted as actual negative, i.e., the positive values were predicted as negative.

From the confusion matrix we can find out the following values:

Precision: It is the proportion of positive cases that were correct. The formula calculates it:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

Recall: It is the number of positive cases that were correctly identified. It is calculated by using the formula:

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

F1 score: It is the harmonic mean of precision and recall.

$$F1\ Score = \frac{2 \times (precision \times recall)}{(Precision + recall)} \quad (8)$$

Accuracy: It is the proportion of the total number of correct predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

Whereas *TP* = True Positive
FP= False Positive
TN= True Negative
FN = False Negative

The experimental results were given in the figure below:

Performance Evaluation of Gradient Boosting Techniques

The results were shown in the following table.

Table 4: Classification report of the classifiers after 10 – fold cross validation

Name of the classifier	Precision	Recall	F1-score	Accuracy
LGBM	98	95	96	97
CAT BOOST	97	89	93	96
XGB	97	92	95	96

The accuracy of LGBM, CATBOOST and XGB were 97%, 96% and 96%, respectively. Hence gradient boosting techniques are the best techniques as they perform gradient descent and yield highly accurate values. The common features generated from these models were considered as an

optimal subset of features and were given to the Naive Bayes classifier. The results for the optimal feature subset when tested with the Naive Bayes classifier was shown below:

The confusion matrix for Naive Bayes before feature selection was shown below:

```
cm = confusion_matrix(y_test.astype(int), y_pred.astype(int))
cancer_cm = pd.DataFrame(cm, range(2), range(2))
plt.figure(figsize=(5,5))
sns.set(font_scale=1.4)  #for label size
cm_plot = sns.heatmap(cancer_cm, annot = True, cmap =
                      'Pastel1', fmt = 'n', annot_kws = {"size": 25})
```

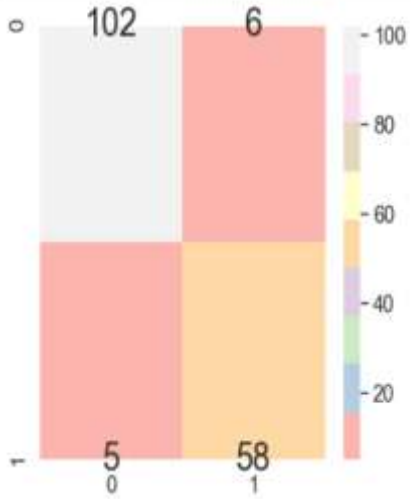

Figure 7: Confusion matrix for Naive Bayes Classifier before feature selection

The confusion matrix for naive bayesafter feature selection was shown below:

```
cm = confusion_matrix(y_test.astype(int), y_pred.astype(int))
cancer_cm = pd.DataFrame(cm, range(2), range(2))
plt.figure(figsize=(5,5))
sns.set(font_scale=1.4)  #for label size
cm_plot = sns.heatmap(cancer_cm, annot = True, cmap =
                      'Pastel1', fmt = 'n', annot_kws = {"size": 25})
```

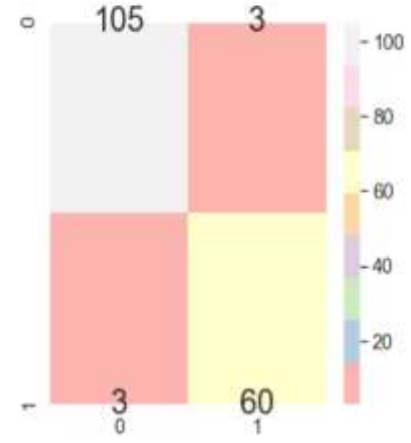

Figure 8: Confusion matrix for Naive Bayes Classifier after feature selection

Table 5: Comparison between the Naive Bayes classifierbefore and after feature selection

Naive Bayes classifier	Precision	Recall	F1-score	Accuracy
Before feature selection	91	92	91	94
After feature selection	97	95	95	96

7. Conclusion and Future Work

In this paper we have found the optimized parameters of Wisconsin Breast Cancer Diagnostic data by using the machine learning techniques LGBM, CATBBOOST and XGB. We compared the performance of Naïve Bayes classifier by giving all the features and optimized feature. The increase in the performance of the classifier after feature selection was shown in table 9. The future work is to implement a hybrid feature selection mechamismfor better classification of breast cancer data.

References

[1] BiplobDey, Arun Kumar, "A Review Article on Breast Cancer", *International Journal of Pharmacy & Pharmaceutical Research,* January, Vol.11, No. 2, pp. 284 – 298, 2018.

[2] RozillaJamiliOskouei,NasroallahMoradiKorandSaeidAbbasiMakeki, "Data mining and medical world: breast cancers diagnosis, treatment, prognosis and challenges", *American Journal of Cancer Research,*Vol.7, No.3,pp610-627,2017.

[3] Ajay Kumar, R. Sushil, A.K. Tiwari, "Comparative Study of Classification Techniques for Breast Cancer Diagnosis", *International Journal of Computer Sciences and Engineering*, Vol.7, No.1, pp.234-240, 2019.

[4] AnimeshHazra, Subrata Kumar Mandal, AmitGupta, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms*", International journal of Computer Applications,*Vol.145, No.2, pp.39-46, 2016.

[5] Ganesh N. Sharma, Rahul Dave, JyotsanaSanadya, Piush Sharma and K. K Sharma,"Various Types And Management Of Breast Cancer: An Overview", *Journal of Pharmaceutical technology and research,* Vol. 1, No. 2, pp.109-126, 2010.

[6] Puneetyadav, rajatVarshney, Vishan Kumar Gupta,"Diagnosis of Breast Cancer using Decision Tree Models and SVM", *International Journal of Engineering and Technology*, Vol. 5, No.03, pp.2845-2848, 2018.

[7] PouriaKaviani, Mrs. SunitaDhotre, "Short Survey on Naïve Bayes Algorithm", *International journal of Advanced Engineering and Research Development*, Vol.4, No.11, pp. 607-611, 2017.

[8] MedisettyHari Krishna, Dr. Kunjam. NageswaraRao, "Prediction of Breast Cancer Using Machine Learning Techniques", *International Journal of Management, Technology And Engineering*, Vol. 8, No. 6, pp. 5261-5269, 2018.

[9] BazilaBanuA, P.subramanian, "Comparison of Bayes Classifiers for Breast Cancer

Classification", *Asian Pacific Journal of Cancer Prevention*, Vol.19, No.10, pp. 2917-2920,2018.

[10] Ch.Shravya,Pravalika, ShaikSubhani,"Prediction of Breast Cancer Using Supervised Machine Learning Techniques", *International Journal of Innovative Technology and Exploring Engineering*, Vol.8, No. 6, pp.1106-1110, 2019.

[11] DestaMulatu, Rupali R. Gangarde, "Survey of Data Mining Techniques for Prediction of Breast Cancer Recurrence", *International Journal of Computer Science and Information Technologies*, Vol. 8, No. 6, pp. 599-60l, 2017.

[12] Ms. Sindhuja C., Ms. Kavitha , "Empirical Study on Data Mining Techniques For Breast Cancer Diagnosis", *Journal of Emerging Technologies and Innovative Research*, , Vol. 6, No. 9, pp. 50-55, 2019.

[13] Subrata Kumar Mandal, "Performance Analysis Of Data Mining Algorithms for Breast Cancer Cell Detection Using Naive Bayes, Logistic Regression and Decision Tree", *International Journal Of Engineering and Computer Science*, Vol. 6, No. 2, pp.20388-20391, 2017.

[14] Akhilesh Kumar Shrivas, Ankur Singh, "Classification of Breast Cancer using Data Mining Techniques", *International Iournal of Engineering Science Invention*, Vol.5, No.12, pp. 62-65, 2016.

[15] Ajay Kumar, R. Sushil, A.K. Tiwari, "Comparative Study of Classification Techniques for Breast Cancer Diagnosis", *International Journal of Computer Sciences and Engineering*, Vol.7, No. 1, pp. 234-240, 2019.

[16] NehaKumari and KhushbooVerma, "A Survey On Various Machine Learning Approaches Used for Breast Cancer Detection", *International Journal of Advanced Research in Computer Science*, Vol.10, No.3, pp.76-79, 2019.

[17] Ms. ShwetaSrivastava, Ms. Nikita Joshi, Ms. Madhvi Gaur, "A Review Paper on Feature Selection Methodologies and Their Applications", *International Journal of Engineering Research and Development*, Vol.7, No. 6, pp.57-61, 2013.

[18] K Sutha, Dr.J. JebamalarTamilselvi, "A Review of Feature Selection Algorithms for Data Mining Techniques", *International Journal of Computer Science and Engineering*, Vol.7 No.6, pp. 63-67, 2015.

[19] B. Senthil Kumar, R. Gunavath, "An enhanced model for diabetes prediction using improved firefly feature selection and hybrid random forest algorithm", *International journal of engineering and advanced technology*, Vol.9, No.1, pp. 3765-3769, 2019.

[20] Ms. ManjiriMahadevMastoli, Dr. Urmila R. Pol, Rahul D. Patil,"Machine Learning Classification Algorithms for Predictive Analysis in Healthcare", *International Research Journal of Engineering and Technology*, Vol.6, No.12, pp. 1225-1229, 2019.