Hybrid Feature Selection and Bayesian Optimization with Machine Learning for Breast Cancer Prediction

Yash Mate
Computer Department, Vivekanand Education Society's Institute Of Technology
Mumbai, Maharashtra
2017.yash.mate@ves.ac.in

Neelam Somai
Computer Department, Vivekanand Education Society's Institute Of Technology
Mumbai, Maharashtra
2017.neelam.somai@ves.ac.in

2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) | 978-1-6654-0521-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICACCS51430.2021.9441914

***Abstract*—Breast Cancer is one of the most ubiquitous type of cancer among women and rarely found in men. According to the World Health Organization (WHO) ,Cancer is defined as an uncontrollable abnormal growth of cells in any organ or tissue of the body. Neoplasm or Malignant tumor are common words that describe cancer. According to a survey conducted by the World Health Organization (WHO), cancer accounts for deaths of approximately 9.6 million people globally, and is the second most prevalent cause of death in humans. Cancer is responsible for one out of every six deaths across the world.The good thing is cancer if diagnosed at an early stage, is likely to be treated successfully resulting in high survival rate and low treatment cost. However, it is quite difficult to diagnose it at an early-stage. Therefore, there is a necessity for an efficient cancer prediction model that can predict breast cancer at an early stage. The proposed model highlights the finest set of features required for the detection of breast cancer. It uses Bayesian optimization technique along with hyper parameter tuning and feature selection techniques to decrease the number of parameters by almost 40% while maintaining high accuracy. The best accuracy of 96.2% is obtained with Extra tree classifier algorithm by using feature selection technique along with bayesian optimization and hyperparameter tuning.**

Keywords*—Breast Cancer, Machine Learning, Feature Selection, Hyperparameter Tuning, WHO, Gradient Boosting, Ensemble Learning, Bayesian Optimization

I. INTRODUCTION

According to the World Health Statistics-2020 report[1] Noncommunicable disease mortality rate is 41 million people all over the globe, that is equivalent to 71% of deaths. Among this , 9 million deaths are due to cancer. The prime cause of premature deaths in developed countries is cancer. Among all the types of cancers, breast cancer is the typical type of cancer detected among women and accounts for a huge number of deaths worldwide. [2] In the United States, breast cancer is the second most common type of cancer detected among females.[3] The study says that death and incidence rates associated with breast cancer among males are highest for men above 80 years of age. [4] In 2017, the United States reported 250,520 breast cancer cases resulting in the death of 42000 women. The survival chances increase when breast cancer is diagnosed at a primary stage. Machine Learning Algorithms along with Optimization and Hyperparameter Tuning techniques can contribute significantly to identify the best features to predict breast cancer accurately resulting in lower treatment cost and higher possibility of survival. This paper gives a brief idea about performances of different algorithms like logistic regression, decision tree classifier, k nearest neighbor, random forest, naive bayes, SVM, linear and quadratic discriminant analysis and boosting algorithms with and without feature selection techniques along with bayesian optimization technique.

II. LITERATURE REVIEW

The paper[5] succinctly describes a comparative study of Machine Learning Classification algorithms namely K Nearest Neighbors, Decision Trees (C4.5), Naive Bayes and Support Vector Machines on Wisconsin Breast Cancer Dataset. The simulation was carried out by the authors on the Weka Data Mining Tool and the highest accuracy of 97.13% was achieved by SVM. In another research [6], Particle Swarm Optimization for Feature Selection in algorithms like Fast Decision Tree Learning, KNN, and Naive Bayes is implemented to consider only the most influential features in predicting Breast Cancer and reduce the training time required. The accuracy of the model with PSO was much greater than without it.The paper[7] compares the accuracies of three different algorithms namely REPTree, naive Bayes and IBK with and without the feature selection algorithm, particle swarm optimization (PSO). Naive Bayes outperforms with and without feature selection The author of paper [8] uses decision trees along with feature selection to build white box models that can distinguish between malignant and benign tumors. Accuracy of 96% in predicting breast cancer is obtained with regression trees and classification algorithms.

978-1-6654-0521-8/21/$31.00 ©2021 IEEE

III.METHODOLOGY

A. CONCEPTUAL DIAGRAM

Fig.1 represents the conceptual diagram of the proposed model. The dataset obtained is Wisconsin Breast Cancer Dataset.The data has been cleaned by removing missing values, normalizing the values of the variables and checking for class imbalance. Oversampling using ADASYN is used to increase the number of instances of the minority class so that they become equal to the number of minority classes. A hybrid feature selection algorithm is implemented by taking a vote of six different feature selection algorithms. Machine Learning Classification Models are applied to the dataset. Bayesian Optimization is used to tune the hyperparameters to further enhance the performance. To interpret the model and see which features influence the target, Shapely Additive Values (SHAP) is used.

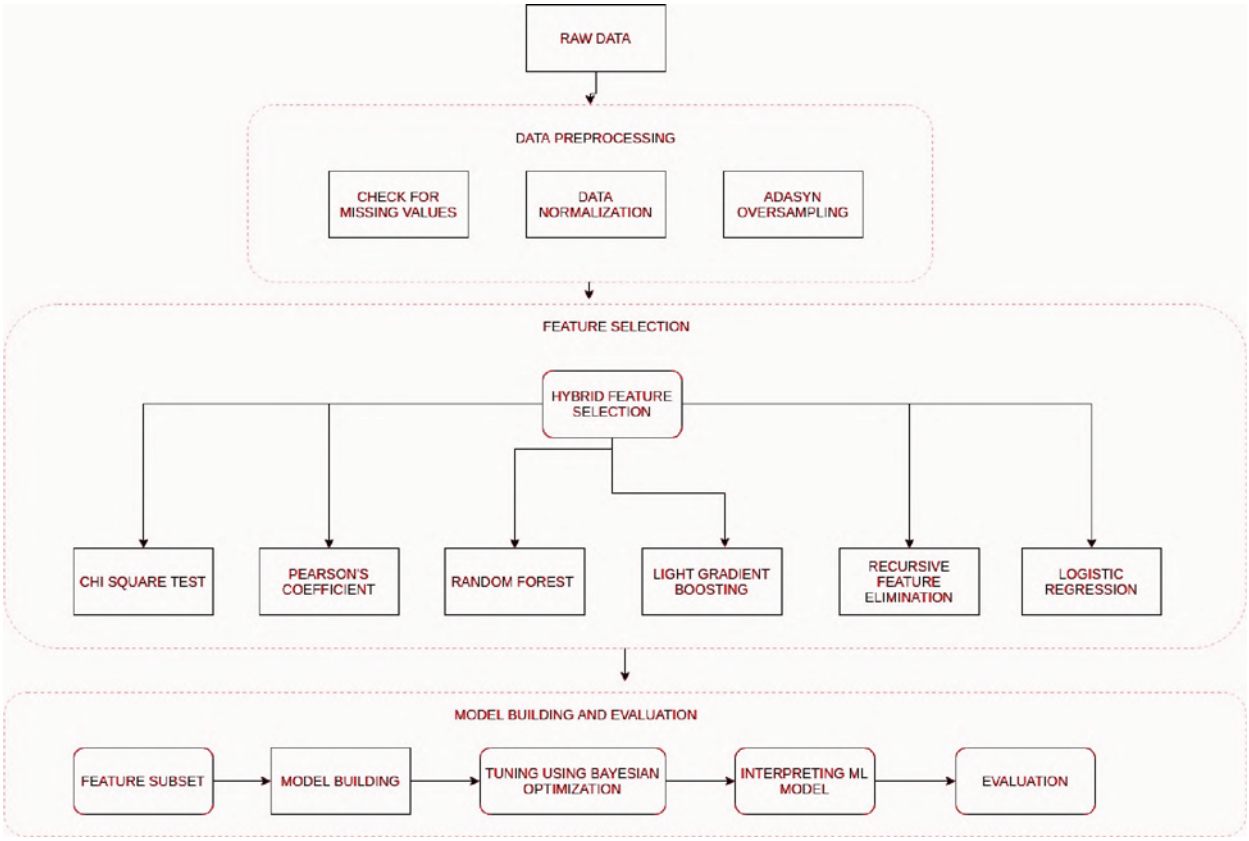

Fig.1 Block Diagram

B. DATA DESCRIPTION

Dataset through which research methodologies are conducted is obtained from the Wisconsin Breast Cancer (WBC) study . The dataset contains 30 different attributes with Class distribution as 357 benign, 212 malignant tuples.The different parameters are calculated from images of fine needle aspirate (FNA) of the mass of the organ. The digitized version of those images give a detailed description of the cell nuclei.

C. PREPROCESSING

Data cleaning is an essential step in order to get rid of the missing values present in the dataset to counter compatibility issues faced during the Machine Learning Model building step . The research highlights and solves the class imbalance problem. Additionally, data normalization techniques are used for the data preprocessing step.715 records were obtained after this. The detailed steps involved in preprocessing of Data with techniques used are given below:

Balanced Dataset- The dataset is unbalanced as it contains 357 benign and 212 malignant tuples. Clearly, One class dominates the other resulting in a model that is highly under fitted. In order to make the data balanced we have used an over sampling technique ADASYN (adaptive-synthetic-sampling-method) which makes use of density distribution, to choose the number of synthetic samples to be generated

for a particular point. We use ADASYN data augmentation technique from imblearn in order to generate synthetic samples from minority class i.e malignant diagnosis.

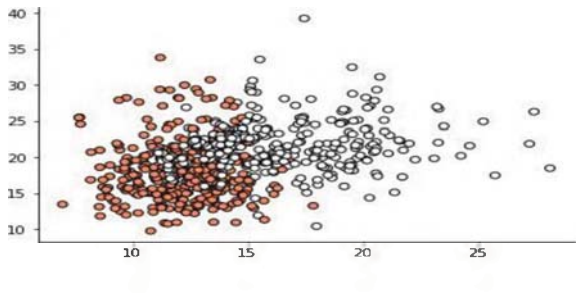

Fig.2 Visual representation of ADASYN

Normalization Using MinMaxScaler - Normalization rescales data from original range to the range of 0 to 1 [0,1], where we have most precision. We have normalized data using scikit learn object MinMaxScaler. The MinMaxScaler First fits the scalar using training data. Here, training data is used in order to determine maximum and minimum values. Scale is applied to the data using the transform function and all the values are scaled in the range of 0 to 1.

Feature Selection - Feature Selection is the mechanism through which the independent features that serve as the biggest determiners of the predictor variable are retained whereas all the features having no effect on the output variable are eliminated. Feature Selection forms an integral part of Machine Learning Models as we can draw inferences and establish a cause-effect relationship between the dependent and independent attributes.The hybrid feature selection approach that we are employing consists of taking a vote from six feature selection methods which are a perfect blend of filter, wrapper and embedded approaches. They are as follows : Pearson's Correlation, Chi Square Test, Recursive Feature Elimination, Light Gradient Boosting Machine (LGBM) Logistic Regression and Random Forest.

1. **Chi Square Test-** Chi Square test is a statistical measure that determines if two variables are dependent or independent of each other. It carries out feature selection by analyzing the relationship between the features and the target. Some important terminologies in Chi Square Test are Chi Square distribution and Degrees of Freedom.
When an attribute is independent of another, then the observed count in the above formula would be approximately close to the expected count, which would yield a smaller Chi Square Value. Higher Chi Square value indicates that the features are dependent on each other.

2. Pearson's Correlation Coefficient - Pearson's Coefficient is the mathematical metric through which we measure the magnitude of association between two attributes. The value of Pearson's coefficient can range from -1 to 1. A value close to one indicates a positive correlation between the two attributes. A positive correlation means that an increase in value of one variable causes a subsequent increase in the value of another variable. On the other hand a value close to -1 indicates a strong negative correlation. In this, as the value of one variable increases the value of the other decreases and vice-versa. If there is absolutely no linear relationship between the two variables (i.e they are independent) then the Pearson's correlation coefficient is zero.

3. Random Forest for Feature Selection - Feature Selection using Random Forest falls in the category of Embedded Method. This method is a hybrid mixture of both wrapper and filter based methods. It relies on the fact that some Machine Learning algorithms like Random Forests have inbuilt feature selection mechanisms. A random forest is made up of numerous decision trees. At each node the data is split based on entropy or information gain criterion. Each split/container holds a set of observations which are very similar with each other and very different from the ones in the other container. And hence, the feature importance is derived on the basis of the purity of the container.

4. Light Gradient Boosting - It uses decision trees and gradient boosting for its implementation. LightGBM differs from other gradient boosting models as it's more efficient and takes up less memory in it's usage. The two additional techniques used in LightGBM are Exclusive Feature Bundling (EFB) and Gradient Based One Side Sampling (GOSS). The training data instances that are not trained properly or are less trained relative to the other samples have larger gradients, and according to the GOSS framework will contribute much larger to the information gain. Hence these instances are kept, and some instances with smaller gradients are retained and others are discarded. This ensures that all the data leads to a larger information gain.

5. Recursive Feature Elimination - It is a kind of backward selection of features. It initially starts with building a model on all the features available in the dataset.

It then computes an information score for each and every attribute. The features with the least feature scores are eliminated. With the remaining set of attributes the model is constructed again and the feature importance scores are re-calculated. A hyperparameter called subset size can be adjusted to evaluate all the subsets with that specific size. The subset with optimal size is then used for evaluation.

6. Logistic Regression - The coefficients in the equation of logistic regression are used to get an insight of the relative importance of the features. It requires the training/ input data to be normalized. Feature importance is a technique used to provide a score for every feature. An attribute having greater feature importance is more relevant to the model and contributes the most in predicting the target variable. Lesser important features are discarded thereby reducing the dimensionality. Therefore an analysis of the coefficients in the final equation of logistic regression is used to get the feature importance of variables.

The fig.3 given below represents the output Dataframe of Feature Selection. A value of True indicates that the feature in the row was selected by the feature selection algorithm represented in the column. For example, the attribute texture_worst was selected by all the six algorithms and hence it has a total of 6. The total represents the summation of True values of all the columns in a particular row. The DataFrame is arranged in the descending order of their impact on the target variable.

	Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	LightGBM	Total
1	texture_worst	True	True	True	True	True	True	6
2	texture_mean	True	True	True	True	True	True	6
3	radius_worst	True	True	True	True	True	True	6
4	perimeter_worst	True	True	True	True	True	True	6
5	perimeter_mean	True	True	True	True	True	True	6
6	concavity_mean	True	True	True	True	True	True	6
7	concave points_worst	True	True	True	True	True	True	6
8	concave points_mean	True	True	True	True	True	True	6
9	radius_mean	True	True	True	True	True	False	5
10	concavity_worst	True	True	True	False	True	True	5
11	compactness_worst	True	True	True	False	True	True	5
12	area_worst	True	True	True	False	True	True	5
13	area_se	True	True	True	False	True	True	5
14	smoothness_worst	True	True	True	True	False	False	4
15	compactness_mean	True	True	True	False	False	True	4
16	area_mean	True	True	True	False	False	True	4
17	texture_se	True	True	True	False	False	False	3
18	symmetry_worst	True	True	True	False	False	False	3
19	symmetry_se	True	True	True	False	False	False	3
20	symmetry_mean	True	True	True	False	False	False	3
21	smoothness_se	True	True	True	False	False	False	3
22	smoothness_mean	True	True	True	False	False	False	3
23	radius_se	True	True	True	False	False	False	3
24	perimeter_se	True	True	True	False	False	False	3
25	fractal_dimension_worst	True	True	True	False	False	False	3
26	fractal_dimension_se	True	True	True	False	False	False	3
27	fractal_dimension_mean	True	True	True	False	False	False	3
28	concavity_se	True	True	True	False	False	False	3
29	concave points_se	True	True	True	False	False	False	3
30	compactness_se	True	True	True	False	False	False	3

Fig.3 Feature Selection

Hyperparameter Tuning using Bayesian Optimization- Setting up the right hyperparameters for the Machine Learning Model are paramount in giving accurate results. When it comes to deciding them, relying on experience or intuition isn't enough. We need a concrete method which ensures that we always end with the most optimal results. While some exhaustive search methods like Grid Search and Randomized Search do suffice our purpose to some extent, they aren't the best for practical scenarios as they consume a lot of time and require computational results. This is where Bayesian Optimization comes into the picture. This tuning method concentrates mostly on the regions where there is a larger probability of finding better results. The next set of hyperparameters for the model is set by taking into account the history of the previously tested hyperparameters in the search space. Having the knowledge of the previous setting

reduces the cost to a great extent. Gaussian Process Models work on the assumption that similar inputs give similar outputs. Rather than predicting a single value, Gaussian process models predict a region/ distribution of values based on the previous experiments that have yielded consistent results. We keep a track of the current best set of parameters. In the next iteration, if the new parameters give better results then we replace the current best with the new value. The algorithm builds a surrogate model, which is the probabilistic model of the objective function. It calculates the probability of getting a score given the set of hyperparameters. The surrogate function that we have used in the Tree Parzen Estimator (TPE). TPE is based on Bayes Theorem. It creates a probability distribution for every hyperparameter. The next parameters are selected on the basis of Expected Improvement.

Steps :

1. Select an objective function and build it's surrogate model.
2. Find the best performing hyper-parameters on the surrogate model
3. Test the hyperparameters on a real objective function
4. Make updations in the surrogate model based on the results
5. Repeat Steps 2 to 4.

D. ALGORITHMS

A) Logistic Regression

Logistic Regression assumes that the training data has a predictor variables and hence it falls under the category of Supervised Machine Learning Algorithm Used to Solve Categorial Problems.The output of Logistic regression is classification of data in binary categories. In our problem either the diagnosis result is M indicating Malignant tumor or B indicating Benign tumor. They use sigmoidal functions which give an S-shaped curve and map the value between 0 and 1.

B) K nearest neighbor

KNN is a widely used ML algorithm because of its simplicity. KNN works on the assumption that the similar objects lie in close proximity. K is initialized which is the nearest neighbour to the data point under classification problem. The closest class for every data point is decided using any distance measure technique like euclidean distance or manhattan distance. Deciding proper K is the most important thing in KNN as the efficiency of algorithms greatly depends on it.

C) Naives Bayes

Naive Bayes works on Bayes Theorem of probability to predict classes of unknown data points . It assumes that a particular feature is independent of other features of a class. Naive Bayes perform well for multi-class prediction. It is easy to build models and known to outperform highly great classification models.

D) Decision Tree Classifier

Decision Trees Classifiers are hierarchical structures that contain root nodes, internal nodes and leaf nodes. Leaves are associated with class labels. While, other nodes contain attribute test conditions to separate records. It uses different techniques like gini index , entropy to properly split attributes. It tries to increase the homogeneity of sub-nodes. Decision trees perform better for large datasets. The most important feature of decision trees is to capture decision making knowledge from given datasets.

E) SVM - Linear Kernel

Support Vector Machines (SVM) is a class supervised machine learning algorithm mainly used for classification and regression tasks. Linear Kernel is used for linearly separable data. It is most commonly used when there are lots of features in a dataset just like in our case there are 30 different attributes. Training SVM with a linear kernel is a much faster option.

F) Ridge Classifier

The Ridge Classifier uses the concept of Ridge Regression for its operation. It normalizes the data values between -1 and 1 and eliminates multi-collinearity in the data.It uses L2 Regularization and hence does not result in a sparse model.

G) Random Forest Classifier

Random Forest uses Ensembling Methods where there are a lot of decision trees that operate together. The decision tree with best accurate output is chosen by voting. The hyperparameters in Random Forest classifier are used to enhance prediction accuracy or to make the model much faster. The random forest model avoids overfitting if there are enough trees in the forest. They usually outperform decision trees.

H) Quadratic Discriminant Analysis

Quadratic Discriminant Analysis is another version of Linear Discriminant Analysis that follows non-linear data separation. Linear Discriminant Analysis is used as a classifier as well as for dimension reduction. QDA estimates covariance matrix for all individual class labels. QDA can not be used for dimensional reduction.

I) AdaBoost Classifier

Boosting Algorithms combine numerous lower accuracy models to form high accuracy models. Boosting algorithms usually provide higher accuracies to model. It is an ensemble machine learning approach. Adaptive Boosting(Adaboost) classifier is one of the iterative

ensemble techniques. It iteratively updates the weights in order to obtain better accuracies for unusual observations.

J) Gradient Boosting Classifier

It makes robust and accurate Machine Learning Models It works by optimizing a loss function, weaker learners to make predictions and a model to to add weak learners. It is a greedy algorithm and can overfit models easily.

K) Linear Discriminant Analysis

Linear Discriminant Analysis is a classification technique that separates linearly separable data . It makes assumption that the data is gaussian data and each feature has the same variance. It makes predictions by estimating probabilities of new inputs belonging to each class . It makes use of the Bayesian theorem for probability calculations. It is also used as dimensionality reduction technique.

L) Extra Trees Classifier

It falls into the category of ensemble learning technique. It aggregates the output obtained from multiple decision trees which are not correlated with each other. It is very similar to Random Forest Classifier but differs from it in the way in which these trees are constructed.

M) Extreme Gradient Boosting

Errors from the previously created models are rectified to form new models in XGBoost. All of the errors are aggregated to form the final output. The algorithm that it uses to minimize the loss is called Gradient Descent. XGBoost (extreme gradient boosting) assumes that the predictor class is numeric not categorical. Hence, techniques like one hot encoding are used to convert characters into numeric 0 and 1. XGBoost generally performs better than random forest.

IV. SHAP FOR MODEL INTERPRETABILITY

The main problem in Machine Learning is that as the model complexity increases the ability to understand and interpret the model decreases. ML Models essentially are like a black-box. One challenge is to interpret the features that contribute to the output. Shapely Additive Explanations (SHAP) are a way to increase the interpretability of ML Models.It is generally used to explain tree-based ML Models.

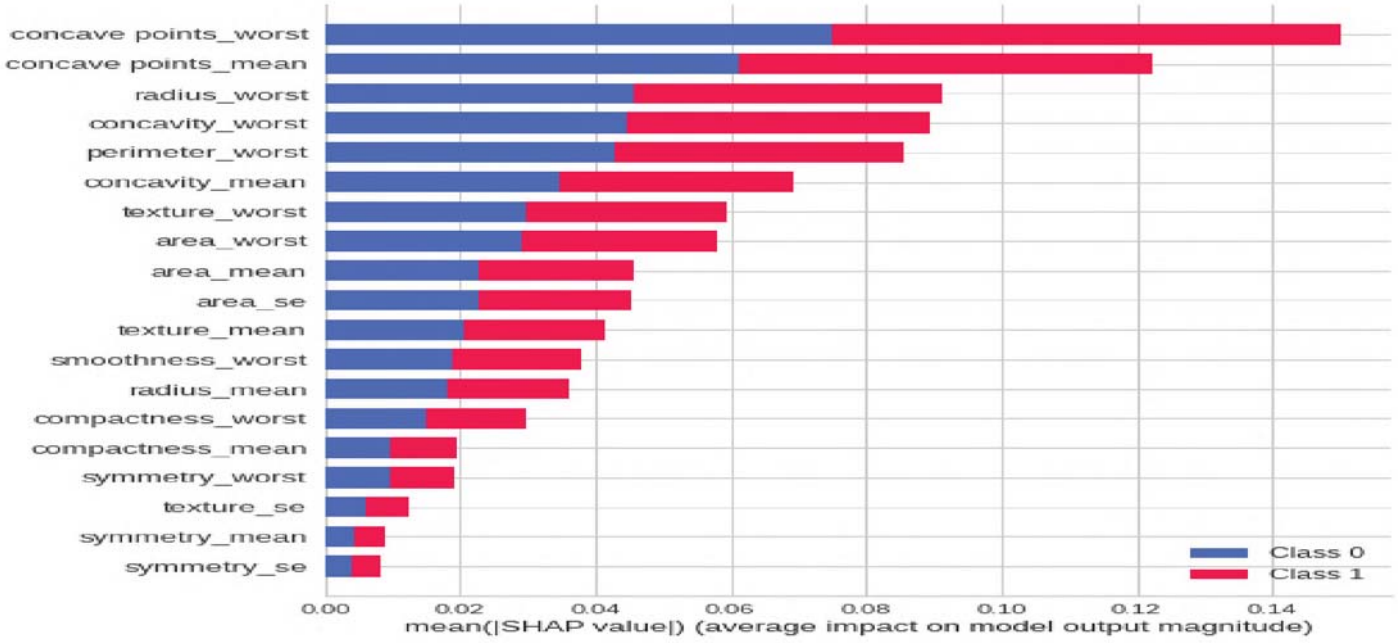

Fig. 4 Model Interpretability

V. RESULTS AND COMPARISON

The Fig.5 below illustrates the comparison of 15 different Machine Learning Algorithms for Binary Classification. It has been observed that tree based ensemble models like Gradient Boosting, Extra Tree Classifier, Light GBM etc perform better than others. The comparison is made between models below without Feature Selection. Adaptive Gradient Boosting Classifier performs the best and gives an accuracy 97%. Other metrics like AUC, F1-Score, recall, and precision are also used for evaluation.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ada	Ada Boost Classifier	0.970	0.9939	0.9723	0.9688	0.9703	0.9400	0.9405	0.155
lightgbm	Light Gradient Boosting Machine	0.968	0.9947	0.9723	0.9649	0.9684	0.9360	0.9364	0.146
catboost	CatBoost Classifier	0.968	0.9965	0.9763	0.9612	0.9685	0.9360	0.9366	8.861
et	Extra Trees Classifier	0.966	0.9970	0.9720	0.9612	0.9663	0.9320	0.9325	0.660
rf	Random Forest Classifier	0.964	0.9936	0.9682	0.9607	0.9641	0.9280	0.9286	0.517
gbc	Gradient Boosting Classifier	0.960	0.9954	0.9603	0.9606	0.9601	0.9200	0.9206	0.556
xgboost	Extreme Gradient Boosting	0.960	0.9960	0.9643	0.9575	0.9605	0.9200	0.9207	2.038
lr	Logistic Regression	0.958	0.9907	0.9523	0.9652	0.9580	0.9160	0.9174	0.312
ridge	Ridge Classifier	0.956	0.0000	0.9365	0.9764	0.9553	0.9120	0.9140	0.020
lda	Linear Discriminant Analysis	0.954	0.9880	0.9365	0.9719	0.9531	0.9080	0.9099	0.062
knn	K Neighbors Classifier	0.948	0.9916	0.9602	0.9398	0.9492	0.8959	0.8975	0.118
svm	SVM - Linear Kernel	0.948	0.0000	0.9403	0.9590	0.9468	0.8960	0.9006	0.018
qda	Quadratic Discriminant Analysis	0.942	0.9842	0.9322	0.9528	0.9411	0.8840	0.8863	0.020
dt	Decision Tree Classifier	0.936	0.9361	0.9445	0.9312	0.9369	0.8720	0.8738	0.024
nb	Naive Bayes	0.900	0.9797	0.8651	0.9323	0.8958	0.8003	0.8046	0.019

Fig. 5 Comparative study without hybrid feature selection

Now, in fig.6 we have made a comparative study of different algorithms after carrying out hybrid feature selection followed by hyperparameter tuning using Bayesian Optimization. Initially there were 30 features. After applying a feature selection algorithm, top 20 most influential features are selected. It means that more than 33% of the redundant data is eliminated. Lesser data also means lesser training time, lesser dimensionality and lesser chance of overfitting. The highest accuracy obtained in this case is 96.2% by Extra Tree Classifier. Although there has been a decrease in accuracy by 0.8%, we would still prefer a simpler model having lesser dimensions and more interpretability than a model that gives only a marginally better accuracy but is difficult to interpret.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.962	0.9963	0.9677	0.9581	0.9620	0.9240	0.9258	0.467
catboost	CatBoost Classifier	0.960	0.9966	0.9637	0.9601	0.9603	0.9200	0.9230	5.365
xgboost	Extreme Gradient Boosting	0.956	0.9954	0.9677	0.9502	0.9571	0.9120	0.9159	0.244
lightgbm	Light Gradient Boosting Machine	0.956	0.9950	0.9677	0.9502	0.9571	0.9120	0.9157	0.095
gbc	Gradient Boosting Classifier	0.954	0.9941	0.9595	0.9543	0.9546	0.9079	0.9126	0.240
ada	Ada Boost Classifier	0.950	0.9874	0.9637	0.9411	0.9509	0.9000	0.9031	0.132
rf	Random Forest Classifier	0.946	0.9938	0.9475	0.9491	0.9460	0.8919	0.8964	0.510
lda	Linear Discriminant Analysis	0.946	0.9862	0.9275	0.9635	0.9442	0.8919	0.8941	0.020
dt	Decision Tree Classifier	0.942	0.9421	0.9558	0.9364	0.9436	0.8840	0.8892	0.022
svm	SVM - Linear Kernel	0.942	0.0000	0.9035	0.9794	0.9382	0.8839	0.8890	0.019
knn	K Neighbors Classifier	0.940	0.9856	0.9595	0.9239	0.9407	0.8800	0.8821	0.122
ridge	Ridge Classifier	0.940	0.0000	0.9033	0.9746	0.9370	0.8799	0.8831	0.017
lr	Logistic Regression	0.938	0.9821	0.9235	0.9528	0.9366	0.8759	0.8785	0.337
qda	Quadratic Discriminant Analysis	0.930	0.9838	0.9077	0.9511	0.9282	0.8600	0.8620	0.019
nb	Naive Bayes	0.904	0.9726	0.8757	0.9310	0.8997	0.8080	0.8138	0.013

Fig. 6 Comparative study with hybrid feature selection

VI. CONCLUSIONS

The paper emphasizes the importance of optimization and a hybrid feature selection technique on traditional and ensemble classification algorithms in breast cancer prediction. The accuracies are improved for most of the machine learning algorithms by using feature selection techniques like pearson's coefficient, chi square test ,logistic regression, random forest, RFE, and light gradient boosting to identify important attributes. The results demonstrated by our experiment on 15 traditional and boosting(ensemble) classification algorithms, along with bayesian optimization techniques are presented in this paper. To conclude , Adaboost Classifier (boosting algorithm) has attained the highest accuracy of 97% without feature selection and Extra Tree Classifier with feature selection has attained an accuracy of 96.2%

by considering 20 important features for breast cancer prediction.

References

[1] World health statistics 2020: monitoring health for the SDGs, sustainable development goals. Geneva: World Health Organization; 2020. Licence: CC BY- NC-SA 3.0 IGO.

[2] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.

[3] Centers for Disease Control and Prevention. Male Breast Cancer Incidence and Mortality, United States—2013–2017. USCS Data Brief, no 19. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2020.

[4] Centers for Disease Control and Prevention. *United States Cancer Statistics Breast Cancer Stat Bite*. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2020.

[5] Asri, Hiba & Mousannif, Hajar & Al Moatassime, Hassan & Noël, Thomas. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science. 83. 1064-1069. 10.1016/j.procs.2016.04.224.

[6] S. B. Sakri, N. B. Abdul Rashid and Z. Muhammad Zain, "Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction," in IEEE Access, vol. 6, pp. 29637-29647, 2018, doi: 10.1109/ACCESS.2018.2843443

[7] Quang, Nguyen & Do, Trang & Wang, Yijing & Heng, Sin & Chen, Kelly & Ang, Wei & Philip, Conceicao & Singh, Misha & Pham, Hung & Nguyen, Binh & Chua, Matthew. (2019). Breast Cancer Prediction using Feature Selection and Ensemble Voting. 250-254. 10.1109/ICSSE.2019.8823106.

[8] Brito-Sarracino, Tamires & Santos, Moisés & Freire Antunes, Eric & Santos, Iury & Kasmanas, Jonas & Carvalho, Andre. (2019). Explanable Machine Learning for Breast Cancer Diagnosis. 681-686. 10.1109/BRACIS.2019.00124.

[9] Supriya, M., Deepa, A.J. A novel approach for breast cancer prediction using an optimized ANN classifier based on a big data environment. *Health Care Manag Sci* 23, 414–426 (2020) https://doi.org/10.1007/s10729-019-09498-w

[10] S. Laghmati, A. Tmiri and B. Cherradi, "Machine Learning based System for Prediction of Breast Cancer Severity," 2019 International Conference on Wireless Networks and Mobile Communications (WINCOM), Fez, Morocco, 2019, pp. 1-5, doi: 10.1109/WINCOM47513.2019.8942575.

[11] Kramer, I & Hooning, Maartje & Mavaddat, Nasim & Hauptmann, M & Keeman, Renske & Steyerberg, EW & Giardiello, D & Antoniou, Antonis & Pharoah, PDP & Canisius, S & Abu-Ful, Z & Andrulis, IL & Anton-Culver, H & Aronson, KJ & Augustinsson, Annelie & Becher, H & Beckmann, MW & Behrens, Sabine & Benítez, Javier. (2020). Breast cancer polygenic risk score and contralateral breast cancer risk. The American Journal of Human Genetics.

[12] Mohammed S.A., Darrab S., Noaman S.A., Saake G. (2020) Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. In: Tan Y., Shi Y., Tuba M. (eds) Data Mining and Big Data. DMBD 2020. Communications in Computer and Information Science, vol 1234.Springer,Singapore.https://doi.org/10.1007/978-981-15-7205-0_10

[13] Lu, Xunxi & Li, Xiaoguang & Ling, Hong & Gong, Yue & Guo, Linwei & He, Min & Sun, Hefen & Hu, Xin. (2020). Nomogram for Predicting Breast Cancer-Specific Mortality of Elderly Women with Breast Cancer. Medical science monitor : international medical journal of experimental and clinical research. 26. e925210. 10.12659/MSM.925210.

[14] Hussain, Omead. (2020). Predicting Breast Cancer Survivability. Cihan University-Erbil Journal of Humanities and Social Sciences. 4. 17-30. 10.24086/cuejhss.v4n1y2020.pp17-30.

[15] Hadi, Gul Anar. (2020). Benign and Malignant Breast Cancer Features Based on Region Characteristics.