Journal of Theoretical and Applied Information Technology
31st August 2022. Vol.100. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN OPTIMIZED EXTREMELY RANDOMIZED TREE MODEL FOR BREAST CANCER CLASSIFICATION

TINA ELIZABETH MATHEW[1]

[1]Assistant Professor, Government College Kariavattom, Department of Computer Science, Thiruvananthapuram, Kerala, India

Email: [1]tinamathew04@gmail.com

ABSTRACT

Breast Cancer is a non-communicable disease seen primarily in women population. As per the statistics published by the World Health Organization, it is presently ranked, globally, as number one in incidence. It can principally affect women of any age, and can be diagnosed in any of the five stages of the disease, but chances of cancer survival become more difficult when diagnosis is made in advanced stages of the disease. Mortality rate of cancer is seen to be high in developing countries than in developed countries. Owing to this fact breast cancer prediction, diagnostic and therapeutic facilities need to be urgently improved in this extent. Henceforth, development of clinical decision support systems for early and precise detection of the disease gains significance and is the need of the hour. The study aims in building a model for precise classification of breast tumors with minimum misclassification of labels. In this paper the potential of extra tree classifiers for breast cancer classification into malignant or benign tumors is examined. A model for breast cancer classification is proposed using extremely randomized tree classifier. Hyperparameter optimization is applied. Identification of important features aids in model performance. Features relevant to disease detection are identified and ranked by importance using 3 techniques- impurity based, permutation based and Shap values. The most important four features identified are Size Uniformity, Shape Uniformity, Bare Nucleoli and Normal Nucleoli. Performance of the optimized model is analyzed using training-testing partitions and k fold stratified cross validation with k as 5 and it was observed that they produced an accuracy of 99.27% on the test set and 97.3 % on the cross validated model respectively. The study reveals the suitability of the extra tree classifier for breast cancer classification. The model is compared with other state of art models and it was seen to be superior in performance. Furthermore, extremely randomized tree classifiers are perceived to be suitable in developing models for breast cancer classification with minimal misclassification of instances.

Keywords: *Breast Cancer, Classification, Extremely Randomized Tree Classifier, Feature Importance, Machine Learning*

1. INTRODUCTION

The statistics on cancer published by the World Health Organization delineates the top priority that should be ascertained in cancer eradication, which is yet to be attained. A predominant cancer affecting women, globally, being Breast Cancer requires urgent attention. Presently, it has moved to top position among all cancers as the one mostly affecting women [1]. It is also a life threatening and second leading cause of deaths in females [2]. Early detection of the disease provides a higher likelihood for survival and the patients would have a better chance of recovery. Hence, significant focus is to be provided for strategies involving early detection of the disease. Breast Cancer is the uncontrolled growth of cells in the ducts, lobules or connective tissues of the breast. Primarily seen in the ducts or lobules it manifests as clumps of cells that grow uncontrollably and are denoted as tumors. Tumors can be malignant, (harmful) or benign, (not harmful). The principal concern is to identify malignant tumors precisely. The tumors are graded by stages that are determined by the characteristics of the cancer, size of the cancer, involvement of lymph nodes, tumour grade, involvement of Her2 protein, oestrogen- progesterone receptor status and metastasis. The stages are graded based on the American TNM staging and are indicated by a number on a scale of 0 to 4 ranging from stage 0 as non- invasive cancer to stage 4, invasive cancers [3]. The more advanced the stage is, less are the chances of survival. In countries like India where breast cancer incidence is high and, on the rise, almost

more than 50% of the women detected with breast cancer are perceived to suffer from stage 3 and 4 of breast cancer. Breast cancer is seen to be high in Indian urban women in cities, albeit cases in rural areas are also on the rise. **T**he survival rates of breast cancer in India are low owing to late detection [4]. Hence early **i**dentification of the disease is a vital factor for better survival. Cancer is usually identified by screening which is done using digital mammography, digital breast tomosynthesis, breast ultrasonography, magnetic resonance imaging, clinical breast examination and thermography [5]. Most of these techniques identify cancers but can occasionally be inconclusive or indecisive besides being painful, stressful and involving harmful radiation, which itself can be a cause for cancer. Fear of undergoing these procedures are reasons for women to shy away from breast cancer screening techniques. Hence to reduce these hitches, along with the available medical modalities, clinical decision systems can be utilized.in the detection and identification process. A major framework for clinical decision systems is provided by Artificial Intelligence (AI) technologies. Subdomains of AI which have recently gained popularity and interest is Machine learning (ML) and Data Mining (DM) [6]. Both of these areas have interlapping sections and coexist by utilizing their techniques. Machine learning consists of several types of learning techniques. Three broad categories of learning techniques are Supervised, Unsupervised and Reinforced Learning techniques. A major role is played by supervised learners in development of clinical decision systems. Machine learning techniques and data visualization techniques provide significant benefits and have much impact in the decision-making process of cancer detection [7], [8]. ML is seen to be successful in modelling systems in several domains [9]. Literature shows a wide range of classifiers being used for risk factor prediction, disease identification and prediction, diagnosis and classification, prediction of recurrence, survival prediction, categorization of various types of cancers and many more [10]. The major groupings of classifiers used are, Support Vector Machines [11], Logistic Regression [12], Artificial Neural Networks, Bayesian Techniques, Decision Trees [13], Lazy learners such as *k*-NN [14] and so on. Each of these types of classifiers comprise of further numerous subcategories. Each of these classifiers are seen to be suitable for numerous problems with vivid range of application [15]. The study done in this paper involves a decision tree classifier denoted as Extremely Randomized Tree Classifier for detection and classification of breast tumours into either benign or malignant tumours.

A major problem identified with various breast classification models is the misclassification error, leading to imprecise outputs. The performance of the models needs to be enhanced. This can be envisaged by identifying the chief features that have significant influence on the target variable, thus serving to improve performance and also by finetuning hyperparameters with optimal values by implementing simpler techniques such as cross validation.

The major contributions of this study are

- A model for breast cancer tumour classification into either malignant or benign based on Extra tree Classifier using feature importance modelling with significantly increased accuracy and improved hyperparameters.
- Identifying important features relevant for breast cancer identification,
- Reduction of misclassification of class labels and thus achieving classification reliability.

The rest of the paper is organized as follows. Section 2 handles the relevant work involved in this area. Section 3 comprises of Materials and Methods used. Section 4 discusses the results and finally section 5 concludes the paper.

2. LITERATURE SURVEY

Machine learning is a data analysis technique which enables an intelligent system to identify the precise output with different algorithms. Decision tree, k-nearest neighbours, SVM, and neural networks are the most common algorithms for machine learning applications [**16**]. While there is no better way to diagnose breast cancer, early diagnosis is considered as the primary step of treatment. With the evolution of machine learning techniques and availability of data, numerous techniques have been explored and exploited and research associated with this area is outlined in brief as follows.

[17], used whale optimization for feature selection with various classifiers such as SVM, LR, Extra Trees, k-NN, Naïve Bayes and found that Extra Trees performed the best in terms of accuracy producing 99% accuracy. [18], conducted a

comprehensive comparative study on performance of machine learning algorithms in breast cancer prediction in terms of performance, effectiveness, and efficiency on big data. They used Spark and WEKA platforms. The performance of the classifiers SVM, RF and DTs were compared on DNA Methylation (DM), Gene Expression (GE), and mixed combined datasets. The results illustrated that SVM outperformed DT and RF on all the datasets with 99.68%, 98.73%, and 97.33% accuracy respectively. [19], applied feature selection to many machine learning classifiers such as Support Vector Machines, Logistic Regression, Adaboost, Random Forest, Extra Trees and Stacking models. They concluded that applying the classifiers on a feature subset helped models to gain an accuracy above 90%. [21], proposed an extra tree classifier model for breast cancer classification and applied features selection. They used genetic programming technique to select the best features and perfect parameter values of various machine learning classifiers. Extra trees were seen to better performance in breast cancer classification with this technique. Even though metaheuristic techniques such as genetic programming, cuckoo search optimization, whale optimization and so on helps in speeding up the decision-making process they are still prone to errors. [21], used Random Forest (RF) and Extremely Randomized Trees (ET) algorithms to classify breast cancer and were able to obtain high diagnostic performance. They identified that varying the number of trees in the classifiers impacted the performance of extra trees. [22], demonstrated comprehensive reviews on five classifiers- Support Vector machines, Random Forest, k Nearest Neighbours, Artificial Neural Networks and Logistic regression. The performance of the classifiers was measured using various metrics of accuracy, Specificity, recall, Precision and so on. The highest accuracy was recorded by Artificial Neural Networks with an accuracy of 98.57%. They concluded that machine learning classifiers can assume the role of clinical assistants and aid medical practitioners in breast cancer diagnosis. [23] proposed an improvised Random Forest Model in combination with cost sensitive learning. Random Forest Trees are Decision Trees similar to Extra Trees and the model displayed an accuracy of 97.5%. [24], constructed learning curves for five classifiers -AdaBoost, Gradient Boost, extra trees, bagging, and random forest classifiers to find the best fit models. The performances of models were evaluated using accuracy on 10-fold cross validation (CV), and leave one out cross validation (LOOCV). The Extra trees classifier was seen to outperform the other four ensemble classifiers with an accuracy of 96.3% and 95.5% with LOOCV and CV, respectively. Parameter optimization was not considered in this study,

Besides being a classifier on its own, extra trees can be utilized as feature selection techniques itself to select the most important features of a dataset [25], [26]. The feature selection using this classifier is seen to be superior and helps in enhancing classifier performance. [27] in their work used metaheuristics-based feature selection methods and employed extra-tree classifier to classify emails into spam and not spam. The proposed model has accuracy of 95.5%, specificity of 93.7%, and F_1-score of 96.3%, which improved the classification when compared to other strategies in the field. [28], in their proposed work compared different classification such as an SVM, KNN, NB, RF, DT, ET and AdaBoost algorithms based on various performance measures such as accuracy, execution time, kappa statistic error, mean absolute error, mean squared error, root mean squared error, true positive, true negative, false positive and false negative. Based on these performance measures, among all the classifiers Extra trees was seen to illustrate superior performance. Parameter optimization was not considered during classification.

Optimization of parameters is a vital factor for better performance of classifiers several techniques utilizing metaheuristic techniques can be seen used while exploring literature. Even though they are good in decision making and help in better results they are seen prone to errors and in terms of quality give inadequate solutions and are prone to be trapped in the local minima or maxima as per the problem and they also involve several parameters that need optimization. With increased optimization misclassification of labels can be avoided. Hence other techniques can be explored for parameter optimization. To improve hyperparameters in this study cross validation scores with k fold validation is implemented.

3. MATERIALS AND METHODS

3.1. Dataset Used

The Wisconsin Breast Cancer databases which is publicly available in the UCI repository **is** used in this study. It comprises of 699 instances of benign and malignant cancer. 11 features are available and the target variable is denoted as class [29]. The dataset was collected and produced in the present form by Dr William H Wolberg of the University of Wisconsin Hospitals and used in his study with statistical and machine learning techniques [30].

The first attribute is Id number which is omitted in the study since it has no significant role in classification. The rest 9 attributes values range between 1-10. Sixteen instances of the dataset have missing values which are discarded during data pre-processing. Hence there are 444 benign and 239 malignant instances in the used dataset.

A comparison of the attributes of the dataset is displayed in Table 1

Table 1 Attribute List

Attribute Name	**Values**	**Comparison of malignant and benign cells**	
		Malignant	Benign
Id number	Numeric	-	-
Feature 0-CT Clump thickness	1-10	Seen in Multilayers	Seen in monolayers
Feature 1-SU Size uniformity	1-10	Size differs	Uniform size
Feature 2-ShU Shape uniformity	1-10	Shape differs	Uniform Shape
Feature 3-MA Marginal adhesion	1-10	Cells do not stick together	Cells stick together
Feature 4-ES Epithelial size	1-10	Enlarged	Small
Feature 5-BN Bare nucleoli	1-10	Have bare Nucleoli	No Bare Nucleoli
Feature 6-BC Bland chromatin	1-10	Coarse in texture	Uniform texture
Feature 7-NN Normal nucleoli	1-10	Nucleus is bigger	Nucleus is small
Feature 8-MT Mitosis	1-10	More Mitosis	Not so
Class	2-Benign 4-Malignant		

3.2. **Extremely Randomized Trees**

Extremely Randomized Trees Classifier or Extra Trees Classifier in short belongs to the category of ensemble decision tree learning techniques. The Extra-Trees classifier produces a group of unpruned decision trees. It constructs trees by randomizing both attribute and cut-point selection strongly while splitting a node of a tree. Extra trees function by aggregating the results of the multiple de-correlated decision trees collected together as a forest and output its classification result obtained by applying the majority voting technique, Conceptually, it is very similar to the Random Forest Classifier, being itself a bagging decision tree ensemble, but differs in the manner in which it constructs the decision trees in the forest. Extra trees contrast with Random Forest on two counts. The original training sample is used for constructing each Decision Tree in the Extra Trees Forest. For every test node, k features from the feature set are selected randomly and each tree has to select the best feature to split. the data which is typically based on a mathematical construct such as the Gini Index, Entropy or Information Gain criteria. Random sampling of features in this

manner, creates multiple de-correlated decision trees.

In terms of computational cost and execution time, the Extra Trees classifier is seen to be much faster [30]. This algorithm saves time as it randomly chooses the split point instead of calculating the optimal one. Since all original training samples are used instead of bootstrap replicas bias is also reduced. Besides, variance is also observed to be reduced. A major strength of the classifier is its computational efficiency. On examining literature, it is seen to have extensive and diverse applications such as in intrusion detection systems [31], land cover classification [32], disease classification and many more.

The extra trees algorithm outputs ensembles of decision trees, which are state-of-the-art for many supervised Machine Learning tasks [33]. The forte of extra trees classifiers is that they can be trained on data having continuous attribute values without the need of sorting it.

The extra tree classifier uses three main hyperparameters in tuning the algorithm; - the number of decision trees in the ensemble denoted usually as M, the number of input features to randomly select and consider for each split point, represented as k, and the minimum number of samples required in a node to create a new split point, denoted as n_min.

3.3. Proposed Model

In the proposed model an extra tree classifier is used to build the model. The Wisconsin dataset with 683 instances is applied on the model. The performance of the model is analysed on four training-testing partitions 80-20, 70-30, 60-40, 50-50 partitions and on a 5-fold cross validation set using accuracy, Precision, F1- score and confusion matrix. The feature set comprises of 9 features and the influence of each of the features is assessed. The features considered important for the accurate classification of the disease are identified using three techniques the impurity-based criteria – gini, permutation-based importance and it is counterchecked with Shap summary plots. Selection of parameter values is a tricky process and applying appropriate values help in enhancing the model performance. The best parameters needed are chosen using k- fold stratified cross validation with value of k as 10

The accuracy scores of three major hyperparameters- number of trees in the forest, minimum number of features for a split and minimum number of samples for a split are used as examination criteria. The values are chosen randomly and these grids of values are analysed against cross validation scores using 10-fold based on accuracy score and the outputs are depicted using box plots.

The workflow is depicted in Figure 1. The model was developed using Python 3.97 in the Spyder IDE environment.
The criteria used to measure the performance is Accuracy, Precision, Recall, and F1 Score.

Accuracy is a measure that provides the percentage of correct predictions made from the total number of examples.

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP/(TP + FP) \quad (2)$$

$$\text{Sensitivity} = TP/(TP + FN) \quad (3)$$

$$\text{F1 Score} = 2 * (Precision * Recall)/(Precision + Recall) \quad (4)$$

Studies show that F1 score is a good measure for imbalanced datasets.

Precision gives the percentage of truly positives and recall also known as True Positive Rate, TPR is the percentage of predicted positives.

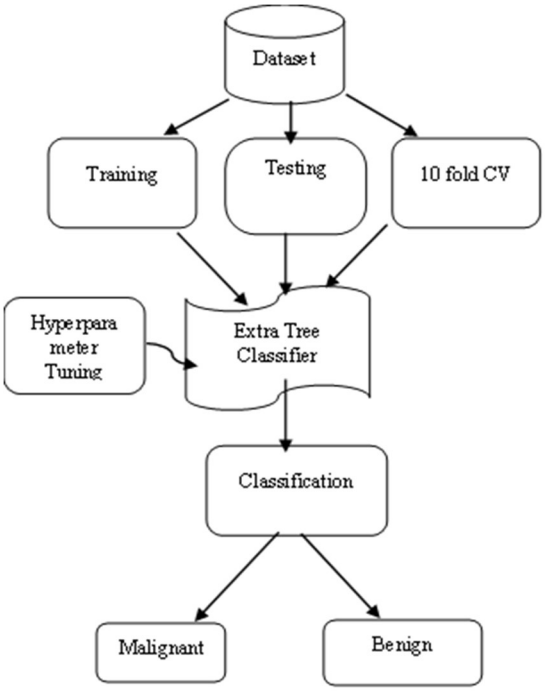


Figure. 1 Proposed Workflow

4. RESULTS AND DISCUSSIONS

The proposed model selects importance of features based on Gini Importance and classification is done based on it. Table 1 displays the results of the classification. Training- Testing splits are varied and four partitions are used- 80-20, 70-30, 60-40, 50-50 on the test sets. Besides using partitioning, model training is performed using stratified 5-fold cross validation. The stochastic nature of the algorithm causes the output values to vary on each run. Hence, a few runs are performed and the average values are taken into account. The best hyperparameters for the model are used and these are obtained by using stratified 10-fold cross validation scores based on accuracy. Feature importance and their effect on the model performance is computed based on Impurity based feature importance-Gini Importance. But since this method can be misleading with high cardinality data the results are counterchecked with feature importance generated by permutation-based metrics. Both techniques reveal that the calculated feature importances have similarities. At the end SHAP summary is used to verify the consistency of the results of feature importance obtained. The precision, recall and F1 score and accuracy values of each test set is illustrated and the best value of 99.2 was attained by the 80-20 partition. With the 5 -fold cv 98.53% accuracy was obtained. Cross validation is adopted so that overfitting issues do not occur. The results of parameter optimization have overall improved the performance of the model and is seen to be superior when compared with literature.

Table 2 Performance Metrics

Partition	*Precision*	*Recall*	*F1 Score*	*Accuracy*
80-20.	*0. 99*	*0.99*	*0.99*	*99.2*
70-30	*0 .97*	*0.98*	*0.975*	*98*
60-40	*0 .98*	*0.98*	*0.98*	*98*
50-50	*0 .99*	*0.99*	*0.99*	*99*
5-fold cv	*0.99*	*0.983*	*0.985*	*98.53*

The confusion Matrix of the test set of the 80-20 split is displayed in Figure 2. The True Negatives are 90 and True Positives are 46 instances. 1 misclassification of the positive class occurred. No false positives occurred. It can be inferred that Extra Tree classifiers produce less misclassifications compared to other Decision Trees where misclassification is seen high [13]

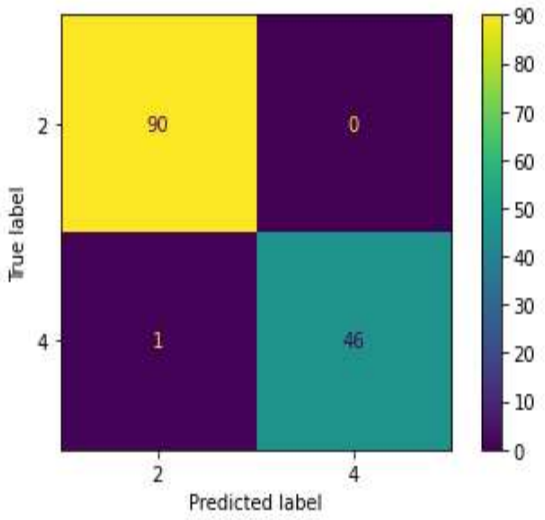


Figure. 2 Confusion Matrix Of 80-20 split

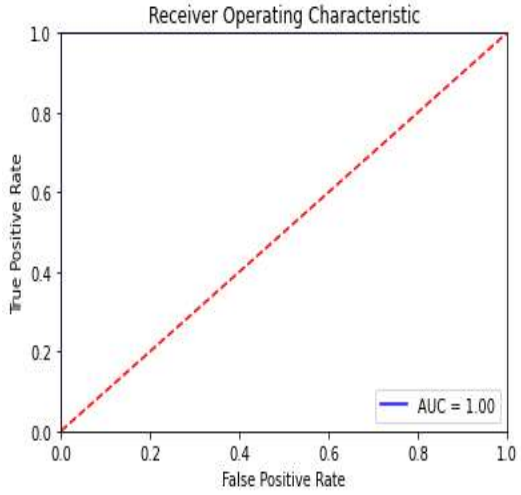


Figure 3 ROC 80-20 Split

The ROC plot is shown in Figure 3. The ROC Area Under the Curve of the model is 1.00. It helps in determining the capability of a model in distinguishing the classes. Higher the value of ROC AUC, better is the performance of the classification model.

The importance of the 9 features of the dataset to the outcome is illustrated using a bar diagram as in Figure 4.

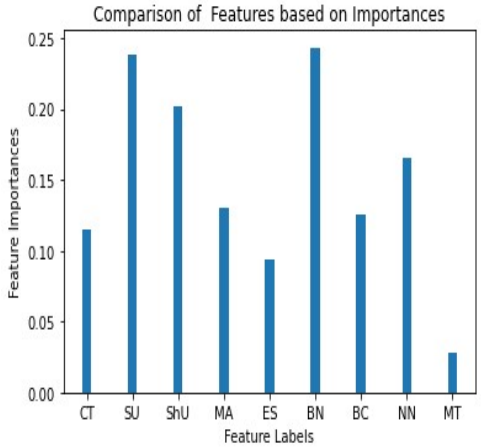


Figure 4 Features vs Importance

The best hyperparameters for the model are obtained by using stratified 10-fold cross validation scores based on accuracy. Three main hyperparameters are evaluated – number of trees in the forest, number of features to be used and number of features to be considered per split.

Box and whisker plots are created to display the distribution of accuracy scores for each of the configured maximum number of trees in the forest. The number of trees is varied from 10 to 500 and the effect is evaluated. The best value of 0.986 was seen between 10 and 50 trees. The box plot in Figure 5 depicts the distribution of accuracy against number of trees.

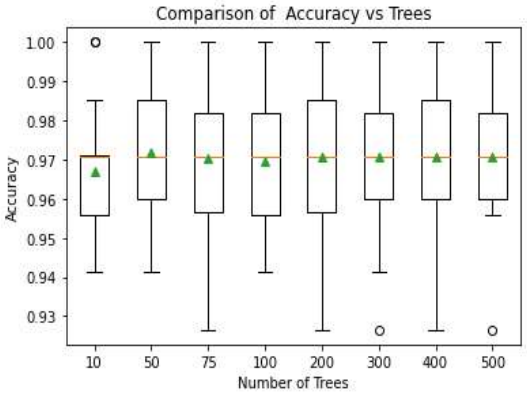


Figure. 5 Accuracy vs Number Of Trees In Forest

The number of features that is randomly sampled for each split point is another hyperparameter to be tuned. The box and whisker plot in Figure 6 displays the distribution of accuracy scores for each feature set size. For all the sets the median is seen to be varying around 0.97, with best value of 0.976 around 3 to 5 features. Number of features does not seem to affect accuracy much. This could be because the classifier is an ensemble of trees,

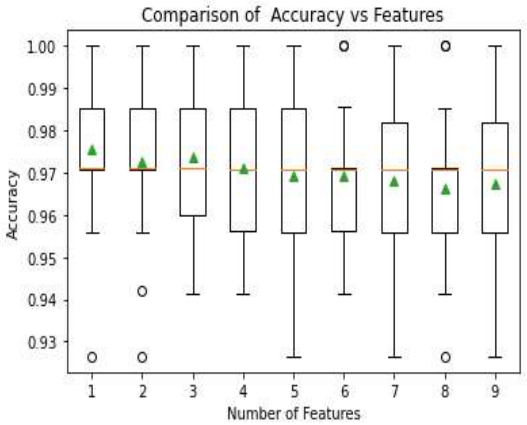


Figure. 6 Accuracy vs No. Of Features

The next hyperparameter evaluated is the number of samples in a node of the decision tree before adding a split represented by the min_samples_split parameter. New splits get added to the decision tree only if the number of samples is equal to or exceeds this forementioned value The box and whisker plot created displays the distribution of accuracy scores for each of the configured maximum tree depth. Figure 7 illustrates the box plot. Accuracy is best for splits with 8, 9 and 14. Keeping the value lower will be help to improve performance as more deeper trees will be generated.

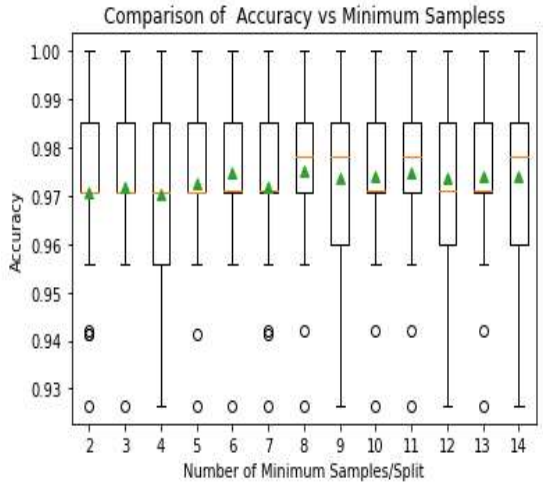


Figure 7 Accuracy vs Minimum SamplesTtaken

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature. Impurity based feature importance uses Gini importance to depict feature importance.

The box plot of Impurity based feature importances is depicted in Figure 8. The highest value decrease in node impurity is shown by feature 5- Bare Nucleoli, followed by Size Unifirmity, Shape Uniformity, Normal Nucleoli, bland chromatin, clump thickness, marginal adhesion Epithelial size and mitoses. The least important feature is Mitoses.

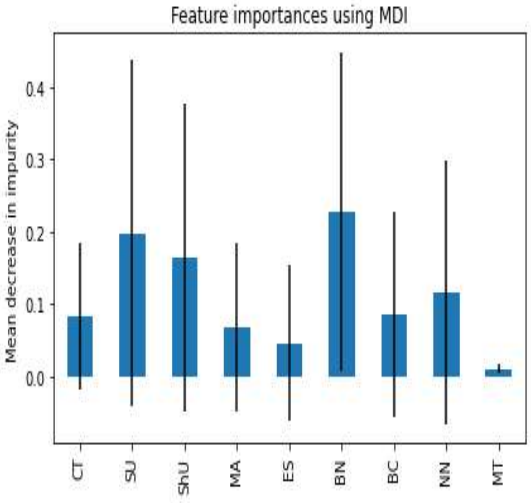
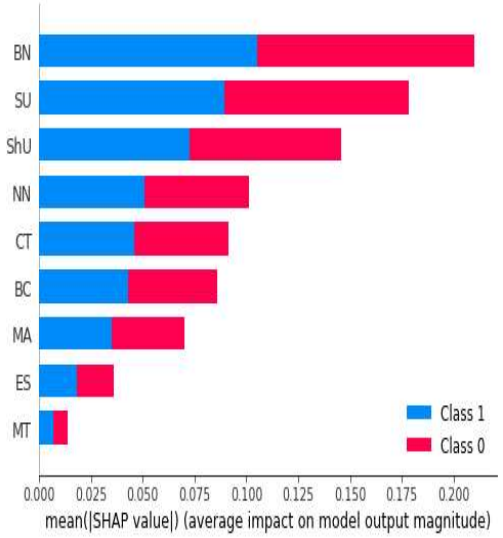


Figure 8 Feature Importance Using MDI

An issue with Impurity based feature importances is that for high cardinality features,



specifically features with many unique values, it can be misleading. Hence to resolve this issue permutation based importance is taken into account. Figure 9 depicts the box plot of the Feature permutation based importances.

Permutation feature importance overcomes the limitation of the impurity-based feature importance, they do not have a bias toward high-cardinality features. Here it computes the feature importance on permuted out-of-bag samples based on the mean decrease in accuracy. It utilizes the model and validation or test data for computation. Each feature is randomly shuffled and the change in the model's performance is computed. The features

which impact performance the most are the most important ones. Here the order of importance is Bare Nucleoli, Shape uniformity, Size Uniformity, Clump Thickness, marginal Adhesion, Normal Nucleoli, the relative importance vary slightly with MDI and the least significant feature is omitted in the plot.

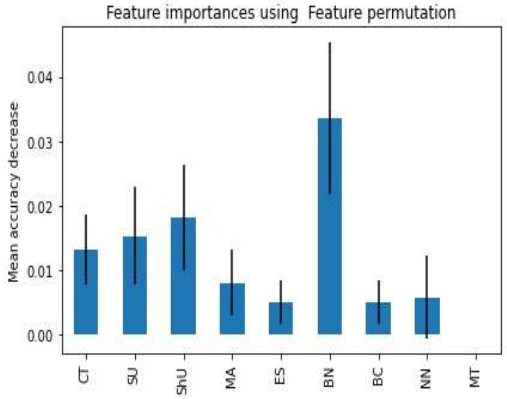


Figure 9 Feature ImportanceUusing Permutation

Figure 10 displays the mean Shap summary plot. Shapley value or Shapley Additive explanation [34] is the average marginal contribution of a predictor considering all the possible combinations. It provides explanation about the output of the machine learning model. It measures the impact of features by taking into consideration its interaction with the other features. From SHAP it can be comprehended how much each feature in a model contributes to the prediction. Mean Shap values help to aggregate and understand how the models makes the predictions as well as, identify and visualize

Figure 10 Shap Values Showing Impact Of Features On Model Output

important relations between the features and outputs. Consequently, it helps to comprehend the working of the model. The y axis gives the variable names and x axis the shap values. The red and blue colours show the impact of the corresponding features on each class. The positive shap values of the features indicated positive impact on the output. The Shap values are consistent with that of the MDI feature importance. Each feature is ordered according to importance.

Hyper Parameter optimization and identification of the best features have helped in better performance of the model. Using cross validation ensures that overfitting does not occur in the model. Better feature selection and hyper optimization thus improved the classification process and misclassifications were less. Most prior works utilize metaheuristics and other conventional methods for parameter optimization and good results have been realized. Nevertheless, from the user's perspective better performance in terms of accuracy may be satisfactory yet simpler methods can be utilized from a developer's perspective which helps to avoid complex processing as well as issues in interpreting the solutions. This study highlights this objective where the model illustrates significant performance improvement by utilizing cross validation.

Extremely Randomized Trees or Extra Trees are a category of Decision Tree Learners. Hence to evaluate its performance in comparison with other Decision Trees is substantial. Decision Trees are non-parametric methods which help to capture nonlinear relationships and are seen useful for exploration of data. The proposed model is compared with 10 other decision tree classifiers (DT's)- J48, Hoeffding Trees, Naïve Bayes Trees, Repeated pruning Trees, Simple CART, Logistic Model Trees, Alternating Decision Trees, Random trees, Decision Trees, Functional Trees and it was seen significantly better in performance.

The comparison with the various DT's is illustrated in Figure 11. The best performers after the proposed Model are Hoeffding Trees, Random Forests, Naïve Bayes Trees, and so on. The least performer was Decision Stump. Random Forest is a bagging ensemble decision Tree similar to Extra Trees. Random Forest uses replacement while using subsamples while extra trees use the whole data and to select the split node it makes use of optimal split while extra trees does it randomly. Random Forest achieved an accuracy of 97.38%. Thus, comparing both the bagging ensembles illustrates the superior performance of Extra Trees.

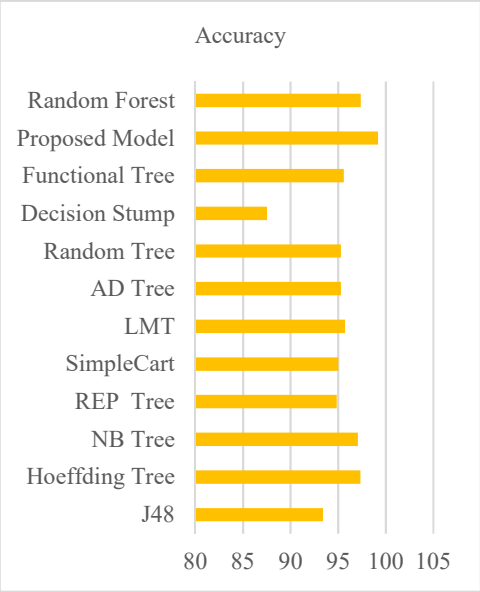


Figure 11 Comparison With DT's

On comparison with state of art techniques available in literature the proposed model was seen to be superior in performance. Table 3 depicts the comparison made.

Table 3 Comparison With Literature

Author	Model	Accuracy
[17]	Extra trees + Whale Optimization	99.0
[24]	Random Forest+ cost sensitive learning	97.9
[30]	Extra Trees	96.3
[13]	Decision Trees	98.5
[35]	Deep Learning and Adaboost	97.2
[36]	MLP homogenous ensembles	98.79
Proposed Model	Extra Trees with parameter Tuning and features importance	99.27

A random forest model with cost sensitive learning was proposed by [24] for Breast cancer classification using the Wisconsin Dataset A cost matrix that penalized wrong classification of the positive or minority class was used. The model obtained an accuracy of 97.9%. An Extra Tree classifier was used by [30] Several Decision Trees Classifiers were used by [13] for breast cancer classification into Benign and Malignant tumours on the Wisconsin Dataset The highest classification was shown by NaiveBayes Tree with 97.07 % accuracy. A Deep Learning model with Adaboost was proposed by [35]. The model illustrated an accuracy of 97.2%. An MLP homogenous ensemble utilizing optimization techniques such as genetic Algorithms, PSO, for parameter optimization was proposed by [36] and an accuracy of 98.79% with best parameters were obtained. Compared to these the proposed model illustrated a superior performance with a better accuracy score of 99.22% and 98.5% with cross validation. Besides, the misclassifications of the two classes were seen least with the proposed model when compared with those in literature.

The proposed work illustrated an accuracy of 99.27% and it is also compared with present high impact work. As per [38] the most utilized ML classifiers for BC diagnosis and classification are SVMs, Decision Trees, and Artificial Neural networks. Among the various categories of Decision Trees Extra trees are seen to be one of the least exploited when considering available quantity of literature. Available studies [37] illustrate the suitability for its implementation in cancer diagnosis and classification. Application of various ML techniques in the diagnostic field is envisaged to enhance human capability and intelligence [39], hence implementation of systems that are easy to use, yet, that are precise in prediction and interpretable, is required. A factor influencing interpretability is feature importance [39]. The current trend is implementation of classification models with deep learning techniques but a major drawback is these models are not suitable for small sized datasets. [36] used deep learning models on the FNAC image dataset with four classifiers- MLP, DT, SVM and KNN and concluded that they could achieve 99% accuracy in classification with MLP with feature extraction using DenseNet 201 architecture. In their work [17] developed an extra trees-based model implemented with feature selection using whale optimization and depicted an accuracy of 99.3%. The proposed model also achieved a similar accuracy score.

5. CONCLUSION

The paper evaluated the performance of Extra Tree classifiers for Breast Cancer tumour Classification into malignant or benign. The model was compared with other Decision Trees and other state of art techniques in literature and was seen to illustrate superior performance with an accuracy of 99.27%. The study proposes a model using Extremely randomized trees for classification of breast cancer tumours. Features which are relevant to the outcome of the models are identified and hyperparameter tuning based on simple techniques such as cross validation with accuracy scores was performed to obtain the best arguments for the model parameters. In addition, the model was able to reduce the misclassification of instances effectively than other state of art models.

The proposed model was evaluated on the Wisconsin dataset of 683 instances. As a future work, the model can be evaluated on larger datasets as well as datasets from different domains to interpret the consistency of the model. Other feature importance techniques can be implemented for enhancing the performance of the model.

ACKNOWLEDGMENT

The author acknowledges the creator of the dataset Dr William H Wolberg of the University of Wisconsin Hospitals, Madison for the dataset used in this study.

REFERENCES

[1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209-249.

[2] Kamel, S.R., YaghoubZadeh, R. & Kheirabadi, M. Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer. *J Big Data* **6,** 90 (2019). https://doi.org/10.1186/s40537-019-0247-7

[3] Plichta, J. K., Ren, Y., Thomas, S. M., Greenup, R. A., Fayanju, O. M., Rosenberger, L. H., ... & Hwang, E. S. (2020). Implications for breast cancer restaging based on the 8th edition AJCC staging manual. *Annals of surgery*, *271*(1), 169.

[4] Mathur P, Sathishkumar K, Chaturvedi M, Das P, Sudarshan KL, Santhappan S, Nallasamy V, John A, Narasimhan S, Roselind FS; ICMR-NCDIR-NCRP Investigator Group. Cancer Statistics, 2020: Report From National Cancer Registry Programme, India. JCO Glob Oncol. 2020 Jul;6:1063-1075. doi: 10.1200/GO.20.00122. PMID: 32673076; PMCID: PMC7392737.

[5] Hakim, A., & Awale, R. N. (2020). Thermal imaging-An emerging modality for breast cancer detection: a comprehensive review. *Journal of Medical systems*, *44*(8), 1-18.

[6] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, *8*, 150360-150376.

[7] Ak, M. F. (2020, June). A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. In *Healthcare* (Vol. 8, No. 2, p. 111). Multidisciplinary Digital Publishing Institute.

[8] Mathew, T. E., Anil Kumar, K. S., 2020, 'A Logistic Regression based hybrid model for Breast Cancer Classification', *Indian Journal of Computer Science and Engineering (IJCSE)*, Vol. 11, Issue 6, pp. 899-903

[9] Ayon, S. I., Islam, M. M., & Hossain, M. R. (2020). Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. *IETE Journal of Research*, 1-20.

[10] Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., & Decker, S. (2021). Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, *22*(1), 393-415.

[11] Mathew, T. E., 2019, 'A comparative study of the performance of different Support Vector machine Kernels in Breast Cancer Diagnosis', *International Journal of Information and Computing Science (IJICS),* Vol. 6, Issue 4, pp. 432-441.

[12] Mathew, T. E., 2019, 'A Logistic Regression with Recursive Feature Elimination, for Breast Cancer Diagnosis', *International Journal on Emerging Technologies (IJET),* Vol. 10, Issue 3, pp. 55-63.

[13] Mathew, T. E., 2019, 'Simple and Ensemble Decision tree Classifier based detection of Breast Cancer', *International Journal of*

Scientific & Technology Research (IJSTR), Vol. 8, Issue 11, pp. 1628-1637.

[14] Mathew, T. E., Anil Kumar, K. S., 2021, 'A Modified- Weighted- K -Nearest Neighbour and Cuckoo Search Hybrid Model for Breast Cancer Classification, *Indian Journal of Computer Science and Engineering (IJCSE),* Vol.12, Issue 1, pp.166-177

[15] Abubaker, H., Ali, A., Shamsuddin, S. M., & Hassan, S. (2020). Exploring permissions in android applications using ensemble-based extra tree feature selection. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(1), 543-552.

[16] Mangasarian, O.L.; Street, W.N.; Wolberg, W.H. Breast cancer diagnosis and prognosis via linear programming. *Oper. Res.* **1995**, *43*, 570–577.

[17] Abbas, S., Jalil, Z., Javed, A. R., Batool, I., Khan, M. Z., Noorwali, A., ... & Akbar, A. (2021). BCD-WERT: a novel approach for breast cancer detection using whale optimization based efficient features and extremely randomized tree algorithm. *PeerJ Computer Science*, *7*, e390

[18] S. Alghunaim and H. H. Al-Baity, "On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context," in IEEE Access, vol. 7, pp. 91535-91546, 2019, doi: 10.1109/ACCESS.2019.2927080.

[19] Chaurasia, V., & Pal, S. (2020). Applications of machine learning techniques to predict diagnostic breast cancer. SN Computer Science, 1(5), 1-11, https://doi.org/10.1007/s42979-020-00296-8

[20] Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated breast cancer diagnosis based on machine learning algorithms. Journal of healthcare engineering, 2019, https://doi.org/10.1155/2019/4253641

[21] Ghiasi, M. M., & Zendehboudi, S. (2021). Application of decision tree-based ensemble learning in the classification of breast cancer. Computers in Biology and Medicine, 128, 104089., https://doi.org/10.1016/j.compbiomed.2020.104089.

[22] Islam, M., Haque, M., Iqbal, H., Hasan, M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1(5), 1-14, https://doi.org/10.1007/s42979-020-00305-w .

[23] Mathew, Tina Elizabeth, An Improvised Random Forest Model for Breast Cancer Classification, *NeuroQuantology* | May 2022, Volume 20, Issue 5, Page 713-722, doi: 10.14704/nq.2022.20.5.NQ22227

[24] Senthilkumar, B., Zodinpuii, D., Pachuau, L., Chenkual, S., Zohmingthanga, J., Kumar, N. S., & Hmingliana, L. (2022). Ensemble Modelling for Early Breast Cancer Prediction from Diet and ifestyle. IFAC-PapersOnLine, 55(1), 429-435, ttps://doi.org/10.1016/j.ifacol.2022.04.071.

[25] Samieinasab, M., Torabzadeh, S. A., Behnam, A., Aghsami, A., & Jolai, F. (2022). Meta-Health Stack: A new approach for breast cancer prediction. *Healthcare Analytics,* 2, 100010., https://doi.org/10.1016/j.health.2021.100010

[26] Kharwar, A. R., & Thakor, D. V. (2022). An Ensemble Approach for Feature Selection and Classification in Intrusion Detection Using Extra-Tree Algorithm. *International Journal of Information Security and Privacy (IJISP)*, *16*(1), 1-21.

[27] Sharaff, A., & Gupta, H. (2019). Extra-tree classifier with metaheuristics approach for email classification. In *Advances in computer communication and computational sciences* (pp. 189-197). Springer, Singapore.

[28] Patel, T., Patel, D., & Patel, T. (2022). Breast Cancer Prediction Analysis Using Data Mining Techniques. In ICDSMLA 2020 (pp. 623-631). Springer, Singapore, https://doi.org/10.1007/978-981-16-3690-5_56.

[29] Breast Cancer Wisconsin (Original) Data Set, Online]. https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data. Accessed 5 May 2022

[30] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." Machine learning 63.1 (2006): 3-42, DOI 10.1007/s10994-006-6226-1.

[31] Sharma, J.; Giri, C.; Granmo, O.C.; Goodwin, M. Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation. EURASIP J. Info. Secur. 2019, 2019, 15, https://doi.org/10.1186/s13635-019-0098-y. Access, 8, 96946- 96954. https://doi.org/10.1109/ACCESS.2020.299353

[32] Zafari, A.; Zurita-Milla, R.; Izquierdo-Verdiguier, E. Land Cover Classification Using Extremely Randomized Trees: A Kernel Perspective. IEEE Geosci. Remote Sens. Lett. 2019, 1–5, https://doi.org/10.1109/LGRS.2019.2953778.

[33] . Melanson, D. (2020). Extremely Randomized Trees with Multiparty Computation (Doctoral dissertation, University of Washington Tacoma). Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[34] Zheng, J., Lin, D., Gao, Z., Wang, S., He, M., & Fan, J. (2020). Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis. IEEE

[35] Zhiqiang Guo, Lina Xu & Nona Ali Asgharzadeholiaee (2022) A Homogeneous Ensemble Classifier for Breast Cancer Detection Using Parameters Tuning of MLP Neural Network, Applied Artificial Intelligence, DOI: 10.1080/08839514.2022.2031820

[36] Zerouaoui, H., Idri, A., Nakach, F.Z., Hadri, R.E. (2021). Breast Fine Needle Cytological Classification Using Deep Hybrid Architectures. In: , *et al.* Computational Science and Its Applications – ICCSA 2021. ICCSA 2021. Lecture Notes in Computer Science(), vol 12950. Springer, Cham. https://doi.org/10.1007/978-3-030-86960-1_14

[37] Abdulkareem, S. A., & Abdulkareem, Z. O. (2021). An evaluation of the Wisconsin breast cancer dataset using ensemble classifiers and RFE feature selection. *Int. J. Sci., Basic Appl. Res.*, *55*(2), 67-80,

[38] ElOuassif, B., Idri, A., Hosni, M., & Abran, A. (2021). Classification techniques in breast cancer diagnosis: a systematic literature review. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *9*(1), 50-77.

[39] Hakkoum, H., Idri, A., & Abnane, I. (2021). Assessing and comparing interpretability techniques for artificial neural networks breast cancer classification. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *9*(6), 587-599.