An Improvised Random Forest Model for Breast Cancer Classification

Dr. Tina Elizabeth Mathew[1*]

Abstract

Breast Cancer is considered as the most common cancer in females with high incidence rate. The evolution of modern facilities has helped in reducing the mortality rate, yet the incidence is still the highest among all cancers affecting women. Early diagnosis is a predominant factor for survival. Hence techniques to assist the current modalities are essential. Machine learning techniques have been used so as to produce better prediction and classification models which will aid in better and earlier disease diagnosis and classification. Random Forest is a supervised machine learning classifier that helps in better classification. Random Forests are applied to the Wisconsin breast cancer dataset and the performance of the classifier is evaluated for breast cancer classification. Here in this study an improvised random forest model which uses a cost sensitive learning approach for classification is proposed and it is found to have a better performance than the traditional random forest approach. The model gave an accuracy of 97.51%.

Key Words: Cost Matrix, Decision Trees, Breast Cancer, Classification, Improved Random Forest Classifier Approach (IRFC).

DOI Number: 10.14704/nq.2022.20.5.NQ22227 **NeuroQuantology 2022; 20(5):713-722**

Introduction

Breast Cancer, the neoplasm of the breast is a prevalent disease among women which is of great concern to women of all ages, countries and ethnicity. Recent statistics published by the World Health Organization in 2021 show that it is ranked first in incidence and first or second in mortality rate in almost all countries of the world. Besides the modern medical modalities available supplementary techniques can be used for early detection of the disease which is a key factor for survival. Several medical modalities like Mammograms, MRI, CT scan, Thermography, USG Scan and many hybrid state of art techniques are available for disease diagnosis, but each of these modalities have their own pros and cons. A major issue is that these techniques use detrimental x rays and this is harmful for the patient. Besides the diagnosis process can be painful as well as stressful to the patients. The diagnostic accuracy can also at times be inconclusive and incorrect, Hence, to avoid these issues alternative techniques can be used to assist the medical practitioner. Literature studies indicate the use of various machine learning techniques for better identification, prediction and classification of the disease. (Algehyne et al, 2022), (Balaraman, 2020), (Mathew, 2019a), (Mathew, 2019b). Machine learning is a subdomain of artificial intelligence and to unravel complex patterns from the heterogenous biological data, machine learning methods can be employed [Qusit et al, 2021).

Corresponding author: Dr. Tina Elizabeth Mathew
Address: [1*]Assistant Professor in Computer Science, Government College Kariavattom, Thiruvananthapuram, Kerala, India.
[1*]E-mail: tinamathew04@gmail.com
Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
Received: 15 April 2022 **Accepted:** 10 May 2022

Machine learning techniques can be classified into supervised and unsupervised techniques and each category has a plethora of methods. These techniques have seen to provide effective classification of diseases as per (Algehyne et al, 2022), (Khourdifi, Bahaj 2018), (Mathew, 2019c), (Mathew, Kumar, 2020), (Mathew, Kumar, 2021). Decision Trees are techniques that belong to the category of supervised techniques. Decision Trees are considered to be simple methods yet powerful enough for classification. Two broad categories of Decision Trees are classification trees and regression trees. Classification trees are further subdivided into various types such as ID3, C4.5, CART, CHAID, MARS and so forth. A major drawback with decision trees is that they can overfit.so usually to solve this concern ensembles or groups of decision trees are taken as classifiers. Four such categories are Bagged trees, Random Forest, Boosted Trees and Rotation Forest. Each category exhibits suitability for classification (Mashudi et al, 2021).

Random Forest is considered to be robust than individual decision trees. Random forest takes a forest of trees and trains them using the bootstrap aggregation technique. The prediction of each tree is taken and the majority votes of all these predictions is produced as the final output. Bagging takes different subsets of samples instead of using one single set. The Random Forest classifiers are seen to provide a moderately high accuracy without the need of normalization of any dataset values (Ara et al, 2021). Albeit, an issue is that they are black box techniques and reasons behind the diagnosis or classification cannot be explained (wang et al, 2020).

Objectives of the Study

- To develop an improved Random Forest Model using cost sensitive learning.
- To compare the performance of the traditional Random Forest Model with the proposed improved model.
- To identify aptness of the proposed model for breast cancer classification.

Significance of Study

- The intent is to identify and develop a supervised machine learning model for breast cancer classification so that medical practitioners can use this model as a diagnostic assistive technology.
- These assistive models are non-invasive, painless and harmless techniques. By utilizing these for the detection of breast cancer the patients need not suffer the painstaking disease detection processes which make use of harmful radiation.
- The paper is organized as follows, Section 1 represents the introduction, Section 2 represents the literature survey, while Section 3 proposes the materials used and methods applied, Section 4 represents the results and discussions, and the final Section 5 proposes the conclusion followed by references.

Literature Review

Various literature reviewed based on random forest models and its application in the breast cancer domain as well as other fields are given in the following section. In their work (Alam et al, 2019) used a feature ranking and selection strategy along with Random Forest and it was seen to provide better performance. The model was tested with 10 datasets and the random forest model was seen to perform consistently across datasets when compared to other classifiers such as SVM and Bayes Networks. In his work (Al-Quraishi et al, 2018) proposed methods where Random Forest was utilized in conjunction with other techniques for breast cancer risk prediction, assessing recurrence probability and development of a prediction model for forecasting survivability status. (Ara et al, 2021) in their work developed an automatic breast cancer diagnostic system using SVM and Random Forest and the model was seen to produce an accuracy of 96.5%. (Buttan et al, 2021), proposed an RF model with grid search to predict breast cancer and analysed the outcomes. In their work (Chaudhary et al, 2016), used feature selection using three attribute evaluators. The classifier is improved using an attribute evaluator method and an instance filter method. For risk assessment of breast cancer (Housseinpour et al, 2022), used an improved random forest algorithm and obtained promising results using the model. (Jackins et al, 2021), compared performance of random forests, naïve bayes, k-means and DBSCAN on various datasets and they concluded that random forests worked well with all datasets. (Jadhav et al, 2019), illustrated that random forests performed better than logistic regression and decision tree models. In their work (Kaur et al, 2019), proposed an IOT based smart health system

with random forest classifier and the results of the study were seen promising with the classifier giving good accuracy for various disease datasets. (Keles, 2019),] in their comparative study on various classifiers illustrated that random forest could achieve an accuracy above 90% for disease classification. Work proposed by (Li, Chen, 2018), also indicated that random forests can be suitable for breast cancer classification. (Macaulay, 2021) proposed a risk prediction model for breast cancer in African women and their proposed model used random forest to identify risk factors and it was found suitable for identifying them. Combining feature selection with the models were seen to improve model performance. (Rohan, 2019), combined random forest with Adaboost and illustrated that this helped in performance enhancement of the model. (Shaik, Srinivasan, 2019), in their work highlighted the need for accuracy improvement using Random Forests. Their work was solely focused on random forests. (Shahhoseini, Hu, 2020), improved the random forest classifier performance by using a weighted model. In their work (Sharma et al, 2018), compared various machine learning techniques for Breast cancer classification. [Zhu et al, 2018] produced a forecasting model for breast cancer prediction using a Random Forest- Adaboost combination.

A major limitation identified in the various literature is that as the number of trees used in the forest grows the model gets slower. And this affects the performance in real time predictions. Even then it is considered to be much better than other classifiers, specifically other decision trees as it can handle large datasets and also takes care of overfitting which is a factor that usually affects the performance of models. It helps tackle the problem of variance.

A main challenge faced in classification is class imbalance. There is no unique judgement about the degree of the imbalance that exist between class cardinalities. Some researchers have studied datasets that have not so severe imbalance where one class is few times smaller than other class, while others have considered more severe imbalance ratios such as 1:100, 1:1000 and so forth. A class with few examples does not help to find the data regularities.

There are several techniques to deal with this. Two broad classifications are databased and algorithmic based sampling methods. Besides this, literature also recommends feature selection, and ensembling of techniques to deal with data imbalance.

In databased techniques the imbalanced data ratio is reduced by either adding more minority instances, known as oversampling or discarding some of the majority instances called under sampling. (Yap, et al, 2014). These are usually applied at the data pre-processing phase. This does not affect the algorithms used. In algorithmic based approach algorithms sae modified internally to achieve this (Napierała, 2012).

Materials and Methods

About the Dataset

The WBCD dataset used in this study taken from the University of Wisconsin Hospitals contains 699 instances, reported till July 15th, 1992. This dataset contains sample code number and 10 attributes for each instance of disease. The independent attributes of the dataset are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses. All of them are represented by values within the range of 1 and 10. The dependent attribute is class, which is represented by integer value 2 and 4, where 2 stands for benign tumours and 4 stands for malignant tumours.

Random Forest

Random Forest is a popular supervised machine learning algorithm that is used for several sorts of classification problems (Huljanah et al, 2019), (Islam et al, 20200), (Jayaraj, Sathiamoorthy, 2019). It is an ensemble of tree-structured classifiers. Each tree of the forest outputs a vote, and it assigns each instance of the input to the most probable class label. RF which basically uses the Classification and Regression Tree (CART) algorithm is seen to employ a number of decision trees as weak classifiers (Ghiasi, Zendehboudi, 2021)]. But in contrast with CART it does not employ a greedy algorithm (Kumar, Poonkudi, 2019). Many studies show that Random Forest is a fast method, robust to noise and it is an efficacious ensemble that identifies non-linear patterns in data. It can handle both numerical as well as categorical data easily. One of the major advantages of Random Forest is that it does not suffer from over fitting, even when more trees get appended to the forest. Random Forests are considered to be a powerful technique for classification, pattern recognition so on and so

forth (Fruend, Mason, 1999), (Ganggayah et al, 2019). Its parallel architecture makes it a fast classifier [31]and it takes care of data imbalance (Khoshgoftaar et al, 2007).

The Parameter used in random forests is the n_trees, number of trees that constitute the forest. Here it is takena s 100. Since a random forest creates ensembles of multiple decision trees, hyperparameters are used to control the number of trees (Gupta, Garg, 2020) and they are:

1. max_features (number of attributes to be selected from data for randomisation)
2. max_depth (for pre-pruning of trees)
3. max_features = sqrt(n_features) (for classification)

Applying Ensemble learning techniques enhances the performance of predictive models by improving their accuracy (Yifan, 2021).

Random Forest builds a tree as an ensemble of decision trees that uses bagging together with random sampling of training points. It consists of an uncorrelated forest of trees and builds multiple decision trees and merges them which helps the classifier in getting a more accurate and stable prediction. It is capable of handling missing values but tends to show bias towards the majority class. Hence to improve the classification of the minority class improvement is to be effected. Besides, even though Random Forests methods can handle imbalanced data as the imbalance rate increases classification ability can decrease so methods to counteract this can be used (zhu et al, 2018). Hence to overcome this an improved random Forest approach is proposed in the next section.

Improved Random Forest Approach

Improved-RFC approach uses the Random Forest algorithm, and a cost matrix that penalizes each misclassification of the positive class or minority class twice thus to give a cost sensitive Random Forest model. That provides better classification for the minority class which is the positive class in this dataset.

Cost Sensitive classification using a cost matrix can be categorized as an algorithmic based approach**.** Misclassification cost is associated with classes**.** A cost matrix proposes a means to differentiate between and highlight the importance of the two classification error categories -Type I and Type II.

The cost matrix used in the IRFDC approach is represented in Figure 1. The cost weights that lead to optimal performance in the classification process is learned by using grid search technique.

0	1
2	0

Figure 1. Cost matrix

The aim of the approach is to improve classification accuracy of the traditional Random Forest algorithm for binary classification and reduce misclassification of the minority class aka. positive class.

The pseudo code of the Improved Random Forest Classifier approach is given below. Figure 2 gives the working architecture of the proposed model.

Algorithm of IRFC
Input: $D_{Train} = \{x_1, x_2 \ldots x_n\}$ // Training dataset which includes a set of training examples with their class labels.
Output: Classification + Performance measures
Method:
Step 1: Partition dataset into training and testing sets.
Step 2: Select Random Forest Classifier for classification.
Step 3: Train the classifier. Use 10-fold CV. Check Accuracy measure. Apply grid search to get the optimal value for the cost matrix **Step 4:** Apply cost matrix for each misclassified positive sample. **Step 5:** Apply classifier on test data
Step6: Output Performance measures- Accuracy, MCC, F Score, Kappa Statistic

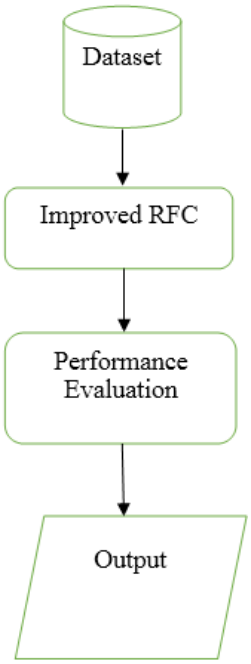

Figure 2. Architecture of IRFC Approach

Results & Discussion

The results obtained are shown in Table 1 The Improved Random Forest Approach enhanced accuracy to 97.51% from 96.63%. The misclassification of the positive class was much reduced from 12 instances to 5 instances.

The Matthews Correlation Coefficient MCC which is a measure of the quality of the binary classifier has shown a significant improvement. The IRFC model gave a value of 0.946 over 0.926. MCC is considered to be a more reliable measure than Accuracy.

Kappa Statistic of IRFC also displays a better value than that of the traditional approach. F measure also improved by a value of 0.009.

By applying the strategy of 10-fold cross validation overfitting is avoided.

P-R AUC and ROC AUC values of both models are also displayed.

Table 1. Performance Metrics of IRFC

Performance Metrics	Standard Random Forest	Improved Random Forest
Accuracy	**96.6325**	**97.51**
Kappa statistic	**0.9259**	**0.9457**
Confusion Matrix	**433 11 12 227**	**432 12 5 234**
F measure	**0.966**	**0.975**
MCC	**0.926**	**0.946**
ROC	**0.992**	**0.992**
PRC	**0.990**	**0.991**
Cost Matrix used Time to build model(sec)	**- 0.04**	**0 1 2 0 0.03**

A comparison of the proposed and traditional model in terms of Precision, Recall and Specificity is illustrated in Figure 3. For Precision and Recall the IRFC approach is superior in performance. While in Specificity the traditional model has a slightly higher value.

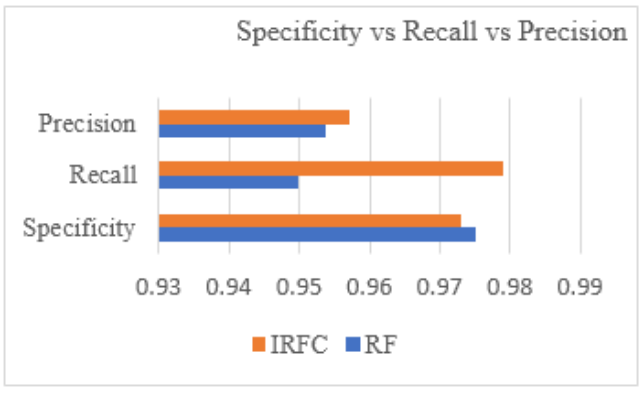

Figure 3. Specificity vs Recall vs Precision

Comparison of FPR and FNR of the models are illustrated in Figure 4. The false negative rate also termed as the miss rate. It denotes the probability that a true positive will be missed by the test. FNR of IRFC is superior than that of RF method. In case of FPR that of the RF method is slightly better than IRFC approach. An ideal model should have very low scores on FNR and FPR, while a practical model often has to make a trade-off between these two scores. The proposed IRFC has a better value for FNR.

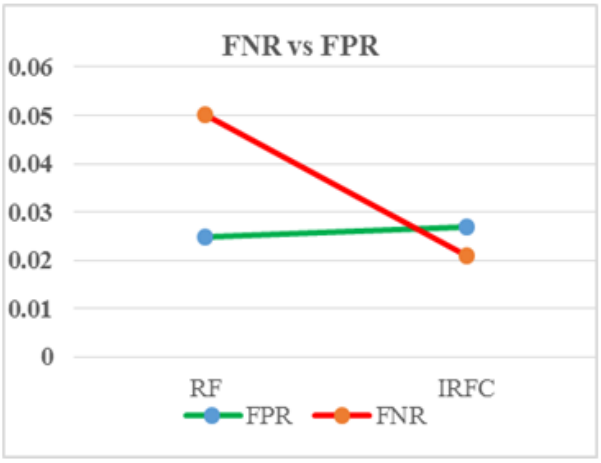

Figure 4. FNR vs FPR

The confusion matrix of the IRFC model depicts that the number of false negatives is reduced drastically, yet number of false positives is to be improved. The false negatives were much reduced from 12 instances to 5 instances. However, the false positives were not improved.

Besides this, the time taken to build the proposed model was seen to be lesser than the time taken by the standard model.

The experiment demonstrates that by integrating cost-sensitive learning to random forests effectively improves the classification performance.

The IRFC model is compared against other decision tree classifiers and figure 5 illustrates this comparison.

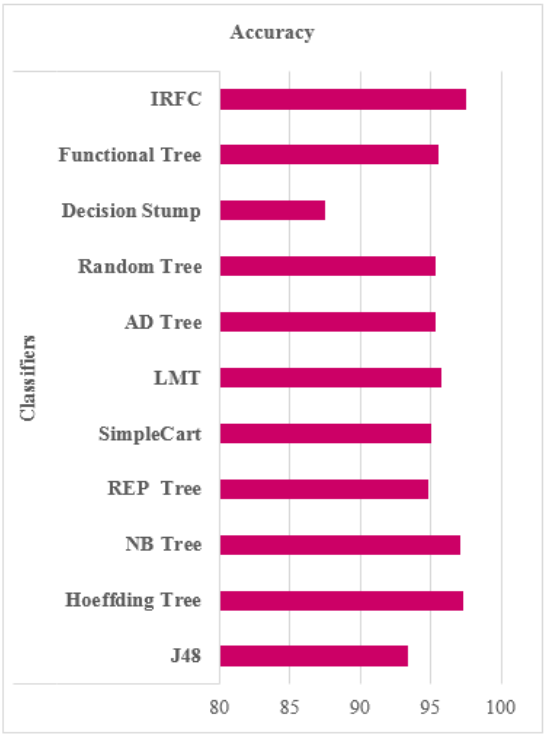

Figure 5. Comparison of Decision Trees

From the figure we can compare the accuracy measure obtained for various types of Decision Trees. The least accuracy is shown by the Decision Stump Model with a value of 87.5%. The proposed model has an accuracy of 97.51%. Other models like Hoeffding Tree and BB Tree show accuracy of 97.36% and 97.07% respectively. This highlights the superior performance of the proposed model amongst various decision Tree Classifiers.

The proposed IRFC model is compared against few other classifiers- Support Vector Machines, k-Nearest Neighbours, Logistic Regression and figure 6 depicts the comparison done. The proposed method was seen to be superior in accuracy when compared with these classifiers.

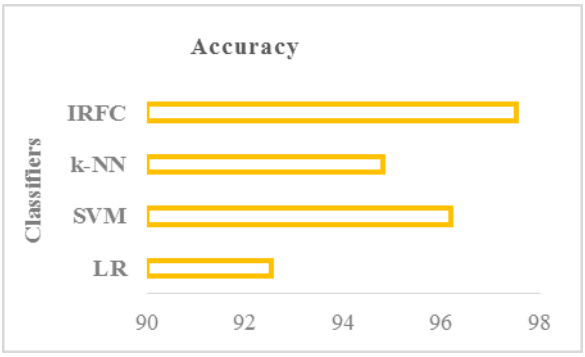

Figure 6. Comparison of Classifiers

The accuracy of the logistic Regression classifier was seen to be the lowest at 92.58%. The k-Nearest Neighbour classifier produced ana accuracy of94.8% and the Support Vector machine classifier gave an accuracy of 96.19%, The results highlight that the proposed IRFC model with accuracy of 97.515 was much superior in performance.

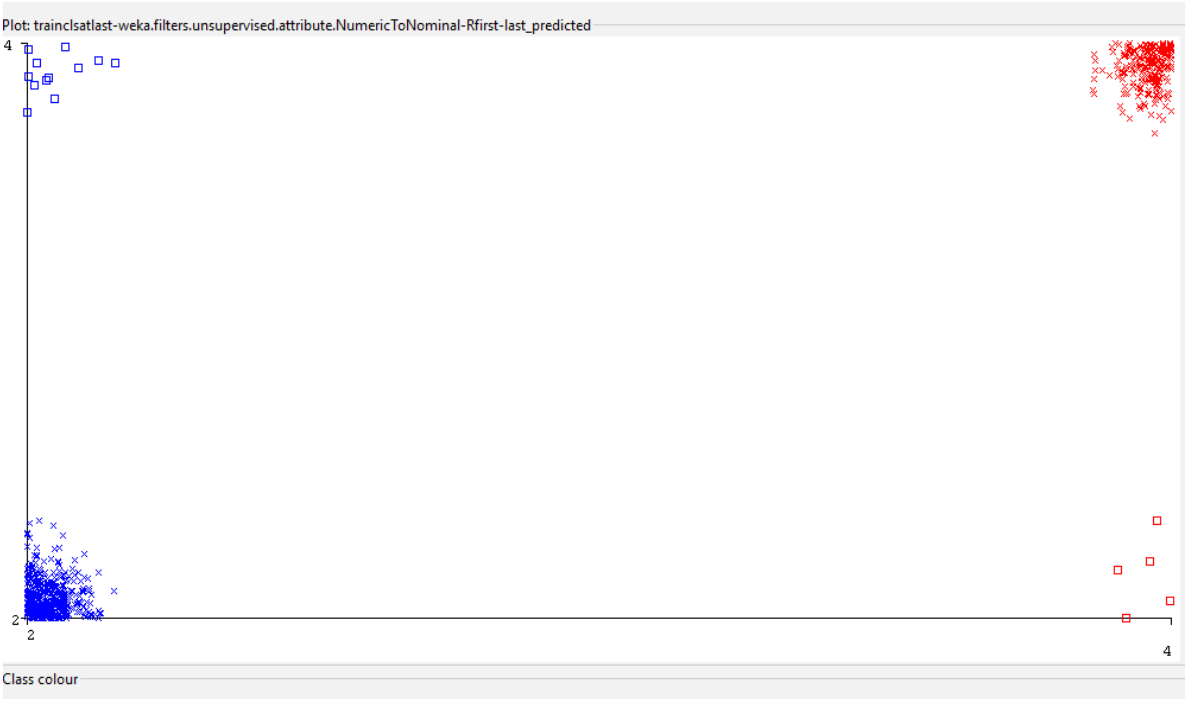

Figure 7. Visualization of Model Misclassification

Figure 7 gives the visualization of the classification errors made. Along the x axis the right most end shows the misclassified 5 positive instances. Similarly, along the y axis at the top end we can see the misclassified actual negative 12 instances.

The P-R curve of the malignant class of the proposed Model is depicted in figure 8. The P R curve uses recall on x-axis and precision on y axis. It is a helpful metric in cases where the data is imbalanced. Large P-R AUC values indicate better performance of the model. This is illustrated by the curve moving up towards the top right to left corner.

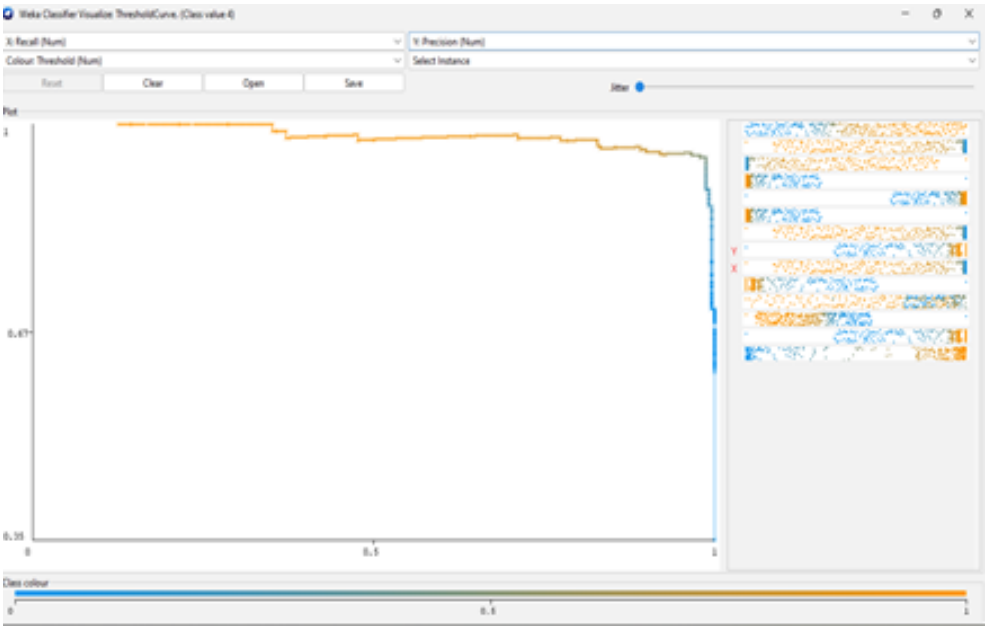

Figure 8. P-R Curve

Figure 9 depicts the ROC curve of the proposed IRFC model for the malignant class. ROC plots FPR on x axis and TPR on y axis. It is a useful metric to illustrate the diagnostic capability of the binary classifier. An ideal ROC curve will be found along the upper left to right corner of the plot. The AUC ROC obtained is 0.9917

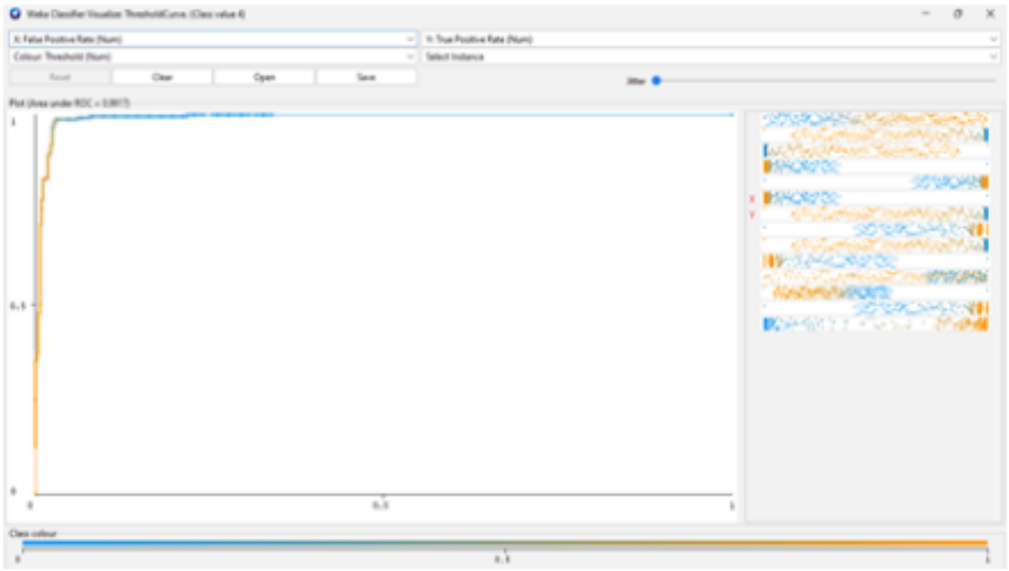

Figure 9. ROC Curve

The proposed model is compared with state of art technologies available in literature in Table 2. The proposed model is seen to outperform the models. (Imran et al, 2022) used an improved Random Forest Model and obtained an accuracy of 96%. (Zheng et al, 2020) used a hybrid Deep learning and Adaboost model and produced an accuracy of 97.2%. The proposed IRFC model produced an accuracy of 97.51%.

Thus, it can be comprehended that the proposed model is superior to these methods and can be used for breast cancer classification with a reduced misclassification of the positive class.

Table 2. Comparison with Literature

Author	Method	Accuracy
Imran et al, 2022	Random Forest	96
Zheng, 2020	Deep learning _ Adaboost	97.2
Proposed Model	IRFC	97.51

Conclusion and Outlook

Classification techniques have gained more traction with the availability of various kinds of clinical data, genomics data, omics data and many more. The paper discusses an approach for better classification accuracy on clinical data. The proposed IRFC approach improved the accuracy of classification over the traditional approach. The false negative rate was much reduced. However, the false positives produced by the approach is still a concern and other approaches are to be applied to provide better performance. In this context a suggestion will be to utilize optimization techniques, feature selection methods to improve the diagnostic accuracy. A problem with machine learning techniques is that they are data dependent and as data changes the structure and parameters required can change. To overcome this problem the models can be employed over different datasets of different data sizes so as to be updated. accordingly.

Acknowledgements

I am grateful to Dr William H Wolberg of the University of Wisconsin Hospitals, Madison for the dataset provided.

Abbreviation List

BC- Breast Cancer
CV- Cross Validation
FNR- False Negative Rate
FPR- False Positive Rate
IRFC- Improved Random Forest approach
MCC- Matthews Correlation Coefficient
P-R- Precision- Recall
RF- Random Forest
ROC- Receiver Operating Characteristics

References

Alam, M.Z., Rahman, M.S., and Rahman, M.S. 2019. A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, *15*, 100180.

Algehyne, E.A., Jibril, M.L., Algehainy, N.A., Alamri, O.A., and Alzahrani, A.K. 2022. Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia. *Big Data and Cognitive Computing*, *6*(1), 13.

Al-Quraishi, T., Abawajy, J.H., Chowdhury, M. U., Rajasegarar, S., and Abdalrada, A.S. 2018, February. Breast cancer recurrence prediction using random forest model. *In International Conference on Soft Computing and Data Mining,* 318-329.

Ara, S., Das, A., and Dey, A. 2021. Malignant and benign breast cancer classification using machine learning algorithms. *In 2021 International Conference on Artificial Intelligence (ICAI),* 97-101.

Balaraman, S. 2020. *Comparison of Classification Models for Breast Cancer Identification using Google Colab.* IEEE.

Buttan, Y., Chaudhary, A., and Saxena, K. 2021. An improved model for breast cancer classification using random forest with grid search method. *In Proceedings of Second International Conference on Smart Energy and Communication,* 407-415.

Chaudhary, A., Kolhe, S., and Kamal, R. 2016. An improved random forest classifier for multi-class classification. *Information Processing in Agriculture*, *3*(4), 215-222.

Freund, Y. and Mason, L. 1999, The Alternating Decision Tree Learning Algorithms S.N. www1.Cs.Columbia.Edu/ Compbio/Medusa/Non_Html_Files/Freund_Atrees.pdf

Ganggayah, M.D., Taib, N.A., Har, Y.C., Lio, P., and Dhillon, S.K. 2019. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC medical informatics and decision making*, *19*(1), 1-17.

Ghiasi, M.M., and Zendehboudi, S. 2021. Application of decision tree-based ensemble learning in the classification of breast cancer. *Computers in Biology and Medicine*, *128*, p 104089.

Gupta, P., and Garg, S. 2020. Breast cancer prediction using varying parameters of machine learning models. *Procedia Computer Science*, *171*, 593-601.

Hosseinpour, M., Ghaemi, S., Khanmohammadi, S., and Daneshvar, S. 2022. A hybrid high-order type-2 FCM improved random forest classification method for breast cancer risk assessment. *Applied Mathematics and Computation*, *424*, 127038.

Huljanah, M., Rustam, Z., Utama, S., and Siswantining, T. 2019, June. Feature selection using random forest classifier for predicting prostate cancer. In *IOP Conference Series: Materials Science and Engineering,* 546(5), 052031.

Imran, B., Hambali, H., Subki, A., Zaeniah, Z., Yani, A., & Alfian, M. (2022). Data Mining Using Random Forest, Naïve Bayes, and Adaboost Models for Prediction and Classification of Benign and Malignant Breast Cancer. *Jurnal Pilar Nusa Mandiri,* 18(1), 37-46. https://doi.org/10.33480/pilar.v18i1.2912

Islam, M., Haque, M., Iqbal, H., Hasan, M., Hasan, M., and Kabir, M.N. 2020. Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, *1*(5), 1-14.

Jackins, V., Vimal, S., Kaliappan, M., and Lee, M.Y. 2021. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing*, *77*(5), pp 5198-5219.

Jadhav, M., Thakkar, Z., and Chawan, P.M. 2019. Breast cancer prediction using supervised machine learning algorithms. *International Research Journal of Engineering and Technology (IRJET),* 6.

Jayaraj, D., and Sathiamoorthy, S. 2019, November. Random Forest based Classification Model for Lung Cancer Prediction on Computer Tomography Images. *In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT),* 100-104.

Kaur, P., Kumar, R., and Kumar, M. 2019. A healthcare monitoring system using random forest and internet of things (IoT). *Multimedia Tools and Applications*, *78*(14), 19905-19916.

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

Khourdifi, Y., and Bahaj, M. 2018, December. Applying best machine learning algorithms for breast cancer prediction and classification. In *2018 International conference on electronics, control, optimization and computer science (ICECOCS),* 1-5.

Khoshgoftaar, T.M., Golawala, M., and Van Hulse, J. 2007, October. An empirical study of learning from imbalanced data using random forest. *In 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007),* 2, 310-317.

Kumar, A., and Poonkodi, M. 2019. Comparative study of different machine learning models for breast cancer diagnosis. *In Innovations in soft computing and information technology,* 17-25.

Li, Y., and Chen, Z. 2018. Performance evaluation of machine learning methods for breast cancer prediction. *Appl Comput Math*, *7*(4), 212-216.

Macaulay, B.O., Aribisala, B.S., Akande, S.A., Akinnuwesi, B.A., and Olabanjo, O.A. 2021. Breast cancer risk prediction in African women using Random Forest Classifier. *Cancer Treatment and Research Communications*, *28*, 100396.

Mashudi, N.A., Rossli, S.A., Ahmad, N., and Noor, N.M. 2021, March. Comparison on Some Machine Learning Techniques in Breast Cancer Classification. *In 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES),* 499-504.

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),* 6(4), 432-441.

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

Mathew, T.E., 2019, 'Simple and Ensemble Decision tree Classifier based detection of Breast Cancer', *International Journal of Scientific & Technology Research (IJSTR),* 8(11), 1628-1637.

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),* 11(6), 899-903.

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),* 12(1), 166-177.

Napierała, K., 2012. Improving rule classifiers for imbalanced data. *Poznan University of Technology*.

Paul, A., Mukherjee, D.P., Das, P., Gangopadhyay, A., Chintha, A.R., and Kundu, S. 2018. Improved random forest for classification. *IEEE Transactions on Image Processing*, *27*(8), 4012-4024.

Quist, J., Taylor, L., Staaf, J., and Grigoriadis, A. 2021. Random forest modelling of high-dimensional mixed-type data for breast cancer classification. *Cancers*, *13*(5), p 991.

Raj, S., Singh, S., Kumar, A., Sarkar, S., and Pradhan, C. 2021. Feature selection and random forest classification for breast cancer disease. *Data Analytics in Bioinformatics: A Machine Learning Perspective*, pp 191-210.

Rohan, T.I., Siddik, A.B., Islam, M., and Yusuf, M.S.U. 2019. A precise breast cancer detection approach using ensemble of random forest with AdaBoost. *In 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2),* 1-4.

Shaik, A.B., and Srinivasan, S. 2019. A brief survey on random forest ensembles in classification model. In *International Conference on Innovative Computing and Communications,* 253-260.

Shahhosseini, M., and Hu, G. 2020, August. Improved weighted random forest for classification problems. In *International Online Conference on Intelligent Decision Science,* 42-56.

Sharma, S., Aggarwal, A., and Choudhury, T. 2018, December. Breast cancer detection using machine learning algorithms. *In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS),* 114-118.

Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y., and Jin, Y. 2020. An improved random forest-based rule extraction method for breast cancer diagnosis. *Applied Soft Computing*, *86*, 105941.

Yap, B.W., Rani, K.A., Rahman, H.A.A., Fong, S., Khairudin, Z. and Abdullah, N.N., 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. *In Proceedings of the first international conference on advanced data and information engineering (DaEng-2013),* 13-22.

Yifan, D., Jialin, L., and Boxi, F. 2021, May. Forecast Model of Breast Cancer Diagnosis Based on RF-AdaBoost. *In 2021 International Conference on Communications, Information System and Computer Engineering (CISCE),* 716-719.

Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., and Ning, G. 2018. Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access*, 6, 4641-4652.

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 Access,* 8, 96946-96954. https://doi.org/10.1109/ACCESS.2020.2993536