Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm

Sajib Kabiraj[1], M. Raihan[2], Nasif Alvi[3], Marina Afrin[4], Laboni Akter[5], Shawmi Akhter Sohagi[6] and Etu Podder[7]
Department of Computer Science and Engineering, North Western University, Khulna, Bangladesh[1-4,6]
Computer Science and Engineering Discipline, Khulna University, Khulna, Bangladesh[3]
Electronics and Communication Engineering Discipline, Khulna University, Khulna, Bangladesh[7]
Department of Biomedical Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh[5]
Emails: kabirajsajib@gmail.com[1], raihanbme@gmail.com[2], mraihan@nwu.edu.bd[2], nasif.cse12@gmail.com[3], afrin.marina123@gmail.com[4], laboni.kuet.bme@gmail.com[5], sohagishawmi0@gmail.com[6] and e2.120926@gmail.com[7]

***Abstract*—Breast cancer is as one of the common and serious cause of death among women globally. This is a disease where the cells grow out of control inside the breast. Family History of cancer disease, physical inactivity, psychological stress, increase in breast size are the risk factors of breast cancer. In this research paper, breast cancer dataset was analyzed to predict breast cancer using popular two ensemble machine learning algorithms. Random Forest and Extreme Gradient Boosting (XGBoost) were used to predict breast cancer. A total of 275 instances with 12 features were used for this analysis. With Random forest algorithm 74.73% accuracy and 73.63% using XGBoost had obtained in this analysis.**

***Keywords*—Breast cancer, Risk factors, Machine learning, Prediction, Ensemble Learning, Random Forest, XGBoost.**

I. Introduction

Breast cancer is the type of cancer that is extensive among women worldwide and it is a prevalent cause of death. It is frequently diagnosed life-threading cancer in women where in the breast a malignant tumor is developed from cells. In Bangladesh, among women breast cancer is one of the most common cancer. Breast cancer is the second leading cause of cancer death among women after lung cancer. In biological study it is confirmed that, a collection of a large number of separate genetic mutations that collectively change elements of the complex internal signaling system of a cell results into breast cancer [1]. Breast cysts are noncancerous lumps found in one or both the breasts. They are common and occur naturally due to changes in breast with aging and hormonal changes [2].

Some stages of breast cancer are found. Stage 0 which is known as ducal carcinoma in situ (DCIS), the cells have not invaded surrounding tissues which are limited to within the ducts [3]. Then at stage 1, measurement of the tumor grows up to 2 centimeters (cm) across. Either there remains small clusters of cancer cells in the lymph nodes or any lymph nodes have not been affected. In stage 2, the tumor has started to expand to nearby nodes or is 2 to 5 cm across and has not expand to the lymph nodes. In stage 3, several lymph nodes are spread by the tumor or the tumor is bigger than 5 cm. In stage 4, the cancer has expanded to distant organs like the bones, liver, brain, or lungs [4]. Among women, breast cancer is the most common spreading cancer. Adult women are the primary patients of breast cancer who have already reached or are approaching menopause. In terms of new cases, in U.S. 90 percent of women were identified at age 45 or older in 2009 [2].

Survival rates of breast cancer vary worldwide, starting from 80% or over in North America, Sweden and Japan to around 60% in middle-income countries and below 40% in low-income countries (Coleman et al., 2008). Because of the lack of early detection late-stage disease is found within a large proportion of women and the lack of sufficient diagnosis and facilities of treatment there remains low survival rates in less developed countries [2]. In under developed regions, breast cancer represents 14.3% among all cancer deaths which makes this disease the primary cause of cancer death. In more developed regions, after lung cancer breast cancer is the second cause of female cancer death. Among world region breast cancer mortality rates also vary. Such as 20 per 100,000 cancer patients died in Western Africa compared to 6 per 100,000 in Eastern Asia. Early prediction of breast cancer can reduce the risk and people can be aware of this disease [5].

The goal of our analysis is to classify of having recurrence and no-recurrence events accurately using two popular machine learning (ML) approach named Random Forest (RF) and Extreme Gradient Boosting (XGBoost).

The other part of the article is arranged as follows: in section II, III the related works and methodology have been elaborated with a distinguishing destination to the justness of the classier algorithms respectively. In section IV the outcome of this analysis has been clarified with the impulsion to justify the novelty of this exploration work. Finally, this research paper is terminated with section V.

II. Related Works

A system was proposed [6] based on non-contact noninvasive screening device to find breast cancer. Their system used four (Best of Features) features and the support vector machine (SVM) classifier. Many researchers worked on developing a wearable and bearable breast cancer screening (BCS) device that can be replaced the CBE's or regular in hospital screening tests required to detect earlier. The proposed model achieved the accuracy of 83.3%. Baku Bektas et al. [7] found that early prediction of breast cancer is crucial for saving life and taking appropriate measure to control the disease. Breast cancer happens mostly in females and is one of the leading cause of the women's death. They used 139 features of patients

mainly. Random forest, k-stars, selective sensor neural network methods were used in this paper. The result showed 61.85% accuracy. Anusha Bharat et al. [8] proposed a system and predicted breast cancer by using ML algorithm. Multiple tests including a mammogram, ultrasound, MRI and biopsy can diagnose breast cancer. In this paper, Researchers suggest that women have a higher risk of breast cancer who have low levels of vitamin D. They used machine learning based algorithms such as Support Vector Machine (SVM), Decision Tree (CART), Naive Bayes(NB) and k-Nearest Neighbors(kNN). The outcome result showed 90.01% accuracy. Tanaya Padhi et al. [9] analyzed the prediction of breast cancer using the data mining tool WEKA. This paper defined several phases of data mining that was applied on the dataset .The system processed about 6 features of patients mainly. For classification they applied J48, REP Tree and Naive Bayes algorithms. The outcomes result showed 72.02% accuracy for REP Tree.

Here we can see that, in previous researches Random Forest was used but XGBoost wasn't. In our work, we wanted to show a comparison using these two algorithms in this research field. We also changed some parameters to compare the efficiency of the analysis.

III. Methodology

This study can be divided in the following sections:

- Data Collection
- Data Preprocessing
- Data Training
- Application of Algorithms
- Simulation Environment and Tools

The flowchart in Fig. 1 shows the whole working procedure of the study.

A. Data Collection

The data was collected from the UCI Machine Learning Repository and available in online. The dataset contains 275 instances with 12 features and one class of having recurrence or non-recurrence event of breast caner. In the feature table, we see that there are two categories, one is the feature name and the other is the sub category. The features list have shown in Table I along with mean and standard deviation of each feature.

B. Data Preprocessing

Trimmed mean and mode had been used for data preprocessing. Trimmed Mean is the removal of a small fixed percentage of the largest and smallest values before calculating the average [10]. It involves trimming the percentage observations from both ends.

C. Data Training

Around 67% data had been used for training the model and 33% data to be used to test model.

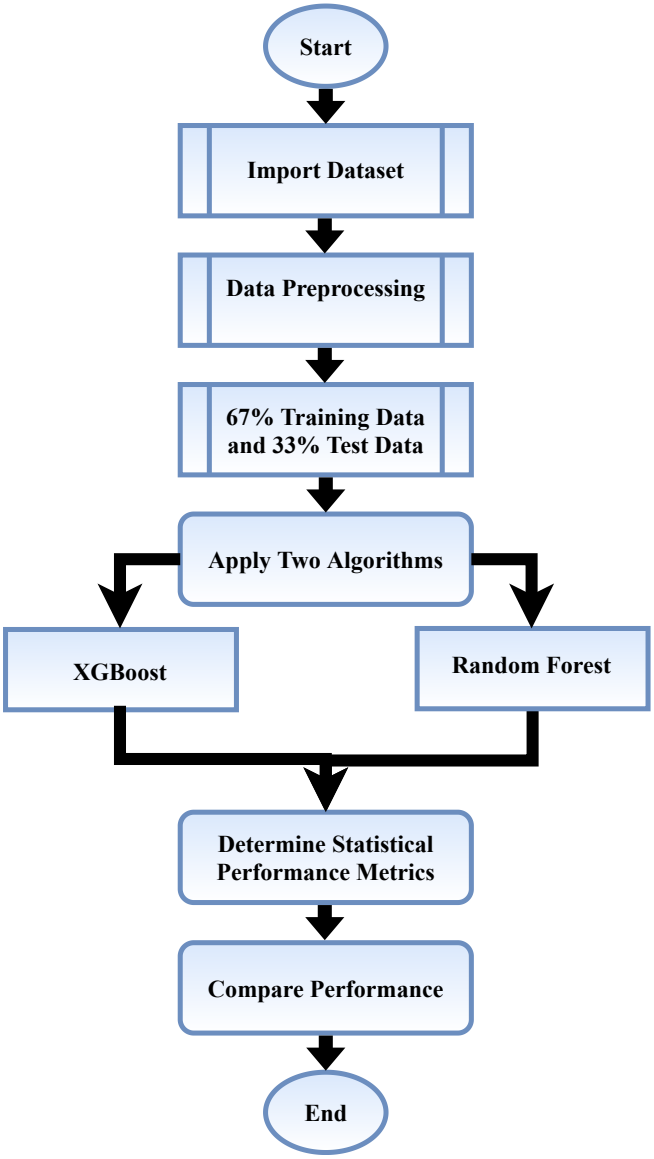

Fig. 1: Work-flow of analysis

D. Application of Algorithms

We implemented two Ensemble ML algorithms:

1) Random Forest (RF): RF is made of many decision trees. There is a direct concern between the outcome and the number of trees in the forest. The higher the number of trees, the more exact results we will get [11]. RF uses C4.5 or J48 as its classifier. In 2001 RF was introduced by Breiman, which combines Bagging with random feature selection for decision trees. RF is a supervised classification algorithm.

2) Extreme Gradient Boosting (XGBoost): XGBoost algorithm is a decision-tree-based ensemble ML algorithm that is used in the gradient boosting framework. Decision trees, in their general form, are easy-to-visualize and fairly interpretable algorithms, but it can be tricky to build intuition for the next-generation of tree-based algorithms [12].

TABLE I: Features List

Features	Subcategory	Data Distribution
Start Age	Minimum: 20	$Mean \pm SD$
	Maximum: 70	46.76 ± 10.22
End Age	Minimum: 29	55.764 ± 10.22
	Maximum: 79	
Start tumor size	Minimum: 0	24.41 ± 10.66
	Maximum: 50	
End tumor size	Minimum: 4	28.41 ± 10.66
	Maximum: 54	
Start_inv_nodes	Minimum: 0	1.38 ± 3.27
	Maximum: 24	
End_inv_nodes	Minimum: 4	3.38 ± 3.27
	Maximum: 26	
Deg-malig	Minimum: 1	2.04 ± 0.75
	Maximum: 4	
Menopause	Premeno:142	51.63%
	Ge40:126	45.81%
	It40:7	2.55%
Node-caps	Yes:53	19.27%
	No:216	78.54%
Breast	Left:148	53.81%
	Right:127	46.18%
Breast-quad	Left_low:10	37.45%
	Right_up:33	12%
	Left_up:94	34.18%
	Right_low:21	7.66%
Irradiate	No:217	78.90%
	Yes:58	21.09%
Class	No-recurrence:193	70.18%
	Recurrence:82	29.81%
SD = Standard Deviation		

E. Simulation Environment and Tools

- Anaconda Python Repository
- Python Machine Learning Libraries

IV. Outcomes

The results of the analysis was analyzed based on a number of statistical metrics given below:

A. Accuracy:

Correctly classified instances is known as accuracy[13]. Here TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

B. Sensitivity (SEN)/ TP Rate/ Recall:

The sensitivity refers to the ability of the test to correctly identify those patients with the disease [13].

$$SEN = \frac{T_p}{P}$$

C. Specificity (SPE)/ TN Rate:

The specificity refers to the ability of the test to correctly identify those patients without the disease [13].

$$SPE = \frac{T_n}{F_p + T_n}$$

TABLE II: Outcome of XGboost and Random Forest

Evaluation Metrics	Algorithms Names	
	XG Boost	Random Forest
Accuracy	74%	75%
Average Precision(%)	70.68	72.44
Average F1-Score(%)	61.55	63.80
Sensitivity	0.94	0.94
Specificity	0.29	0.32
Confusion Matrix	**[[59 4] [20 8]]**	**[[59 4] [19 9]]**

D. F1 Score:

It is also known by F-Measure which can be denoted as [14],

$$FM = 2 \times \frac{PRE \times REC}{PRE + REC} = \frac{2 \times T_p}{2 \times T_p + F_p + F_n}$$

Where, FM = F1 Score / F-measure

E. Precision (PRE):

The ratio of TP divided by summation of TP and FP [14] is called precision [14].

$$PRE = \frac{T_p}{T_p + F_p}$$

F. Explanation of the Analysis:

The outcomes of this study have been described below:

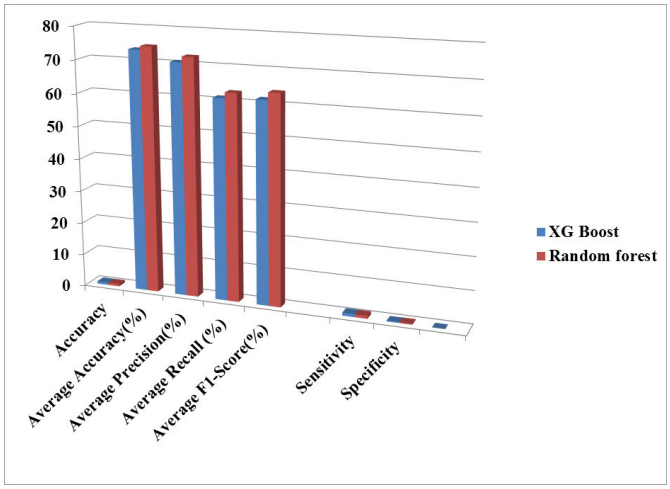

Fig. 2: Bar-chart of performance parameters between XGBoost and Random Forest

In Fig.2 the Bar-chart of performance parameters between XGBoost and RF had been shown.

In Table II, the Random Forest algorithms showed the best result compare to XGboost. On the other hand the Table III showed the comparison between our proposed system and several existing systems based on accuracy as well as the number of instances and attributes were used. The table clarify that our proposed system with Random Forest algorithm is better than the existing systems in terms of performance.

TABLE III: Comparison with Existing Systems

Reference	Sample	No. of Features	Algorithms Name	Average Accuracy
[7]	256	5	SVM	83.3%
[8]	244	139	RF, K-stars, neural network	61.85%
[9]	357	10	SVM, Decision Tree Naive Bayes (NB),KNN	90.01%
[10]	951	6	The J48, REP Tree and Nave Bayes	72.02%
Our Proposed System	275	12	XGBoost and Random Forest	74.18%

V. Conclusion

Most women in Bangladesh suffer from breast cancer due to their unconsciousness. We used two ML algorithms here. By applying the algorithms we have got accuracies for example 74.73% accuracy of Random forest and 73.63% accuracy for XGBoost. We compared the results of our study with other existing systems and found that our system performed better than the existing system. The datasets and techniques were very important for our research paper. We faced several limitations to perform this study. The study was performed based one on a dataset which contains 275 instances and each instance had 12 features. It is one of our limitations. We have seen that a few of the previous works have used more features and samples. We need to use more efficient methods for training and preprocessing the dataset to get a more efficient outcome. In Future, we will increase the sample size of the dataset to make the analysis more reliable, accurate and effective.

References

[1] M. Raihan, Muhammad Muinul Islam, Promila Ghosh, Shakil Ahmed Shaj, Mubtasim Rafid Chowdhury, Saikat Mondal, Arun More, "A Comprehensive Analysis on Risk Prediction of Acute Coronary Syndrome Using Machine Learning Approaches", in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2018, pp. 1 - 6.

[2] "What Is Breast Cancer? — CDC", *CDC*, 2020. [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm. [Accessed: 01- Feb- 2020].

[3] "Breast Cancer", *Wikipedia*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Breast_cancer. [Accessed: 02- Feb- 2020].

[4] "Everything You Need to Know About Breast Cancer", *Healthline*, 2020. [Online]. Available: https://www.healthline.com/health/breast-cancer#prevention. [Accessed: 08- Feb- 2020].

[5] "Breast Cancer - Stages", *Cancer.Net*, 2020. [Online]. Available: https://www.cancer.net/cancer-types/breast-cancer/stages. [Accessed: 08- Feb- 2020].

[6] A. JT, "Breast cancer in sub-Saharan African women", *PubMed - NCBI*, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/7839882. [Accessed: 08- Mar- 2020].

[7] B. Majeed, H. T. Iqbal, U. Khan and M. A. Bin Altaf, "A Portable Thermogram based Non-contact Non-invasive Early Breast-Cancer Screening Device"in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Cleveland, OH, 2018, pp. 1-4.

[8] B. Bekta and S. Babur, "Machine learning based performance development for diagnosis of breast cancer", in *2016 Medical Technologies National Congress (TIPTEKNO)*, Antalya, 2016, pp. 1-4.

[9] A. Bharat, N. Pooja and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis", in *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*, Bangalore, India, 2018, pp. 1-4.

[10] T. Padhi and P. Kumar, "Breast Cancer Analysis Using WEKA"in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2019, pp. 229-232.

[11] "How the Trimmed Mean Is Used", *Investopedia*, 2020. [Online]. Available: https://www.investopedia.com/terms/t/trimmed_mean.asp. [Accessed: 14- Mar- 2020].

[12] "How Random Forest Algorithm Works in Machine Learning", *Medium*, 2020. [Online]. Available: https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674. [Accessed: 14- Mar- 2020].

[13] "XGBoost Algorithm: Long May She Reign!", *Medium*, 2020. [Online]. Available: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d. [Accessed: 14- Mar- 2020].

[14] "Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on?", *Medium*, 2020. [Online]. Available: https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124. [Accessed: 14- Mar- 2020].

[15] M. Islam, M. Raihan, S. Akash, F. Farzana and N. Aktar, "Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques", *Advances in Computational Intelligence, Security and Internet of Things*, vol. 1192, pp. 453-467, 2020. Available: 10.1007/978-981-15-3666-3_37 [Accessed 10 April 2020].