2023-10

A Comparative Study of Machine Learning Algorithms for Detecting Breast Cancer

Khan, Razib Hayat

Independent University, Bangladesh

Draft Version

A Comparative Study of Machine Learning Algorithms for Detecting Breast Cancer

Razib Hayat Khan
Department of Computer science and Engineering
Independent University, Bangladesh
Dhaka, Bangladesh
rkhan@iub.edu.bd

Jonayet Miah
Department of Computer Science
University of South Dakota
South Dakota, USA
Jonayet.miah@coyotes.usd.edu

Md Minhazur Rahman
Department of Physics
University of South Dakota
South Dakota, USA
minhazur.rahman@coyotes.usd.edu

Maliha Tayaba
Department of computer Science
University of South Dakota
South Dakota, USA
Maliha.tayaba@coyotes.usd.edu

ABSTRACT — Breast cancer poses a major hazard to women, with high morbidity and fatality rates, because there is a lack of reliable prognostic models, clinicians find it challenging to develop a treatment regimen that could increase patient life expectancy. There are required to detect breast cancer early stages so the necessary steps should be taken as early as possible to stop this disease first we need more research in this field. So, in this work, we aim to build a machine-learning model which can detect the type of breast cancer whether benign or malignant. Through the detection, we proposed the best model which can detect this outbreak efficiently. In our study, we examined the performance of five machine learning algorithms (XGBoost, Naïve Bayes, Decision Tree, Random Forest, and Logistic Regression) in predicting human health behavior. Among these algorithms, XGBoost had the highest accuracy (95.42%) and performed well in terms of sensitivity (98.5%), specificity (97.5%), and F-1 score (99%). Our findings suggest that XGBoost has promising potential in predicting breast cancer, but further research is needed to develop and apply it for commercial use in the healthcare industry.

Keywords — Breast cancer, Machine learning, Artificial Intelligence, XGBoost

I. Introduction

Breast cancer is the leading cause of death for women currently in this world. Only 264000 cases of breast cancer in the USA are diagnosed per year according to WHO (World Health Organization). Moreover, Black women died more than White women from this disease per year. The greatest method to handle breast cancer outcomes is early recognition [1]. Machine learning algorithms are important for predicting breast cancer because they can process large amounts of data and identify patterns that might not be evident through traditional methods. By analyzing patterns in medical imaging data, for example, these algorithms can help identify early signs of breast cancer, enabling medical professionals to intervene and provide prompt treatment. Additionally, these algorithms can help improve accuracy and reduce human error, which can be critical in the early detection and diagnosis of breast cancer. Overall, the use of machine learning algorithms has the potential to improve breast cancer diagnosis and treatment, ultimately leading to better patient outcomes. The early detection of breast cancer is greatly facilitated by computer-aided detection or diagnosis (CAD) systems, which can also be utilized to lower the death rate among females. The major goal of this study is to utilize the most recent developments in the creation of CAD systems and associated approaches [11]. There are so many machines learning approaches have done to detect breast cancer, but the problem is in the making decision of whether the model can reliably classify the cancer type and make the decision for early detection. Our study fully worked on early age detection because we are working on the dataset which contains pre-symptoms of breast cancer which help us to give the decision and necessary measure that should take by the physician. Early prediction can save the life effectively. However, Machine learning is nothing more

than teaching computers to learn and function on their own, without the aid of specific software or instruction. Therefore, using the training data, it is possible to determine whether a person has breast cancer. Breast cancer sufferers rank fourth in terms of disease frequency among women between the ages of 20 and 29. Malignant thyroid growth, melanoma, and lymphoma are the top three diseases. Some risk factors, including family ancestry, are impossible to avoid. Other risk factors, like smoking, are modifiable. As implied by the name, artificial intelligence allows robots to learn from massive amounts of information that can be used in calculations to create expectations [10,2]. Machine learning can make the diagnosis easier and more proper for the physician using the machine learning algorithm and statistical analysis which also can reduce the cost of treatment of the disease. Our primary goal is to give a convenient model for healthcare and patients to check the chance of getting breast cancer in one individual. This machine learning takes place great development in the health sector, especially in heart disease and cancer with the help of different model uses of machine learning which is sustainable in the long term to give a proper diagnosis to humankind [9]. AI also adds machine learning which gives a good parallel model and solution for the patient to get an appropriate decision. In our work, we proposed the best machine learning model which can predict breast cancer at an early stage and provide which machine learning model is sustainable for this disease while most of the related work just focused on predicting breast cancer.

II. LITERATURE REVIEW

B.fu et al.[3]. In this task, the author implies that AI techniques were employed to develop a PC-aided platform for lung cancer progression. The architecture consists of three stages: highlight extraction, stage of include selection, and phase of aggregation. Different wavelet capabilities have been employed to include extraction/determination to identify which produced the highest exactness level. K-Closest Neighbor Grouping has been developed or applied for the order. For grouping, accuracy values of over 96% have been attained, demonstrating the suggested approach's advantages. The results that were seen in the previous area demonstrate the strategy's capacity for malignant development grouping. Up to now, there has been the practice with 96.58% precision. To increase the strategy's accuracy, additional experiments will be conducted with larger wavelet capacities.

Mariam et. al. [4] The Author compares the accuracy of Naive Bayes and K Nearest Neighbors, two alternative classifiers, for the classification of breast cancer. KNN obtained 97.51% accuracy with the lowest error rate compared to the Naive Bayes Classifier's 96.19% accuracy.

Mirsadeghi et al. [5] The Author used different types of machine-learning algorithms. This analysis tries to concentrate on the results in several MBCA prognostic and diagnosis-related areas. We begin by introducing the passengers and drivers predicted by SVM, ANN, RF, and EARN. Secondly, biological conclusions were drawn from estimations. We conducted a pathway enrichment analysis (PEA) using the ReactomeFIVIz tool with an FDR of 0.03, focusing on the top 100 genes predicted by EARN. The analysis identified several genes including NCOR1, TBL1XR1, SIRT4, KRAS, CACNA1E, PRKCG, GPS2, SIN3A, ACTB, KDM6B, PRMT1, HDAC3, ABAT, GRIN1, PLCB1, and KPNA2. To further evaluate the results, we examined 983 primary breast invasive carcinoma (BRCA) tumor samples from the Cancer Genome Atlas, they compare MBCA results to other outputs (TCGA). When comparing the results, for MBCA, EARN achieves a ROC-AUC of 99.24%, while for BRCA, it achieves 99.79%. In each example, this statistical finding outperforms three separate classifiers.

Mohammed et al. [6] In a study using the Breast Cancer Wisconsin Diagnostic dataset, five different machine learning algorithms (Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5), and K-Nearest Neighbors (KNN)) were applied to predict and diagnose breast cancer. The performance of each algorithm was evaluated by analyzing their confusion matrices, accuracy, and precision. The main objective of the study was to determine the most effective algorithm. The results showed that support vector machines had the highest accuracy (97.2%) and outperformed all other classifiers in the analysis.

Kurian et al. [7]. The author works on the three main steps of the breast cancer prediction procedure feature extraction, Performance evaluation, and classification using machine learning. For the classification process, ten DNA sequence characteristics were extracted from three types of sequences, including ORF (Open Reading Frame) count, the average count of individual nucleobases A, T, C, G, AT, and GC-content, AT/GC composition, G-quadruplex frequency, and MR (Mutation Rate). Additionally, the sequence type was added to the set of features as a target attribute with the values 0, 1, and 2 for classes 1, 2, and 3 correspondingly. Among all the supervised models, the decision tree machine-learning method had the highest classification accuracy of 94.03%. The performance of the classification model was evaluated using precision, recall, F1-score, and support values. The

F1 score was found to be the most relevant metric to assess the accuracy of the classification.

Milon et al. [8] In this work, the authors propose the Support Vector Machine and K-Nearest Neighbors, in this study, we propose a new approach for predicting breast cancer using supervised machine learning algorithms. The system employs 10-fold cross-validation to ensure accurate results. The Wisconsin breast cancer diagnosis dataset from the UCI machine learning repository was used as the training data. The performance of the system was evaluated using various metrics, including accuracy, sensitivity, specificity, false discovery rate, false omission rate, and Matthew's coefficient of correlation. Our proposed approach yielded better results for both training and testing. Specifically, using the support vector machine and K-nearest neighbors separately, we achieved an accuracy and specificity of 98.57% and 97.14%, respectively.

Kabiraj et al. [9] In this work the authors propose using two well-known ensemble machine learning methods to examine a breast cancer dataset and forecast the development of breast cancer. Breast cancer was predicted using Extreme Gradient Boosting (XGBoost) and Random Forest. For this research, a total of 275 examples with 12 features were used. In this investigation, accuracy rates of 74.73% using the Random Forest method and 73.63% using XGBoost were both achieved.

III. Methodology

Our study involved the analysis and monitoring of the subject health using collected data. We applied various machine learning algorithms to identify patterns and make predictions. The main objective of the study was to develop classification models that can accurately distinguish between benign and malignant cancers. The research also focused on the challenges associated with chest data analysis and highlighted potential areas for future advancements in this field. Data collection is the initial stage of this study's technique, followed by a plan for preprocessing the dataset. The dataset was utilized for both training and testing. a variety of classifiers, including XGBoost, Logistic Regression, Random Forest, Decision Tree, and Naive Bayes. By analyzing the machine learning model, we can decide who has the maximum and minimum chance to get breast cancer. To find the most accurate result we are doing data preprocessing which is in the next section and Figure 1 shows the overall layout of the suggested study.

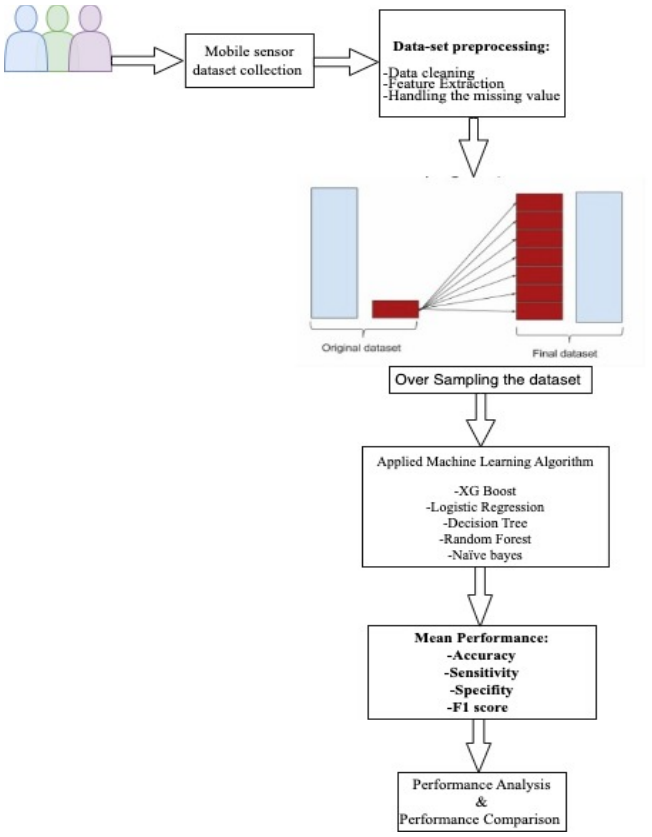

Fig. 1. The overview of the study

A. Dataset collection and Data Preprocessing

In this paper, we will analyze the dataset which contains signs of the body. The breast cancer dataset was acquired from the University of Wisconsin Hospitals. The dataset in this paper consists of 31 attributes and 8670 instances. Data collection is the initial stage of this study's technique, followed by a plan for preprocessing the dataset. Moreover, we are doing feature extraction in our data processing part to reduce the dataset's dimensionality. The dataset was used for both training and testing a variety of classifiers, including XGBoost, Logistic Regression, Random Forest, Decision Tree, and Naive Bayes. By analyzing the machine learning model, we compared which model works perfectly and classify the breast cancer case and non-breast cancer cases. This dataset can be utilized for various purposes such as basic chest x-ray monitoring, testing for various arrhythmias, and examining the effects of exercise on the x-ray, among others. In figure 2 We have shown how the dataset has a relationship with each of the attributes. We are also shown How positive features related to each other cause breast cancer. It has highly positive symptoms and mathematical relationships.

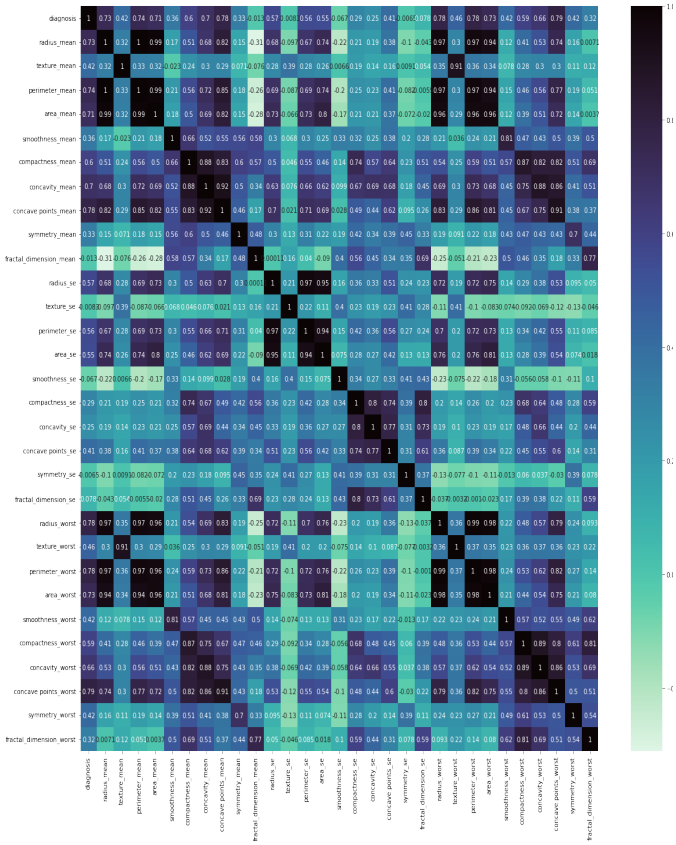

Fig. 2. Correlation between attributes

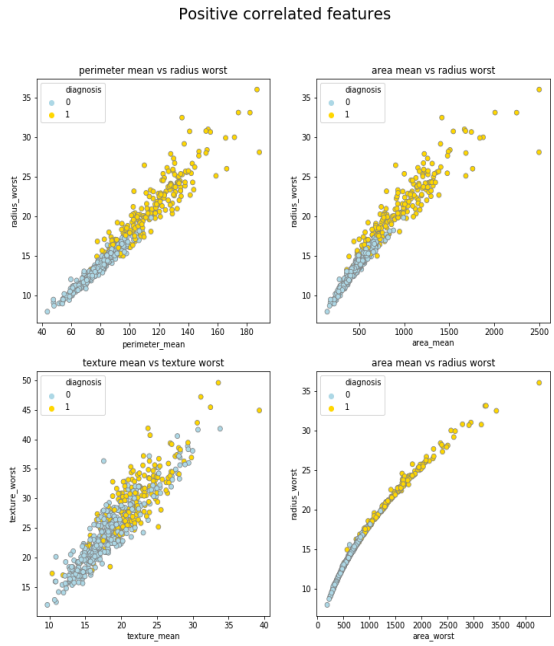

Fig. 3. Positive Symptoms correlated features

B. *Validation Process*

choosing the proper validation technique for specific Datasets is essential. The hold-out validation method is the most effective strategy. we are training 80% of the dataset and testing 30% of it, we applied a holdout validation technique to get good results.

Furthermore, we measured the accuracy, sensitivity, specificity, and F1- Score by the implied confusion Metrix. A thorough examination is provided in the visualization and display of the performance indicators bar graphs.

IV. Results And Discussion

In this paper, Fig 4 depicts the performance analysis of various machine learning models used to predict physical fitness for athletics, including XG Boost, Decision Tree, Logistic Regression, Random Forest, and Naive Bayes. We used different types of performance metrics to evaluate our models such as accuracy, sensitivity, specificity, and F1 score. One of the most important performance metrics to evaluate how accurately the machine learning model performs is accuracy. Although logistic regression has the lowest accuracy of all the models in our chosen model (57.07%), XGBoost demonstrated superior accuracy in comparison to other models with a 94.92% accuracy rating. We can determine the model performance by analyzing the sensitivity and specificity. Sensitivity and specificity for all our selection model's performance show better results and are nearly identical, at 99% and 98%, respectively. The F1 score is another vital metric to evaluate models to analyze performance. In this research, most of the models show promising results except logistic regression (57.07%). After examining all the results from each machine learning model, we determined that XGBoost performed better than other models in terms of accuracy (94.92%), sensitivity (98.5%), specificity (97.5%), and F1 score (99%). For the dataset we have chosen, XGBoost performs better than other models at predicting Breast cancer fitness. Therefore, we may utilize XGBoost to estimate Breast cancer and receive accurate and encouraging results.

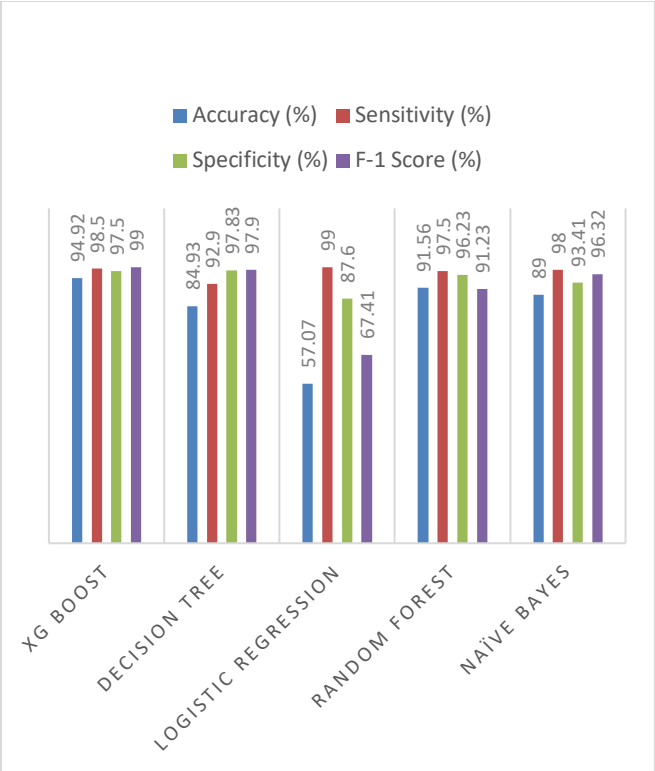

Fig. 4. Performance analysis of machine learning models

V. CONCLUSION AND FUTURE WORK

Our study was done with the data of 8670 Unique ID participants We are using many machines learning models to detect breast cancer, but we aimed to build a sustainable model to predict breast cancer and we got the XGBoost which performs very well on our dataset. Many machine learning projects found XG-Boost to outperform any other machine learning algorithm, this was also true within our study. the use of machine learning algorithms for predicting breast cancer has shown tremendous promise in recent years. By leveraging these algorithms to process and analyze large amounts of medical imaging and patient data, healthcare professionals can better identify early warning signs of breast cancer and intervene with treatment in a timely and effective manner. We were able to get 94.92% accuracy, 98.5% in sensitivity, 99% in specificity, and 99% F-1 score using XGBoost. The high accuracy rates achieved by XGBoost in breast cancer prediction, along with their ability to reduce human error and improve patient outcomes, make them a valuable tool in the fight against this disease. With continued research and development in this area, machine-learning algorithms will undoubtedly play a critical role in improving breast cancer detection, diagnosis, and treatment in the years to come.

REFERENCES

[1] Walberg WH, Mangasarian OL, “Multi-surface method of pattern separation for medical diagnosis applied to breast cytology”, Proc Natl Acad Sci U S A, 1990, 87(23):9193-6, Doi: 10.1073/pnas.87.23.9193

[2] Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader benyettou, “Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules”, International Journal of Computer Applications, Volume 62 - No. 1, January 2013

[3] B. Fu, P. Liu, J. Lin, L. Deng, K. Hu, and H. Zheng, "Predicting Invasive Disease-Free Survival for Early-Stage Breast Cancer Patients Using Follow-Up Clinical Data,", IEEE Transactions on Biomedical Engineering, vol. 66, no. 7, pp. 2053-2064, July 2019.doi: 10.1109/TBME.2018.2882867

[4] Mariam Amrane, Saliha Oukid, Ikram Gagaoua, and Tolga Ensari, "Breast cancer classification using machine learning", Electric Electronics, Computer Science, Biomedical Engineerings, Meeting, 2018

[5] Mirsadeghi, L., Haji Hosseini, R., Banaei-Moghaddam, A.M, “EARN: an ensemble machine learning algorithm to predict driver genes in metastatic breast cancer”, BMC Med Genomics **14**, 122, 2021, https://doi.org/10.1186/s12920-021-00974-3

[6] Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar, Rachida Ait Abdelouhahid, Olivier Debauche, “Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis”, Procedia Computer Science, Volume 191, 2021, Pages 487-492, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.07.062.

[7] Kurian B, Jyothi V, “Breast cancer prediction using an optimal machine learning technique for next generation sequences”, Concurrent Engineering. 2021;29(1):49-57. doi:10.1177/1063293X21991808

[8] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors", IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017, pp. 226-229, DOI: 10.1109/R10-HTC.2017.8288944.

[9] S. Kabiraj et al., "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm”, 11th International Conference on Computing, Communication and Networking Technologies,

2020, pp. 1-4, DOI: 10.1109/ICCCNT49239.2020.9225451.

[10] Kayyum, Salsavil, Miah, jonayet, Shadaab, Anwar., Islam, Minazul Islam., Islam, Majharul., Nipun, shah Ashiul Abed., Rakib Rahat,md Abdur., Faisal, Faiz Al, "Data Analysis on Myocardial Infarction with the help of Machine Learning Algorithms considering Distinctive or Non-Distinctive Features", International Conference on Computer Communication and Informatics, IEEE, 2020.

[11] Islam, Md., nipun, Shah Ashisul Abed., Islam, Majharul, Rakib Raht, Md Abdur., Miah, Jonayet., Kayyum, Salsavil., Shadaab, Anwar., Faisal, Fiaz al, "An Empirical Study to Predict Myocardial Infarction Using K-Means and Hierarchical Clustering", International Conference on Machine Learning, Image Processing, Network Security and Data Sciences, Springer, Singapore, 2020.

[12] M. S. Yarabarla, L. K. Ravi and A. Sivasangari, "Breast Cancer Prediction via Machine Learning," International Conference on Trends in Electronics and Informatics, 2019, pp. 121-124, DOI: 10.1109/ICOEI.2019.8862533.

[13] Ojha, U, Goel, S, "A study on prediction of breast cancer recurrence using data mining techniques", International Conference on Cloud Computing, Data Science & Engineering-Confluence, pp. 527-530

[14] Chaurasia, V, Pal, S., Tiwari, B. B., "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms & Computational Technology, 12(2), 119-126.

[15] J. Miah, M. Mamun, M.M. Rahman, M. I. Mahmud, M. H. B. Nasir, S. Ahmad, "MHfit: Mobile Health Data for Predicting Athletics Fitness using Machine Learning Models", International seminar on machine learning, Optimization, and Data Science (ISMODE), 2022

[16] Yash Amethiya, Prince Pipariya, Shlok Patel, Manan Shah, "Comparative analysis of breast cancer detection using machine learning and biosensor", Intelligent Medicine, Volume 2, Issue 2, 2022, Pages 69-81, ISSN 2667-1026,