International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-11 Issue-3, September 2022

Breast Cancer Prognosis using Machine Learning Ensemble Techniques

Disha H. Parekh, Vishal Dahiya

Abstract*: According to WHO, breast cancer is the disease that affects people the most frequently and most dangerously in the world. Researchers are paying more attention to breast cancer because of how deadly it is and how early detection can prevent it. Since the advent of supervised machine learning algorithms, the early detection of breast cancer has advanced. The usage of several machine learning techniques as well as ensemble algorithms is demonstrated in the study. The outcomes were extremely precise, allowing for the best-possible cancer prediction. The paper's modest goal is to save people suffering from the disease by enabling them to know if the detected tumour is cancerous or non-cancerous, being Malignant. It focuses on early diagnosis of breast cancer. This paper would be useful and aiding for those who are novel researchers in prediction and diagnosis of breast cancer using machine learning.

Keywords*: Breast Cancer Prediction, Machine Learning, Ensemble XG Boost, AdaBoost.

I. INTRODUCTION

According to the statistics of World Health Organization during 2020, breast cancer has been the most prevailing disease of the world. It has mentioned that during 2020, 2.3 million of women across world has been diagnosed with breast cancer and by the end of the month, almost 7.8 million of women had been surviving in the world with the record of past five years [1]. Breast cancer has been an invasive disease since 1930 and is right now an area of attraction for researchers to infringe this invasive disease and bring awareness amongst the population with early detection and diagnosis of the disease. Breast cancer disease can be treated effectively when detected in its early stages. This early detection is an area where many researchers are working today. There are several researchers working on medicine development and its discovery for eliminating the meddlesome disease. Thus cancer biology is found to be gearing up the interest of researchers across the world. Breast cancer is not a contiguous or transferrable disease. It is a disease spreading widely due to mutations in cells. It is not a viral or bacterial infectious disease but is mutant to changes in gene material, particularly protein sequencing. Though breast cancer is mostly observed in females, some 1% of males are also victims to this disease. The disease is caused by a lump in the breast. This lump is painless but it is abnormal and hence should be treated urgently by consulting surgeons. There are basic two genes called Breast Cancer Gene 1 and Gene 2, usually referred as BRCA1 and BRCA2 which produce proteins to remove ruptured DNA. These genes help in suppressing tumors in the body. But any pathogenic disorders or any mutations in the gene sequence of any of these genes, leads to breast cancer. About 13% of women in the general population will develop breast cancer sometime during their lives (N et al., 2020). By contrast, 55%–72% of women who inherit a harmful BRCA1 variant and 45%–69% of women who inherit a harmful BRCA2 variant will develop breast cancer by 70–80 years of age [2], [3]. Breast cancer is represented in two different ways. The lumps in breast cancer can be either cancerous or non-cancerous. All those lumps which are non-cancerous are usually called as Benign type which means there exists no cancer. While all those lumps which are cancerous in nature are termed as Malignant tumors. These malignant tumors need diagnosis using biopsy of the lump mass or can be diagnosed using breast imaging. The objective of the WHO Global Breast Cancer Initiative (GBCI) is to reduce global breast cancer mortality by 2.5% per year, thereby averting 2.5 million breast cancer deaths globally between 2020 and 2040. Reducing global breast cancer mortality by 2.5% per year would avert 25% of breast cancer deaths by 2030 and 40% by 2040 among women under 70 years of age. The three pillars toward achieving these objectives are: health promotion for early detection; timely diagnosis; and comprehensive breast cancer management [4]. There is an organization called National Breast Cancer Coalition (NBCC) which works dedicatedly towards the end of breast cancer through action and advocacy. According to their study carried out in 2022, breast cancer is found to be the most common disease where there are estimated to be 2,87,850 new cases of invasive breast cancer in women and 2710 new cases in men. They have even shown that there will be an additional 51,400 cases of ductal carcinoma in situ diagnosis in women. The claim made by the NBCC is depicted in the below figure which even justifies the rise in mortality rate as age increases.

Figure I.1 An Image Showing the Statistics of Breast Cancer From 1975 to 2022. [5]

Manuscript received on 31August 2022 | Revised Manuscript received on 02 September 2022 | Manuscript Accepted on 15 September 2022 | Manuscript published on 30 September 2022.
* Correspondence Author
Prof. Disha Harshadbhai Parekh *, Department of Computer Science, Indus University, Ahmedabad, India. Email: disha.hparekh213@gmail.com
Prof. Dr. Vishal Dahiya, Department of Computer Science, Indus University, Ahmedabad, India. Email: cs.hod@indusuni.ac.in

Breast cancer diagnosis has been a major concern for decades and thus, there are several research communities working on this area for finding solutions to the cancer, its treatments, or maybe even drug discovery. The diagnosis and treatment of breast cancer has been an area of interest for researchers of the computer science community today. Those interested in biomedical research or life sciences bioengineering are focusing on healthcare industries today and its solvability implementing computer fundamentals like Artificial Intelligence (AI), Machine Learning, Block chain Technology, and Deep Learning. Machine Learning is a recent technology that is used to train machines with various algorithms in order to improve automatically through learning. As discussed in the paper, benign tumors are non-cancerous by nature. The diagnosis of tumors in benign or malignant conditions can prevent a human from unnecessary treatments if found benign. The bifurcation of breast cancer into benign or malignant, just in order to avoid unnecessary surgery and treatment if not cancerous, is much research specific today. Due to its uniqueness in feature categorization, breast cancer diagnosis of complex datasets uses Machine Learning (ML) as one of the most prevailing methods. There are certain ML algorithms classified into supervised and unsupervised learning algorithms and its enhancements as ensemble learning algorithms. Researchers opt for various sets of algorithms and even sometimes use more than one learning algorithm to help in analysis of datasets. In the paper [6], use of Machine Learning supervised algorithms on Wisconsin Diagnostic Breast Cancer Dataset (WDBC) was carried out showing the accuracy and F1 score for several algorithms. The code was done in the R environment using RStudio Framework. The WDBC dataset contains around 569 observations only which was found to be very less for training a machine learning model. Secondly the diagnosis was carried out using Statistical Programming Language R which was found to be visually a bit less in exploration of data analysis as compared to Python. The paper focused on basically three algorithms namely Support Vector Machine (SVM), General Linear Model (GLM) and Single Neural Network (NNET) where the accuracy obtained by SVM was best among all three algorithms. Moving an extra mile ahead, this paper focuses on Python technology and Supervised Machine Learning Algorithms. This paper focuses on the result and analysis between different ensemble classifiers as well. A comparative representation of supervised with ensemble is also justified with the results obtained.

II. MATERIALS AND METHODS

To carry out the prediction of breast cancer occurrence in terms of malignant “M” and benign “B”, a mammography dataset of breast masses known as CBIS which is Curated Breast Imaging Subset of Digital Database for Screening Mammography, usually known as CBIS - DDSM, was used. DDSM is a database containing 2620 scanned film mammography studies. The images here are decompressed and converted in DICOM format, which was then used to get access to the .csv file. Here the dataset used consisted of records of “B” and “M” kinds of breast cancers for exactly 1319 patients. Further the analysis of breast cancer with various machine learning supervised algorithms were carried out. The results obtained have been discussed in chapter 4 of the paper. Though the algorithms gave better accuracy, using ensemble algorithms was recommended. Hence, two ensemble mechanisms, AdaBoost and Xtreme Gradient (XG) Boost were used. The coding of the algorithms for the CBIS-DDSM dataset was done on Python Platform, online, using Google COLAB. The model constructed is depicted in the figure below:

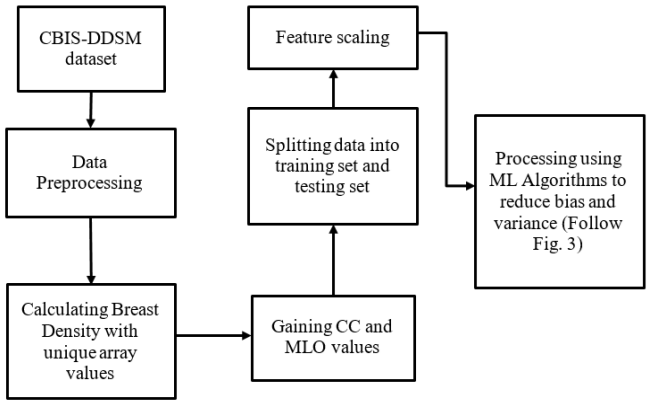


Figure II.1 Processing of CBIS-DDSM dataset

The Algorithms mainly targeted, experimented and analyzed here are Ensemble Random Forest algorithm, Ensemble AdaBoost algorithm and Ensemble XGBoost algorithm. Figure 3 represents the data pre-processing steps and the outcomes targeted after using the mentioned supervised and ensemble techniques.

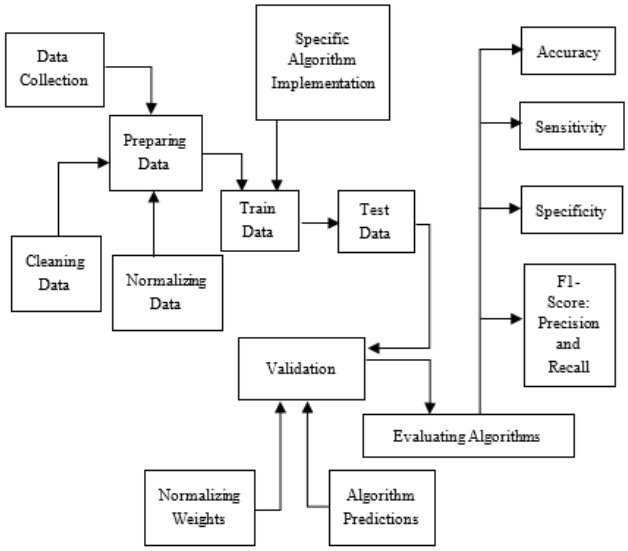


Figure II. 2 Data Processing Stages and algorithm implementation model to achieve desired result

III. RESULTS AND DISCUSSION

The conduction of data analysis for predicting breast cancer was carried out on an unscaled and then feature scaling data. Feature scaling of data is important in Machine Learning as it helps in standardizing all the independent variables or features present in the data for a fixed range. Though, sometimes the problem of outliers may pop up due to this scaling, which may lead to unwanted results.

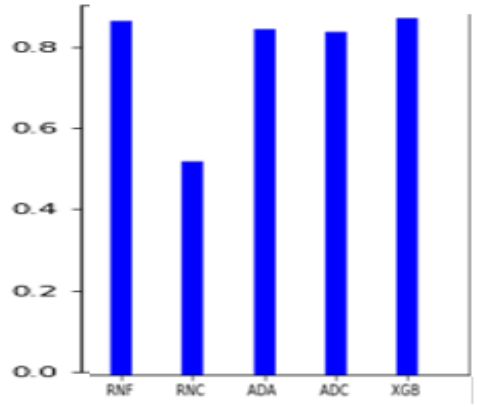
The results were obtained in terms of Accuracy, Precision, Recall and lastly the F1-Score. Table 1 depicts the results of each algorithm.

Table III-1 Comparison of all the supervised algorithms in terms of Accuracy, Precision, Recall and F1-Score

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest Ensemble Classifier	85.05	91.06	84.02	0.8843
AdaBoost Ensemble Classifier	83.24	87.76	85.49	0.8661
XGBoost Ensemble Classifier	85.86	92.30	86.15	0.8912

After fetching the results, the study was concluded with comparative analysis of each of the algorithmic values with scaled / unscaled except the two ensemble algorithms which are boosting algorithms and hence does not require scaling the values. The below figure 4 indicates the bar chart showing the comparative analysis of all the algorithms.

Table III-1 Comparison of ensemble algorithms with the obtained accuracy values



IV. CONCLUSION

Breast Cancer prediction has been a very novel topic for data analysts as it helps diagnosing breast cancer and today many researchers are also working on drug discovery on the basis of the prediction. This paper aims at predicting breast cancer with accuracy ratio of supervised algorithms in machine learning. While carrying out the experiment on the CBIS-DDSM dataset, it was found that supervised learning algorithms gave pretty good results but using ensemble algorithms would enhance the accuracy and finally the F1-score. Hence, Random Forest ensemble, AdaBoost ensemble and XGBoost ensemble were used and to our hypothesis, it proved to be better. As depicted in the result shown in table 1, F1-score for Ensemble methods were found to be comparatively higher than the normal supervised algorithms. From the analysis done, it can be concluded that using XGBoost Ensemble technique would enhance the performance of the model and will lead to better F1-score. F1-score, is basically a harmonic mean between precision and recall and is primarily used to compare the performance of classifiers. Better the F1-Score, better is the classification of observations into perfect classes. F1-score lies between 0 to 1, and the score obtained in XGBoost is 0.8912 which is better than every other algorithm used, which justifies optimal and better classification of the observations in the dataset. This paper can aid researchers in carrying out their study on breast cancer prediction. It may further help in diagnosing the mammography directly. Here the mammographic dataset was converted into a csv file and then the prediction was carried out, but any researcher can directly diagnose mammography also without converting it into a csv. The paper can be enhanced in future to show the AUC scores of predicted probabilities with a mammographic dataset.

REFERENCES

1. Breast cancer. (2021, March 26). WHO | World Health Organization. Retrieved April 18, 2022, from https://www.who.int/news-room/fact-sheets/detail/breast-cancer
2. Chen, S., & Parmigiani, G. (n.d.). Meta-analysis of BRCA1 and BRCA2 penetrance. journal of Clinical Oncology 2007, 25(11), 1329–1333. [CrossRef]
3. Kuchenbaecker, K., JL, H., & DR, B. (n.d.). Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. JAMA 2017, 317(23), 2402–2416. [CrossRef]
4. DeSantis, C., Bray, F., Ferlay, J., Lortet-Tieulent, J., Anderson, B., & Jemal, A. (n.d.). International Variation in Female Breast Cancer Incidence and Mortality Rates. Cancer Epidemiol Biomarkers Prev. 2015, 24(10), 1495-506. [CrossRef]
5. Breast Cancer Statistics | Facts & Figures | NBCC. (n.d.). National Breast Cancer Coalition. Retrieved April 19, 2022, from https://www.stopbreastcancer.org/information-center/facts-figures/
6. Parekh, D. H., & Dahiya, V. (2021, October). Predicting breast cancer using machine learning classifiers and enhancing the output by combining the predictions to generate optimal F1-score. Biomedical and Biotechnology Research Journal (BBRJ), 5(3), 331 - 334. 10.4103/bbrj.bbrj_131_21. [CrossRef]

AUTHORS PROFILE

Prof. Disha Harshadbhai Parekh, is working as an Assistant Professor with Indus University, Ahmedabad. She is passionate for research and carries an extensive ability to perform extraordinarily in research aspects. She is currently motivating her students to write papers during their UG level and also helping several PG students for writing quality papers in the field of computer science. She has completed her MCA and M.Phil in Computer Science. Her area of interest lies with cloud computing, security issues with cloud and cybercrimes, data science and NLP approaches. She is pursuing Ph.D. in the area of Bioinformatics where breast cancer analysis using ML approach is targeted. She is in the education field since 2009 and she has a good databank of research papers under several headings, which can be fetched from scholar and can be contacted on disha.hparekh213@gmail.com

Prof. Dr. Vishal Dahiya, is working as a HOD of Computer Science Department with Indus University. She has wide experience of 18+ Years. She has completed her Ph.D. in computer science from Sardar Patel University and currently is guiding more than 10 research scholars. She is acting as a chair person of research committee of Indus University for Computer Science and Engineering Departments. She is a motivator and a mentor for students and faculties interested in research. Her research area focuses on Image Processing and Big Data. She has been widely renowned for her literally work on image processing. She can be approached at cs.hod@indusuni.ac.in