IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657
Volume 6, Issue 2, 2018

Prediction of Breast Cancer using Machine Learning

Ambala Sreedhar
N Ch Ravi
B Pannalal

Abstract:

Breast cancer is the most common reason for deaths due to cancer. It is very necessary to detect cancer at early stages. There are various Machine Learning techniques available for the purpose of diagnosis of breast cancer data. This paper presents a Machine Learning model to perform automated diagnosis for breast cancer. This method employed CNN as a classifier model and Recursive Feature Elimination (RFE) for feature selection. Also, five algorithms SVM, Random Forest, KNN, Logistic Regression, Naïve Bayes classifier have been compared in the paper. The system was experimented on BreaKHis 400X Dataset. The performance of the system is measured on the basis of accuracy and precision. Activation function such as ReLu have been used to predict the outcomes in terms of probabilities.

Keywords: Artificial Intelligence, Breast Cancer, Computing Methodologies, Genetic Algorithm, Machine Learning

INTRODUCTION

According to the Centers for Disease Control and Prevention (CDC)Trusted Source, breast cancer is the most common cancer in women. Breast cancer survival rates vary widely supported by many factors. Two of the most important factors are the type of cancer women have and the stage of cancer at the time they receive a diagnosis. Breast cancer is cancer that develops in breast cells.

Assistant Professor [1,2]

Nagole institute of Engineering and technology

ISSN 2347–3657
Volume 6, Issue 2, 2018

Typically, the cancer forms in either the lobules or the ducts of the breast. Cancer also can occur within the adipose tissue or the fibrous connective tissue within your breast. The uncontrolled cancer cells often invade other healthy breast tissue and may visit the lymph nodes under the arms Doctors say that breast cancer happened due to abnormal growth of cells in the breast and these cells spread in size like Meta Size from breast to lymph nodes or the other parts of the body also. Hence it is necessary to detect and stop the growth of these unwanted cells as early as possible to avoid the next phase consequences. If a tumor is diagnosed then the first step taken by the doctor is, they check whether the tumor is Benign or Malignant. Because the treatment and prevention methods of both the tumors are different. Benign cells are neither cancerous and nor spread but Malignant cells are cancerous and can spread to other parts of bodies. The problem with this disease is, there is no such proper diagnostic machine is present to detect cancer in the early phase so the person can start the treatment as early as possible and try to stop the growth of unwanted cells or tumors

Breast cancer is considered a multifactorial disease and the most common cancer in women worldwide [1, 2] with approximately 30% of all female cancers [3, 4] (i.e. 1.5 million women are diagnosed with breast cancer each year, and 500,000 women die from this disease in the world). Over the past 30 years, this disease has increased, while the death rate has decreased. However, the reduction in mortality due to mammography screening is estimated at 20% and improvement in cancer treatment is estimated at 60% [5 , 6].

Diagnostic mammography can assess abnormal breast cancer tissue in patients with subtle and inconspicuous malignancy signs. Due to a large number of images, this method cannot effectively be used in assessing cancer suspected areas. According to a report, approximately 50% of breast cancers were not detected in screenings of women with very dense breast tissue [7]. However, about a quarter of women with breast cancer are diagnosed negatively within two years of screening. Therefore, the early and timely diagnosis of breast cancer is crucial [8].

ISSN 2347–3657
Volume 6, Issue 2, 2018

Most mammography-based breast cancer screening is performed at regular intervals - usually annually or every two years - for all women. This “A fix screening program for everyone” is not effective in diagnosing cancer at the individual level and may impair the effectiveness of screening programs [9]. On the other hand, experts suggest that considering other risk factors along with mammography screening can help a more accurate diagnosis of women at risk [9 - 11]. Moreover, effective risk prediction through modeling can not only help radiologists in setting up a personal screening for patients and encouraging them to participate in the program for early detection but also help identify high-risk patients [12 , 13].

Machine learning, as a modeling approach, represents the process of extracting knowledge from data and discovering hidden relationships [14], widely used in healthcare in recent years [15] to predict different diseases [16 - 18]. Some studies only used demographic risk factors (lifestyle and laboratory data) in predicting breast cancer [19 , 20], and several studies predicted based on mammographic stereotypes [21] or used data from patient biopsy [22]. Others showed the application of genetic data in predicting breast cancer [23].

A major challenge in predicting breast cancer is the creation of a model for addressing all known risk factors [24 - 26]. Current prediction models might only focus on the analysis of mammographic images or demographic risk factors without other critical factors. In addition, these models, which are accurate enough for identifying high-risk women, could result in multiple screening and invasive sampling with magnetic resonance imaging (MRI) and ultrasound. The financial and psychological burden could be experienced by patients [27 - 29].

The effective prediction of breast cancer risk requires different factors, including demographic, laboratory, and mammographic risk factors [24 , 25 , 30 , 31]. Therefore, multifactorial models with many risk factors in their analysis can be effective in assessing the risk of breast cancer through more accurate analysis [32 , 33].

ISSN 2347–3657
Volume 6, Issue 2, 2018

The current study aimed to predict breast cancer using different machine learning approaches considering various factors in modeling.

In this analytical study, the database was obtained from a clinical breast cancer research center (Motamed cancer institute) in Tehran, Iran. The research was conducted in 4 stages: data collection, data pre-processing, modeling, and model evaluation.

Data Collection

In the first stage, 5178 records of people, referred to the research center over the past 10 years (2011-2021), were prepared retrospectively. Each record covered 24 features (11 demographic features, 9 laboratory features, and 4 mammography features) (Table 1), all labeled to indicate the presence or absence of breast cancer, of which 1,295 records (25%) were identified as breast cancer.

Proposed system

The lack of prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, time requires developing the technique which gives minimum error to increase accuracy. The available tests to detect breast cancer such as mammogram, ultrasound, and biopsy were timeconsuming, so there was a need for a computerized diagnostic system in which Machine Learning methodology was used. This methodology includes algorithms that help for the classification of the tumor and detect the cells more accurately and take less time as well.

DATA SET

The data used for the experiments was acquired from Kaggle. This dataset is BreakHist_Dataset consisting of four directories representing the magnification of the images respectively i.e. 100X, 200X, 400X and 40X. The dataset consists of 7,858 instances in total which are divided into the four magnification directories. Each magnification directory consists of two directories representing the tumours i.e. Benign and Malignant.

PREPROCESSING

ISSN 2347–3657
Volume 6, Issue 2, 2018

Feature Selection The importance of feature selection in a machine learning model is inevitable. It turns the data to be free from ambiguity and reduces the complexity of the data. Also, it reduces the size of the data, so it is easy to train the model and reduces the training time. It avoids over fitting of data. Selecting the best feature subset from all the features increases the accuracy. Some feature selection methods are wrapper methods, filter methods, and embedded methods.

Recursive Feature Elimination

RFE is a wrapper-type feature selection algorithm. This means that a different machine learning algorithm is given and used in the core of the method, is wrapped by RFE, and used to help select features. This is in contrast to filter-based feature selections that score each feature and select those features with the largest (or smallest) score. Technically, RFE is a wrapper-style feature selection algorithm that also uses filter-based feature selection internally. RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains. This is achieved by fitting the given machine learning algorithm used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model. This process is repeated until a specified number of features remains. Segmentation Splitting operation performed on images in 2X2, 3x3 up to 10X10 patches we called it as segmentation. In this segmentation process we train to the system to identify the close regions of interest which are important to detect the BC.By eliminating unrelated data from the image, it's easy to identify the tumor as early as possible. K-mean clustering algorithm is a method of groups it means similar objects combine in same group. Segmentation operation rely on it for better results and it gives better results when similar objects present in one group. It processes fastly as compare scattereddata

Conclusion

The proposed machine-learning approaches could predict breast cancer as the early detection of this disease could help slow down the progress of the disease and reduce the mortality rate through appropriate therapeutic

ISSN 2347–3657
Volume 6, Issue 2, 2018

interventions at the right time. Applying different machine learning approaches, accessibility to bigger datasets from different institutions (multi-center study), and considering key features from a variety of relevant data sources could improve the performance of modeling.

Refernces

Zhang X, Shengli SU, Hongchao WA. Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. *Big Data Research* . 2019;5(1):2019005. doi: 10.11959/j.issn.2096-0271.2019005. [CrossRef] [Google Scholar]

2. Chen SI, Tseng HT, Hsieh CC. Evaluating the impact of soy compounds on breast cancer using the data mining approach. *Food & function* . 2020;11(5):4561–70. doi: 10.1039/C9FO00976K. [PubMed] [CrossRef] [Google Scholar]

3. Aavula R, Bhramaramba R, Ramula US. A Comprehensive Study on Data Mining Techniques used in Bioinformatics for Breast Cancer Prognosis. *Journal of Innovation in Computer Science and Engineering* . 2019;9(1):34–9. [Google Scholar]

4. Kaushik D, Kaur K. Application of Data Mining for high accuracy prediction of breast tissue biopsy results. 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC); Moscow, Russia: IEEE; 2016. p. 40-5. doi: 10.1109/DIPDMWC.2016.7529361. [CrossRef] [Google Scholar]

5. Mokhtar SA, Elsayad A. Predicting the severity of breast masses with data mining methods. *ArXiv preprint arXiv:1305* . 7057 2013 doi: 10.48550/ARxIV.1305.7057. [CrossRef] [Google Scholar]

6. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology* . 2018;12(2):119–26. doi: 10.1177/1748301818756225. [CrossRef] [Google Scholar]

ISSN 2347–3657
Volume 6, Issue 2, 2018

7. Fan J, Wu Y, Yuan M, Page D, Liu J, Ong IM, Peissig P, Burnside E. Structure-leveraged methods in breast cancer risk prediction. *The Journal of Machine Learning Research* . 2016;17(1):2956–70. [PMC free article] [PubMed] [Google Scholar]

8. Burnside ES, Liu J, Wu Y, Onitilo AA, McCarty CA, Page CD, et al. Comparing Mammography Abnormality Features to Genetic Variants in the Prediction of Breast Cancer in Women Recommended for Breast Biopsy. *AcadRadiol* . 2016;23(1):62–9. doi: 10.1016/j.acra.2015.09.007. [PMC Free Article] [PMC free article] [PubMed] [CrossRef] [Google Scholar]

9. Stephens K. New Mammogram Measures of Breast Cancer Risk Could Revolutionize Screening. *AXIS Imaging News* . 2020 [Google Scholar]

10. Feld SI, Fan J, Yuan M, Wu Y, Woo KM, Alexandridis R, Burnside ES. Utility of Genetic Testing in Addition to Mammography for Determining Risk of Breast Cancer Depends on Patient Age. *AMIA Jt Summits Transl Sci Proc* . 2018;2017:81–90. [PMC Free Article] [PMC free article] [PubMed] [Google Scholar]

11. Guan Y, Nehl E, Pencea I, Condit CM, Escoffery C, Bellcross CA, McBride CM. Willingness to decrease mammogram frequency among women at low risk for hereditary breast cancer. *Sci Rep* . 2019;9(1):9599. doi: 10.1038/s41598-019-45967-6. [PMC Free Article] [PMC free article] [PubMed] [CrossRef] [GoogleScholar]