Machine Learning Method for the Breast Cancer Detection

Vikas
Lovely Professional University
Jalandhar

Vishu Madaan
Lovely Professional University
Jalandhar

ABSTRACT

Bioinformatics can be defined as a science of storing, fetching, arranging, interpreting and using information obtained from biological series and molecules. Prediction can be defined as a statement about future event on the basis of present situation. This work focuses on diabetic prediction with machine learning algorithms. The breast cancer prediction has various steps. A voting-based classifier is devised in this research to predict cancer. The performance for the breast cancer prediction is optimized up to 2 percent using proposed algorithm

KEYWORDS

Breast cancer, Machine learning, Classification, Voting

INTRODUCTION

As the most dominant cancer in women, breast cancer has always had high incidence and mortality rates. As per the most recent cancer insights, BC alone is predicted to be accounted for 25% of all new cancer diagnosis and 15% of all cancer fatalities among ladies around the world [1]. Researchers have known about the threats of BC from right off the bat, in this manner much early exploration has just been executed in the treatment of BC. Because of the endeavours of scientists and early recognition strategies, the death rate has demonstrated a consistent and declining pattern over the previous many years. As indicated by the insights of Cancer Research Institute (CRI) of UK, the five years endurance rate for BC is practically 100% whenever recognized at its soonest stage, yet can be as low as 15% when identified at the most last stage. This chapter introduces the detailed description of data mining-based breast cancer prediction.

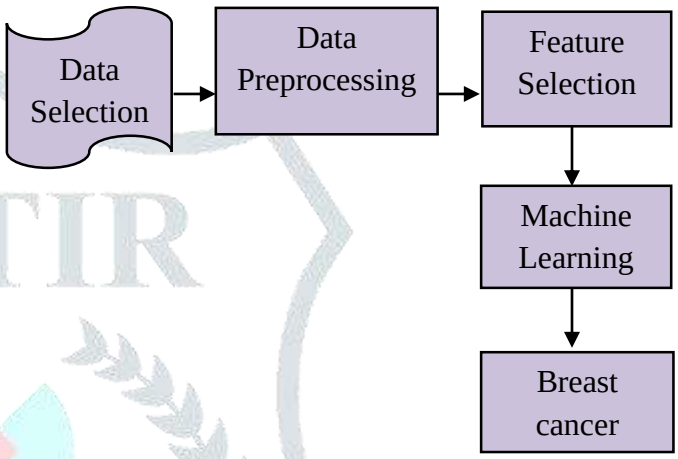


Figure 1: General process of Breast cancer prediction based on Machine Learning

Breast Cancer Prediction

Breast cancer is one of the most well-known dangerous tumours in ladies around the world, and it remains the main source of malignancy demise among females in developing nations. As of late, the frequency and mortality of breast cancer have expanded step by step, this genuinely undermines the lives and wellbeing of ladies and causes incredible monetary, social and family issues. In spite of the fact that the predominance of breast cancer shows the tendency to be in more youthful females, postmenopausal ladies additionally have the danger of breast cancer. Therefore, investigating the qualities of cancer malignancy in postmenopausal ladies, and discovering important data from various information to give clinical analysis and treatment to logical dynamic and clinical exploration are of incredible importance. As of now, some western nations have set up comparing breast cancer risk prediction models for different explicit severity factors. As indicated by the risk factors incorporated in the breast cancer risk prediction model is chiefly separated into two types of models known as genotype and statistical models. Genotype models can compute the danger of breast malignant growth at a particular age, and can likewise figure the likelihood of involving mutations through the incidence of breast cancer and ovarian malignancy in the family.

Artificial intelligence and, specifically, machine learning models have an obvious history in malignancy research and useful execution [11]. The vast majority of these works utilize machine learning strategies for demonstrating the growth of disease and recognize useful variables that are utilized a while later in a classification approach, concerning generally malignancy vulnerability, repeat, and endurance. The utilization of various ML models in malignancy research gives immense space to different applications. Artificial Neural Networks (ANNs) and Decision trees (DTs) have been utilized in disease prediction and detection for almost 30 years. Various models dependent on Support Vector Machine (SVM) applied to cancer prediction issues have been utilized for about a few decades. Different models for forecast of malignancy advancement and result have likewise been utilized for a few examinations. Today, not exactly a portion of information science and bioinformatics techniques are utilized by ML-driven models with a wide scope of uses, from diagnostics to expectation and forecast in malignancy. All this examination contemplates are worried about utilizing ML strategies to recognize, classify, identify, or differentiate tumours and different malignancies, just as to foresee disease growth. Breast cancer prediction works dependent on machine learning models possess a considerable part of the contemporary examination in this domain. Machine learning algorithms are contributing significantly to diagnose and predict breast cancer by implementing classification schemes to recognize individuals with breast cancer, differentiate cancerous from non-cancer our tumours and to predict disease [12]. Precise clustering can additionally help clinicians to recommend the most proper treatment system.

LITERATURE SURVEY

Naveen, et.al discussed that the main concern of medical community was to predict the breast cancer [21]. This research focused on predicting the breast cancer accurately. The most effectual ensemble ML model was constructed using the breast cancer Coimbra dataset that was extracted from UCI. This model assisted in enhancing the system performance with un-biasing. There were 6 ML algorithm implemented namely DT, SVM, MLP, KNN, LR and RF and their predictive analysis was compared with ensemble as well as non-ensemble methods. The accuracy obtained from the DT and KNN was computed 100% as the greatest accuracy. The prediction rate of presented model was quantified with regard to the precision, confusion matrix. The outcomes obtained from the presented model were highly accurate as they were useful in taking exact decision.

Anusha Bharat, et.al analyzed that the breast cancer was a disease due to which a great number deaths were occurred in every year [22]. Various algorithms were available in order to classify and predict the breast cancer such as SVM, DT, NB and KNN. The Wisconsin Breast Cancer dataset was employed to execute the SVM. Other algorithms namely k Nearest Neighbors, NB and CART were deployed for training the dataset. The comparison of the accuracy obtained from every algorithm was done while predicting the breast cancer. The Support Vector Machine utilized with Gaussian kernel had proved as an appropriate method to predict the breast cancer.

Tanishk Thomas, et.al described that the major cause of breast cancer was the division of abnormal cell in the breast itself due to which benign or malignant cancer was developed [23]. Therefore, the earlier prediction of breast cancer was very essential. The life of numerous patients was saved if they got the proper treatment. A variety of ML algorithms including SVM, KNN, NB, DT, K-Means and ANN were deployed for carrying out the comparative study on the public Wisconsin Diagnostic dataset so that the breast cancer was predicted in initial phase. All the utilized algorithms were compared and it was found that the superior predictive rate of 97.85% was obtained from the ANN in comparison with the other algorithms. Thus, the ANN was proved the best model for predicting the breast cancer. In the future, the data size would be maximized for enhancing the accuracy.

Madhuri Gupta, et.al suggested an ensemble model for predicting the breast cancer for which 4 ML schemes were employed namely SVM, LR, DT and KNN [24]. The outcomes demonstrated that the suggested model performed more accurately over the conventional single classification system. The weight was allocated to every classification technique applying Sequential Least Squares Programming technique. The soft voting method was implemented in order to integrate the prediction of every classifier. The outcomes depicted that the suggested model of Machine learning scheme provided superior performance to individual ML scheme. The suggested framework was computed using three different evaluation metrics. The ten-fold cross validation was implemented for the authentication of the model.

Sidharth S. Prakash, et.al stated that the major intend was to develop a DNN algorithm for predicting the malignancy of the breast cancer [25]. The Wisconsin breast cancer data set that was taken from UCI had employed to acquire the data. The optimization of intended algorithm was obtained with the help of early stopping mechanism and dropout layers to deal with the over fitting. The obtained F1 score of this algorithm was computed 98. This CAD model was not the replacement of the expertise of professional doctors and medical practitioners. However, it assisted in performing the diagnosis procedure successfully. The future work would focus on expanding this approach further for predicting the malignancy of breast cancer image data with the utilization of CNNs.

Mamatha Sai Yarabarla, et.al investigated that the CAD systems were acted significantly while recognizing the breast cancer and they were employed to mitigate the death rate

among women [26]. This approach aimed at the implementation of the current developments of CAD systems and its related methods. The RF algorithmic approach was utilized to detect and predict the breast cancer. The major intend of this project was that the person had to be predicted as normal or affected with BC. The machines were trained in the ML for learning and performing without any explicit program. This trained data was deployed to classify that the person had a BC or normal.

Tahreem Shouket, et.al discussed that the major purpose was that the OS and a number of years for the survival of patient who had not any sign of breast cancer had to be predicted [27]. The experiments were conducted considering the dataset of female patients of Pakistan who were diagnosed with breast cancer. The data that was collected from INMOL hospital of Pakistan had utilized for training the ML classification algorithms such as NB, DT, SVM, RF, JRip and AdaBoost. The outcomes of experiment exhibited that the JRip ML algorithm performed more efficiently for overall survival and disease free survival in comparison with the others. The presented approach assisted the patients and doctors in predicting the upcoming situation.

RESEARCH METHODOLOGY

Breast is a crucial part of human body. The smooth functioning of this part is very essential for the healthy lifestyle. If any kind of disease occurs in this organ, the other body parts are also disturbed. The researchers have often utilized the computer aided data which is extracted from the enormous databases. A number of businesses implement the DM systems and methods in extensive manner. The medical domain has made the deployment of these DM methods in order to predict the various diseases. Various risk factors may cause the breast cancer.

There are different stages included while predicting the breast cancer. These stages are mentioned as:

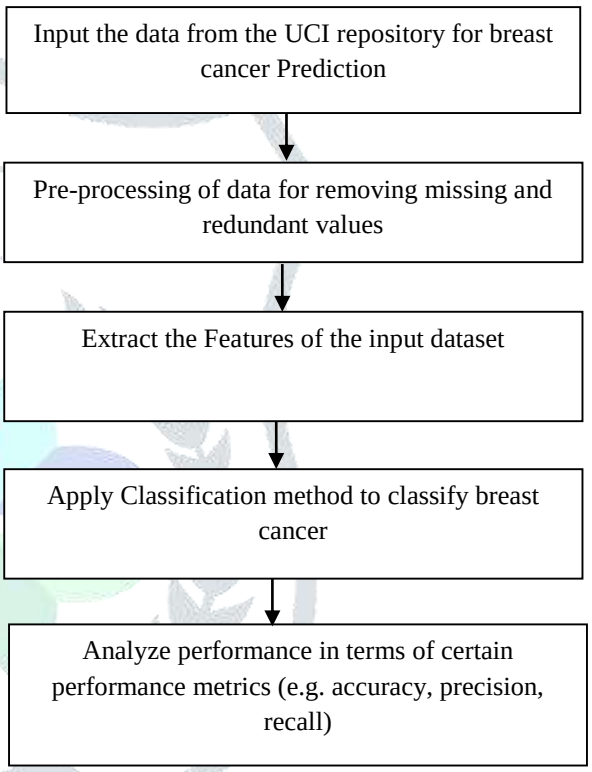
A. Data Acquisition: The data is gathered from different clinical organizations so that the experiments can be conducted.

B. Data pre-processing: The ML techniques can be implemented by establishing the wholeness and attaining a meaningful analysis on the data and data is processed. The redundant attributes is eliminated from the dataset for transmitting the clean and de-noised data so that the effectiveness of the training model can be improved.

C. Feature selection: A subset in which extremely unique attributes are included in order to detect the breast cancer have implemented in this phase. These selective attributes are associated with the existing class of attributes. The suggested technique has employed the RF algorithm to select the feature. The RF algorithm considers 100 as the estimator value and the tree structure of the most relevant attributes are created. The attributes that seems as most suitable or significant are selected using Random Forest to predict breast cancer.

D. Classification: The selected attributes are mapped for the training model so that the provided attributes are classified for making the prediction of breast cancer possible. Every separate class exposes a type of breast cancer. The classification is carried out using a LR algorithm. The input of extracted attributes is taken by the LR. This research work has discussed about the two classes: the breast cancer for the person who has the possibility of breast cancer or normal denotes for the person who has not any possibility of breast cancer.

Figure 2 Proposed Methodology



RESULT AND DISCUSSION

This work applies one of the most common datasets of Cleveland for breast cancer prediction. This dataset comprises 14 features. This work predicts breast cancer by implementing and comparing various classifier models. These classifiers include DT, Multilayer perceptron, NB, and an ensemble classifier consisting random forest, and naïve bayes models. The newly devised models are compared with respect to certain performance metrics.

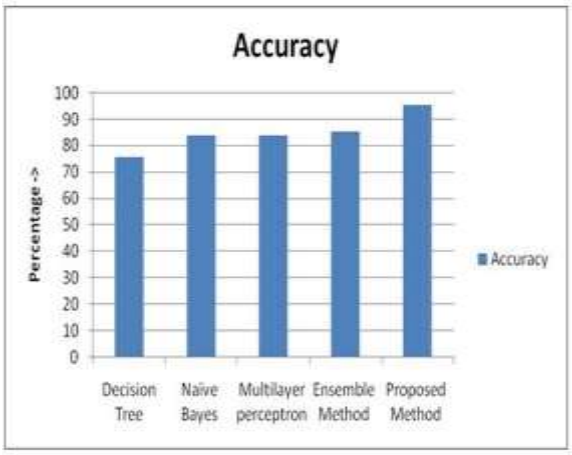


Figure 3: Accuracy Analysis

Figure 3 illustrates accuracy-based comparison amongst many classifier models such as DT, NB, multilayer perceptron, ensemble and proposed models. The analytic results prove that the new model gets maximal accuracy of about 95% and outperforms all other classifiers for breast cancer prediction.

Figure 4: Precision analysis

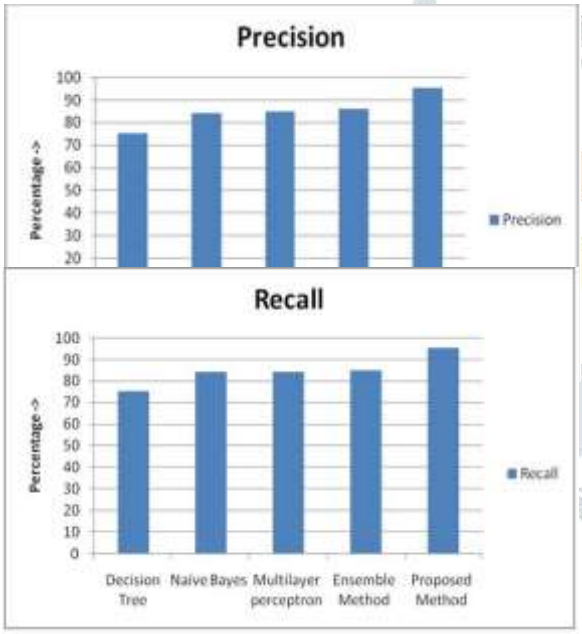


Figure 6.3 illustrates precision-based comparison amongst many classifier models such as DT, NB, multilayer perceptron, ensemble and proposed models. The analytic results prove that the new model gets maximal precision of about 95% and outperforms all other classifiers for breast cancer prediction.

Figure 5: Recall Analysis

Figure 5 illustrates recall-based comparison amongst many classifier models such as DT, NB, multilayer perceptron, ensemble and proposed models. The analytic results prove that the new model gets maximal recall value of about 95% and outperfroms all other classifiers for breast cancer prediction.

CONCLUSION

Breast cancer endurance rates have expanded, and the fatalities related with this infection is consistently declining, generally because of certain elements, for example, timely detection, another customized way to deal with treatment and a superior comprehension of the disease This work reaches to the conclusion that breast cancer prediction is incredibly difficult because of the involvement of multiple attributes. The different classifier models, for example, decision tree, naïve bayes, MLP are tested for predicting breast cancer. The new model integrates random forest and logistic regression for breast cancer prediction. This work uses random forest for feature extraction and logistic regression for classification. The recall, accuracy and precision provided by the devised model is counted as 95%.

References

[1] Gopal K. Dhondalay, Dong L. Tong, Graham R. Ball, “Estrogen receptor status prediction for breast cancer using artificial neural network”, International Conference on Machine Learning and Cybernetics, Volume: 2, Issue: 24, PP: 256-264, 2011

[2] Ir Cath Tee, Ali H. Gazala, “A novel breast cancer prediction system”, International Symposium on Innovations in Intelligent Systems and Applications, Volume: 67, Issue: 9, PP: 984-992, 2011

[3] Chintan Shah, Anjali G. Jivani, “Comparison of data mining classification algorithms for breast cancer prediction”, Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Volume 14, Issue: 30, PP: 3980-3987, 2013

[4] Hon-Yi Shi, Jinn-Tsong Tsai, Wen-Hsien Ho, Ming-Feng Hou, “Application of artificial neural networks for the prediction of quality of life in breast cancer patients”, SICE Annual Conference, Volume: 53, Issue: 10, PP: 734-742, 2011

[5] Xiaoyi Xu, Ya Zhang, Liang Zou, Minghui Wang, Ao Li, “A gene signature for breast cancer prognosis using support vector machine”, 5th International Conference on BioMedical Engineering and Informatics, Volume: 4, Issue: 28, PP: 2712-2720, 2012

[6] Glenn D Francis, Sandra R Stein, Glenn D Francis, “Prediction of histologic grade in breast cancer using an artificial neural network”, The 2012 International Joint Conference on Neural Networks (IJCNN), Volume: 12, Issue: 2, PP: 854-861, 2012

[7] Daphne Teck Ching Lai, Jonathan M. Garibaldi, "Improving semi-supervised fuzzy c-means classification of Breast Cancer data using feature selection", IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Volume: 29, Issue: 23, PP: 1092-1100, 2013

[8] Aida Ali, SitiManyam Shamsuddin, Anca L. Ralescu, "Hybrid intelligent systems in survival prediction of breast cancer", 12th International Conference on Hybrid Intelligent Systems (HIS), Volume: 80, Issue: 4, PP: 787-795, 2012

[9] Gul ShairaBanu, Amjath Fareeth, Nisar Hundewale, "Prediction of breast cancer in mammagram image using support vector machine and fuzzy C-means", Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, Volume: 27, Issue: 56, PP: 639-647, 2012

[10] Peter Adebayo Idowu, Kehinde Oladipo Williams, Jeremiah Ademola Balogun and Adeniran Ishola Oluwaranti, "Breast Cancer Risk Prediction Using Data Mining Classification Techniques", Transactions on Networks and Communications, Volume: 3, Issue: 42, PP: 187-194, 2015

[11] R. Preetha, S. Vinila Jinny, "A Research on Breast Cancer Prediction using Data Mining Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume: 8, Issue:11, PP: 693-701, 2019

[13] Dr. C Nalini, D.Meera, "Breast cancer prediction system using Data mining methods", International Journal of Pure and Applied Mathematics, Volume: 119, Issue: 12, PP: 10901-10911, 2018

[14] G. Ravi Kumar, Dr. G. A. Ramachandra, K. Nagamani, "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques", International Journal of Innovations in Engineering and Technology (IJIET), Volume: 2, Issue: 14, PP: 3677-3685, 2013

[15] Nitasha, "Review on Breast Cancer Prediction Using Data Mining Algorithms", International Journal of Computer Science Trends and Technology (IJCST), Volume: 7, Issue: 25, PP: 732-740, 2019

[16] S. Yuvarani and Dr. C. JothiVenkateswaran, "Breast Cancer Detection In Data Mining: A Review", Journal of Computer Science and Applications, Volume: 7, Issue: 1, PP: 245-252, 2015

[17] A. Priyanga, Dr. S. Prakasam, "The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness", International Journal of Computer Science and Engineering Communications, Volume: 10, Issue: 26, PP: 1740-1748, 2013

[18] Deneshkumar V, Manoprabha M, Senthamarai Kannan K, "Comparison of Datamining Techniques for Prediction of Breast Cancer", International Journal of Scientific & Technology Research Volume: 8, Issue: 08, PP: 268-276, 2019

[19] K. Arutchelvan, Dr. R. Periyasamy, "Cancer Prediction System using Data Mining Techniques", International Research Journal of Engineering and Technology (IRJET), Volume: 2 Issue: 68, PP: 797-804, 2015

[20] Shelly Gupta, Dharminder Kumar, Anand Sharma, "Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis", Indian Journal of Computer Science and Engineering (IJCSE), Volume: 2, Issue: 25, PP: 3782-3790, 2011

[21] Naveen, R. K. Sharma, Anil Ramachandran Nair, "Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models", 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Volume: 3, Issue: 25, PP: 953-961, 2019

[22] Anusha Bharat, N Pooja, R Anishka Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis", 3rd International Conference on Circuits, Control, Communication and Computing (I4C), Volume: 36, Issue:7, PP: 632-640, 2018

[23] Tanishk Thomas, Nitesh Pradhan, Vijaypal Singh Dhaka, "Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey", International Conference on Inventive Computation Technologies (ICICT), Volume: 6, Issue: 50, PP: 760-768, 2020

[24] Madhuri Gupta, Bharat Gupta, "An Ensemble Model for Breast Cancer Prediction Using Sequential Least Squares Programming Method (SLSQP)", Eleventh International Conference on Contemporary Computing (IC3), Volume: 15, Issue: 1, PP: 888-896, 2018

[25] Sidharth S. Prakash, K. Visakha, "Breast Cancer Malignancy Prediction Using Deep Learning Neural Networks", Second International Conference on Inventive Research in Computing Applications (ICIRCA), Volume: 37, Issue: 22, PP:251-259, 2020

[26] Mamatha Sai Yarabarla, Lakshmi Kavya Ravi, A. Sivasangari, "Breast Cancer Prediction via Machine Learning", 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Volume: 7, Issue: 30, PP: 653-660, 2019

[27] TahreemShouket, Sajid Mahmood, Malik Tahir Hassan, AfnanIftikhar, "Overall and Disease-Free Survival Prediction of Postoperative Breast Cancer Patients using Machine Learning Techniques", 22nd International Multitopic Conference (INMIC), Volume: 5, Issue: 56, PP: 565-563, 2019