

Prediction of breast cancer using tools of machine learning techniques.

Smita Parija*

Department of Electronics and Telecommunication Engineering, CV Raman Global University, Bhubaneswar, Odisha, India

Abstract

Machine learning techniques have been shown to support multiple medical prognoses. The purpose of this article is to compare some machine learning techniques to compare the diagnosis of breast cancer (cancerous and noncancerous) using the inputs from five supervised machine learning approaches through the different feature selections to get a correct result. Random Forests and the K-NNs model predict the most significant true positives among the five techniques. In addition, SVC and RFs models predict the most significant number of true negatives and the lowest number of false negatives. The SVC obtains the highest specificity of 96%, and the XGB obtains the lowest specificity of 92.3%. From this study, it is concluded that the Random Forests and K-NN Machine Learning models are the most suitable models for breast cancer diagnosis with an accuracy rate greater than 95%.

Keywords: Machine Learning, Breast Cancer, Mammography, Logistic Regression.

Accepted on April 29, 2022

Description

Breast cancer is caused by the abnormal development of cells in the breast and is the most common cancer globally. There is no evading the main issue about breast cancer; it is the mainly common form of cancer in India, with cervical cancer obsolete. In cities like Bengaluru, Mumbai, Delhi, Kolkata, Bhopal, Ahmedabad, Chennai, breast cancer accounts for 25% to 32% of all cancers in women, more than a quarter of all cancers in women. Younger age groups (25-50) are very often affected these days, and the worst news is that more than 70% of advanced cases have had poor survival and high death rates. A recent report from the Indian Council for Medical Research assumed the breast cancer count is likely to rise to 18.3 lakhs in 2022 [1,2].

Various techniques have been introduced to diagnose breast cancer (BSE, mammography, ultrasound, MRI, and positron emission tomography). However, every technique has some loopholes. Breast Self-Examination (BSE) is effective when done regularly but fails due to the patient's inability to check for changes in the early stages. Mammography is used to examine a woman's breasts through X-rays. In general, due to the small size of the cancer cells, it is almost impossible to detect breast cancer from the outside. Ultrasound is a well-known technique using sound waves to diagnose breast cancer. However, a transducer that emits false sound waves due to ambient noise makes a correct diagnosis more difficult. Positron Emission Tomography (PET) imaging using F-fluorodeoxyglucose is based on detecting radioactively labeled cancer-specific

tracers. However, the majority of patients cannot afford the cost of PET, so it has disadvantages. Dynamic Magnetic Resonance Imaging (MRI) predicts the rate of contrast enhancement using the breast distortion detection method by increasing angiogenesis in cancer [3-6].

Vast amounts of diagnostic data are available on behalf of the dataset through numerous websites around the world. The dataset was created by compiling data from various hospitals, diagnostic centers, and research centers. They hardly need to be organized so that the system can diagnose diseases quickly and automatically. Diagnosing a disease is usually based on medical plotter information and skills in the medical field (Improving Diagnosis in Health Care. Washington (DC): National Academies Press (US). Human error affects unwanted prejudices, wrong circumstances that later delay the accurate diagnosis of the disease. Enlightened by various disadvantages of the various techniques, additional techniques are needed to confirm the existing technology's findings, which will help the physician make the right decision. So this study tried to minimize the gap between doctors and the technologies available to them to make the right decisions through the concept of machine learning. Accurately diagnosing critical information in medicine is a need of the hour and is possible through bioinformatics or machine learning since diagnosing the disease is a vital and tricky task in the medical field. Machine learning techniques have been shown to support multiple medical prognoses. The purpose of this article is to compare some machine learning techniques to compare the diagnosis of breast cancer (cancerous

and noncancerous) using the inputs from five supervised machine learning approaches through the different feature selections to get a correct result. The objective of clustering is to partition a data set into groups according to some decisive element in an effort to systematize data into a more meaningful form. Clustering may advance according to some parametric model or by classifying points according to some distance or resemblance measure as in hierarchical clustering. A natural way to put cluster boundaries is in areas in data space where there is little data, i.e., in "valleys" in the likelihood distribution of the data. With the development of medical research, machine learning techniques for detecting breast cancer have been developed. The confusion matrix of all models is calculated for clarity of the techniques. The confusion matrix of the machine learning strategies used which provide the prediction results of SVC, Logistic Regression, Random Forests, XGBoost, and K-NNs, respectively. This study concluded that the Random Forests and K-NN Machine Learning models are the most suitable models for breast cancer diagnosis with an accuracy rate of greater than 95%. Additionally, this investigation suggests a feature selection method (mode) that uses an overall base classifier accuracy of 99% compared an ensemble model with batch classifiers to classify the instances with all the attributes compared to a reduced subset of data. These studies competitive result with other cutting-edge techniques and can provide radiologists with a valuable second opinion for breast cancer diagnosis.

References

1. Fletcher-Brown J, Pereira V, Nyadzayo MW. Health marketing in an emerging market: The critical role of signaling theory in breast cancer awareness. *J Bus Res* 2018; 86: 416-434.
2. Mathur P, Sathishkumar K, Chaturvedi M, Das P, Sudarsshan KL, Santhappan S. Cancer statistics. 2020: Report from national cancer registry programme, India. *JCO Glob Oncol* 2020; 6: 1063-1075.
3. Azar AT, El-Said SA. Probabilistic neural network for breast cancer classification. *Neural Comput Appl* 2013; 23: 1737-1751.
4. Ayon SI, Islam MM, Hossain MR. Coronary artery heart disease prediction: A comparative study of computational intelligence techniques. *IETE J Res* 2020; 1-20.
5. Horn D, Siegelmann HT, Vapnik V. Support vector clustering. *J Mach Learn Res* 2001; 2: 125-137.
6. Zhang Z. Introduction to machine learning: K-nearest neighbors. *Ann Transl Med* 2016; 4: 218.

*Correspondence to:

Smita Parija
Department of Electronics and Telecommunication Engineering
CV Raman Global University
Bhubaneswar
Odisha
India