International Journal of Thesis Projects and Dissertations (IJTPD)
Vol. 10, Issue 4, pp: (41-45), Month: October - December 2022, Available at: www.researchpublish.com

Apply Machine Learning Techniques to Detect Breast Cancer

Prokash Sharma[1]

[1]Department of Engineering & Technology, University of Calcutta (State University), India

DOI: https://doi.org/ 10.5281/zenodo.7230150

Published Date: 20-October-2022

Abstract: **Breast Cancer (BC) is one of the most extensive diseases worldwide. Proper and earlier diagnosis is a critical stage in treatment. Moreover, it is not easy to detect mammograms due to different uncertainties. Machine Learning (ML) approaches can produce tools for doctors that can be utilized as a valuable system for early identification and diagnosis of BC that will significantly improve the survival rate of patients. This article compares three of the most famous ML approaches typically utilized for BC detection and diagnosis, namely Bayesian Networks (BN), Support Vector Machine (SVM), and Random Forest (RF). The Wisconsin actual BC data set was utilized as a training set to review and compare the execution of the three ML classifiers in terms of main parameters such as precision, accuracy, recall, and area of ROC. The effects gained in this article give a summary of the state of art ML approaches for BC detection.**

Keywords: **Cancer, Breast Cancer, Machine Learning, Deep learning.**

I. INTRODUCTION

Breast cancer (BC) is a malignant tumor that activates in the breast cells. A tumor can disseminate to other body organs [1], [2]. BC is a universal disease that hammers the lives of women typically aged 25–50. The distress is alarming with the potential rise in the number of BC cases in India. During the past five years, the survival rates of BC patients have been about 90% in the USA, whereas in India, the figure reports approximately 60% [3]. BC projection for India during 2020 suggests the number to go as high as two million [4].

Consultant doctors have found hormonal, way of life, and environmental elements that may raise an individual's odds of growing BC. Over 5%–6% of BC patients have been connected to gene modifications that went through the ages of the family. The bulkiness, growing period, and postmenopausal hormonal inequalities are the other elements that cause BC.

BC has no prevention mechanism, but early identification can significantly improve the result. Additionally, this can also broadly decrease the costs of the treatment. Moreover, it is unusual to show cancer symptoms occasionally, so early detection is difficult. It is essential to utilize mammograms and self-breast examinations to detect any early irregularities before the tumor develops [5].

Laboratory tests are performed at a stage where the cancer is known in a patient [11]. This article is to offer a novice technique to detect BC. This article provides a detailed analysis of current cancer detection methods and the exact and efficient outcomes.

II. LITERATURE REVIEW

Soft computing methods play a dynamic role for judgment in request with imprecise and uncertain knowledge. The application of soft computing disciplines is fact developing foe the analysis and forecast in medical application. Between the many soft computing methods, unclear skilled system takes benefit of fuzzy skilled system; information is signified as a set of obvious philological rules. Study of breast cancer worries from uncertainty and fuzziness linked with in accurate input action and incompleteness of information of expects. However, there is several technology-oriented studies described for breast cancer analysis, few studies have been started for the breast cancer forecast.

Fatima et al. [6] identify an uncertain expert system for breast cancer forecast to additional support the procedure of breast cancer analysis. This method is accomplished enough to capture incomplete and imprecise information prevalent in the classification of breast cancer. For this, the paper utilized an uncertain reasoning model, which has high interpretability early diagnosis of the system's accuracy with an average of 95%, which shows the advantage of the system in the forecast process compared to other related work. Breast cancer analysis and forecast were two medical requests, which positioned as great tests for the investigations. Machine learning and data mining methods usage has transformed the entire practice of breast cancer Diagnose and Forecast. Breast cancer Diagnose decides design from breast lump and breast cancer Diagnose and Forecast. Breast Cancer Forecast predicts while Breast Cancer is probable to return in patients that had their cancers removed. Thus, these two problems were mainly within the scope of the organization's problems. This study paper encapsulates various reviews and technical articles on breast cancer diagnosis & prognosis.

Shelly, et al [7] describes to boost the breast cancer diagnosis & forecast. The subjective of our study is to explain the automated breast cancer detection support tool by implanting BBN (Bayesian Belief Networks). That is perceptive of Bayesian Belief Network is engaged as one feasible selection to discover the disease by speaking to the relationship among judgments, physical finding and research centre like Image Processing Experts, Radiologists, Database Professionals, & Applied Mathematicians on a typical stage A pithy available computation tool and stages were labelled.

Sri Hari Nallamala et al. [8] describe an assortment of web use mining practices that can propel exertion on different regions of logical, stimulating, and online networking applications to progress toward the exploration and security joined zone. Vazirani, at all [8], proposes the 2 NN standards, BPNN (Back Propagation Neural Network) and RBFN (Radial Basis Function). The development is finished utilizing a probabilistic total policy. The measured neural method gave a precision of 95.75% over preparing data and 95.22% over testing data, which was resolved to be best to solid neural systems.

III. MACHINE LEARNING TECHNIQUES

We used machine learning techniques to define if a cell is benign or malignant, i.e., Artificial Neural networks [18]. The learning process in ML techniques can be divided into two main classifications, overlooked and neglected to learn. In supervised learning, a set of information instances are used to prepare the machine and are marked to give the proper result. Yet, in unsupervised learning, there are no pre-determined information sets and no notion of the expected outcome, which means that the goal is harder to achieve. Classification is among the most common methods that go under supervised learning. It utilizes marked historical data to design a method that is used then for future projections. Hospitals and clinics maintain extensive databases in the medical sector, including records of patients' symptoms and diagnoses. Moreover, researchers utilize this knowledge to create classification standards to develop inferences based on historical cases. The medical inference has become a much easier task with machine-based help utilizing the sheer amount of medical data available today. It is helpful to note that all the approaches used in this article fall under classification standards.

A. Support Vector Machine (SVM)

One of the supervised ML classification techniques is SVM, which is broadly used in cancer diagnosis and prediction. SVM functions by selecting critical samples from all classes known as support vectors and separating the types by generating a linear process that divides them as broadly as possible using these support vectors. To find out the most suitable hyperplane, a mapping between an input vector and a high dimensionality space is made using SVM that divides the data set into classes [9]. This linear classifier seeks to maximize the distance between the decision hyperplane and the nearest data point, which is called the marginal distance by finding the best-suited hyperplane [10].

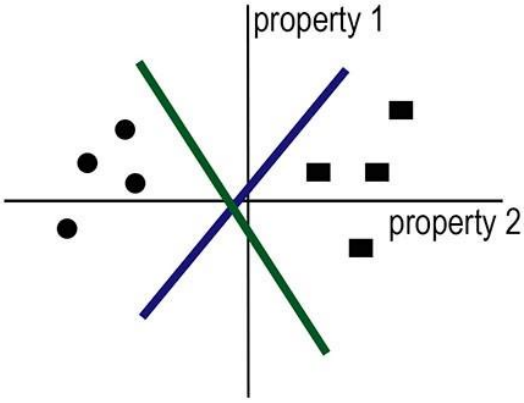

Fig. 1. SVM generated hyper-planes.

Fig 1 displays a dispersed plot of two types with two properties. A linear hyperplane is defined as ax1 + bx2 and the aim is to find a,b and c such that ax1 + bx2 6 c for class 1 and that ax1 + bx2 > c for class 2 [9][12]. Unlike other techniques, SVM depends on the support vectors, which are the data sets closest to the decision boundary, in their algorithms. This is because removing other data points further away from the decision hyperplane will not change the boundary as much as if the support vectors were removed.

B. Random Forest (RF)

Like a jury of people is used to make a court decision, RF combines many decision trees to ensemble a forest of trees. The argument is that having a single decision tree can provide either a simple model or a particular one [13]. It utilizes RF outcomes in improved strength compared to single decision trees. This shows that RF is unsympathetic to the noise of the input data set. One of the primary causes behind utilizing RF in cancer identification is its capacity to control data minorities. For example, a tumor can be classified as either benign or malignant, although the latter class is only 10% of the input data set. The RF technique is based on a recursive process. Every iteration concern picking one random selection of size N from the data set with alternate and another random sample from the visionaries without reserve. Then the information obtained is partitioned. Unnecessary data is removed, and the above steps are repeated many times depending on how many trees are needed. Finally, a calculation is made over the trees that categorize the observation in one category. Cases are then categorized based on a majority vote over the conclusion trees [14].

C. Bayesian Networks (BN)

BN is a subfield of probabilistic illustrated measures used to predict and represent uncertain domains [15]. BN corresponds to a widely used structure in machine learning called the directed acyclic graph (DAG). This graph consists of several nodes, each corresponding to a random variable, and the node edges represent direct dependence among the corresponding nodes in the graph.

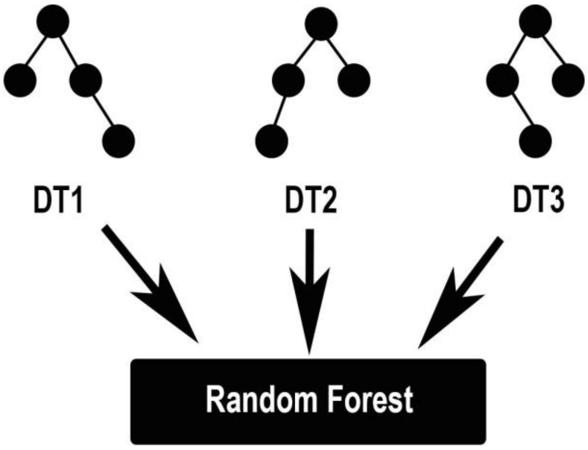

Fig. 2. A visual of how random forests works.

An edge between X1 and X2 represents a direct dependence between these two nodes. Statistical methods are often used to obtain an estimate for these dependencies. Every variable must have a conditional probability table that shows its probability distribution knowing its direct antecedents. Also, all variables are conditionally independent of their non-descendants, given their instant predecessors in the nodal frame. The following function is used to compute the joint probability of the values (x1, x2, xn) assigned to the network variables (X1, X2, XN).

P(x1, x2, , xn) = Yn

j=1

P(xi|Parents(xi)) (1)

The parents of a node represent their direct predecessors in the network, and the conditional probability is obtained from the probability table associated with **each node.**

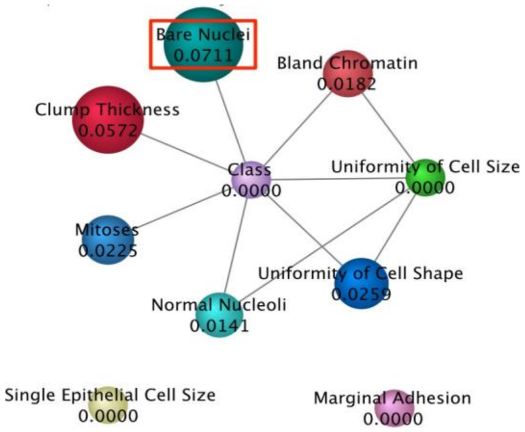

Fig. 3. DAG model combining breast cancer attributes [10]

IV. SIMULATION SETUP

A. Data Set

Our research is based on the Original Wisconsin Breast Cancer Data set acquired from the UCI Machine Learning Repository, an online unrestricted source repository [16]. This data set was gathered periodically over three years by Dr. William H. Wolberg from the University of Wisconsin Hospitals and consisted of 669 specimens, where the circumstances are classified as either malignant or benign. Four hundred fifty-eight of the cases are harmless, and 241 are malignant. The ten attributes are:

- Clump Thickness
- Cell Size Uniformity
- Cell Shape Uniformity
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nuclei
- Mitoses
- Class

All of the above attributes except for the class are numerical, and their values range between 1 and 10. The course has a value of 2 for benign and 4 for malignant.

B. Training Set

The classifier will be experimented utilizing the k − fold cross-validation technique. This validation approach will randomly split the training set into k subsets, where 1 of the k − 1 subsets will be utilized for testing and the rest for training. 10- fold cross-validation is the preferred k value used in most validation in ML and will be utilized in this paper [17][19]. It describes nine subsets utilized for the classifier's training and the remaining 1 for the experiment. This method avoids overfitting the training set, which is likely to occur in small data sets and a large number of attributes.

C. Simulation Software

This article utilized the Analysis of Waikato Environment for Knowledge (WEKA) software as an ML tool. WEKA is a Java based open-source tool that was first released to the public in 2006 under the GNU General Public License [20]. This tool gives different ML approaches and algorithms, containing the classification approaches examined in this article. Other features include data pre-processing, clustering, feature selection evaluation and rule discovery algorithms. Data sets are accepted in several formats, such as CSV and ARFF. Besides being an open-source tool, WEKA is also attractive due to its portability and ease of use GUI.

V. CONCLUSION

The ML methods have been broadly used in the medical sector and have helped as a valuable diagnostic tool that allows doctors to analyze the available information as well as designing medical expert systems. This article offered three of the most famous ML approaches generally utilized for BC detection and diagnosis: Random Forest (RF), SVM, and Bayesian Networks (BN). The main features and methodology of each of the three ML techniques was described. Performance comparison of the investigated techniques has been carried out using the Original Wisconsin Breast Cancer Data set.

The outcomes have confirmed that classification performance is based on the preferred approach. Effects have shown that SVMs have the highest version of precision, particularity, and accuracy. Therefore, RFs have the highest possibility of perfectly classifying tumors.

REFERENCES

[1] http://www.breastcancer.org/symptoms/understand_bc/what_is_bc

[2] Hotko Y.S. Male breast cancer: clinical presentation, diagnosis, treatment Exp. Oncol., 35 (4) (2013), pp. 303-310

[3] https://www.biospectrumindia.com/views/21/15300/statistical-analysis-of-breast-cancer-in-india.html.

[4] Malvia S., Bagadi S.A., Dubey U.S., Saxena S. Epidemiology of breast cancer in Indian women Asia Pac. J. Clin. Oncol., 13 (4) (2017), pp. 289-295

[5] Shallu S., Mehra Rajesh Breast cancer histology images classification: Training from scratch or transfer learning? ICT Express, 4 (2018), pp. 247-254

[6] M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in Systems Engineering (DeSE), Liverpool, 2016, P. 35-39.

[7] C. Deng and M. Perkowski, "A Novel Weighted Hierarchical Adaptive Voting Ensemble Machine Learning Method for Breast Cancer Detection," 2015 IEEE International Symposium on Multiple-Valued Logic, Waterloo, ON, 2015, P. 115-120.

[8] Sri Hari Nallamala, Siva Kumar Pathuri, Dr Suvarna Vani Koneru, “An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records”, International Journal of Engineering & Technology (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7 (2018), SI 7, P. 542 – 545.

[9] G. Williams, “Descriptive and Predictive Analytics”, Data Min. with Ratt. R Art Excav. Data Knowl. Discov. Use R, pp. 193-203, 2011.

[10] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” Comput. Struct. Biotechnol. J., vol. 13, pp. 8-17, 2015.

[11] Md Haris Uddin Sharif,” Cancer detection by machine learning”, International Journal of Computer Science and Information Security (IJCSIS), Vol. 19, No. 2, February 2021

[12] T. J. Cleophas and A. H. Zwinderman, “Machine Learning in Medicine,” pp. 1-271, 2013.

[13] I. Kononenko, “Machine learning for medical diagnosis: history , state of the art and perspective,” vol. 23, 2001.

[14] Y. Yasui and X. Wang, Statistical Learning from a Regression Perspective by BERK, R. A., vol. 65, no. 4. 2009.

[15] B. Networks, F. Faltin, and R. Kenett, “Bayesian Networks,” Encycl. Stat. Qual. Reliab., vol. 1, no. 1, p. 4, 2007.

[16] M. Lichman, UCI Machine Learning Repositry, 2013. [Online]. Available: https://archive.ics.uci.edu/.

[17] T. Fushiki, “Estimation of prediction error by using K-fold crossvalidation,” Stat. Comput., vol. 21, no. 2, pp. 137-146, 2011.

[18] Sharif, M., H., U. (2021). Breast Cancer Detection using Artificial Neural Networks. International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 9(X), 1121-1126.

[19] D. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” J. Mach. Learn. Technol., vol. 2, no. 1, pp. 37-63, 2011.

[20] K. J. Edwards and M. M. Gaber, Astronomy and Big Data.2014