IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11723

Breast Cancer Prediction Using Machine Learning

Sahana S[1], Dr. H K Madhu[2]

Student, Department of MCA, Bangalore Institute of Technology, Bangalore, India[1]

Professor, Bangalore Institute of Technology, Bangalore, India[2]

Abstract: The number of fatalities from breast cancer is rising dramatically every year. It is the most common kind of cancer overall and the leading cause of death for women globally. Any advancement in the identification and prognosis of cancer is crucial to a long and healthy life. Therefore, it's critical to have a high level of accuracy in cancer prognosis in order to update patient survival standards and treatment aspects. Machine learning approaches have shown to be a powerful method, have become a research hotspot, and may significantly contribute to the process of early detection and prediction of breast cancer. Using the Breast Cancer Wisconsin Diagnostic dataset, we ran five machine learning algorithms through this study: Support Vector Machine (SVM), Classification and Regression Tree (CART), Navi Bayes, and K-Nearest Neighbors (KNN). Once the results were in, we compared and evaluated the performance of each classifier. This study paper's primary goal is to identify the most efficient machine-learning algorithms in terms of confusion matrix, accuracy, and precision for the detection and prediction of breast cancer. The Support Vector Machine is shown to have attained the maximum accuracy of 97.2%, outperforming all other classifiers. All of the work is completed in the Anaconda environment using the Scikit-learn package and the Python programming language.

Keywords: Support Vector Machine, Navi Bayes, Classifier, Diagnostic

1. INTRODUCTION

Breast cancer is regarded as the leading cause of cancer fatalities globally, as it is difficult to detect and only treatment options are available for this condition. Early diagnosis is also suggested to help people lead healthy lives.

Machine learning algorithms may nearly diagnose cancer based on the conditions and available datasets before a thorough medical checkup. This has developed into a procedure that is ahead of its time for assessing the condition in addition to helping to anticipate with the necessary precision. Using this so-called computer produced reaction, many lives have been saved.

Anticipate abnormal tumours. Research has been conducted in order to accurately identify and classify individuals as malignant or benign. Magnetic resonance imaging (MRI), trans sonography, diagnostic mammography, and biopsies.

A cancer that originates in the breast tissue is called breast cancer. Breast cancer symptoms include breast lumps, breast shape changes, skin dimpling, milk rejection, fluid emerging from the nipple, recently inverted nipples, and red or scaly skin patches. Those who have the disease spread far may experience yellow skin, loss of breath, enlarged lymph nodes, and bone discomfort.

Obesity, inactivity, alcoholism, hormone replacement the therapy during menopause, ionizing radiation, early menarche age, late or non-existent birth, old age, previous breast cancer history, and a family history of breast cancer are all risk factors for this cancer.

Hereditary genes, including BRCA mutations, lead for five to ten percent the cells that line milk ducts and the lobules that provide milk to them are the most prevalent sources of breast cancer cells. Ductal carcinomas are tumors formed in the ducts, whereas lobular carcinomas form in the lobules. There are over 18 more subtypes of breast cancer. Pre-invasive lesions can lead to some forms of cancer, such as ductal carcinoma in situ. Breast cancer is determined by biopsy of questionable tissue.

2. LITERATURE SURVEY

- According to Anika Singh from UEM, Kolkata in 2020, breast cancer risk may be estimated utilizing both keen and slow learners. But this isn't able to grind the highest accuracy achievable with the current algorithms. According to their findings, the necessary accuracy—roughly 88 percent—can be achieved by employing lazy learners to identify just the cell growth.
- According to Nitasha (2019), eager learners and sluggish learners (about 90% of the population) may both be educated to achieve the necessary accuracy by applying the necessary precision.

- In 2019, Navya Sri published a cross-comparison study comparing Bayesian classifiers with decision tree algorithms. The study measured the accuracy of Bayesian with a decision tree score of 75.875% and a result of 75.27% using the Waikato Environment for Knowledge Analysis.
- Shannon (2011) employed a metaplasticity Artificial Neural Network approach to train an image scanning algorithm, with a correctness close to 95%. It returned a result of 99.26%, which is considered an overfit the quantity by current parlance. However, it succeeded in a renowned breakthrough in that the value continuously encouraged researchers to train and evaluate their data in a way that raised their accuracy requirements above 95%, with capacities above 90%.
- The assertion made by Jiaxin Li in 2020 that training and testing are necessary to obtain the only true accuracy, irrespective of the algorithm, served as a major source of inspiration for this work. As a result, their accuracy standard was raised to more nearly 95%.
- Nam Nhut Phan (2021) proposes employing convolutional neural networks in three stages:
training (50%) and testing (43%), with the remaining percentage going toward validation to produce results free of redundant empty values or correlations.
- N Gupta used an ensemble-based training strategy in 2021 to get the intended results. 96.77 was the notable breakthrough that was obtained without the usage of gradient boosting techniques.
- In 2019, Chang Ming employed BOADICEA and BCRAT in conjunction with eight simulated datasets containing cancer carriers and their relatives free of cancer. The startling discovery was that one of the cancer-free patients had a positive accuracy of 97%, indicating a reasonably high susceptibility to cancer.

3. METHODOLOGY

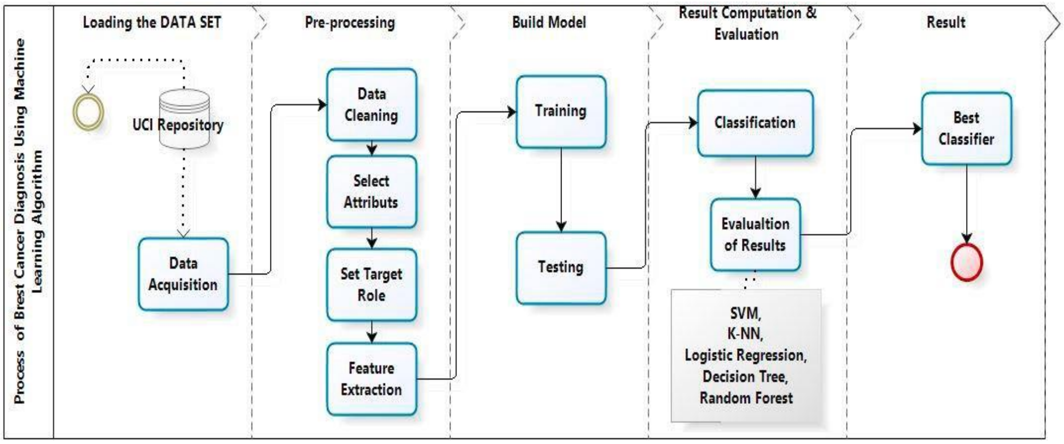

Fig. system design

Our experiment's primary goal is to determine the most accurate and predictive algorithm for breast cancer detection. To this end, we applied machine learning classifiers to the Breast Cancer Wisconsin Diagnostic dataset, including Support Vector Machine (SVM), Classification and Regression Tree (CART), Navi Bayes, and K-Nearest Neighbors (KNN). We then analyzed the results to determine which model has the highest accuracy.

Pre-processing, which consists of four steps: data cleansing, pick characteristics, determine target role, and feature extraction, comes first in our technique. Using the preprocessed data, machine learning algorithms are developed to forecast breast cancer for a fresh set of measurements. We present the model with fresh data that we have labels for in order to assess the algorithms' performances. Typically, we accomplish this by using the Train_test_split method to divide the labelled data we have gathered into two sections. Our machine learning model is constructed using 75% of the data, which is referred to as the training data or training set. A quarter of the total data, referred to as the test data or test set, will be utilized to gauge the model's performance.

Algorithms are used for model building

Classification and Regression Tree (CART)

This approach is applicable to regression as well as classification. The CART algorithm divides a node into sub-nodes using the Gini Index criteria. It begins with the training set as a root node. Once the root node has been successfully split in half, it uses the same logic to split the subsets and then splits the sub-subsets once more. This process is repeated until

DOI: 10.17148/IARJSET.2024.11723

it discovers that splitting any further will result in either a maximum number of leaves in a growing tree or pure sub-nodes. We call this procedure "tree pruning."

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classifier that uses the closest data points to find a maximum marginal hyper plane (MMH) by dividing datasets into classes.

Classifying data is a typical machine learning duty. The objective is to determine which of the two classes the information points supplied belong to, and say that a new data point has to be added in one of groups. In support vector machines, a data point is referred to by the term p {\displaystyle p}-dimensional vector. Our goal is to see if we may divide such places using a (p -1)-dimensional hyperplane. We term this a linear classifier. Different hyperplanes might be used for sorting the data. The hyperplane containing the greatest margin, or distance, between the two classes is an excellent contender for best hyperplane.

k-Nearest Neighbors (K-NN)

A supervised classification technique is called k-Nearest Neighbors (K-NN). It utilizes a large number of labelled points as input and learns how to categorize new input. In order to designate a new point, its nearest neighbors—the labelled points closest to the new point—are considered, and their vote is solicited.

Among machine learning's most fundamental but crucial classification algorithms is KNN. Pattern recognition, data mining, and intrusion detection are three major applications for this supervised learning domain member.

Naïve Bayes

The classification techniques that use Bayes' Theorem as a foundation are called naive Bayes classifiers. The idea that each pair of characteristics under categorization is independent unites a group of algorithms that make up this approach rather than a single one. Let us examine a dataset to get things going. Apply the Naïve Bayes classifier for quickly creating machine learning models with swift prediction abilities. It is one of the most efficient classification algorithm design.

4. RESULTS

after the application of algorithms for learning on the Wisconsin Diagnostic Dataset for Breast Cancer Kind As measures of performance, Confusion Matrix, Accuracy, Precision, Sensitivity, F1 Score, and AUC were employed in order to contrast the models to determine the best algorithm for the most accurate cancer prediction.

A confusion matrix is a method to determine the performance of a classification issue when the output might belong to two or more different types of classes. A table with the dimensions "Actual" and "Predicted" together with the columns "True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)" makes up a confusion matrix. For classification algorithms, accuracy is the most often used performance parameter.

It may be defined as the ratio of all forecasts made to the number of right ones. The quantity of accurate documents given in our machine learning model may be used to determine proficiency in document retrievals. You may define sensitivity as the quantity of favorable results that your machine learning model produces. We can find the harmonic mean of accuracy and sensitivity using the F1 score. The weighted average of accuracy and sensitivity is the mathematical representation of the F1 score.

Table 4.1: Accuracy

Algorithms	Accuracy Training Set	Accuracy Testing
	(%)	Set (%)
SVM	98.4%	97.2%
CART	99.8%	96.5%
KNN	95.5%	95.8%
NB	98.8%	95.1%

DOI: 10.17148/IARJSET.2024.11723

5. CONCLUSION

We have used four algorithm’s such as SVM, Classification and Regression Tree, Navi Bayes, K-NN—on the breast cancer dataset. We calculated, evaluated and compared on various forms of results based upon the confusion matrix, accuracy, sensitivity, precision, and AUC to estimate which machine learning algorithm is the most accurate, dependable, and precise.

For programming each algorithm python is used, with the package called sklearn in the Anaconda environment we found that SVM overcomes all the other techniques and achieves a efficiency of 97.2%, precision of 97.5%, and AUC of 96.6% after a precise model comparison between them.

In conclusion, SVM provides best accuracy and precision and have proved to be effective in the detection and prediction of breast cancer. As a result, we can apply for future work with the same algorithms and techniques on other databases to confirm the results obtained

Additionally, in our future work, we can plan to use same or other machine learning algorithms using new parameters on a huge dataset with large disease classes to obtain good accuracy, it should be noticed that all the results obtained are related only to the breast cancer dataset, which can be considered as drawback of our work.

REFERENCES

[1] WHO | Breast cancer’, WHO. http://www.who.int/cancer/prevention/diagnosisscreening/breast-cancer/en/ (accessed Feb. 18, 2020).
[2] Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015.
[3] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.
[4] B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.
[5] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, ‘Using Machine Learning Algorithms for Breast Cancer Risk Prediction and
[6] Y. khoudfi and M. Bahaj, Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification, 978-1-5386- 4225- 2/18/$31.00 ©2018 IEEE.