Available online at
www.mdl.mazedan.com

MCET

www.mazedan.com

www.mazedan.com/mcet

BREAST CANCER CLASSIFICATION USING MULTIPLE MACHINE LEARNING TECHNIQUES

VAISHNAWI PRIYADARSHNI*, SANJAY KUMAR SHARMA

MAZEDAN COMPUTER ENGINEERING TRANSACTIONS
e-ISSN: 2583-0414
Article id: MCET0202001
Vol-2, Issue-2
Received: 09 Aug 2021
Revised: 25 Sep 2021
Accepted: 8 Oct 2021

Citation: Priyadarshni, V., & Sharma, S. K. (2021). Breast Cancer Classification using Multiple Machine Learning Techniques. *Mazedan Computer Engineering Transactions*, 2(2), 1-3.

Abstract

Breast cancer (BC) is one of the most frequent cancers in women around the world, accounting for the majority of new cases and cancer-related deaths. It is a form of cancer that develops in the breast cells and may grow as a result of irregular cell division in the breast, resulting in a tumor that can be diagnosed and detected early, resulting in an increase in prognosis and chances of survival. It can also help patients receive better health care. Early cancer detection may save a patient's life. this paper, a comparison between multiple machine learning algorithms: Support Vector Machine, Decision Tree and Multi-Layer Perceptron. Support Vector Machine Provide better results than Multi-Layer Perception and Decision Tree. Support Vector Machine provides 95 % accuracy.

Keys words- Machine Learning, Support Vector Machine, Decision Tree, Multi-Layer perceptron, Breast Cancer

1. INTRODUCTION

Cancer is a prominent cause of death for both men and women. In the United States, cancer is the second leading cause of death, accounting for 21.8 percent of all deaths. Breast cancer is the tenth leading cause of death among cancer kinds, and it is more typically encountered in women. Early identification of cancer can aid in the total cure of the disease. Early discovery can save people's lives because the tumor is easier to treat and prevent from spreading. Tumors are abnormal cell growths. For many years, detecting breast cancer was to utilize an X-ray. Many additional was used. [1]

Every woman can undertake a monthly self-exam using her hand to check for any abnormally growing cells, or she can go to a specialist doctor for a mammography test. Mammography is defined as "the examination of the human breast utilizing low-dose X-rays as a diagnostic and screening technique. "Machine learning is described as a computational method that may be used to identify optimal solutions to a given problem without having to be explicitly written into a computer program by a programmer or an experimenter. Machine learning algorithms have been used in a multiple of sectors, including medicine, over the previous decades. Machine learning algorithms may now be used to assess medical data that is difficult to analyze manually, thanks to enhanced processing units.Over the last decade, there has been a growth in such studies, and every day, more and more effective methods for analyzing medical data are published in the academic literature.[2]

Classification is the most often used data mining technique, it employs a set of pre-classified examples to create a model that can categorize the entire document population.[3]

The classification technique's principal goal is to correctly anticipate the target class for each example in the data. Medical classification approaches are used in this study. It classifies the data before determining the optimum algorithm for breast cancer detection and prediction.[4]

Prediction begins with the identification of symptoms in patients, followed by the selection of sick individuals from a large group of sick and healthy patients. As a result, the primary goal of this study is to use a classification technique to examine data from a breast cancer data set in order to reliably forecast the class in each case. [5]

The classification technique's goal is to forecast with accuracy. This paper's primary goal is to examine data from using a classification method on the breast cancer dataset to reliably forecast the class using medical bioinformatics. It categorizes the data set before determining what to do with it. [6]

The goal is to figure out whether a patient's tumour is benign or malignant. In this study, machine learning techniques for breast cancer categorization. Each classifier's performance will be measured in terms of accuracy, as well as the training and testing processes.[7] Several experiments using several classifiers and feature selection strategies are undertaken on medical data sets. The literature contains a lot of information about breast cancer datasets. Aruna et al. used the decision tree classifier (CART) to reach an accuracy of 69.23% in breast cancer data sets [8].

School of Information and Communication Technology, Gautam Buddha University, Greater Noida, India
*Corresponding author email: vaishnawipriyadarshni02@gmail.com

On the Wisconsin Diagnostic Breast Cancer dataset, Chen et al. used the Support Vector Machine classification technique. The training and testing groups in the study were split 50-50, 70-30, and 80-20 percent respectively. Accuracy values have been estimated based on various training/testing percent's [9].

Liu et al. used the C45 method to predict breast cancer survivorship by creating new data for training from the original set using combinations with repetitions up to construct multiple sets of the same size as the original data [10].

Delen et al. gave 18 202,932 breast cancer patient records, which were subsequently divided into two "survivor" groups (93,273). & "did not make it" (109,659). Survivability predictions were accurate to within a margin of error of 93 percent [11]

Latchoumiet al. suggested a weighted particle swarm optimization (WPSO) combined with a smooth support vector machine (SSVM) to achieve a classification rate of 98.42 percent [12].

2. PROPOSED METHODOLOGY

Pre-processing: The dataset was obtained for free from the UCI Machine Learning Repository website, which is a public database. The dataset was converted to .csv file format. The first attribute in the dataset has been erased after successful import because it just contains the dataset's ID numbers. In addition, the class property was updated from "numerical" to "nominal" using a text editor to allow for proper classification [13].

Feature Extraction: Feature extraction is a dimensionality reduction technique that breaks down a vast amount of raw data into smaller chunks for processing. The vast number of variables in these massive data sets needs a significant amount of computer power to process. Principal component analysis (PCA), a dimensionality reduction approach, was applied in the feature extraction.[14]

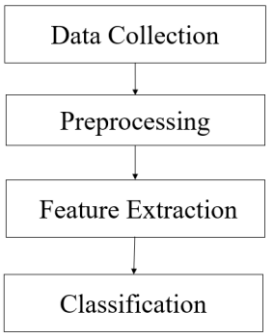

Figure 1 Flow Chart

Classification

The final procedure is classifications, which are critical for categorizing input data and producing output data. Two datasets were employed in the classification: training datasets and testing datasets. By extracting a feature from the input data, a training dataset is created. A feature that is utilized to train data that is being trained. The features are extracted in a similar way for the testing dataset. 80:20 ratio used to train/test data respectively. In classification used support vector machine, decision tree and multilayer perceptron classifier

Anaconda, an open-source software, was used to perform the classification. At the time of classification, the most recent version of the programme is used. [15].

Result

This section explains the parameters and examines the outcomes of the evaluation of the machine learning methods that were used.

Accuracy

The accuracy of detection is defined as the ratio of the number of correct detections to the total number of correct detections. This quantity is measured as a Percentage of incidents accurately detected. The total number of incidents is divided by the number of successful forecasts. It's worth noting that the accuracy is greatly dependent on the threshold established by the classifier, and hence may differ between various sets of tests. As a result, while this is not the best way for comparing different classifiers, it can provide a general picture of the class. The following equation is written as

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} * 100\%$$

Where: TP = True positive,

FN= False negative,

FP= False positive,

TN = True negative.

Recall

The rate of positive observations that are accurately predicted as positive is known as recall, also known as sensitivity. This is a useful metric, particularly in the medical industry, because it shows how many observations are accurately diagnosed (the sensitivity or true positive rate).

$$\text{Recall} = \frac{TP}{TP+FN} * 100\%$$

Support vector machine provides 95.61 % accuracy. The following values are given below.

	precision	recall	f1-score	support
0	0.95	0.99	0.97	71
1	0.97	0.91	0.94	43

Decision tree gives 93.85 % accuracy. The following values are given below.

	precision	recall	f1-score	support
0	0.93	0.97	0.95	71
1	0.95	0.88	0.92	43

Multilayer perception provides 93 % accuracy. The following values are given below.

	precision	recall	f1-score	support
0	0.92	0.99	0.95	71
1	0.97	0.86	0.91	43

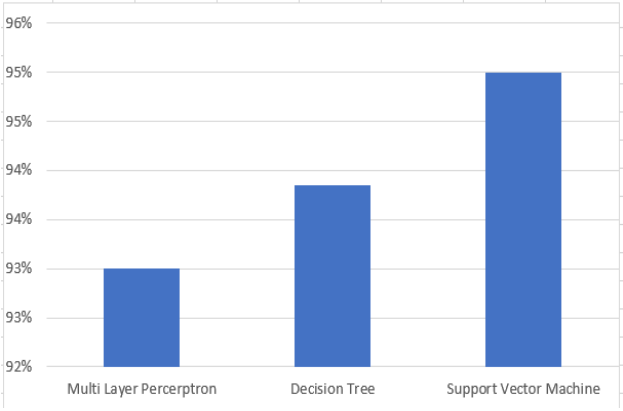

Figure 2 Accuracy of Classifiers

3. CONCLUSION

Our goal and challenge from breast cancer classification are to produce precision and dependability in classifiers. Support vector machine provide better accuracy than others. Following a thorough investigation, when comparing algorithms, discovered that Ada boost is the most efficient. In the future, we intend to dig deeper into these datasets by combining machine learning techniques with deep learning models and applying more complex deep learning architectures to increase performance. In addition, we test our in-depth learning approach on larger data sets with new ailment classes to increase accuracy. Another possible research direction is to use machine learning techniques to limited medical E-health applications. The resulting findings will be reported in a later study.

REFERENCES

[1] Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), 570-577.

[2] Lundin, M., Lundin, J., Burke, H. B., Toikkanen, S., Pylkkänen, L., & Joensuu, H. (1999). Artificial neural networks applied to survival prediction in breast cancer. Oncology, 57(4), 281-286.

[3] Vacek, P. M., Geller, B. M., Weaver, D. L., & Foster Jr, R. S. (2002). Increased mammography use and its impact on earlier breast cancer detection in Vermont, 1975–1999. Cancer, 94(8), 2160-2168.

[4] Brown, M. L., Houn, F., Sickles, E. A., & Kessler, L. G. (1995). Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. AJR. American journal of roentgenology, 165(6), 1373-1377.

[5] Mccarthy, J. F., Marx, K. A., Hoffman, P. E., Gee, A. G., O'neil, P., Ujwal, M. L., & Hotchkiss, J. (2004). Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. Annals of the New York Academy of Sciences, 1020(1), 239-262.

[6] Tan, A. C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification.

[7] Martens, H. A., & Dardenne, P. (1998). Validation and verification of regression in small data sets. Chemometrics and intelligent laboratory systems, 44(1-2), 99-121.

[8] K. R. A. Padmanaban and G. Parthiban (2016). Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease, Indian J. Sci. Technol., vol. 9, no. 29.

[9] Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert systems with applications, 38(7), 9014-9022.

[10] Salekin, A., & Stankovic, J. (2016, October). Detection of chronic kidney disease and selecting important predictive attributes. In 2016 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 262-270). IEEE.

[11] Latchoumi, T. P., & Parthiban, L. (2017). Abnormality detection using weighed particle swarm optimization and smooth support vector machine. Biomedical Research, 28(11), 4749-4751.

[12] Osman, A. H. (2017). An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. Int. J. Adv. Comput. Sci. Appl, 8(4), 158-165.

[13] Khourdifi, Y., & Bahaj, M. (2018, December). Applying best machine learning algorithms for breast cancer prediction and classification. In 2018 International conference on electronics, control, optimization and computer science (ICECOCS) (pp. 1-5). IEEE.

[14] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4). IEEE.

[15] Turgut, S., Dağtekin, M., & Ensari, T. (2018, April). Microarray breast cancer data classification using machine learning methods. In 2018 Electric Electronics, Computer Science, Biomedical Engineering's' Meeting (EBBT) (pp. 1-3). IEEE.