International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

Assessing the Effectiveness of Machine Learning Algorithms in Breast Cancer Classification

Nisharani Bhoi [1]

[1] Research Scholar, Department of Computer Application, Dr. A. P. J. Abdul Kalam University, Indore, Madhya Pradesh

Dr. Sandeep Singh Rajpoot [2]

[2] Supervisor, Department of Computer Application, Dr. A. P. J. Abdul Kalam University, Indore, Madhya Pradesh

Abstract

By using massive datasets and sophisticated computational methods to discover patterns and correlations that may not be visible to human eyes, machine learning algorithms have tremendous promise for enhancing breast cancer risk assessment and detection. This research compares the performance of four machine learning algorithms on the original datasets for Wisconsin breast cancer: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k-Nearest Neighbors (k-NN). The primary goal is to evaluate the efficacy of each algorithm in terms of data classification accuracy, precision, sensitivity, and specificity. In terms of accuracy (97.13%) and error rate (lowest), experimental findings demonstrate that SVM delivers the best performance. The trials are carried out using the WEKA data mining tool in a simulated setting.

Keywords: Machine learning, Data mining, Breast cancer, Effectiveness, Accuracy

I.INTRODUCTION

Breast cancer continues to be a very frequent and lethal illness that impacts women on a global scale. Timely identification and precise assessment of potential risks are essential for enhancing patient outcomes and minimizing fatality rates. Machine learning algorithms have become more valuable for predicting and diagnosing breast cancer. They have the potential to improve the accuracy and efficiency of screening programs and clinical decision-making. Machine learning algorithms may use extensive datasets of patient demographics, clinical factors, and imaging results to spot patterns and associations that may not be easily discernible to human observers. This enables more accurate risk assessment and early identification of breast cancer.

Machine learning algorithms use computational methods to autonomously acquire knowledge from data and generate predictions or judgments without the need for explicit programming. When it comes to predicting and diagnosing the risk of breast cancer, these algorithms may be trained using a wide range of datasets that include different sorts of information such as demographic characteristics, family history, genetic markers, mammographic pictures, and histological results. Machine learning methods may use key characteristics and trends from these datasets to create prediction models that accurately categorize patients into distinct risk groups or identify breast tumors as either benign or cancerous.

A key benefit of machine learning algorithms in breast cancer risk prediction and diagnosis is their capacity to incorporate diverse data sources and discern intricate connections among factors. Conventional risk assessment models often depend on a restricted range of variables, such as age, family history, and hormonal state, which may not comprehensively include the diversity of risk factors for breast cancer. On the other hand, machine learning methods may encompass a diverse set of factors and consider how they interact with each other, resulting in more thorough and customized risk prediction models.

Furthermore, machine learning algorithms possess the ability to adjust and enhance their performance when they encounter fresh data, rendering them flexible and multifunctional instruments for evaluating the risk of breast cancer. Machine learning systems may adapt to shifting trends and improve their predictions by regularly updating their predictive models using input from clinical outcomes and new research results. This allows them to better represent the developing landscape of breast cancer risk factors and diagnostic criteria.

Recent research have shown that machine learning algorithms have the ability to detect and diagnose breast cancer with great promise. Researchers have created models that use mammographic pictures to detect minor characteristics linked to early-stage breast cancer. These models have achieved levels of accuracy that are equivalent to, or even beyond, those of expert radiologists. Additional research has investigated the use of genetic markers and molecular profile data to categorize patients into various risk groups and direct individualized screening and preventative approaches.

Although there have been some positive advancements, there are still a number of obstacles that need to be overcome when it comes to using machine learning algorithms for predicting and diagnosing breast cancer. An essential obstacle is the need for extensive, top-notch datasets that accurately reflect various patient groups and cover a broad spectrum of therapeutic circumstances. Gaining access to such datasets might pose challenges owing to privacy considerations, restrictions on data sharing, and variations in data gathering methods across healthcare establishments.

Moreover, the comprehensibility of machine learning models is a notable obstacle, especially in healthcare environments where clear decision-making is crucial. Several machine learning algorithms, including deep learning neural networks, are sometimes referred to as "black-box" models because their decision-making processes are not readily understandable by humans. Consequently, healthcare professionals may be reluctant to depend only on machine learning predictions without a comprehensive comprehension of the underlying rationale behind the model's suggestions.

II.REVIEW OF LITERATURE

Filali, Sanaa et al., (2021) The annual mortality rate due to breast cancer is seeing a significant surge. It is the most prevalent form of cancer and the primary cause of mortality among women globally. Advancements in cancer prediction and detection are crucial for maintaining a healthy life. Therefore, achieving a high level of accuracy in cancer prediction is crucial for improving treatment outcomes and the overall survival rate of patients. Machine learning approaches have the potential to significantly enhance the prediction and early detection of breast cancer. This area of study has gained much attention and has been shown to be a powerful approach. This study utilized five machine learning algorithms, namely Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5), and K-Nearest Neighbours (KNN), to analyze the Breast Cancer Wisconsin Diagnostic dataset. Subsequently, a performance evaluation and comparison were conducted to assess the effectiveness of these classifiers. The primary aim of this study work is to use machine-learning algorithms to forecast and diagnose breast cancer, while determining the most efficient approach based on the confusion matrix, accuracy, and precision. Support Vector Machine demonstrated superior performance compared to other classifiers, with the greatest accuracy of 97.2%.The whole of the work is completed inside the Anaconda environment, using the Python programming language and the Scikit-learn module.

Kumar, Pawan et al., (2021) Breast cancer is widely recognized as one of the most prevalent diseases affecting women globally. This kind of melanoma is the most prevalent and is the main cause of the rising death rate. Precise and prompt identification of this lethal illness is crucial in order to enhance the likelihood of patient survival. Multiple implementations have been developed in this field, with a plethora of machine learning and soft computing algorithms offered for the analysis and detection of cancer. These algorithms assist clinicians in promptly prescribing appropriate treatments. This study presents a comprehensive comparison of important machine learning approaches, assessing their performance. The breast cancer detection and diagnosis approaches include support vector machines (SVM), random forest, and k-nearest neighbor (k-NN). All tests were performed using the R programming language, which is a data mining tool. The results showed that the k-NN algorithm exhibited the best level of accuracy (97.32%) when compared to SVM and RF.

Houfani, Djihane et al., (2020) Cancer of the breast is one of the most significant diseases that take the lives of a significant number of women. In the medical procedure, early diagnosis is a crucial activity that should be performed in order to reduce these numbers. The application of machine learning (ML) techniques serves as an efficient approach for categorizing data, particularly in the field of medicine, where these techniques are extensively utilized in the process of diagnosis and decision making. The purpose of this study is to provide a review of the most current papers that focus on the application of Machine Learning techniques in the detection of breast cancer. Several machine learning approaches were used to a variety of datasets in order to develop the classification models that are being presented here.

B.M, Gayathri et al., (2013) In today's world, breast cancer has emerged as a prevalent risk factor. Despite the fact that mammograms are used to identify breast cancer, not all general hospitals have the facilities necessary to perform the procedure. Waiting for a diagnosis of breast cancer for an extended period of time may raise the likelihood that the disease may spread to other parts of the body. As a result, a computerized breast cancer diagnostic has been created in order to cut down on the amount of time required to detect breast cancer and to lower the mortality rate associated with the disease. The purpose of this study is to provide a

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

summary of the survey that was conducted on breast cancer detection utilizing a variety of machine learning algorithms and methodologies. These approaches assist in improving the accuracy of cancer prediction. In addition, this survey might assist us in gaining knowledge regarding the quantity of papers that are utilized in the process of diagnosing breast cancer.

Aloraini, Adel (2012) Currently, a significant number of people are afflicted with breast cancer. This illness is brought on by a multitude of variables, none of which can be identified with relative simplicity. Additionally, the procedure of diagnosis, which is what decides whether the cancer is benign or malignant, demands a significant amount of effort from the doctors and physicians who are administering the treatment. When several tests are involved in the process of diagnosing breast cancer, such as clump thickness, uniformity of cell size, uniformity of cell shape, etc., the final result may be difficult to get, even for those who are considered to be considered to be specialists in the field of medicine. This has resulted in an increase in the utilization of machine learning and artificial intelligence in general as diagnostic tools throughout the course of the more recent few years. The purpose of this study was to evaluate and contrast several categorization learning algorithms in order to make a substantial prediction regarding the difference between benign and malignant cancer in the Wisconsin breast cancer dataset. The Wisconsin breast cancer dataset was utilized to evaluate and contrast five distinct learning algorithms, namely Bayesian Network, Naïve Bayes, Decision trees J4.8, ADTree, and Multi-layer Neural Network. Additionally, a t-test was conducted to determine which method exhibited the highest level of accuracy in terms of prediction. Based on the results of the experiment, it has been demonstrated that Bayesian Network is much superior to the other methods.

III.RESEARCH METHODOLOGY

Weka machine learning environment19 libraries were used to do all experiments involving the classifiers detailed in this work. Data pre-processing, classification, regression, clustering, and association rules are all part of WEKA's set of machine learning techniques. Weka applies machine learning approaches to a wide range of real-world challenges. The software provides a clear structure for developers and experimenters to construct and assess their models. This research makes use of the original datasets20 for Wisconsin breast cancer that are housed at the UCI Machine Learning Repository.

IV.RESULTS AND DISCUSSION

Time to create the model, instances properly categorized, instances wrongly classified, and accuracy are the metrics used to assess the performance of each classifier in this section. You can see the outcomes in Table 1.

Table 1: Performance of the classifiers

Evaluation criteria	**Classifiers**			
	C4.5	**SVM**	**NB**	**k-NN**
Time to build a model (s)	0.05	0.09	0.07	0.02
Correctly classified instances	665	678	671	666
Incorrectly classified instances	34	21	28	33
Accuracy (%)	95.19	97.09	96.01	95.31

From Table 1, we can see that k-NN only takes 0.02 seconds to construct its model, whereas SVM takes around 0.09 seconds. This is because, in contrast to other classifiers, k-NN is a sluggish learner and doesn't exert much effort during training. On the other side, SVM achieves a higher accuracy (97.09%) compared to C4.5, Naïve Bayes, and k-NN, whose accuracy ranges from 95.19% to 95.31%. Additionally, compared to the other classifiers, SVM clearly has the best ratio of properly classified instances to total instances, while having the lowest ratio of wrongly classified instances.

In this work, we also take simulation error into account so that we can quantify classifier performance better. The efficiency of our classifier is assessed in order to achieve this goal. Quantitative values are used for KS, MAE, and RMSE. Percentages represent RAE and RRSE. The findings may be seen in Table 2.

Table 2: Training and simulation error

Evaluation criteria	**Classifiers**			
	C4.5	**SVM**	**NB**	**k-NN**
Kappa Statistic (KS)	0.89	0.96	0.92	0.89
Mean Absolute Error (MAE)	0.08	0.05	0.02	0.07
Root Mean Square Error (RMSE)	0.19	0.21	0.17	0.23
Relative Absolute Error (RAE) %	14	6.30	8.61	10.50
Root Relative Squared	45	35.60	40.98	44.79

Error (RRSE) %				

We can observe from Table 2 that SVM produces the greatest classification chance (0.96%) with the lowest warning error rate (0.05). The optimum compatibility between the dependability and validity of the acquired data is also shown by SVM. The huge number of misclassified examples for both C4.5 and k-NN can be explained by their greatest error rates.

V.CONCLUSION

The application of machine learning algorithms has the potential to revolutionize the detection and treatment of breast cancer, which might ultimately result in better patient outcomes and lower death rates. There is a wide variety of data mining and machine learning techniques that may be utilized for health care data analysis. In the fields of data mining and machine learning, one of the most significant challenges is the development of classifiers that are both accurate and computationally economical for use in medical applications. SVM, NB, k-NN, and C4.5 were the four primary algorithms that we utilized in this investigation. These algorithms were applied to the Wisconsin Breast Cancer (original) datasets.

REFERENCES: -

1. Jena, Lambodar & Ammoun, Lara & Patra, Bichitrananda. (2022). Machine Learning Model for Breast Cancer Tumor Risk Prediction. 10.1007/978-981-16-9873-6_47.
2. Filali, Sanaa & Aarika, Kawtar & Naji, Mohammed & Benlahmar, EL Habib & Ait Abdelouahid, Rachida & Debauche, Olivier. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Computer Science. 191. 487-492. 10.1016/j.procs.2021.07.062.
3. Ellamey, Mazen & M. Eid, Mohab & Gamal, Muhammad & Bishady, Nour & Wagdy, Ali. (2021). Using Machine Learning Algorithms for Breast Cancer Diagnosis. International Journal of Applied Metaheuristic Computing. 12. 117-137. 10.4018/IJAMC.2021100107.
4. Harinishree, M. & C R, Aditya & Sachin, D.. (2021). Detection of Breast Cancer using Machine Learning Algorithms – A Survey. 1598-1601. 10.1109/ICCMC51019.2021.9418488.
5. Kumar, Pawan & Bhatnagar, Ashutosh & Jameel, Roshan & Mourya, Ashish. (2021). Machine Learning Algorithms for Breast Cancer Detection and Prediction. 10.1007/978-981-16-0695-3_14.
6. Houfani, Djihane & Slatnia, Sihem & Kazar, Okba & Zerhouni, Noureddine & Merizig, Abdelhak & Saouli, Hamza. (2020). Machine Learning Techniques for Breast Cancer Diagnosis: Literature Review. 10.1007/978-3-030-36664-3_28.
7. Sengar, Prateek & Gaikwad, Mihir & Nagdive, Dr-Ashlesha. (2020). Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction. 796-801. 10.1109/ICSSIT48917.2020.9214267.
8. Bharat, Anusha & Pooja, N & Reddy, R. (2018). Using Machine Learning algorithms for breast cancer risk prediction and diagnosis. 1-4. 10.1109/CIMCA.2018.8739696.
9. Sharma, Shubham & Aggarwal, Archit & Choudhury, Tanupriya. (2018). Breast Cancer Detection Using Machine Learning Algorithms. 114-118. 10.1109/CTEMS.2018.8769187.
10. Khourdifi, Youness & Bahaj, Mohamed. (2018). Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification. 1-5. 10.1109/ICECOCS.2018.8610632.
11. B.M, Gayathri & Sumathi, C.P. & T, Santhanam. (2013). Breast Cancer Diagnosis Using Machine Learning Algorithms - A Survey. International Journal of Distributed and Parallel systems. 4. 105-112. 10.5121/ijdps.2013.4309.
12. Aloraini, Adel. (2012). Different Machine Learning Algorithms for Breast Cancer Diagnosis. International Journal of Artificial Intelligence & Applications. 3. 21-30. 10.5121/ijaia.2012.3603.