Journal of Basic and Applied Engineering Research
p-ISSN: 2350-0077; e-ISSN: 2350-0255; Volume 5, Issue 7; October-December, 2018, pp. 603-614
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Comparative Analysis of Different Machine Learning Algorithms For Breast Cancer Classification.

Shabana Nargis Rasool[a]

[a]*Department Of Computer Science, Islamic University of Science and Technology Kashmir India, shabana.nargis@islamicuniversity.edu.in*

Abstract

Breast cancer is a significant public health issue globally, with early detection playing a critical role in increasing diagnosis, treatment, and survival rates. Machine Learning (ML) techniques have been used to predict the malignancy or benignancy of breast cancer, allowing for the development of predictive models to facilitate effective decision making. This paper presents a Comparative analysis of ML techniques, including Logistic Regression, Support Vector Machines, Decision Trees, and Random Forest and their use in breast cancer classification. The results show that these algorithms produce competitive results with SVM performing better than other algorithms with an accuracy of 97.66% , F1 score of 98% and an AUC of 97.8%. and have the potential to be used in detecting breast cancer.

Keywords: Machine Learning, Breast Cancer, Logistic Regression, SVM

1. Introduction

Cancer, also known as malignant neoplasm, covers a wide range of diseases characterized by the unusual proliferation of cells that can spread to the adjacent tissues in the body [1]. This insidious disease is a leading cause of death globally, with breast cancer being the second leading cause of cancer deaths among women in more developed countries and the leading cause in less developed countries [2]. Recent statistics indicate that one in eight women in the United States will develop breast cancer at some point in their lives[3].

Early detection of breast cancer is a crucial area of research as it has the potential to increase the rates of diagnosis, treatment, and survival [4]. Given

the high cost of treatment and widespread prevalence of the disease, early diagnosis is considered the most effective way to minimize its health and social implications.There are numerous techniques and processes for detecting this kind of cancer, each having its unique benefits and drawbacks.

Regrettably, cancer is often detected in its later stages, when the chances of metastasis and successful treatment are greatly reduced. This is often due to a lack of self-testing, with most breast cancers being discovered as a lump or mass through self-examination or mammography [5].

Machine Learning, a subfield of Artificial Intelligence, enables machines to learn and perform tasks without explicit programming through exposure to data sets. Over the years, these methods have been widely adopted in the development of predictive models, facilitating effective decision making [6]. In the field of cancer research, ML techniques can be utilized to identify patterns in data sets, thereby predicting the malignancy or benignancy of a cancer. The effectiveness of these techniques can be measured based on accuracy, recall, precision, and the area under the Receiver Operating Characteristic (ROC) curve[7].

In this paper, we present a detailed comparative analysis of four different machine learning algorithms on the classification of breast cancer on Wisconsin breast cancer dataset.Different evaluation metrics are used to evaluate the performance of these algorithms. The highlights of this study is that SVM performed better than all other algorithms with F1 score of 98% and Accuracy of 97.66%.

The rest of the paper is organised as follows: Literature survey and Machine learning algorithms are discussed in section 2, Experiments are presented in section 3 and finally paper is concluded in section 4.

2. Background

In this section, we will commence with a literature survey, followed by an introduction to various machine learning methods employed in breast cancer classification.

2.1. Literature Survey

Numerous studies have been conducted in the area of breast cancer classification. Authors in [8]focused on the comparison of three popular machine learning algorithms for predicting breast cancer in Indian women: Random Forest, kNN (k-Nearest-Neighbor), and Naive Bayes. The study uses the

Wisconsin Diagnosis Breast Cancer dataset as a training set to evaluate the performance of the algorithms in terms of accuracy and precision. The results of the study show that the algorithms produce competitive results and can potentially be used for the detection and treatment of breast cancer among Indian women, where it is one of the most frequently occurring cancers and has a fatality rate of 50%.

Hussain et al.[9] concentrated on the necessity for alternative techniques for the early detection of breast cancer, which is the most prevalent cause of death among women across the world. The research proposed a hybrid model that incorporates multiple machine learning algorithms, such as Support Vector Machine, Artificial Neural Network, K-Nearest Neighbor, and Decision Tree, for the efficient identification of breast cancer. The study also examined the various datasets utilized for breast cancer detection and diagnosis, emphasizing the importance of methods that are less expensive, safer, and produce more dependable results than conventional methods like mammogram images, which sometimes generate false positives. The proposed hybrid model is adaptable and can be used with various data types, including images and blood samples.

A new approach for detecting breast cancer with improved accuracy was proposed in [10]. The approach contains two stages: the initial stage involves using image processing methods to process mammography images for feature extraction, and the subsequent stage involves utilizing the extracted features as input for two supervised learning models (Back Propagation Neural Network and Logistic Regression). The models' effectiveness was assessed, and the findings revealed that the Logistic Regression model employed more features than the Back Propagation Neural Network model.

Researchers in [11] present a comprehensive survey of the literature on the application of machine learning algorithms and techniques for enhancing the accuracy of breast cancer predictions in diagnosis. The survey provides a valuable overview of the current state of the field and the number of studies conducted in this area.

Authors in [12]compared three prevalent machine learning classifiers: Support Vector Machine, Random Forest, and Bayesian Networks. The performance of these techniques was evaluated using the Wisconsin original breast cancer dataset, and key performance indicators such as accuracy, recall, precision, and the area under the receiver operating characteristic curve. The results of this study provide a comprehensive overview of the state-of-the-art in machine learning techniques for breast cancer detection and hold signifi-

cant implications for the field of medical diagnosis and patient care.

The paper [13] presented the application of Deep Learning for the diagnosis of breast cancer using the Wisconsin Breast Cancer Database. The dataset was pre-processed and 11 features were selected for the diagnosis. The Deep Learning algorithm achieved an accuracy of 99.67% and was compared to other machine learning algorithms, demonstrating superior performance. The results of this study highlight the potential of Deep Learning technology in the early detection and diagnosis of breast cancer.

A comparative study of six machine learning algorithms applied to the Wisconsin Diagnostic Breast Cancer dataset was presented in [14]. The algorithms used were GRU-SVM, Linear Regression, Multilayer Perceptron, Nearest Neighbor, Softmax Regression, and Support Vector Machine. The dataset was divided into a training set (70%) and a testing set (30%), and the algorithms' effectiveness was assessed in terms of classification accuracy, sensitivity, and specificity. The results showed that all the algorithms performed well with an accuracy greater than 90%. The best performing algorithm was the Multilayer Perceptron, achieving a test accuracy of approximately 99.04%.

Authors of [15] reviews the use of machine learning (ML) techniques for the diagnosis and prognosis of breast cancer (BC), a significant public health issue among women worldwide. The early diagnosis of BC can significantly improve the chances of survival, hence the accurate classification of benign and malignant tumours is the subject of much research. This paper provides an overview of popular ML techniques including ANNs, SVMs, DTs, and k-NNs and their applications in BC diagnosis and prognosis. The study uses the Wisconsin breast cancer database (WBCD) as the primary data source and compares the results of different algorithms. A healthcare system model of the authors' recent work is also presented.

Researchers in [16] aimed to evaluate the accuracy and robustness of a deep learning-based method for the detection of invasive tumor on digitized whole slide images of breast tissue. The method employs a convolutional neural network to classify the presence of invasive tumor in the images. The classifier was trained on nearly 400 samples from multiple sources and validated on 200 cases from The Cancer Genome Atlas. The results showed a high level of accuracy, with a Dice coefficient of 75.86%, a positive predictive value of 71.62%, and a negative predictive value of 96.77% compared to manual annotations. This approach holds promise as a decision support tool in the management of breast cancer patients, providing a more efficient and

consistent method for identifying the extent of invasive tumor.

2.2. Machine learning algorithms

Logistic Regression: In machine learning, Logistic Regression is a supervised learning algorithm used for binary classification problems. A linear method is used to construct a model that describes the correlation between a reliant variable and one or multiple autonomous variables. The algorithm outputs a probability value that represents the likelihood of a particular binary outcome.The Logistic Regression algorithm is trained using labeled data, where the target variable is binary. The training process involves estimating the coefficients for the independent variables that best predict the dependent variable. These coefficients are then used to forecast outcomes on novel, unobserved data.The predictions made by the Logistic Regression model can be evaluated using metrics such as accuracy, precision, recall, and F1 score. The model can also be fine-tuned by adjusting the parameters and/or transforming the features.It's important to note that Logistic Regression presupposes a linear association between the autonomous variables and the logarithm of the probability ratio of the dependent variable. In cases where this assumption is not met, alternative methods such as non-linear logistic regression or decision trees may be used[17].

SVM: Support Vector Machines (SVMs) are a type of supervised machine learning algorithm used for classification and regression analysis. They are particularly useful in cases where the data has a clear margin of separation between the classes. The SVM (Support Vector Machine) algorithm identifies the hyperplane that can most effectively divide the data into categories by optimizing the margin, which is the separation distance between the hyperplane and the nearest data points from each group. These nearest data points, known as support vectors, are essential for determining the hyperplane. In the case of binary classification, the SVM algorithm outputs a binary class label, assigning each data point to one of the two classes. In the case of multi-class classification, the SVM algorithm can be used in combination with other techniques to make predictions. SVMs can also be used for regression tasks by finding the hyperplane that best fits the data. To make predictions for a new data point, the point is projected onto the hyperplane

SVMs are highly effective in high-dimensional spaces and can manage data that is not linearly separable by utilizing kernels. They are also robust to outliers, but require a lot of computational resources, especially when working with large datasets[18].

Decision Tree: A decision tree is a popular machine learning algorithm that is used for both regression and classification problems. It is a tree-like model that can be used to represent a series of decisions and their possible consequences.The decision tree is formed by repeatedly dividing the data into smaller subsets using the most effective features for separating the data. In each node of the tree, the algorithm selects the feature that offers the highest information gain, or reduction in entropy, and splits the data into two or more child nodes. This process is repeated until a halt criterion is fulfilled, such as a minimum number of samples in a terminal node or the highest level of the tree. The final result is a tree of decisions that can be used to make predictions by following the branches to the leaves, where the predictions are made based on the class labels or regression outputs associated with the data in that leaf node.Decision trees are straightforward to comprehend and interpret, and can process both categorical and numerical features, making them a widely used method for both simple and complex problem[19].

Random Forest: Random Forest is an ensemble learning method for classification and regression that builds multiple decision trees and combines their predictions. The idea behind using multiple trees is that they can provide a more accurate and stable prediction compared to a single decision tree, which can be prone to overfitting to the training data. In a Random Forest, the algorithm randomly selects a subset of the features for each split in each decision tree, instead of always using the best feature as in a normal decision tree. This decorrelates the trees and makes the model more robust and less susceptible to overfitting. To make a prediction, the Random Forest combines the predictions of each individual tree by taking the average (for regression) or majority vote (for classification) of the predictions. This combination of multiple decision trees can lead to improved performance and better generalization to unseen data. Random Forest is widely used in many applications due to its simplicity, interpretability, and strong performance on a wide range of tasks. It is also a good choice for high-dimensional and complex data, as the combination of many trees can help mitigate overfitting and improve the model's ability to capture complex relationships in the data[20].

3. EXPERIMENTS

3.1. Dataset

The Wisconsin breast cancer dataset contains information on 569 patients with breast cancer, including 30 numerical features that describe the characteristics of their tumors. Each feature represents a different measure, such as the radius of the tumor, the texture of its tissue, or the smoothness of its boundaries. The target variable is a binary label indicating whether the tumor is benign (0) or malignant (1).

3.2. Model selection

Four Machine learning algorithms were selected namely Logisctic Regression, SVM, Decision Classifier and Random Forest. The dataset was split into 70:30 that is 70% of data was selected as training data while as 30% of data was selected as test data. All the machine learning models were trained on the training data and then evaluated on the test data.

3.3. Evaluation Metrics

The methods and metrics employed to assess the model outcomes are presented in this subsection. The chosen evaluation metrics take into consideration both a classifier's general classification performance and its ability to accurately categorise minority data. The models were evaluated on the basis of Accuracy, F1 score and ROC Curve.

3.4. Experimental Results

The objective of this study is to classify breast cancer with different machine learning algorithms. It was ensured that the train data and test data stayed the same throughout the experimentation, regardless of the machine learning algorithm employed.

3.4.1. Comparative analysis of Machine learning algorithms on Breast Cancer Dataset

The effectiveness of different machine learning algorithms on the Breast cancer dataset was evaluated by comparing their F1 Score and Accuracy metrics. The algorithms assessed in this study included Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest.

The results in Table 1 indicate that SVM achieved the highest performance, with an F1 Score of 98% and an Accuracy of 97.66%. This finding

Table 1: Comparison of machine learning algorithms on Breast cancer dataset.

Algorithm	**F1 Score**	**Accuracy**
Logistic Regression	97%	97.07%
SVM	98%	97.66%
Decision Treee	92%	92.39%
Random Forest	94%	94.15%

is consistent with previous research that has identified SVM as a powerful machine learning algorithm for cancer classification. Logistic Regression also demonstrated strong performance, achieving an F1 Score of 97% and an Accuracy of 97.07%.

The comparatively lower scores of Random Forest and Decision Tree algorithms suggest that they may not be the optimal choice for breast cancer classification using this dataset. However, further research is needed to investigate the potential benefits and limitations of these algorithms in other contexts.

Overall, these findings provide valuable insights into the performance of different machine learning algorithms on the Breast cancer dataset and contribute to the ongoing efforts to develop effective tools for cancer diagnosis and treatment.

The ROC curve of different algorithms are shown in Fig 1. SVM has the highest performance among the evaluated algorithms, with an AUC of 0.978. This means that the SVM model has a higher overall ability to distinguish between positive and negative classes across all possible classification thresholds compared to the other models.

Logistic Regression also demonstrated strong performance with an AUC of 0.971, indicating that it is a close second to the SVM model in terms of classification accuracy.

The comparatively lower AUC values of Decision Tree (0.930) and Random Forest (0.944) suggest that they may not be the optimal choice for binary classification tasks using this dataset.

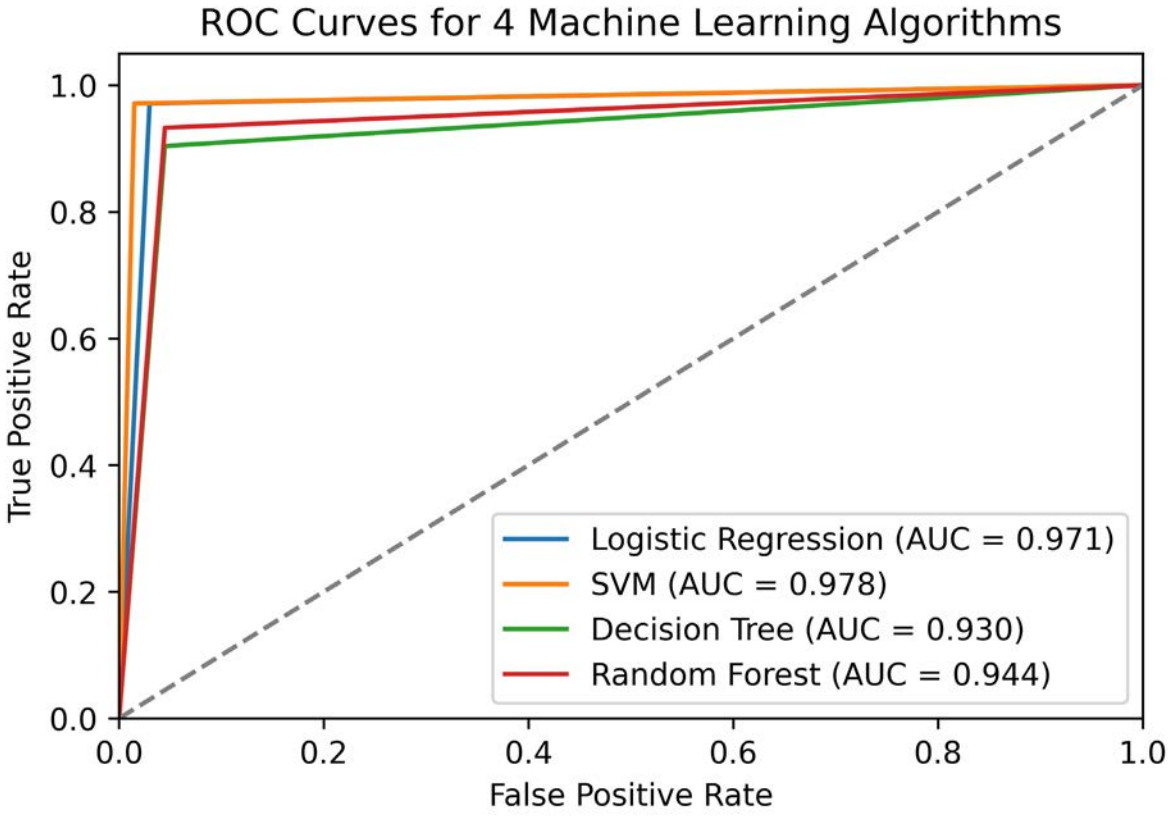

Figure 1: ROC curve of four different Machine learning algorithms

4. Conclusion

In conclusion, the early detection of breast cancer through the use of ML techniques is an important area of research. These techniques can assist in the identification of patterns in data sets, thus predicting the malignancy or benignancy of breast cancer. The Logistic Regression, Support Vector Machines, and Decision Trees algorithms are powerful tools in the classification of breast cancer. The literature survey conducted shows that Random Forest, kNN, and Naive Bayes algorithms can be used for predicting breast cancer.In this paper four different ML algorithms were evaluated on the Breast cancer dataset and it was observed that SVM performed better than others with an accuracy of 97.66% and an F1 score of 98%. With further development and application of these techniques, the battle against breast cancer can be improved.

References

[1] T. Klonisch, E. Wiechec, S. Hombach-Klonisch, S. R. Ande, S. Wesselborg, K. Schulze-Osthoff, M. Los, Cancer stem cell markers in common cancers–therapeutic implications, Trends in molecular medicine 14 (2008) 450–460.

[2] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, A. Jemal, Global cancer statistics, 2012, CA: a cancer journal for clinicians 65 (2015) 87–108.

[3] R. Chandrasekar, V. Palaniammal, M. Phil, Performance and evaluation of data mining techniques in cancer diagnosis, IOSR Journal of Computer Engineering (IOSR-JCE) 15 (2013) 39–44.

[4] M. Cohen, F. Azaiza, Early breast cancer detection practices, health beliefs, and cancer worries in jewish and arab women, Preventive medicine 41 (2005) 852–858.

[5] P. Etingov, N. Voropai, Application of fuzzy logic pss to enhance transient stability in large power systems, in: 2006 International Conference on Power Electronic, Drives and Energy Systems, IEEE, 2006, pp. 1–9.

[6] C. Taylor, Machine learning, neural and statistical classification: Neural and statistical classification, 1994.

[7] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artificial Intelligence in medicine 23 (2001) 89–109.

[8] S. Sharma, A. Aggarwal, T. Choudhury, Breast cancer detection using machine learning algorithms, in: 2018 International conference on computational techniques, electronics and mechanical systems (CTEMS), IEEE, 2018, pp. 114–118.

[9] L. Hussain, W. Aziz, S. Saeed, S. Rathore, M. Rafique, Automated breast cancer detection using machine learning techniques by extracting different feature extracting strategies, in: 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), IEEE, 2018, pp. 327–331.

[10] A. Alarabeyyat, M. Alhanahnah, et al., Breast cancer detection using k-nearest neighbor machine learning algorithm, in: 2016 9th International Conference on Developments in eSystems Engineering (DeSE), IEEE, 2016, pp. 35–39.

[11] B. Gayathri, C. Sumathi, T. Santhanam, Breast cancer diagnosis using machine learning algorithms–a survey (2013).

[12] D. Bazazeh, R. Shubair, Comparative study of machine learning algorithms for breast cancer detection and diagnosis, in: 2016 5th international conference on electronic devices, systems and applications (ICEDSA), IEEE, 2016, pp. 1–4.

[13] N. Khuriwal, N. Mishra, Breast cancer diagnosis using deep learning algorithm, in: 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), IEEE, 2018, pp. 98–103.

[14] A. F. M. Agarap, On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset, in: Proceedings of the 2nd international conference on machine learning and soft computing, 2018, pp. 5–9.

[15] W. Yue, Z. Wang, H. Chen, A. Payne, X. Liu, Machine learning with applications in breast cancer diagnosis and prognosis, Designs 2 (2018) 13.

[16] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. Shih, J. Tomaszewski, F. A. González, A. Madabhushi, Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent, Scientific reports 7 (2017) 1–14.

[17] M. P. LaValley, Logistic regression, Circulation 117 (2008) 2395–2399.

[18] V. Jakkula, Tutorial on support vector machine (svm), School of EECS, Washington State University 37 (2006) 3.

[19] C. Kingsford, S. L. Salzberg, What are decision trees?, Nature biotechnology 26 (2008) 1011–1013.

[20] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2016) 197–227.