Performance Analysis of Random Forest (RF) and Support Vector Machine (SVM) Algorithms in Classifying Breast Cancer

FHA. Shibly[1], Uzzal Sharma[2], HMM. Naleer[3]

South Eastern University of Sri Lanka, Oluvil, Sri Lanka[1,3]

Assam Don Bosco University, Guwahati, Assam, India[2]

Corresponding author: FHA. Shibly, Email: shiblyfh@seu.ac.lk

Breast cancer is a severe illness cause of mortality among females. In cancer diagnosis, accurate classification of breast cancer data is critical, and the distinction between malignant and benign tumors can help patients avoid unnecessary procedures. The categorization of breast cancer may also be used to determine the appropriate treatment choices. The categorization of patients into benign and malignant categories is a well-known medical research issue. Machine learning in Artificial Intelligence is commonly employed in predicting these types of cancer because it has the lead of finding important aspects from a medical data gathering. Several empirical types of research have used machine learning and soft computing techniques to treat breast cancer. Many people claim that their algorithms are better than others because they are faster, easier, or more accurate. Therefore, which algorithm is more accurate in classifying breast cancer was the research question. Furthermore, the major purpose of this research project is to calculate and evaluate the performance of SVM and RF algorithms in detecting breast cancer more correctly. The Wisconsin Breast Cancer Data Set (WBCD) is adopted for the empirical analysis. There are a total of 699 instances and 10 qualities to be examined. Based on the Accuracy, Reminder, Precision and F1 values, RF has the higher ratios in all four measurement scales with 92.98%, 93.65%, 88.05% and 90.67%, respectively. Therefore, RFs have the best probability of successfully diagnosing breast cancer.

Keywords: Machine Learning, Algorithms, Breast Cancer, Performance.

2021. In Raju Pal & Praveen K. Shukla (eds.), *SCRS Conference Proceedings on Intelligent Systems*, 253–259. Computing & Intelligent Systems, SCRS, India.
https://doi.org/10.52458/978-93-91842-08-6-25

1 Introduction

Tumors are the uncontrolled development of cells in an organ that may be malignant. Tumors are classified as benign or malignant. Benign tumors do not grow and do not pose a threat to one's health. Malignant or cancerous tumors, on the other hand, are growing and posing a life-threatening danger [1]. Regular breast cancer screenings, accompanied by adequate cancer care, will help to minimize the risk of developing the disease. Every 4-6 weeks, a tumor assessment test is recommended. As a result, detecting benign and malignant tumors using classification features is important [2]. Malignant breast cancer is characterized as the presence of proliferating cells in the breast tissue. Breast cancer is the second biggest cause of death in women over the age of 40, and the highest among those aged 40 to 55. But, early detection and thorough diagnosis have been found to minimize the mortality rate from breast cancer [3]. Medical personnel occasionally may make tiny errors while diagnosing a condition, depending on their degree of expertise. With the use of technology like data mining and machine learning, diagnosis may be more accurate (91.1 %) than a diagnosis conducted by an experienced doctor (79.9 %) (79.9 %) [4]. However, there are no proper mechanisms to mitigate and prevent BC, detecting it in the early stage can considerably recover the prognosis. Furthermore, as a result of this, care expenditures will be greatly lowered. Early detection, however, can be challenging due to the rarity of cancer signs. Mammograms and self-breast examinations are critical for recognizing early anomalies before a tumor advance [5].

There are some studies available on automatic detection of breast cancer [5, 6, 7]. Mainly Artificial Intelligence and its branches have taken some key roles. Specially Machine Learning is one of the efficient ways to detect many things automatically with structured datasets. Machine Learning is an Artificial Intelligence approach that generally utilizes a broad variety of probabilistic, optimization, and statistical methods to boost efficiency based on previous experiences and new data. Disease prediction has been extensively used utilizing several machine learning approaches. It is easier to separate patients with Breast Cancer from others by applying the branch of artificial intelligence like machine learning algorithms these days. Clinicians will be able to identify cancer at an early stage with the support of correct categorization. Classification is a supervised and tough task to tackle. To classify cancer data, numerous classification algorithms such [6].

During the preceding several decades, artificial intelligence, deep learning, and machine learning algorithms have already been utilized in the creation of prediction models to enhance effective decision-making. These ML algorithms are great to employ on cancer investigations to find many findings in a data collection and thus determine whether a tumor is malignant or benign. The area under the ROC, recall, precision, and classification accuracy may all be used to quantify the success of such techniques [7].

The major purpose of this essay is to see how effective machine learning is at diagnosing breast cancer. This study compares two common machine learning techniques utilizing a breast cancer data set. These techniques could be incorporated in medical research especially in cancer-based studies.

This work focuses on the usage of such machine learning algorithms to classify breast cancer and the successful classification method in them. This study will help to determine the best performing algorithms, especially on the Support Machine and Random Forrest algorithms in classifying breast cancer. This contribution will be useful for medical purposes for taking timely decisions to save human lives. The outline of this research is a literature review, methodology, proposed approach, results, discussion, and conclusions.

2 Literature Review

This section offers a review of the literature. For breast cancer detection studies, relevant literature from several sources is studied. In addition, the authors looked at data from regional and national cancer registries.

Medical diagnosis, like a breast cancer diagnosis, is increasingly relying on classification schemes. A key consideration is the process of assessment and decision-making based on expert medical diagnosis. An intelligent classification algorithm, on the other hand, could help doctors, particularly when it comes to reducing errors caused by some practitioners who have a lack of experience in detecting such diseases [2]. Many mechanisms were carried out to anticipate and discover relevant trends in breast cancer diagnosis. Ryua [8] invented the isotonic separation method for data classification. The effects of assistance, and other techniques were compared using data from two breast cancer patients.

Sahan [9] employed a mixed machine learning technique to detect breast cancer. In this technology, a fuzzy-artificial immune system was merged with the k-nearest neighbor algorithm. In the Wisconsin Breast Cancer Dataset, the hybrid strategy worked well (WBCD) (WBCD). It can also be used to screen for additional breast cancer diagnostic challenges, according to the researchers.

The SVM classifier procedure [10] combines RFE with SVM. RFE is a recursive approach to selecting dataset features based on the lowest feature value. As a result, the wrong characteristics (characteristic with the lowest weight) are deleted in all rounds of SVM-RFE. In [11] the authors used four ML algorithms to breast cancer data. The analyses were applied with the WEKA tool. The SVM classifier achieved an accuracy of 97.13 percent of four machine learning methods.

In [12], the authors examined the suitability of algorithms for categorizing the disease. The above data set is used for classifying purposes widely. ML and DL algorithms were used to evaluate the data set. In addition to the ANN, numerous ML algorithms have dominated the findings on breast cancer [13-15].

An outline of the part of machine learning in breast cancer diagnosis is listed by Agarap et al. given [16]. GRU-SVM, MLP, NN, linear regression, SVM, and Softmax regression were among the six approaches used for machine learning. The Wisconsin Diagnostic data collection for breast cancer was used in the empirical study. The data set is divided into two sections: training (70 %) and testing (20%) (30%). (30%). With 99.04 %, MLP had the highest accuracy of all classifiers. The Random Forest (RF) strategy is based on a recursive technique according to Yasui and Wang (2003), in this each iteration consists of taking an arbitrary sample of size N from the data set with replacement and an additional random sample. There is no replacement for the predictors. The information attained is then employed to architect a module. Breast cancer classification model based on three machine learning algorithms Murugan et al [19] applied RF, LR, and DT, with LR achieving an accuracy of 84.14 % and random forest achieving an accuracy of 88.14 %.

Several machine learning algorithms [20] were implemented to predict breast cancer in the investigation described above, but most methods failed to attain outstanding metrics in breast cancer diagnosis. This study, based on priorworks, evaluates the efficacy of SVM and RF algorithms in categorizing breast cancer data.

3 Methodology

To classify breast cancer, SVM and RF algorithms have been used in this research. For the experimental analysis, the Wisconsin Breast Cancer Data Set (WBCD) is employed. There are a total of 699 instances and 10 qualities to examine. WBCD can be found at [21]. Two individual algorithms will be discussed in the subsections. (See figure 1)

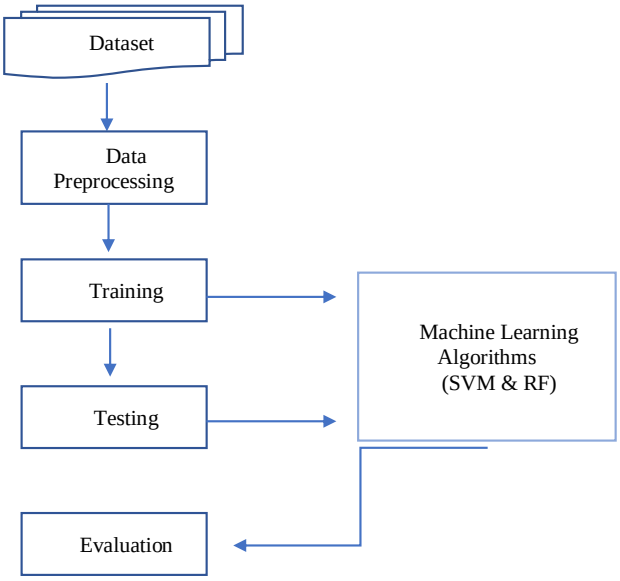

Fig. 1. Research Method

Support Vector Machine (SVM)

SVM is an abbreviation for supervised machine learning classification and is commonly used in cancer diagnosis and prognosis. SVM works by gathering key examples from all of the classes known as support vectors and then producing a linear function that separates them as much as possible. As a result, it can be claimed that SVM is used to transfer an input vector into high-dimensional space to discover the ideal hyperplane for categorizing the data [22]. By choosing the most appropriate hyperplane, this linear classifier seeks to maximize the distance between the decision hyperplane and the next data point, also known as marginal dispersion [23].

A scatter plot of two classes, each with two properties, is demonstrated in Figure 2. The sake of a linear hyperplane is to find a, b, additionally c implies such that for class 1 ax1 + bx2 6 and c suggest and for class 2 ax1 + bx2> c [22] [24]. In contrast to other approaches, SVM algorithms depend on support vectors, which are the data sets closest to the decision boundaries. This is the case because of the weaker influence on the boundary than deleting data points that are added from the decision hyperplane than removing supporting vectors [18].

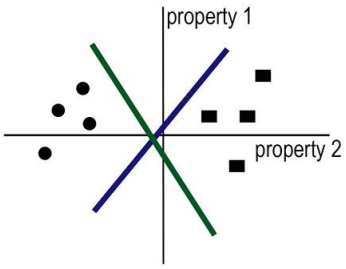

Fig. 2. Workflow of SVM

Random Forest (RF)

RF is not affected by the noise in the input data set. The ability to regulate data minorities is one of the main reasons for RF's usage in cancer diagnosis. Although just 10% of the input data set is used in the later class, the tumor may be categorized as either benign or malignant.

The RF methodology comprises picking a random N-sample from a replacement dataset and another random N-sample from the predictors without replacement in iteration [Fig 3]. The RF technique is used to describe each process. The data is subsequently divided. The outside data is then emptied, and the earlier processes are performed as often as necessary, depending on the number of trees necessary. Finally, two groups of trees split the observation. The decision-making bodies are afterward utilized to classify circumstances by majority vote [17].

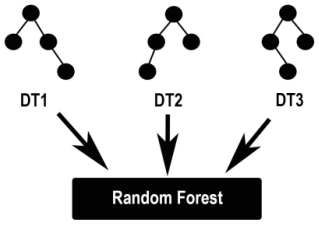

Fig. 3. Workflow of RF

Proposed Approach

This research uses the WBCD dataset and evaluates the values before being tested. RF and SVM algorithms were used to calculate the performance of both recording, precision, precision, and F1 score algorithms to effectively classify breast cancer for the performance assessment purpose. The methodology proposed is shown in Figure 4 for this research.

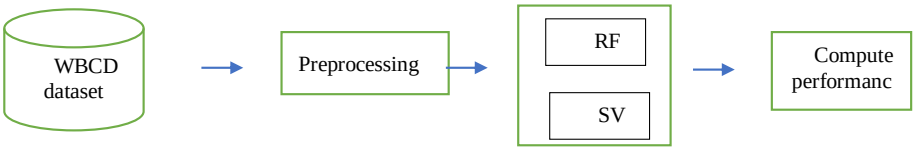

Fig. 4. Proposed Approach

4 Results

This section explains the settings and gives the data that helps the two classifiers being studied here.

A. Accuracy

The precision of a classifier is to determine how correctly instances can be classified into the relevant category. This is the total cases in numbers in the given dataset divided by the right predictions. It's worth noting that the classifier's threshold, which varies for different testing sets, has a significant impact on accuracy. As a result, while it can provide a basic view of the class, it is not the ideal tool for comparing different classifiers. The accuracy of a classifier is a measure of how well it can accurately classify situations into the appropriate category. It's determined by multiplying the total number of instances in the data set by the number of right forearms.

B. Recall

The rate of properly predicted positive observations is defined as the percent of favorable notes that are properly anticipated as positive, also known as recall [12]. This is a crucial metric, particularly in the field of medicine, because it shows how many observations are accurately classified. Correctly detecting a dangerous tumor is more significant in this study than wrongly diagnosing a benign tumor.

C. Precision

The proportion of real positives and real negatives that have been flagged as real positives is known as precision, also known as confidence. This demonstrates that the classifier can accept positive input while rejecting negative data.

D. F1 Score

F1 is a metric for a model's accuracy that combines precision and recall, similar to how addition and multiplication combine two ingredients to make a new dish. A high F1 score suggests that you have a low number of false positives and false negatives, meaning that you are accurately assessing significant threats and are unaffected by false alarms. The F1 score of 1 implies that the model is flawless, while a score of 0 indicates that the model is a total failure.

Based on the analysis, the accuracy, recall, precision, and F1 scores are given below for both algorithms (Table 1).

Table 1. Performance comparison of SVM and RF

Algorithm	Accuracy	Recall	Precision	F1 Score
SVM	91.81 %	92.06 %	86.56%	89.23%
RF	92.98 %	93.65 %	88.05 %	90.76 %

5 Discussion and Conclusion

In terms of measurement matrices, the findings are shown in Table I reveal that Random Forest (RF) has the best performance. Support Vector Machine (SVM) algorithm, on the other hand, has good results and identifies breast cancer somewhat lower than RF. This means that RF has a better probability of distinguishing between benign and malignant instances.

Machine learning techniques are frequently employed in the scientific field as a beneficial diagnostic tool to help medical people evaluate data and construct medical expert systems. Support Vector Machine (SVM) and Random Forest, two of the most widely used machine learning algorithms for breast cancer detection and diagnosis, were presented in this paper. The major features and techniques of each of the two ML algorithms were discussed. The performance of the investigated approaches was compared utilizing the Original Wisconsin Breast Cancer Data collection. According to the findings, RF has better values and best practices in classifying breast cancer data somewhat than SVM.

References

[1] Subashini, T. S., Ramalingam, V. and Palanivel, S. (2009). Breast mass classification based on cytological patterns using RBFNN and SVM. *Expert System with Applications*, 36(3):5284–5290.

[2] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert System with Applications*, 36(2):3240–3247.

[3] West, D. et al. (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research*, 162(2):532–551.

[4] Brause, R. W. (2001). Medical Analysis and Diagnosis by Neural Networks. In *Proceedings of the Second International Symposium on Medical Data Analysis*, 1–13.

[5] Shallu and Mehra, R. (2018). Breast cancer histology images classification: Training from scratch or transfer learning?. *ICT Express*, 4:247–254.
[6] Jabbar, M. (2021). Breast cancer data classification using ensemble machine learning. *Engineering and Applied Science Research*, 48(1):65-72.
[7] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89-109.
[8] Ryu, Y. U., Chandrasekaran, R. and Jacob, V. S. (2007). Breast cancer prediction using the isotonic separation technique. *European Journal of Operational Research*, 181(2):842–854.
[9] Sahan, S. et al. (2007). A new hybrid method based on the fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. *Computers in Biology and Medicine*, 37(3):415–23.
[10] Reddy, V. A. and Soni, B. (2020). Breast cancer identification and diagnosis techniques. *Machine Learning for Intelligent Decision Science*, 49-70.
[11] Asri, H. et al. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, 83:1064-1069.
[12] Shailaja, K., Seetharamulu, B. and Jabbar, M. A. (2018). Machine learning in healthcare: a Review. *Second International Conference on Electronics, Communication and Aerospace Technology*, 910-914.
[13] Shailaja, K., Seetharamulu, B. and Jabbar, M. A. (2018). Prediction of breast cancer using big data analytics. *International Journal of Engineering and Technology,* 7(4(6)):223-226.
[14] Jabbar M.A., Samreen, S. and Aluvalu, R. (2018). The future of healthcare: machine learning. *International Journal of Engineering and Technology,* 7(4(6)):23-25.
[15] Douangnoulack, P. and Boonjing, V. (2018). Building minimal classification rules for breast cancer diagnosis. *10th International Conference on Knowledge and Smart Technology,* 278-281.
[16] Agarap, A. F. M. (2018). On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset. *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, 5-9.
[17] Yasui, Y. and Wang, X. (2009). Statistical Learning from a Regression Perspective. Springer.
[18] Bazazeh, D. and Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. *5th International Conference on Electronic Devices, Systems and Applications*, 1-4.
[19] Murugan, S., Kumar, B. M. and Amudha, S. (2017). Classification and prediction of breast cancer using linear regression, decision tree, and random forest. *International Conference on Current Trends in Computer, Electrical, Electronics and Communication,* 763-766.
[20] Kumar, U. K., Nikhil, M. B. S. and Sumangali, K. (2017). Prediction of breast cancer using voting classifier technique. *IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy, and Materials*, 108-114.
[21] Wolberg, W. H., Street, W. N. and Mangasarian, O. L. (2020). Breast Cancer Wisconsin (Diagnostic) Data Set. *https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)*.
[22] Williams, G. (2011). Descriptive and Predictive Analytics. *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!)*.193-203.
[23] Kourou, K. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8-17.
[24] Cleophas, T. J. and Zwinderman, A. H. (2013). Machine Learning in Medicine. Springer.