Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 1957-1964
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5965

# Comparison of classification techniques based on medical datasets

Alyaa Abdulhussein Al-Joda[a], Enas Fadhil Abdullah[b,], Suad A. Alasadi[c]

[a]*Engineering Technical College of Al-Najaf, Al-Furat Al-Awsat Technical University(ATU), Al-Najaf, Iraq*
[b]*Faculty of Education for Girls, University of Kufa, Al- Najaf, Iraq*
[c]*College of Information Technology, University of Babylon, Babil, Iraq*

*(Communicated by Madjid Eshaghi Gordji)*

**Abstract**

Medical data mining has been a widespread data mining area of late. Mainly, diagnosing cancers is one of the most important topics that many researchers studied to develop intelligent decision support systems to help doctors. In this research, three different classifiers are used to improve the performance in terms of accuracy. The classifiers are Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), and Random forests (RF). Two machine learning repository datasets are used to evaluate and verify the classification methods. Classifiers are trained using the 10-fold cross-validation strategy, which splits the original sample into training and testing sets. In order to assess classifier efficiency, accuracy (AC), precision, recall, specificity, F1, and area under the curve are used (AUC). The Experiments showed that the AdaBoost classifier's achieved an accuracy of 100% which is superior in both datasets in comparison with SVM and RF with AC of 97%. The accuracy is also compared with another study from the previous work that uses the same datasets, and the results demonstrated that the current research has better accuracy than the other study.

*Keywords:* Classifier, AdaBoost, SVM, RF, ROC, Breast Cancer.

## 1. Introduction

Breast cancer is the second cause of the death for women [18]. Near 268600 new instances in 2019 had been probable to be identified in women in the United States for invasive breast cancer, along with 62930 new cases of breast cancer that are non-invasive [25]. To enlarge the chance of therapy and survivability, early detection is the best satisfactory way. The significant development in the discovery

*Email addresses:* dr.alyaa@atu.edu.iq (Alyaa Abdulhussein Al-Joda), inasf.alturky@uokufa.edu.iq (Enas Fadhil Abdullah), suad.alsady@uobabylon.edu.iq (Suad A. Alasadi)

*Received:* October 2021 *Accepted:* December 2021

of knowledge especially the methods of data mining helped in medicine and in several other fields such as finance marketing, and social science [7, 22]. Many classifiers have been developed in recent years to perform prediction evaluations on patient's medical prognosis using scientific datasets. For example, the use of machine learning methods to evaluate breast cancer patient's tumor behavior [16, 26, 6]. Clustering methods are used to anticipate particular classification labels in order to handle large amounts of data. Classification models are used to assign newly available data to a category label. Classification is the process of developing a model that adequately defines and differentiates distinct data classes or ideas. These methods are capable of classifying both category and numerical characteristics. There are many classification methods including Naive Bayesian, k-nearest neighbor, SVM, SMO, and random forest [21]. This paper compares the accuracy of three different classifiers including Support Vector Machine (SVM), Random Forests (RF), and the AdaBoost in the diagnosis of breast cancer. To enhance the classifier's performance, the dataset is preprocessed with an appropriate technique for managing missing values and the dataset's imbalance.

## 2. Literature review

In recent years, data mining methods such as classification on different medical datasets are used for breast cancer by many researchers. These algorithms provide accurate classification results and they are often used to predict and classify abnormal events, to get a better understanding of incorrigible diseases like cancer. The discoveries of data mining-based classification are promising for the detection of breast cancer [6]. Table 1 summarizes some of the researches that have been conducted on this technique. These studies used the Breast Cancer Dataset (BCD) and the Wisconsin Breast Cancer (WBC) Dataset [19], respectively.

Table 1: Some methods for breast cancer diagnosis

| Paper Reference | Datasets | Classification Algorithms and accuracy | | | |
|---|---|---|---|---|---|
| [21], 2019 | BC | NB | SVM | GRNN | J48 |
| | | 89% | 89% | 91% | 91% |
| [14], 2017 | WPBC | KNN | NB | C5.0 | SVM |
| | | | | 81% | 81% |
| [15], 2016 | WPBM | NB | C4.5 | SVM | |
| | | 67.17% | 73.73% | 75.75% | |
| [16], 2016 | WBC | SVM | C4.5 | NB | KNN |
| | | SVM : 97.13% | | | |
| [11], 2016 | WDBC | NB | SVM | Ensemble | |
| | | 97.3% | 98.5% | 97.3% | |
| [12], 2020 | BC | NB | SMO | J48 | |
| | | 76.61% | 95.32% | 98.20% | |
| | WBC | NB | SMO | J48 | |
| | | 99.12% | 99.56% | 99.24% | |

The classification techniques that are used to create prediction models and evaluate their accuracy results in this paper are (SVM), (RF), and (AdaBoost) algorithm.

## 3. Background

In the knowledge discovery process, data mining is the main step. Some of the disciplines of data mining are data science, database technology, statistics, visualization, and machine learning

[8]. Data classification is an issue with multiple applications in a wide range of mining applications. Classification is used because the challenge tries to figure out how a set of feature variables relates to a target variable of interest. There is a wide range of models that can be utilized because many practical problems may be described as relationships between feature and goal variables [8]. A well-defined specification of the classes and a training collection of pre-classified cases characterize the classification problem. The goal is to create a model that can be applied to unclassified data to classify it. Classification is the process of developing a model that adequately explains and differentiates various data classes or ideas. SVM, SMO, fuzzy set method, decision tree, and Bayesian inference are all examples of machine learning techniques [17]. Support vector machine is a kind of statistical learning theory that is based on the concept of structural risk reduction. Support vector machines and neural network approaches are collections of supervised learning methods for classifying both linear and nonlinear data, performing regression, and detecting outliers. SVM classifies data into two groups along a hyperplane while avoiding overfitting by raising the hyperplane separation margin. [9, 1]. The (RF) Method is a supervised learning method that may be used for classification and regression. It is also the easiest to use and most flexible algorithm around. Trees form a forest, and the more trees there are in a forest, the stronger it is. RF creates decision trees from random data samples, gets predictions from each tree, and then votes on the best one. Additionally, it serves as a fairly accurate indication of the feature's significance [1, 23].

The ( AdaBoost) method, which is often employed in a variety of areas, where several weak classifiers are combined into a robust classifier [24]. AdaBoost was created by Robert Schapire and Yoav Freund in 1995. This algorithm uses updated sample of the old weight to train a new frail classifier. Each time the weak classifier is trained on the sample population, new sample weights are generated and the process continues until the target error rate is attained or the maximum number of iterations is reached [12]. Imbalanced data occurs while there is a substantial variance among the classes in the dataset. Prediction models made from imbalanced datasets are regularly biased in the direction of the majority concept, while there is a better value of misclassifying for diagnosing uncommon diseases [5]. The k-fold Cross-validation (CV) is a widespread technique to assess classification performance [20].

The CV is split data into k subgroups where each time is used one of the k subsets. A training set is formed from the validation/test set and the other $(k-1)$ subsets. The mistakes estimation is an average of overall (k) trials to get the entire usefulness of the model. It significantly decreases bias to use of a maximum of the data for fitting [27].

## 4. Datasets

Two datasets used in this study are available at the University of California, Irvine's Machine Learning Repository [19]. Breast Cancer Dataset (BC) and Wisconsin Breast Cancer (WBC) are the datasets in question. The BC dataset contains 286 instances where 201 examples containing no recurrence events and 85 containing recurrence occurrences. The BC dataset contains ten characteristics, for this dataset are derived from a high-quality digital picture for needle aspirate of a breast tumor. In the target feature records, the prediction is made (i.e., malignancy or benign) [27]. A total of 699 instances and 11 features were found in the WBC dataset, with 458 being benign and 241 being malignant. The characteristics (Bare Nuclei) status was missing of 16 submissions. This dataset requires extensive data preparation to deal with missing values and imbalanced data [28, 3, 10, 2, 4, 13].

## 5. Research methodology

Before applying any classifier, preprocessing is done on the data to remove missing and unbalanced data and improve the classifier's performance. All occurrences with missing values are deleted to control the missing points. The issue of asymmetric data requires regulation of both the training set and the classifier. To do this, tenfold cross-validation is used, followed by an assessment of the three classifiers. To begin, the preprocess removes any missing values from the dataset. Second, to mitigate the preference associated with a random selection of training data, 10 fold cross-validation was used, and then research was conducted on the three classifiers SVM, RF, and AdaBoost, as shown in Figure 1.

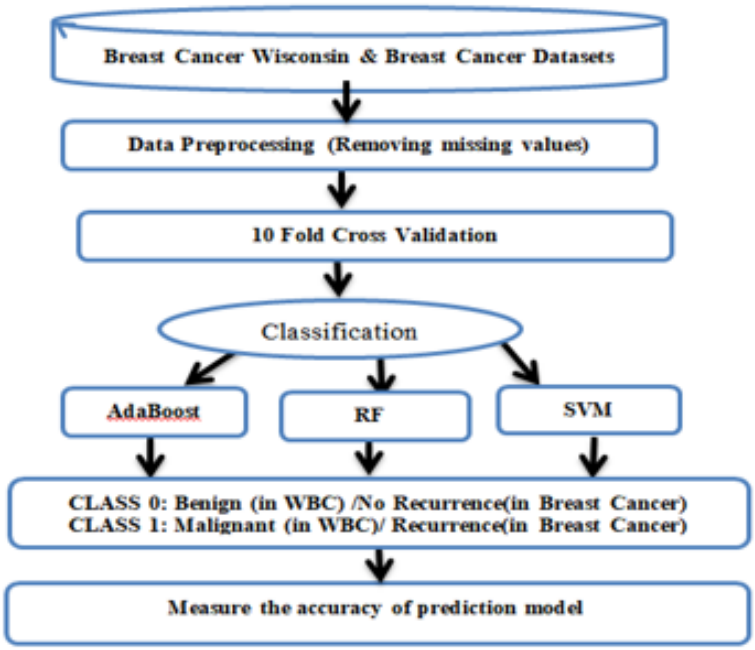


Figure 1: Breast cancer detection model

Six performance metrics are used in this study to assess all classifiers. They are as follows: Precision, the area under curve (AUC), specificity and accuracy (AC), F1, and recall [3, 10].

## 6. Experimental results

In the presented study, Breast Cancer and WBC were used to validate the classification methods. The classification efficiency has been evaluated and displayed in Table 2 for the BC dataset. The classification accuracy for the AdaBoost classifier is 0.981, which is better than SVM that has an accuracy of 0.860. The RF is the lower accuracy of 0.786.

The scatter plot of the classes for classifications algorithms for the Breast Cancer dataset is displayed in figure 2.

Table 2: Classification evaluation measurements for breast cancer dataset

| Classification Evaluation for Breast Cancer dataset | | | | | | |
|---|---|---|---|---|---|---|
| Model | AUC | AC | F1 | Precision | Recall | Specificity |
| AdaBoost | 0.999 | 0.981 | 0.981 | 0.981 | 0.981 | 0.976 |
| SVM | 0.914 | 0.860 | 0.845 | 0.878 | 0.860 | 0.655 |
| RF | 0.767 | 0.786 | 0.757 | 0.782 | 0.786 | 0.527 |

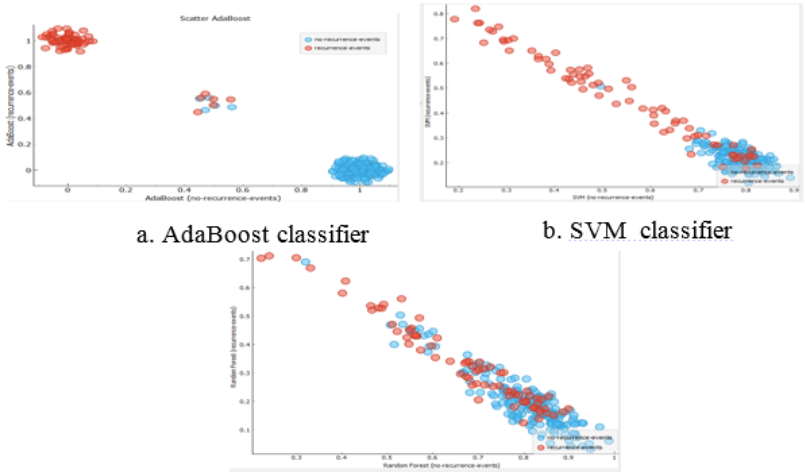


Figure 2: Classifiers scatter plot for Breast Cancer dataset

The AdaBoost classifier success in detecting classes as shown in figure 2. The misclassified classes are shallow and shown in the middle of the scatter. The SVM can classify no recurrence-events class very well rather than recurrence events where some classes are inaccurate. The RF classifier is the lowest accuracy in detecting classes where there are overlapping between classification results.

The ROC curve for the no recurrence-events and recurrence events for Breast Cancer dataset is shown in figure 3. The ROC is explained the relation between the (sensitivity) against the (specificity) at various threshold settings.

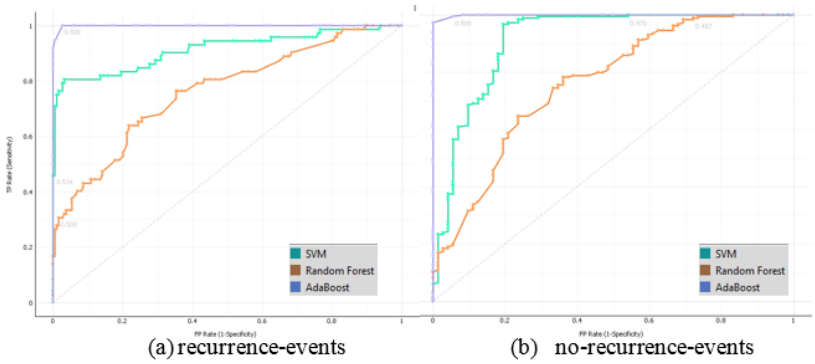


Figure 3: The RC in Breast Cancer dataset

The classification efficiency for the WBC dataset has been displayed in Table **??**.

The scatter plot of the classes for classifications algorithms for the WBC dataset is displayed in figure 4.

The performance measures for the classifiers is compared with the research proposed in [12] in which three classifiers including J48, SMO, and NB, for the same datasets are used.

Table 3: Classification Evaluation measurements for Breast Cancer dataset

| Classification Evaluation for WBC dataset | | | | | | |
|---|---|---|---|---|---|---|
| Model | AUC | AC | F1 | Precision | Recall | Specificity |
| AdaBoost | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| SVM | 0.994 | 0.976 | 0.976 | 0.976 | 0.976 | 0.973 |
| RF | 0.994 | 0.976 | 0.976 | 0.977 | 0.976 | 0.985 |

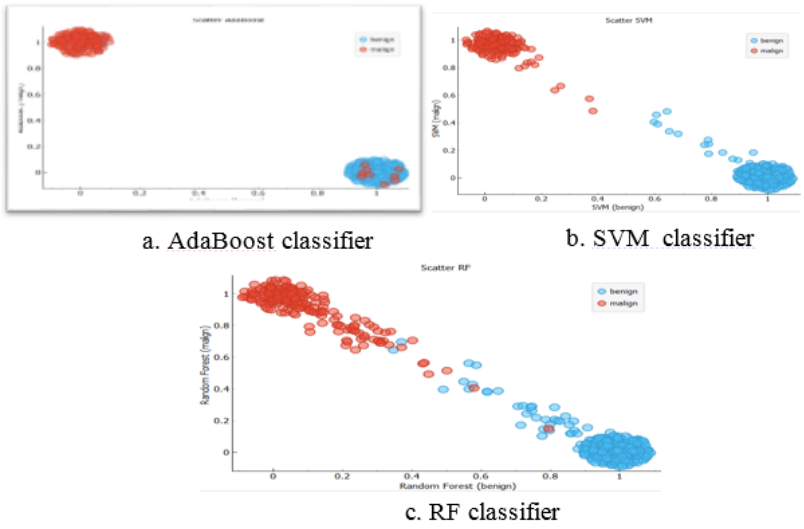


Figure 4: Classifiers scatter plot for the WBC dataset

The ROC curve for benign and malignant for WBC dataset is shown in figure 5. The ROC is explained the relation between the (sensitivity) against (specificity) at various threshold settings.

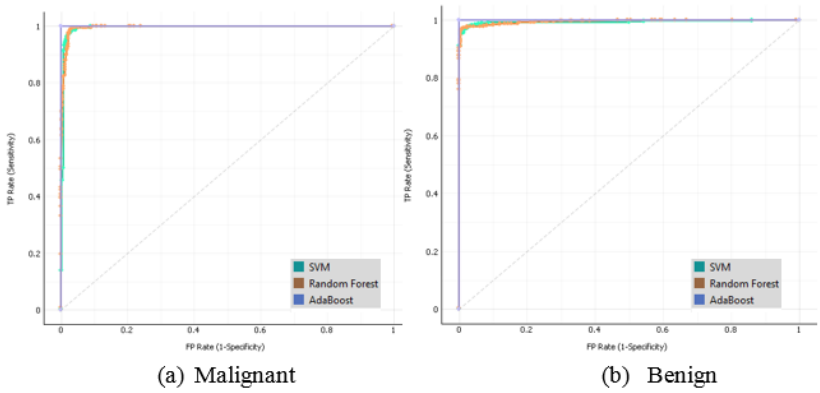


Figure 5: The ROC in WBC dataset

According to the findings shown in Table 4, all classifiers in the proposed method achieve a high level of accuracy when compared to other classifiers. [12].

## 7. Conclusions

Among women, breast cancer is considered most significant cause of death. Women's survival depends on the early detection of breast cancer. Current machine learning techniques make it feasible to identify breast cancer. Three classifier algorithms including the AdaBoost, SVM, and RF

Table 4: Comparison of the accuracy measures for BC and WBC Datasets

| Dataset | Paper [12] | used classifiers |
|---|---|---|
| BC | NB: 75.53%<br>J48: 74.82%<br>SMO: 72.66% | AdaBoost: 98.1%<br>SVM: 86%<br>RF: 78.6% |
| WBC | NB: 97.37%<br>SMO: 96.78%<br>J48: 95.91% | AdaBoost: 100%<br>SVM: 97.6%<br>RF: 97.6% |

are applied on two breast cancer benchmarking datasets were employed in this research to identify breast cancer. The AdaBoost classifier is the beast one such that it classified the breast cancer with accuracy near to 100% for the WBC dataset and 98% for the BC dataset. For SVM and RF, the accuracy is about 97% in WBC dataset, while SVM has 86% and RF has 78% as accuracy for the BC dataset. The same tests will be repeated in the future with other classifiers and datasets. The results of this study may be utilized as a guide for malignant tumor centers to improve the consistency of breast cancer diagnosis.

## References


[1] E. Abdullah, A. Lafta and S. Alasadi, *Information gain-based enhanced classification techniques*, Next Generation of Internet of Things (2021) 499–511.

[2] E. Abdullah, S. Alasadi and A. Al-Joda, *Text mining based sentiment analysis using a novel deep learning approach*, Int. J. Nonlinear Anal. Appl. 1(12) (2021) 595–604.

[3] N. Al-Aaraji, E. Al- Shamery and A. Abdulhussein, *ARNN for enhancing driftdetection of data stream based on modified page hinckley model*, J. Engin. Appl. Sci. 13(10) (2018) 8281–8291.

[4] T.A. Al-Asadi, A.J. Obaid and A.A. Alkhayat, *Proposed method for web pages clustering using latent semantic analysis*, J. Engin. Appl. Sci. 12(8) (2017) 8270–8277.

[5] H. Alghodhaifi, A. Alghodhaifi and M. Alghodhaifi, *Predicting invasive ductal carcinoma in breast histology images using convolutional neural network*, IEEE IEEE National Aerospace . Electr. Conf. 2019, pp. 374-378.

[6] V. Chaurasia and P. Saurabh, *A novel approach for breast cancer detection using data mining techniques*, Int. J. Innov. Res. Computer Commun. Engin. 2 (2017).

[7] M. Goyani and N. Patel, *Multi-level haar wavelet based facial expression recognition using logistic regression*, Int. J. Next-Generation Comput. 1 (2018) 51–131.

[8] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kauffman. 2011.

[9] J. Han, J. Pei and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier. 2011.

[10] A. Janssens and F. Martens, *Reflection on modern methods: revisiting the area under the ROC curve*, Int. J. Epidemio. 49(4) (2020) 403–1397.

[11] D. Lavanya and D. Rani, *Analysis of feature selection with classification: Breast cancer datasets*, Indian J. Comput. Sci. Engin. 2(5) (2011) 756–63.

[12] S. Mohammed, S. Darrab, S. Noaman and G. Saake, *Analysis of breast cancer detection using different machine learning techniques*, InInternational Conference on Data Mining and Big Data, 2020, pp. 108–117.

[13] J. Obaid, T. Chatterjee and A. Bhattacharya, *Semantic Web and Web Page Clustering Algorithms: A Landscape View*, EAI Endorsed Transactions on Energy Web. 8(33) (2020).

[14] U. Ojha and S. Goel, *A study on prediction of breast cancer recurrence using data mining techniques*, IEEE 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, 2017, pp. 527–530.

[15] A. Pritom, M. Munshi, S. Sabab and S. Shihab, *Predicting breast cancer recurrence using effective classification and feature selection technique*, IEEE 19th Int. Conf. Comput. Inf. Technol. (2016) pp.310–314.

[16] A. Pritom, M. Munshi, S. Sabab and S. Shihab, *Predicting breast cancer recurrence using effective classification and feature selection technique*, IEEE 19th Int. Conf. Comput. Inf. Technol. 2016, pp. 310–314.

[17] A. Saabith, E. Sundararajan and A. Bakar, *Comparative study on different classification techniques for breast cancer dataset*, Int. J. Comput. Sc. Mob. Comput. 3(10) (2014) 91–185.

[18] G. Salama, M.B. Abdelhalim and M.A. Zeid, *Experimental comparison of classifiers for breast cancer diagnosis*, IEEE. In2012 Seventh Int. Conf. Comput. Engin. Syst. 2012 pp. 180–185.
[19] G. Salama, M. Abdelhalim and M. Zeid, *Breast cancer diagnosis on three different datasets using multi-classifiers*, Breast Cancer (WDBC) 32(569) (2012).
[20] M. Santos, J. Soares, P. Abreu, H. Araujo and J. Santos, *Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches*, IEEE Comput. Intell. Mag. 13(4) (2018) 59–76.
[21] J. Silva, O. Lezama, N. Varela and L. Borrero, *Integration of data mining classification techniques and ensemble learning for predicting the type of breast cancer recurrence*. Int. Conf. Green, Pervasive, and Cloud Computing. 2019, pp. 18–30.
[22] T. Simon, I. Gambo, R. Ikono and H. Soriyan, *A multi-nodal implementation of apriori algorithm for big data analytics using MapReduce framework*, Int. J. Appl. Inf. Syst. 12(31) (2020).
[23] R. Srinivas, *Managing Large Data Sets Using Support Vector Machines*, University of Nebraska at Lincoln, 2010.
[24] A. Taherkhani, G. Cosma and T. McGinnity, *AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning*, Neurocomput. 3(404) (2020) 66–351.
[25] W. Wolberg, W. Street and O. Mangasarian, *Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates*, Cancer lett. 77(2-3) (1994) 71–163.
[26] Z. Xiong, Y. Cui, Z. Liu, Y. Zhao, M. Hu and J. Hu, *Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation*, Comput. Materials Sci. 1(171) (2020).
[27] *Dataset Description*, Available at: UCI Machine Learning Repository.
[28] *Breast Cancer Wisconsin Dataset*, Available at: UCI Machine Learning Repository.