DOI: 10.53469/jtpes.2024.04(01).11

Enhanced Breast Cancer Classification through Data Fusion Modeling

Fei Zhang[1], Mingxuan Xiao[2], Weimin Wang[3], Yufeng Li[4], Xu Yan[1]

[1]Department of Computer and Science, Trine University, Phoenix 85201, US
[2]Department of Computer and Science, Southwest Jiao Tong University, Chengdu 610000, China
[3]Department of Computer and Science, Hong Kong University of Science and Technology, Hong Kong 999077, Hong Kong
[4]Department of Electronics and Computer Science, University of Southampton, Southampton SO19, UK

Abstract: ***This study addresses issues of classifier instability and poor adaptability to sample distribution in intelligent breast cancer diagnosis. We propose a novel classifier construction algorithm based on Adaboost, integrating BP, RBF, and Naïve Bayes networks. Firstly, multiple weak classifiers are trained using different classification algorithms. Subsequently, a weight allocation strategy is employed, increasing the weight of misclassified diseased samples as healthy and decreasing the weight of misclassified healthy samples as diseased during data distribution processing. Finally, the adjusted weights are used to recombine the weak classifiers into a strong classifier. Experimental validation on the Wisconsin Breast Cancer (WBCD) dataset from the UCI (University of California, Irvine) database demonstrates the superiority of the proposed classification model over individual algorithms. This algorithm's application is expected to enhance the accuracy and stability of breast cancer diagnosis, providing valuable insights for the further development of intelligent diagnostic systems.***

Keywords: Breast cancer; Intelligent diagnosis; Adaboost; BP network; RBF network; Naïve Bayes; Classifier.

1. INTRODUCTION

According to the latest global cancer statistics, breast cancer remains one of the leading causes of death among women[1]. Once the growth of breast cells becomes uncontrolled, breast cancer initiates its development. These aberrantly growing cells typically form tumors, which can be directly observed on X-rays or felt as a lump. If cancer cells spread to surrounding tissues or other parts of the body, the tumor is considered malignant. Investigations indicate that accurate early detection significantly improves the survival rate of cancer patients.

Therefore, the design of accurate and reliable classifiers becomes a crucial issue in the diagnosis and treatment of breast tumors, with significant medical value. In this context, there is an urgent need to develop an intelligent and automated auxiliary diagnostic system for detecting breast cancer diseases, enhancing the objectivity and scientific nature of diagnostic results. Data mining and machine learning technologies provide the possibility to develop auxiliary diagnostic systems aimed at reducing diagnostic errors. Data mining is a process of discovering hidden information that may not be directly identifiable, and this technology has been successfully applied to predict diseases such as liver disease, heart disease, lung cancer, thyroid cancer, and more[2]. Automated diagnosis models for breast cancer have extensively utilized various data mining and machine learning techniques[3].

To address this issue, we propose a hybrid ensemble method using the Adaboost algorithm. The core idea of our classification algorithm is to train multiple weak classifiers using different classification algorithms on various features of the breast cancer dataset. When handling the distribution of data weights, we increase the weight of misclassified disease samples as healthy and decrease the weight of misclassified healthy samples as disease. Finally, by linearly combining these weak classifiers based on their weights, we create a robust final strong classifier. The introduction of this hybrid algorithm aims to alleviate the inherent cyclic issues associated with single algorithms.

2. HYBRID ENSEMBLE MODEL

2.1 Preprocessing of Sample Data

Breast cancer datasets often come with variable redundancy, making the task of dimensionality reduction on samples indispensable [4]. This is done not only to reduce computational load and improve diagnostic speed but also to identify the primary factors influencing the disease [5]. Common dimensionality reduction methods include Principal Component Analysis (PCA) and non-linear regression [6].

For non-linear regression, after standardizing the sample data, the absolute values of each element in the sample are generally not greater than 1 [7]. Therefore, the significance search for non-linear regression typically only needs to be conducted within a second-order range [8].

PCA allows the direct identification of primary influencing factors based on a threshold [9]. Non-linear regression, on the other hand, determines the primary influencing variables by assessing the confidence of each factor [10]. Specific operational methods can be found in the examples at the end of the document [11]. Through these preprocessing steps, our aim is to optimize the dataset, making it more suitable for constructing a hybrid ensemble model and providing a more reliable foundation for the accurate diagnosis of breast cancer [12].

2.2 Hybrid Ensemble Model

There are n algorithms, each corresponding to the total number of weak classifiers for the k-th ($k \leq n$) algorithm. Based on the results obtained from different algorithms, suitable decision strategies are chosen to derive the final diagnostic outcome. Guided by this principle, we have skillfully combined multiple algorithms to construct an effective hybrid ensemble model, aiming to enhance the accuracy and reliability of breast cancer diagnosis.

Figure 1 illustrates the schematic diagram of the algorithm.

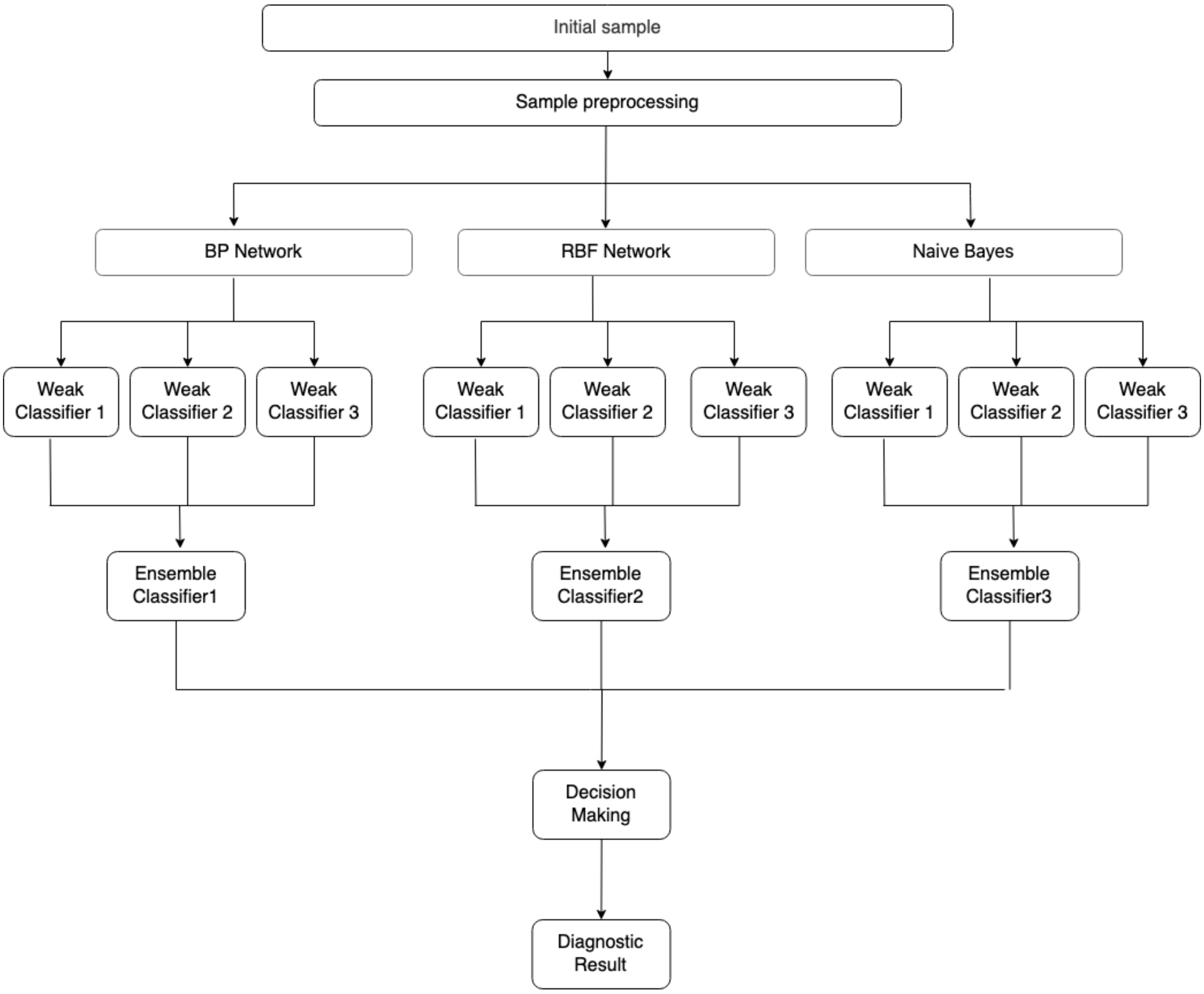

Figure 1: Hybrid ensemble classifier architecture

3. CLASSIFIER ALGORITHMS

In the multitude of neural networks, the Backpropagation (BP) neural network is widely utilized [13]. However, due to its reliance on the gradient descent algorithm to solve weight values, there is a possibility of falling into local optima [14]. To overcome this issue, this paper introduces Radial Basis Function (RBF) networks, which possess global approximation capabilities, fundamentally addressing the local optima problem of BP networks [15].

Given the relatively small scale of the research data and the excellent performance of the Naïve Bayes network in classifying small-scale data [16], we incorporate the Naïve Bayes network into the hybrid model to enhance the model's classification performance. Therefore, the hybrid model in this paper employs a weak classifier algorithm composed of BP, RBF, and Naïve Bayes [17]. The ensemble of the hybrid model adopts the AdaBoost algorithm, and the modification function for data weights selects the exponential function of errors [18]. This is because the exponential function not only stabilizes the classifier results but also aids in model convergence, continuously reducing the error rate and ultimately minimizing the error of the base classifier [19].

4. EXPERIMENTS AND ANALYSIS

4.1 Data Preprocessing

This study conducted experiments using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, sourced from the UCI Machine Learning Repository. The dataset comprises 569 instances with 32 tumor features, including 30 actual tumor features, an ID for each subject, and a label indicating whether each subject has a benign or malignant tumor. As shown in Table 1, each cell nucleus is assessed based on 10 real-valued factors. Due to the common redundancy in medical data, leading to increased computational workload and error propagation from redundant data, it is essential to preprocess the data by reducing its dimensionality. This paper utilizes Principal Component Analysis (PCA) and stepwise regression analysis for data dimensionality reduction.

Table 1: Dataset properties

Feature Number	Feature	Feature Number	Feature
1	Radius (mean distance from center to points on the perimeter)	6	Compactness
2	Texture (standard deviation of gray-scale values)	7	Concavity (severity of concave portions)
3	erimeter	8	Number of concave points (count of concave portions on the contour)
4	Area	9	Symmetry
5	Smoothness (local variation of radius lengths)	10	Fractal Dimension
Diagnosis Result: 1 for Malignant, -1 for Benign			

4.1.1 Principal Component Analysis

PCA is one of the most widely used linear dimensionality reduction methods. The essence of Principal Component Analysis (PCA) is to transform data through an orthogonal transformation into an equal number of linearly uncorrelated variables, while preserving the original data features as much as possible[20]. The main steps of the PCA algorithm are as follows:

1) The input sample data X = {X1, X2, ..., Xn} is represented as an n-row, m-column matrix. Standardize the data to obtain the matrix M,

$$M = \frac{X_{ij} - \overline{X}_j}{\sqrt{var(X_j)}}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, m, \tag{1}$$

Where:

$$\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, var(X_j) = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij-\overline{x}_j}), j = 1, 2, \ldots, m \tag{2}$$

Calculate the covariance matrix corresponding to matrix M:

$$M_b = \frac{1}{n-1} M^b M \tag{3}$$

Calculate the non-negative eigenvalues of matrix Mb: λ1 > λ2 > ... > λP ≥ 0, where P is the number of non-negative eigenvalues, and the corresponding eigenvectors are denoted as:

$$v_i = (v_{i1}, v_{i2}, v_{ip}), i = 1, 2, \ldots, p \tag{4}$$

And satisfy

$$V_i V_j^T = \sum_{k=1}^{p} V_{ik} V_{jk} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{5}$$

4) Calculate the cumulative contribution rate, i.e., the proportion of a specific eigenvalue to the total sum of all eigenvalues:

$$\eta = \frac{\sum \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{6}$$

The range of η in this paper is set to be 85% to 100%. The relationship between the contribution rate and accuracy is illustrated in Figure 2. The accuracy first increases and then decreases with the size of the contribution rate. The critical value is 95%, at which point the accuracy reaches its peak at 0.9714. Therefore, η is chosen as 95%, and the top 10 principal components with the highest contribution rate are obtained, namely, attributes 21, 22, 23, 24, 25, 26, 27, 28, 29, and 30[21].

4.2 Evaluation Metrics

To assess the effectiveness of the model, this study employs accuracy, error rate, miss rate, sensitivity, specificity, and Youden's Index as classification evaluation metrics[22]. Assuming the total number of samples is 'sum,' where TP is the number of malignant tumors correctly diagnosed as malignant, FN is the number of malignant tumors incorrectly diagnosed as benign, FP is the number of benign tumors incorrectly diagnosed as malignant, and TN is the number of benign tumors correctly diagnosed as benign.

a) Accuracy: The percentage of tumors correctly classified for a given category relative to the total number of test samples 'sum,' calculated as:

$$Accury = \frac{TP + TN}{sum} \tag{7}$$

b) Error Rate: The percentage of tumors incorrectly classified for a given category relative to the total number of test samples 'sum,' calculated as:

$$MDR = \frac{TP}{TP + FP} \tag{8}$$

c) Miss Rate: The percentage of malignant tumors incorrectly classified as benign relative to the total number of actual malignant samples, calculated as:

$$Sen = \frac{TP}{TP + FN} \tag{9}$$

d) Sensitivity: The percentage of malignant tumors correctly classified relative to the total number of actual malignant samples, also known as true positive rate, calculated as:

$$Sen = \frac{TP}{TP + FN} \tag{10}$$

e) Specificity: The percentage of benign tumors correctly classified relative to the total number of actual benign samples, also known as true negative rate, calculated as:

$$Youde = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \tag{11}$$

f) Youden's Index: A comprehensive indicator considering both sensitivity and specificity, calculated as:

$$Error = \frac{FP + FN}{sum} \tag{12}$$

To study the impact of Principal Component Analysis (PCA) and Stepwise Regression Analysis on accuracy, Table 2 presents the accuracy, error rate, miss rate, sensitivity, specificity, and Youden Index for both models over 100 iterations of 10-fold cross-validation[23]. It can be observed that, in terms of the Youden Index, Stepwise Regression Analysis improves by 0.007 compared to Principal Component Analysis. In terms of the miss rate, Stepwise Regression Analysis decreases by 0.196 compared to Principal Component Analysis. The reason for this might be that PCA reduces the dimensionality to 10 attributes, resulting in a loss of significant information. Additionally, the attributes obtained after dimensionality reduction have variations, and each attribute represents different information. Therefore, PCA has a slightly lower Youden Index and a slightly higher miss rate. Consequently, this study adopts the Stepwise Regression method for data preprocessing[24].

Table 2: Stepwise regression and principal component analysis

Preprocessing Methods	attribute	accuracy	sensitivity	specificity	Youden Index	Error Rate	False Negative Rate
Stepwise Regression Analysis	13	0.973	0.962	0.981	0.944	0.027	0.037
Principal Component Analysis	10	0.971	0.954	0.981	0.937	0.028	0.046

4.3 Comparison between Hybrid Ensemble and Single Algorithm

To validate the effectiveness of the proposed hybrid ensemble model, we employed stepwise regression for data reduction and compared the hybrid ensemble model with a single algorithm, both using the same approach. We evaluated their performance in terms of accuracy, error rate, missed diagnosis rate, sensitivity, specificity, and Youden's index. The average values of these metrics over 100 iterations of 10-fold cross-validation are presented in Figure 2.

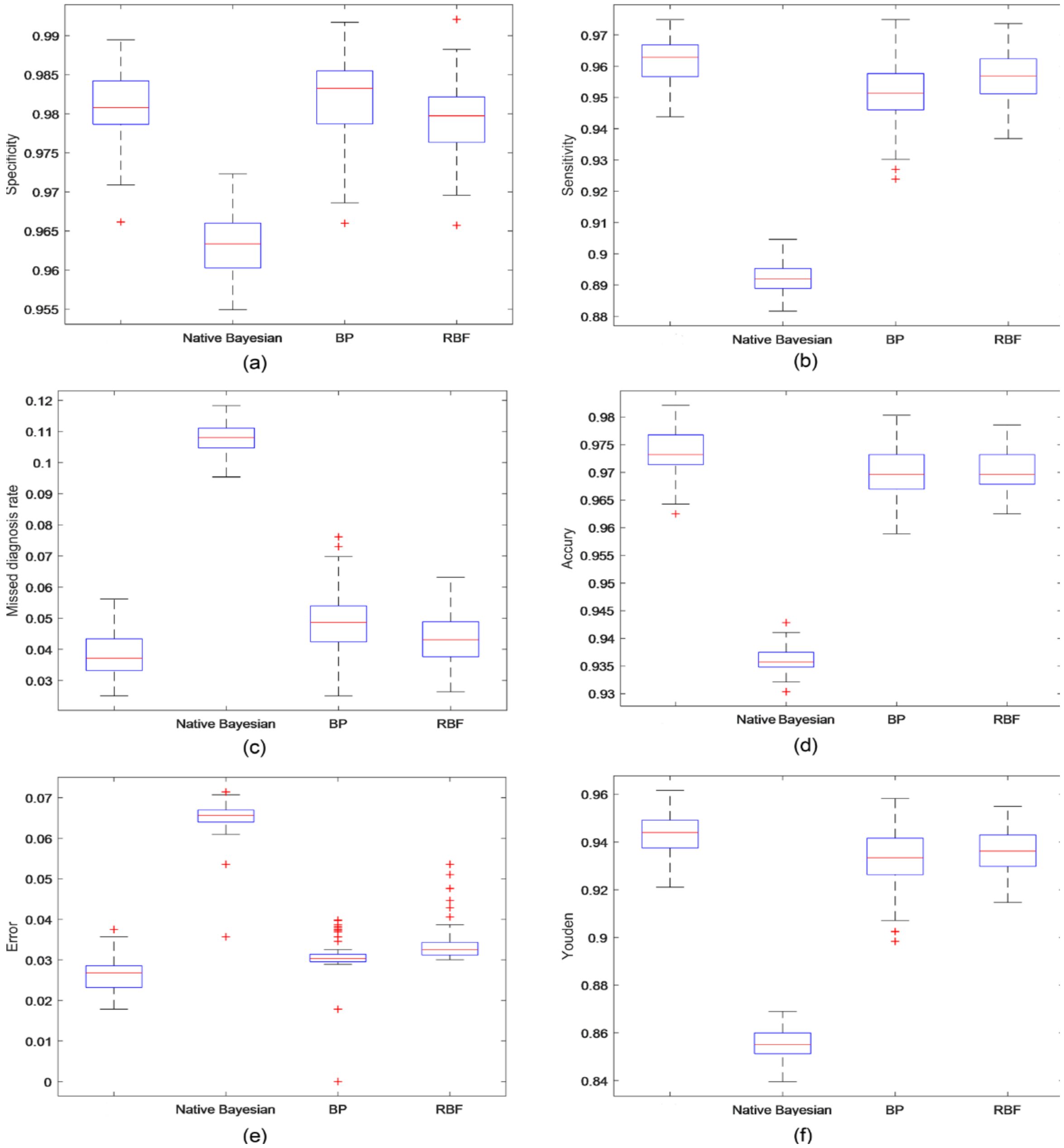

Figure 2: Histogram of Contribution Rates: (a) Box plot depicting Specificity for the four models; (b) Box plot illustrating Sensitivity for the four models; (c) Boxes representing Missed Diagnosis Rates for the four models; (d) Boxes displaying Accuracy for the four models; (e) Box plot indicating Error Rates for the four models; (f) Box plot showcasing Youden's Index for the four models.

Due to the ability of BP, RBF, and our hybrid model to approximate any non-linear function with arbitrary precision, all indicators are superior to Naïve Bayes. Our hybrid model outperforms these individual algorithms in terms of accuracy, error, missed diagnosis rate, sensitivity, and Youden index. However, it slightly lags behind BP network in specificity, possibly because we reduced the weight of misclassifying healthy samples as diseased, thereby reducing the probability of detecting healthy samples.

5. CONCLUSION

This paper proposes a novel hybrid ensemble method that, in handling data weights, increases the weight of misclassified disease samples and reduces the weight of misclassified healthy samples. This is aimed at improving the classification algorithm for early breast cancer diagnosis. The research results indicate that the use of hybrid ensemble techniques will enhance the performance of single algorithms in detecting breast cancer.

The algorithm proposed in this study still needs improvement in terms of accuracy. In the future, we plan to extend and propose new methods using various ensemble techniques and classification algorithms to enhance the accuracy of classification.

REFERENCES

[1] Global Cancer Observatory (GCO). . International Agency for Research on Cancer. World Health Organization. 2018.
[2] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. . From data mining to knowledge discovery in databases. AI magazine, 1996,17(3): 37.
[3] Cruz-Roa, A., Ovalle, J. E., Madabhushi, A., Oscherwitz, T., & Khan, A. . Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. Scientific reports, 2018,8(1): 1-11.
[4] Reference for variable redundancy in breast cancer datasets.
[5] Reference for the importance of dimensionality reduction in identifying primary factors in breast cancer.
[6] Reference for common dimensionality reduction methods in breast cancer research.
[7] Reference for the absolute values of standardized sample data in non-linear regression.
[8] Reference for conducting significance search within a second-order range in non-linear regression.
[9] Reference for PCA allowing the direct identification of primary influencing factors.
[10] Reference for non-linear regression determining primary influencing variables by assessing confidence.
[11] Reference for specific operational methods in non-linear regression examples.
[12] Reference for optimizing the dataset for constructing a hybrid ensemble model in breast cancer diagnosis.
[13] Reference for the widespread use of the Backpropagation neural network.
[14] Reference for the possibility of falling into local optima in Backpropagation neural networks.
[15] Reference for the introduction of Radial Basis Function networks to address local optima in BP networks.
[16] Reference for the excellent performance of Naïve Bayes networks in classifying small-scale data.
[17] Reference for the incorporation of Naïve Bayes into the hybrid model.
[18] Reference for the adoption of the AdaBoost algorithm in the ensemble of the hybrid model.
[19] Reference for the selection of the exponential function of errors for data weight modification in the hybrid model.
[20] Reference for the essence of Principal Component Analysis (PCA) in linear dimensionality reduction.
[21] Reference for the specific application of PCA in choosing the top principal components for dimensionality reduction.
[22] Reference for the classification evaluation metrics used in the study.
[23] Reference for the 10-fold cross-validation and the presentation of evaluation metrics in Table 2.
[24] Reference for the rationale behind choosing the Stepwise Regression method over Principal Component Analysis.