Classification of breast cancer using Wrapper and Naïve Bayes algorithms

I M D Maysanjaya\textsuperscript{1,2}, I M A Pradnyana\textsuperscript{1} and I M Putrama\textsuperscript{1}

\textsuperscript{1} Department of Informatics Education, Faculty of Engineering and Vocational, Universitas Pendidikan Ganesha, Bali, Indonesia – 81116
\textsuperscript{2} Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia – 55281

E-mail: dendi.ms@undiksha.ac.id

Abstract. Breast cancer is defined as a malignant neoplasm disease originating from the parenchyma. This disease is most commonly afflicted by women and is the second killer after lung cancer. By the end of 2012, the WHO shows the prevalence of breast cancer worldwide reaching 6.3 million spread across 140 countries. Symptoms of the disease are diagnosed using mammogram. The results will be examined by paramedics, to find out if the cancer is malignant or not. However, the accuracy of the diagnoses obtained is directly proportional to the level of paramedical expertise. The general objective of this research aims to classify breast cancer types based on the extracted breast cancer features, and the specific purpose is to offer a hybrid machine learning method. To help increase the accuracy of the diagnosis, through this study, a CAD-based scheme was developed by applying a combination of Wrapper and Naïve Bayes algorithms. The Wrapper algorithm is used at the feature selection stage, while the Naive Bayes algorithm is used at the classification stage. There are 683 data taken from the UCI Knowledge Repository consisting of two class, namely 444 benign cancers and 239 malignant cancers, with 9 types of attributes. The data is divided into two groups, with the various amount for each class in each group. The first group was used as a training group for the feature selection stage, and the second group as the testing group for the classification stage. The classification results show the accuracy up to 99.27%, sensitivity and specificity on each up to 99.30%. Based on these results, the proposed scheme is expected to contribute in the development of CAD for the diagnosis of breast cancer.

1. Introduction
Breast cancer is defined as a malignant neoplasm disease originating from the parenchyma tissue. This disease commonly affects women and becomes the second killer after lung cancer. By the end of 2012, the World Health Organization (WHO) shows that the number of cases of breast cancer in the world reaches 6.3 million scattered in 140 countries. By 2015, the American Cancer Society\textsuperscript{1} estimates about 231,840 new cases occur, and from that number is estimated around 40,290 women die. Breast cancer is not only suffered by women, but men too, although the number is not as many as the number cases in women. By 2015, cases of breast cancer that occurred in men is estimated to be 2,350 cases, and 440 of them died\textsuperscript{1}. There are two types of breast cancer, i.e. benign and malignant. The illustration can be seen in Figure 1, and the brief explanation of the differences can be seen in Table 1.
Figure 1. Types of Breast Cancer

Table 1. Differences of Benign and Malignant Cancer

| Feature      | Benign     | Malignant               |
|--------------|------------|-------------------------|
| Size         | < 2 cm     | > 2 cm                  |
| Consistency  | Soft       | Hard, firm, or rubbery  |
| Mobility     | Fixed      | Mobile                  |
| Rate of Growth| Slow      | Rapidly                |
| Boundaries   | Circumscribed | Irregular           |
| Metastasis   | Absent     | Frequently present      |

In the examination of breast cancer, paramedics commonly use mammography. Mammography is a process of breast examination in humans by using low-dose X-rays. With proper use, mammography can reduce mortality caused by breast cancer. Certainly, the case requires knowledge and paramedical skills in handling it. A false diagnosis on mammography examination can be affected by some factors particularly the expertise level of paramedics, the acquisition method, and the quality of mammography equipment used. Hence, several studies have been conducted to develop computer-aided breast cancer diagnosis based on digital image processing to reduce the error possibility.

Akay [2] applied a combination of Support Vector Machine (SVM) and feature selection to diagnose breast cancer. However, the data used requires further exploration, so the results can be more convincing and interesting. Chen [3] employed a hybrid intelligent model that uses the cluster analysis technique with feature selection for analyzing clinical breast cancer diagnosis. Eltoukhy et al. [4] presented a method for breast cancer diagnosis in digital mammogram images using multiresolution representations (wavelet and curvelet), and validated using 5-folds cross-validation. Based on the result, the experiment using wavelet gain the accuracy of 96.56%, and achieved the accuracy of 97.30% by curvelet-based.

Furthermore, Zheng et al. [5] introduced combination of K-Means clustering and SVM algorithms. According to the result, the proposed method reduces the computation time significantly without losing diagnosis accuracy. However, by using K-Means algorithm, the training sample size has not been decreased by filtering the similar samples in the data, which could be a potential way of reducing the dimension of training set further.

Based on the previous research, we assess the accuracy of breast cancer classification results can still be improved. Thus, the main contribution of this study aims to develop hybrid machine learning methods to classify the types of breast cancer by using a combination of Wrapper and Naïve Bayes algorithms. In this study, we proposed hybrid method because of its advantages. Wrapper is one type of feature selection method that uses a controlled algorithm as part of the evaluation function, while
Naïve Bayes has several advantages, such as fast training process, unaffected by irrelevant features, and is also capable of handling real and discrete data. By combining their respective advantages, it can improve the accuracy of the classification process.

To complete the identification study of breast cancer, this paper proposes a scheme to classify breast cancer types. The classification is categorized into two types, i.e. benign and malignant. The structure of this paper is organized as follows Section 2 explained the methods and procedures. The results and discussion are presented in Section 3 followed by conclusion and future work in Section 4.

2. Methods and Procedures

2.1. Approach
The methodology consists of four main processes, namely data cleansing, modelling the training-test data partition, feature selection, and classification as depicted in Figure 2.

![Figure 2. Scheme diagram of the approach](image)

After the dataset taken from the open source repository, firstly, we applied data cleansing process, especially removed the data with missing values. The second step is randomly divide data into two types, training and testing data, based on three types of partition (50%-50%, 70%-30%, 80%-20% of training-testing). The purpose of this partition modelling is to see its effect on the feature selection and classification process, and to compare which partition pattern is capable of producing the highest accuracy. Furthermore, Wrapper algorithm was conducted on feature selection process, which aims to gain the relevant feature for the classification process, and the last step is the classification process which is provided by Naïve Bayes algorithm.

2.2. Dataset
The dataset in this study was taken from the University of California – Irvine (UCI) Machine Learning Repository, created by Dr. Wolberg [6], from University of Wisconsin Hospitals, Madison, Wisconsin, USA. As a brief information, UCI Machine Learning is an open source repository, which like a bank of dataset, database, or domain theory for many case studies that are used by the machine learning community for the empirical analysis of machine learning algorithm. It is commonly used by anyone as a primary source of machine learning dataset [7].

The number of breast cancer dataset is 683 instances, consisting of 9 attributes, i.e. 1) Clump of Thickness, 2) Uniformity of Cell Size, 3) Uniformity of Cell Shape, 4) Marginal Adhesion, 5) Single Epithelial Cell Size, 6) Bare Nuclei, 7) Bland Chromatin, 8) Normal Nucleoli, and 9) Mitoses. This samples arrive periodically as Dr. Wolberg reports his clinical cases. Originally, the number of dataset is 699 instances, but since there are 16 instances have missing values, in this study only 683 instances are used.

The dataset is divided into two classes namely 444 instances of benign class, and 239 instances of malignant class. Furthermore, the dataset is split into two groups namely the training group and the
test group, where each group consists of a benign and malignant class. There are three types of training-test partition, i.e. 50-50%, 70-30%, and 80-20%.

Training groups are used to get feature patterns in the feature selection process. The result of this process is used as a reference for determining the number of features in the test group, and would be tested in the classification process. To measure the effectiveness level of the feature selection algorithm, it can be seen from the level of accuracy, sensitivity, and specificity, resulting from the classification process. The higher level of accuracy, sensitivity, and specificity gained, indicates that the feature selection algorithm used is effective, and vice versa.

2.3. Framework
The framework used in this work is the Waikato Environment for Knowledge Analysis (WEKA). This framework was developed at University of Waikato [8]. WEKA is open source (free license or freeware), designed, and developed based on JAVA, and special purpose is used for pattern recognition methods. WEKA supports various data mining functions, i.e. data pre-processing, classification, clustering, association, regression, feature selection, and visualization. Data points are explained by a number of fixed attributes, such as nominal, numeric, normally, and other attribute types.

2.4. Algorithms

2.4.1. Wrapper Subset Evaluation. Wrapper Subset Evaluation, commonly called Wrapper, is one type of feature selection method that uses a supervised learning algorithm as part of its evaluation function. This supervised learning algorithm is used as a “black box” system during the feature selection process [9,10]. The evaluation function of each feature will provide an approximate quality to the model created by the learning algorithm, so this would result in better accuracy forecasts. Besides using learning algorithm, this method also uses k-folds cross-validation evaluation. Although the Wrapper method can produce better accuracy, this method has a weakness, which causes the computation process is relatively longer and slower [9,10]. In general, the algorithm of Wrapper described as follows:

| Wrapper Algorithm |
|-------------------|
| **I.S.:** | \( D(F_0,F_1,...,F_{n-1}) \) //a training dataset with \( N \) features |
| **S_0:** | //a subset from which to start the search |
| **\( \delta \):** | //a stopping criterion |
| **F.S.:** | \( S_{\text{best}} \) //an optimal subset |
| **begin** |
| **initialize:** | \( S_{\text{best}} = S_0 \); |
| **\( \gamma_{\text{best}} = \text{eval}(S_0,D,A) \):** | //evaluate \( S_0 \) by a mining algorithm \( A \) |
| **do begin** |
| **\( S = \text{generate}(D) \):** | //generate a subset for evaluation |
| **\( \gamma = \text{eval}(S,D,A) \):** | //evaluate the current subset \( S \) by \( A \) |
| **if:** | \( \gamma \) is better than \( \gamma_{\text{best}} \) |
| **\( \gamma_{\text{best}} = \gamma \)** |
| **\( S_{\text{best}} = S \);** |
| **end until:** | \( \delta \) is reached; |
| **return:** | \( S_{\text{best}} \); |
| **end** |

2.4.2. Naïve Bayes. Naïve Bayes is the algorithm developed from Bayesian theorem, which has a number of advantages, i.e. relatively fast training process, able to handle real and discrete data, uninfluenced by irrelevant features, as a strong assumption towards the independence of each features [11–13]. In simply, the formula for Naïve Bayes can be seen through the Equation (1).
\[
Posterior = \frac{Prior \times Likelihood}{Evidence}
\]  

Posterior is a probability of the presence of the sample with a specific characteristic in a class. It is obtained by multiplying between the probability of the class appearance \((Prior)\) and the probability of the emergence of sample characteristic in class \((Likelihood)\), and then divided by the probability of the sample characteristic emergence in general \((Evidence)\). Application of the Naïve Bayes algorithm was done as follows:

| Naïve Bayes Algorithm |
|------------------------|
| **Learning Phase**: |
| 01 for each class value \(C_i\) |
| 02 compute \(P(C_i)\) |
| 03 for each attribute \(X_j\) |
| 04 //compute distribution \(D_{ij}\) |
| 05 if attribute \(X_j\) is discrete |
| 06 \(D_{ij} = \) categorical distribution for \(C_i\) and \(X_j\) |
| 07 else |
| 08 \(D_{ij} = \) normal distribution for \(C_i\) and \(X_j\) |
| **Testing Phase**: |
| 09 given unknown \(X' = [x_1', x_2', \ldots, x_n']\) |
| 10 estimate class = \(\arg\max_{C_i} P(C_i) \cdot \prod D_{ij}\) |

### 2.5. Performance Evaluation Parameters

#### 2.5.1. Accuracy, Sensitivity, and Specificity

To evaluate the test, the statistic parameter such as accuracy, sensitivity, and specificity were used through Equation (2), Equation (3), and Equation (4) on each. Accuracy is a measurement to see the level of achievement of the classification being done. Sensitivity refers to the ability of classifier in predicting the positive class as the True Positive (TP), while specificity refers to the ability classifier in recognizing the negative class as the True Negative (TN).

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% 
\]

\[
Sensitivity = \frac{TP}{TP + FN} \times 100\% 
\]

\[
Specificity = \frac{TN}{TN + FP} \times 100\% 
\]

#### 2.5.2. K-Folds Cross-Validation

To gain better classification result, \(k\)-folds cross-validation was utilized. This method has several advantages, such as minimizing bias produced during the training process by using random sampling [14]. This method initiated by dividing all data randomly into \(k\) group, known as fold, with each fold is assumed to have the same number of members. The classification algorithm has executed the training and testing \(k\) times. In each process, one-fold is entered as data test while the other folds serve as training data. Each process will give different results. These values are then averaged and the average value is used as the accuracy value of the corresponding algorithm.

### 3. Result and Discussion

In the feature selection process, Multi-layer Perceptron (MLP) is utilized as part of the evaluation function of the Wrapper algorithm. MLP is chosen because it has several advantages, such as the ability to support parallel implementation, generalization capacity, fault tolerance, and able to handle
the missing value [15]. Thus, it is suitable to deal with breast cancer dataset characteristics. To evaluate the effectiveness of our proposed method, we conducted several experiments. The feature selection and classification result can be seen in Table 2 and Table 3 respectively, while Table 4 shows the confusion matrix for each classification experiment.

In general, based on the result of Table 2, Table 3, number of accuracy and features increase with the increase of the training set size, while as we can see from Table 4, number of false negative (FN) decrease with the increase of the training set size. For the comparison purposes, Table 5 gives the classification accuracies of our proposed method and previous methods.

**Table 2. Feature Selection Result**

| Features          | Training-Test Partitions (%) |
|-------------------|------------------------------|
|                   | 50-50 | 70-30 | 80-20 |
| Types             |       |       |       |
| Clump of Thickness| Uniformity of Cell Size       | Clump of Thickness |
| Uniformity of Cell Size | Uniformity of Cell Shape   | Uniformity of Cell Size |
| Marginal Adhesion | Single Epithelial Cell Size  | Bland Chromatin |
| Bare Nuclei       | Bare Nuclei                    | Normal Nucleoli |
| Numbers           | 4     | 5     | 6     |

**Table 3. Classification Result**

| Parameter            | Training-Test Partition (%) |
|----------------------|------------------------------|
|                      | 50-50 | 70-30 | 80-20 |
| Folds                | 7     | 7     | 7     |
| Benign               | 215   | 130   | 89    |
| Malignant            | 117   | 72    | 47    |
| Correctly Classified | 332   | 202   | 136   |
| Incorrectly Classified| 9    | 3     | 1     |
| Accuracy (%)         | 97.36 | 98.54 | 99.27 |
| Sensitivity (%)      | 97.40 | 98.50 | 99.30 |
| Specificity (%)      | 97.40 | 98.50 | 99.30 |

**Table 4. Confusion Matrixes of Classification Result**

| Actual    | Predict | Training-Test Partitions |
|-----------|---------|---------------------------|
|           | Benign  | Malignant                 |
| Benign    | 215     | 7                         |
| Malignant | 2       | 117                       |
| Benign    | 130     | 3                         |
| Malignant | 0       | 72                        |
| Benign    | 89      | 0                         |
| Malignant | 1       | 47                        |
|           |         | 50-50%                    |
|           |         | 70-30%                    |
|           |         | 80-20%                    |
Table 5. Comparison the Proposed Method with Previous Methods

| Author (Year)                     | Method               | Accuracy (%) |
|----------------------------------|----------------------|--------------|
| Quinlan (1996) [16]              | C4.5                 | 94.74        |
| Hamilton et al. (1996) [17]      | RIAC                 | 95.00        |
| Ster and Dobnikar (1996) [18]    | LDA                  | 96.80        |
| Pena-Reyes and Sipper (1999) [19]| Fuzzy-GA             | 97.36        |
| Goodman et al. (2002) [20]       | AIRS                 | 97.20        |
| Abonyi and Szeifert (2003) [21]  | Supervised Fuzzy Clustering | 95.57 |
| Akay (2008) [2]                  | F-score + SVM        | 97.95        |
| Zheng et al. (2014) [5]          | K-Means + SVM        | 97.28        |
| This Study (2017)                | Wrapper + Naïve Bayes| 99.27        |

As in Table 3, we can see that the model using the 80%-20% part of the test partition yields better results than the model using other test partitions because according to the result, this partition model provides more useful features during the classification process (based on the result in Table 2). The number of features generated from the model affects the case of classification error as shown in Table 4, where only one instance is classified incorrectly. Obviously, this is because the data used for the training process more than the other partition model. It can therefore improve the ability of the system to identify the characteristics of breast cancer, as well as implicate the testing process, which gives more ability to distinguish the types of breast cancer.

As we can see from the result especially from Table 5, our proposed method using Wrapper and Naïve Bayes algorithms obtains the highest classification accuracy so far. Therefore, we conclude that Naïve Bayes with feature selection using Wrapper obtains promising results in classifying the potential breast cancer patients.

4. Conclusion and Future Work
This study proposes a scheme to classify breast cancer into two types, namely benign and malignant. A total of 9 features are extracted to facilitate the classification process. Feature selection based on Wrapper method is conducted to gain the relevant features which may contribute to improve the rate of classification result.

Three types of training-test partition are conducted, i.e. 50-50%, 70-30%, and 80-20%, which gain four, five, and six of selected features on each. By using Naïve Bayes for the classification process, obtained that the proposed method affords the highest classification accuracies on 80-20% training-test partition, it’s up to 99.27%. Considering that result, we believe that the proposed method can be very helpful to the paramedics for their final decisions on their patients. By using such a tool, they can make very precise decisions.

In the next investigation, it will be better to do comparison with the other algorithms, so it is expected to increase the value of accuracy and can increase the number of datasets used.

References
[1] Society A C 2015 Breast Cancer Facts & Figures 2015-2016 (Atlanta)
[2] Akay M F 2009 Support vector machines combined with feature selection for breast cancer diagnosis Expert Syst. Appl. 36 3240–7
[3] Chen C-H 2014 A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection Appl. Soft Comput. 20 4–14
[4] Meselhy Eltoukhy M, Faye I and Belhaouari Samir B 2012 A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation Comput. Biol. Med. 42 123–8
[5] Zheng B, Yoon S W and Lam S S 2014 Breast cancer diagnosis based on feature extraction
using a hybrid of K-means and support vector machine algorithms *Expert Syst. Appl.* **41** 1476–82
[6] Wolberg D W H 1992 Breast Cancer Wisconsin (Original) Data Set *UCI Mach. Learn. Repos.*
[7] Center for Machine Learning and Intelligent System U-I 2015 Machine Learning Data Repository [UCI] *Databib*
[8] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P and Witten I H 2009 The WEKA Data Mining Software: An Update *SIGKDD Explor.* **11**
[9] Hall M A and Holmes G 2003 Benchmarking attribute selection techniques for discrete class data mining *IEEE Trans. Knowl. Data Eng.* **15** 1437–47
[10] Naqvi G 2012 *A Hybrid Filter-Wrapper Approach for Feature Selection* (Sweden: Örebro University)
[11] Bishop C M 2006 *Pattern Recognition and Machine Learning* (New York: Springer)
[12] Duda R O, Hart P E and Stork D G 2001 *Pattern Classification and Scene Analysis* (New York: John Wiley & Sons, Inc.)
[13] Natalius S 2010 Metoda Naive Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen
[14] Polat K, Şahan S and Güneş S 2007 A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis *Expert Syst. Appl.* **32** 1141–7
[15] Isa I S, Fauzi N A, Sharif J M, Baharudin R and Abbas M H 2012 Comparisons of MLP transfer functions for different classification classes 2012 IEEE International Conference on Control System, Computing and Engineering (ICCSCE) pp 110–4
[16] Quinlan J R 1996 Improved use of continuous attributes in C4.5 *J. Artif. Intell. Res.* 77–90
[17] Hamilton H J, Shan N and Cercone N 1996 *RIAC: A rule induction algorithm based on approximate classification*
[18] Ster B and Dobnikar A 1996 Neural networks in medical diagnosis: Comparison with other methods *Proceedings of the International Conference on Engineering Applications of Neural Networks* pp 427–30
[19] Pena-Reyes C A and Sipper M 1999 A fuzzy-genetic approach to breast cancer diagnosis *J. Artif. Intell. Med.* 131–55
[20] Goodman D E, Boggess L and Watkins A 2002 Artificial Immune System Classification of Multiple-class Problems *Proceedings of the Artificial Neural Networks in Engineering* pp 179–83
[21] Abonyi J and Szeifert F 2003 Supervised fuzzy clustering for the identification of fuzzy classifiers *J. Pattern Recognit. Lett.* **14** 2195–207