Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Development of Statistically Modelled Feature Selection Method for Microwave Breast Cancer Detection

V.Vijayasarveswari[1,2,*], Norfadila Mahrom[1,3,4], Rafikha Aliana A. Raof[1,3,4], Phak Len Al Eh Kan[1,4], Muhammad Amiruddin Ab Razak[1], Bavanraj Punniya Silan[1], Ahmad Ashraf Abdul Halim[1], Mohd Wafi Nasrudin[1,4], Nuraminah Ramli[1], Yusnita Rahayu[5]

[1] Faculty of Electronic Engineering Technology, Universiti Malaysia Perlis, Kangar 01000, Perlis, Malaysia
[2] Advanced Communication Engineering (ACE), Centre of Excellence, Universiti Malaysia Perlis, Kangar 01000, Perlis, Malaysia
[3] Sports Engineering Research Centre (SERC), Centre of Excellence, Universiti Malaysia Perlis, Kangar 01000, Perlis, Malaysia
[4] Advanced Computing, Centre of Excellence, Universiti Malaysia Perlis, Kangar 01000, Perlis, Malaysia
[5] Department of Electrical Engineering, Universitas Riau, Pekan Baru, Indonesia

ARTICLE INFO

Article history:
Received 26 September 2023
Received in revised form 26 January 2024
Accepted 15 June 2024
Available online 10 August 2024

Keywords:

Feature extraction method; Feature selection method; breast cancer detection

ABSTRACT

Microwave technology is very promising tool for breast cancer detection. Microwave transmits and receives UWB signals. UWB signals carries information of the breast cancer. UWB signals need to be pre-processed in order to remove irrelevant and redundant features. Feature extraction and feature selection methods are mostly used to remove the unwanted features. In this paper, a statistically modelled feature selection (SMFS) method is proposed for microwave breast cancer detection. Initially, performance of different feature extraction and feature selection method are analysed using Anova test (p-value) and machine learning (SVM, DT, PNN, NB) accuracy. The best feature extraction and feature selection methods are combined and tested. Based on the performance of feature extraction and feature selection method, Combined Neighbour Component Analysis (feature selection) and Statistical features (feature extraction) are combined and tested. This method is able to achieve up to 85%. The result proves two stage methods are able to improve the accuracy compared to single stage method. Therefore, SMFS is able to detect breast cancer efficiently.

1. Introduction

Microwave based technology is a potential technology which can replace the invasive and expensive screening traditional technology (mammography, MRI and ultrasound). Furthermore, this technology is safe, robust, ionizing radiation free and causes lesser physical harm to users [1-3]. There are three types of breast imaging methods in microwave imaging which are passive, hybrid and active [4]. The passive method classifies the detected tumour by measuring the differences of temperature between healthy and unhealthy breast and hybrid method uses more energy to identify the tumour

[*] *Corresponding author.*
E-mail address: vijaya@unimap.edu.my

https://doi.org/10.37934/araset.50.1.250263

and images the breast. Microwave signals are transmitted and received rapidly for breast imaging in active method.

Microwave based technology uses two approaches which are microwave tomography and radar-based imaging. In both approaches, use the received UWB signals to classify breast cancer according to the dielectric properties [3,5]. Basically, the received UWB signals is obtained either in time domain or frequency domain but frequency domain UWB signal is better because it has better signal-to-noise ratio compared to time domain UWB signal [5,6]. These UWB signals need to be pre-processed before fed into the machine learning. The signals' characteristics must be revealed by mathematically and statistically. Feature reduction, data normalization, feature extraction and feature selection are the methods used to preprocess the features effectively [7,8]. Initially the data sample has a huge number of features and some of them redundant and irrelevant. Thus, it is important to identify which features contribute in the greatest extent to the quality for better results and discard some of the irrelevant features to avoid overfitting and high complexity [9]. Feature extraction is a part of optimization and ensures the effectiveness of the machine learning. The feature extraction's aim is to extract robust features without loss of the important information in the signals [10]. Feature selection method is a procedure to select a subset of features from the original set of features and plays a significant role in the optimization of the machine learning. According to the researchers, three main selection methods are widely used which are filter, wrapper and embedded [11,12]. Wrapper method considers a selection of features based on the learning model score. Different combination of feature sets is prepared and then, are evaluated and compared among them using the learning model. The feature set is selected based on the learning model accuracy. Due to high computation time needed by the wrapper method, filter method is more convenient. Filter method selects a set of independent features. Each feature is statistically measured and ranked up based on the measurement score. Features are ranked based on a suitable ranking criterion and a threshold is used to remove the features below the threshold. Embedded method learns which the feature is more suitable while the learning model is created [12]. This method selects features faster and has lower complexity.

Majdi *et al.,* [13] uses sequential backward selection (SBS) method to select features. Three classifiers (Support Vector Machine (SVM), Random Forest (RF) and Decision Tree (DT) are trained using five selected features. Local binary pattern texture features are pre-processed by using hybrid binary BAT algorithm and optimum path forest. The final fused features are used to classify breast cancer using SVM [14]. Hybrid algorithm using meta learning and Artificial Neural Network (ANN) is proposed and tested by using 16 features (selected using correlation method) for breast cancer predicating [15]. Different filter, wrapper and embedded methods are used and tested using classification and regression tree (CART). The improvised CART is proposed by using wrapper recursive Feature Elimination (obtained the highest accuracy) [11]. Rakibul *et al.,* [16] employs wrapper feature selection method by using WEKA tool. Logistic regression (LR), linear SVM and quadratic SVM are used to test the selected features to classify the breast cancer. LR outperform with the selected features compared to other machine learnings. Pre-trained Convulation Neural Network (CNN) and the univariate based method (pearson correlation coefficient, cosine coefficient, Euclidien distance and mutual information) are used to select features [17]. The numbers of features are reduced to increase the accuracy by using generic algorithm (GA) and particle swarm optimization (PSO). Extreme learning machine (ELM) is used with different poly to predict breast cancer with the feature selection method [18].

Two stages of feature selection are proposed where first stage consists of filter-based feature selection method (chi- squared, F-statistic and mutual information) and second stage consists of wrapper- based sequential forward selection. The selected features are fed into SVM, DT, RF and k-nearest neighbour (KNN) separately [12]. Extreme Gradient Boosting (XGBoost) is used to pick weighted features and test them with XGBoost, RF, SVM and LR [19]. Similarly, two stages are proposed, first stage by using filter methods (Chi-squared, Fisher Score, ReliefF and Gini Index) and second stage by using boost diversity methods (least absolute shrinkage and selection operator (LASSO), regression with recursive feature elimination (LR-RFE), mutual information and correlation-based feature selection (CFS). The selected features after these two stages feature selection method, are fed into machine learning [20]. Multi-stage feature selection (MSFS) is proposed by using principal component analysis (PCA), data normalization, feature extraction (statistical features) and feature selection (ranking method using statistical method). Out of three classifiers (Naïve bayes (NB), probabilistic neural network (PNN) and SVM, NB outperforms by using 8 hybridfeatures [8]. Similarly, MSFS- backalgorithm PSO (MSFS-BPSO) is proposed. MSFS-BPSO consists of feature normalization, singular vector decomposition (SVD), feature extraction (statistical features) and feature selection (Anova test and BPSO). This feature selection method selected 30 features [7]. 31 features are selected by using hybrid feature selection (PSO and adaptive local search method [21].

Researches show hybrid or multiple stage of feature selection method is able to increase the accuracy and reduce the misclassification. However, the most of research can be seen to use only feature selection method. Therefore, in this paper, statistically modelled feature selection method (SMFS) is proposed by combining feature extraction method and feature selection method. SMFS is consists of first stage, feature extraction method and second stage, feature selection method. The selection of feature extraction and selection methods is done by using Anova test and accuracy of machine learning.

2. Methodology

In this section, the data collection, preprocessing and statistically modelled feature selection methods are explained. Figure 1 shows the overall experiment workflow conducted in this research. In the first stage, data is collected using breast phantom and the data is converted from analogue to digital values. Data dimension reduction is done by using Principal Component Analysis (PCA) to reduce the dimension of the data in the second stage. In the third stage, the data undergoes statically modelled feature selection method to select optimum features. Finally, the breast cancer is detected by using the selected features.

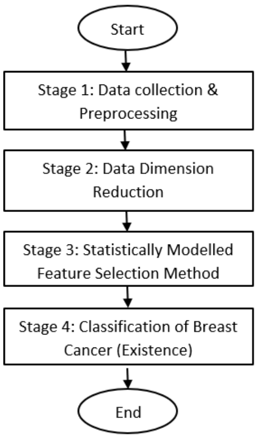

Start
Stage 1: Data collection & Preprocessing
Stage 2: Data Dimension Reduction
Stage 3: Statistically Modelled Feature Selection Method
Stage 4: Classification of Breast Cancer (Existence)
End

Fig. 1. Overall experiment workflow

2.1 Stage 1: Data Collection

Data samples are collected by using breast phantom, a pair of antennae [22,23] and UWB transceivers as shown in Figure 2 [8]. The steps are taken to collect the data samples are taken from the previous studies [7,8,35]. Breast phantom is developed by using a mixture of low-cost materials (petroleum jelly, wheat flour and soy oil with ratio of 100:50:37). Tumour is developed by using the mixture of 10:5.5 of wheat flour and water with 2mm size. Total of 2000 data samples are collected which consists of 1000 data samples with the presence of tumour and 1000 data samples with the absence of tumour. These 2000 analogue signals are converted into digital value by using Discrete Cosine Transform (DCT). Each analogue signals have 1632 digital values.

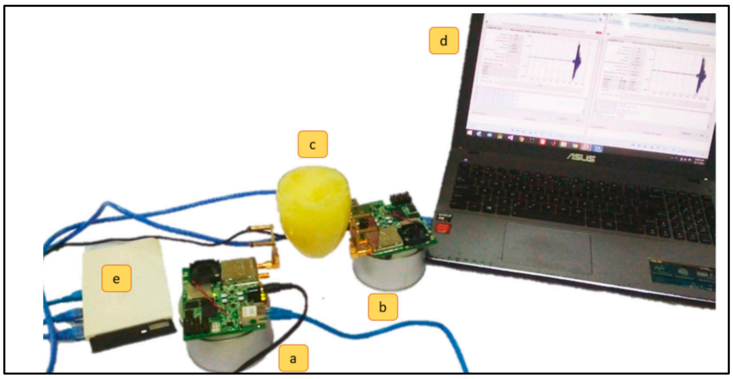


Fig. 2. Experimental setup. (a) transmitter; (b) receiver; (c) breast phantom; (d) computer; (e) router [8]

2.2 Stage 2: Data Dimension Reduction

Principal component analysis (PCA) is a useful statistical technical method to find the pattern in high dimensional data. PCA generates a new set of features called principal component and simplifies the complexity of the high dimensional data by reducing the dimension of the data [24,25]. PCA steps are summarized as below:

Step 1: Assume a sample data matrix of i number of samples which results to k number of characterization method and can be represented by matrix X using Eq. (1).

$$X = \begin{matrix} X_{11} & .. & X_{1k} \\ \vdots & .. & \vdots \\ X_{n1} & .. & X_{nk} \end{matrix} \tag{1}$$

Step 2: Matrix S is developed by subtracting the mean of the data matrix, X from each data point. This process is known as mean- centering the data. The matrix S is as below:

$$S = \begin{matrix} X_{11} - \overline{X_1} & .. & X_{1k} - \overline{X_k} \\ \vdots & .. & \vdots \\ X_{n1} - \overline{X_k} & .. & X_{nk} - \overline{X_k} \end{matrix} \tag{2}$$

where $(\overline{X_k})$ is the mean of the data matrix, X.

Step 3: The covariance value is calculated for the data matrix, S using Eq. (2). The dataset's, covariance matrix, C is constructed as shown in Eq. (3).

$$conv\ (C) = \frac{\sum_{i=1}^{n} s_i - \bar{S}}{n-1} \tag{3}$$

where S is the data matrix.

$$C = \begin{matrix} C_{11} & .. & C_{1k} \\ \vdots & .. & \vdots \\ C_{n1} & .. & C_{nk} \end{matrix} \tag{4}$$

where $C_{ij} = 1/2\ \{(x_i - \bar{X}_i)((x_j - \bar{X}_j)\ (i,j = 1,2,\ldots.,k)$

Step 4: Compute the matrix A of eigenvectors.

$$D = ACA^T \tag{5}$$

where A is a matrix, whose columns are the eigenvectors of D and D is the matrix of eigenvalues of C.

$$C = \begin{matrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & \cdots & \lambda_n \end{matrix} \tag{6}$$

where λ_1 to $\lambda_n$are the eigenvalues of D.

Step 5: Rearrange the eigenvector matrix, V and eigenvalues matrix, D in order of decreasing eigenvalue. The pairing between two matrixs should be correct.

Step 6: Generate PCA component using Eq. (7).

$$\text{FinalData} = V^T * X^T \tag{7}$$

where V^T is the eigenvectors matrix and X^T is the data matrix. Eigenvectors matrix is the output matrix after compute the data matrix, X from the Step 2 until Step 5 while data matrix is the matrix of raw data sample.

2.3 Stage 3: Statistically Modelled Feature Selection Method

In this section statistically modelled feature selection (SMFS) method is explained as shown in Figure 3. Different types of feature extraction and feature selection methods are analysed separately. The best performance of feature extraction method and feature selection method (based on Anova test and machine learning accuracy) are combined together in the third stage for the performance verification.

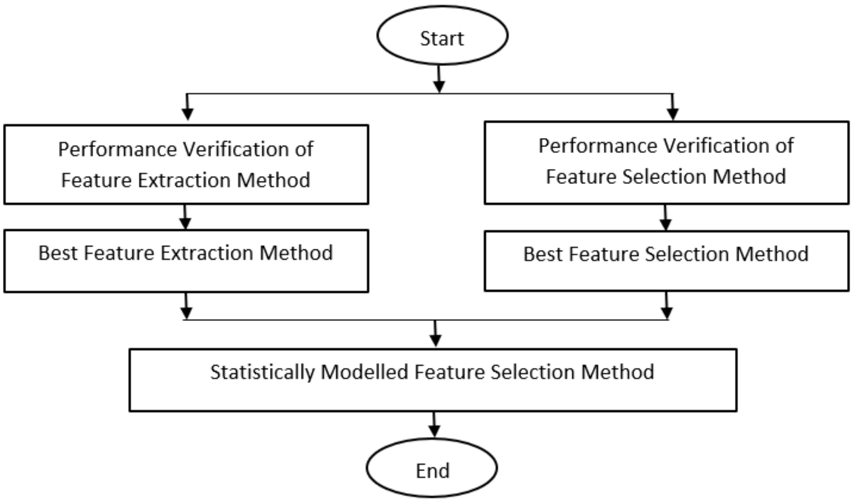

Start
Performance Verification of Feature Extraction Method
Performance Verification of Feature Selection Method
Best Feature Extraction Method
Best Feature Selection Method
Statistically Modelled Feature Selection Method
End

Fig. 3. Workflow of development of SMFS Method

2.3.1 Feature extraction method

Feature extraction method is to reduce the number of features by extracting important features and creating new set of features. Three feature extraction method used are k-means clustering, reconstruction independent component analysis and statistical features.

i. K-means clustering: K-means clustering is an unsupervised learning which helps in finding inherent structure in the data. Here, k-means clustering uses only the input without knowing the labels/target of the particular input [26,27]. It aggregated all similar data points together with the centre points. The centres are used to choose the feature. Let n -class problem with k features. C_1, C_2, … C_n are the centres of the n cluster. Each cluster centre in k dimension is C_{i1} = [C_{i1}, C_{i2}, … C_{in}]. Cluster centres can be calculated using Eq. (8).

$$dist_i = C_{1i} - C_{2i} \tag{8}$$

where $dist$ is the vector contains element for all dimension and i is the attributes. The attributes for each cluster are identified using Eq. (9).

$$i_m = \max\ (dist - dist_j\ (m-1) \text{ for } m = 1, 2, 3 .. j \tag{9}$$

where j is the number of attributes. Once $dist_i$ is calculated, the features are extracted. Most relevant features on top of attribution while least relevant features on the bottom of the attribution.

ii. Reconstruction independent component analysis: Independent component analysis (ICA) has optimization problem, high cost, sensitive to whitening and unable to learn overcomplete features. Therefore, reconstruction independent component analysis (RICA) is introduced to overcome the shortcomings of ICA. RICA uses soft reconstruction penalty to replace the orthogonality condition ($WWT = I$) that in ICA. ICA and RICA are represented as in Eq. (10) and Eq. (11) respectively [28].

$$ICA\text{: } \underset{W}{minimize} \sum_{i=1}^{m} \sum_{j=1}^{k} g(W_j x^{(i)}) \text{ subject to } WWT = I \tag{10}$$

$$RICA\text{: } \underset{W}{minimize} \frac{\lambda}{m} \sum_{i=1}^{m} \left\| W^T W x^{(i)} - x^{(i)} \right\|_2^2 + \sum_{i=1}^{m} \sum_{j=1}^{k} g(W_j x^{(i)}) \tag{11}$$

where g is a nonlinear convex function.

iii. Statistical Features: For statistical features, only two features are considered which are mean and variance. Mean is the ratio of sum of all features value to total number of features as expressed in Eq. (12). Variance is the distance between the value to the mean as expressed in Eq. (13) [8].

$$Mean, \mu_N = \frac{v_1 + v_2 + v_3 + \cdots v_N}{N} \tag{12}$$

$$Variance, \sigma^2 = \sum \frac{(v - \mu_N)^2}{N} \tag{13}$$

where v is the feature value and N is the total number of features.

2.3.2 Feature selection method

Feature selection method is essential in order to reduce the number of features in the data. Feature selection selects a small subset of significant features from the original set of features. Three different feature selection methods are used (Relief-F, Neighbourhood component analysis (NCA) and Particle Swarm Optimization (PSO)).

i. Relief-F: Relief-F [29,30] is robust towards noisy and incomplete data. It basically calculates feature score and selects the top feature score. Feature scoring is done by estimation of feature value differences between nearest neighbour instance pairs. Relief finds for two nearest neighbours (from same class and different class). The same class nearest neighbour is called as nearest hit whereas the different class nearest neighbour is called nearest miss. Relief updates the quality estimation based on the value for selected instances, nearest hit and nearest miss for all attributes as in Eq. (14).

$$W[A] = W[A] - \sum_{j=1}^{k} \frac{diff\ (A, R_i, H_j)}{m * k} + \frac{diff\ (A, R_i, M_j)}{m * k} \tag{14}$$

where $W[A]$ is quality estimation, $R_i$is instance, A is feature, H_j is nearest hit, M_j is nearest miss and k is nearest neighbour. If instances and nearest hit have different values of the attributes, then, the attribute separates two instances within the same class and reduce the quality estimation. If instance and nearest miss have different values of the attributes, then the attribute separates two instances in differen class and increases the quality estimation. The function in Eq. (15) calculates the distance between the values of the feature A for two observations (O_1, O_2) and calculates the distance between the two observations to find the nearest neighbors.

$$diff\ (A, O_1, O_2) = \frac{I_1\ (A) - I_2\ (A)}{\max(A) - \min(A)} \tag{15}$$

ii. Neighbourhood component analysis (NCA): Neighbour component analysis (NCA) [31-33] is a statistical method based on k-Nearest neighbour algorithm. It selects neighbour randomly and searches a vote for each class. NCA learns feature weight by maximizing the leave-one-out (LOO) accuracy and optimized regulation parameter. Let T, set of training samples $\{(x_{1,}\, y_1), \ldots..(x_{i,}\, y_i)\}$, where $x_{i,}$ is feature vector of dimension and y is its corresponding class. The distance between two samples (x_i and x_j) is as in Eq. (16).

$$d_w\,(x_i, x_j) = \sum_{r=1}^{d} w_r^2\,|x_{ir} - x_{jr}| \tag{16}$$

where w_r is rth feature's weight. Based on LOO, the reference point is determined by a probability of x_i selects x_j as defined in Eq. (17).

$$p_{ij} = \begin{cases} \frac{k\,(D_W\,(x_i\,x_j))}{\sum_{k\neq 1} k\,(D_W\,(x_i\,x_j))} & if\ i \neq j \\ 0 & if\ i = j \end{cases} \tag{17}$$

where k is kernel function. Therefore, the x_i's probability is classified correctly in the particular class using Eg. (18) and (19) respectively.

$$p_i = \sum_j y_{ij}\, p_{ij} \tag{18}$$

$$H_i = \{j | y_j = y_i\} \tag{19}$$

where $y_{ij} = \begin{cases} 1, & if\ y_i = y_j \\ 0, & otherwise \end{cases}$. Final optimization can be defined as $f(A) = \sum_i p_i$ .Gradient rule is used to optimized the matrix A.

$$\frac{\partial f}{\partial W} = 2W\ \sum_i p_i \sum_k p_{ik}\, x_{ik}\, x_{ik}^T - \sum_{j\in H_i} p_{ik}\, x_{ik} x_{ik}^T \tag{20}$$

iii. Particle Swarm Optimization (PSO) [21,34]: PSO is a technique based on population or swarm social such as birds in a flock. PSO searches objective function in the landscape adjusting the trajectories in a quasi-stochastic manner. Each particles adjusts velocity and position based on own knowledge and adjoining particles as in Eq. (21) and Eq. (22).

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{21}$$

$$v_{id}^{t+1} = w\ x\ v_{id}^{t} + c_1\ x\ r_{1i}\ x\ (p_{id} - x_{id}^{t}) + c_2\ x\ r_{2i}\ x\ (p_{gd} - x_{id}^{t}) \tag{22}$$

where i is the position of the particle, x is the features, d is the space's dimensionality, v is the particle's velocity, t is the t^th iteration, w is the weight, c_1 and c_2 are the acceleration contants, r_{1i} and r_{2i} are dispersed homogeneously's values and p_{id} and p_{gd} are the values of personal best and global best in the particular dimension.

2.3.3 Statistically modelled feature selection method (SMFS)

K-means clustering, RICA and statistical features method are used to extract features from the data samples. Similarly, Relief-f, NCA and PSO method are used to select features from the data samples. These six data sets undergo performance verification by using statistical analysis (Anova test) and machine learning's accuracy in breast cancer classification. Based on the performance, the best feature extraction method and the best feature selection method are chosen for the development of SMFS method.

2.3.4 Performance verification

For performance verification, statistical analysis (Anova test) and machine learning's accuracy are used to verify the best feature extraction method. Anova test examines the significant difference between the mean of more than two groups which is similar to t-test. The six datasets are prepared by dividing the data samples into two groups (the presence of tumour and the absent of tumour). P-value is identified for each dataset. P-value is calculated using Eq. (23) and then, refers to p-value table for its value. P-value should be less than 0.05 which means the null hypothesis (all group has same mean) is rejected and statistically significant.

$$z = \frac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \tag{23}$$

where $\bar{x}$ is the sample mean, μ_o is the hypothesized mean, σ is sample standard deviation and n is the sample size. For feature selection method, performance of each method is calculated based on the accuracy it provided after the method is applied on the data sample.

2.4 Stage 4: Classification of Breast Cancer

Classification of breast cancer is done by detecting the presence and absence of the tumour by using machine learning Machine learning is used to find the accuracy in classifying breast cancer for all six datasets. Three different machine learning are used to identify the performance of the feature extraction method and feature selection method. For feature extraction method, Decision tree (DT), Naïve bayes (NB) and Probabilistic Neural Network (PNN) are used whereas for feature selection method NB, DT and SVM are used. Together with the machine leaning, k-fold cross-validation is used for training and testing purpose. Tenfold cross-validation is used. The total of 2000 data samples (1000 samples with the presence of tumour and 1000 samples with the absence of tumour) are divided into ten sets. Therefore, each set contains 200 data samples. Confusion matrix is generated. From the confusion matrix as shown in Figure 4, the accuracy of each set is identified by using Eq. (24). Average accuracy is calculated by adding accuracy of ten folds and divided by ten.

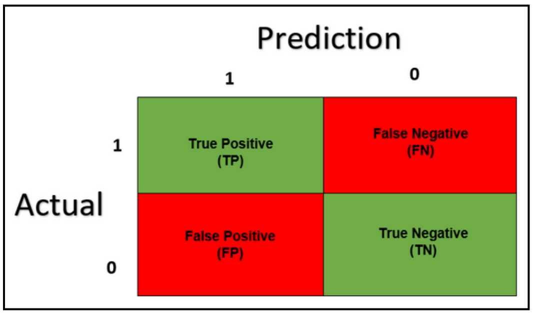


Fig. 4. Figure quality confusion matrix for two classes

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \tag{24}$$

where TN is true negative (predict absence of tumor correctly), TP is true positive (predict presence of tumor correctly), FN is false negative (predict absence of tumor when the tumor is present) and FP is false positive (predict presence of tumor when the tumor is absent).

3. Results and Discussion

Performance of each feature extraction and feature selection methods are as shown in Table 1 and Table 2 respectively. Based on the obtained performance in Table 1 and Figure 5, for feature extraction method, statistical features method outperforms compared to RICA and k-means. Statistical features method obtains p-value of 0.008 where closest to 0 while RICA obtains p-value of 1. If the p-value is less than 0.05, there is difference between the groups and the data is highly significant. Machine learning's accuracy supports this statement by obtaining high accuracy for the statistical features method. The highest average machine learning's accuracy obtained by statistical feature method (75.05%) followed by k-means (73.61%) and RICA (51.78%). Therefore, statistical features method is chosen to develop SMFS framework.

Table 1
Performance verification for feature extraction method

Feature Extraction Method		K-means	RICA	Statistical features
p-value		0.025	1.000	0.008
Machine learning's Accuracy (%)	DT	78.98	51.40	86.28
	NB	78.88	47.80	85.99
	PNN	62.97	56.13	52.87
	Average Accuracy	73.61	51.78	75.05

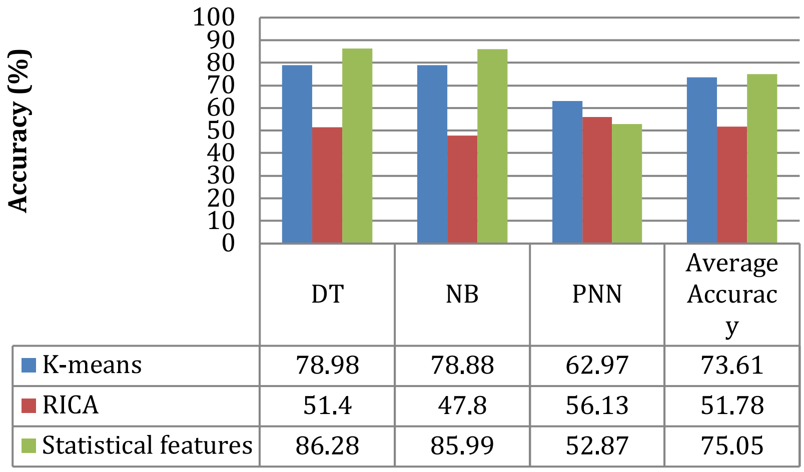
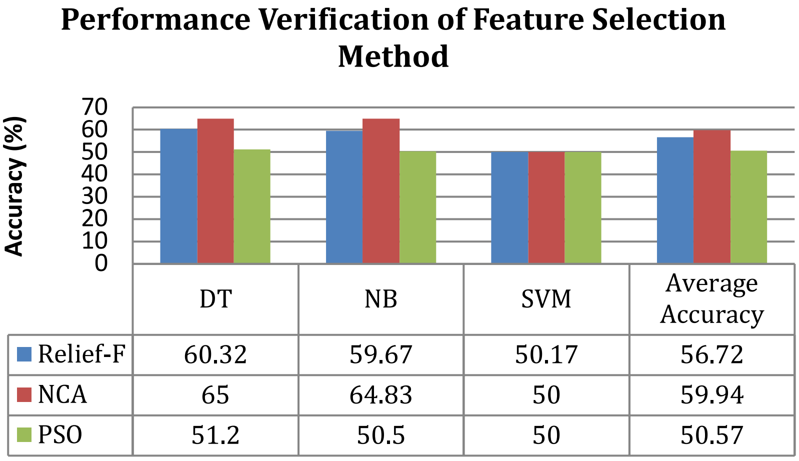


Fig. 5. Performance verification of feature extraction method

For feature selection method as shown in Table 2 and Figure 6, the highest accuracy obtained by using its own accuracy measurement is by PSO (96.75%) followed by Relief-F (94.60%) and NCA (93.80%). Whereas the highest machine learning's accuracy is obtained by NCA (59.94%) followed by Relief-F (56.72%) and PSO (50.57%). Due to two feature selection methods perform better in two different performance verification methods, NCA and PSO are selected to develop statistically modelled feature selection (SMFS) framework.

Table 2
Performance verification for feature selection method

Feature Selection Method		Relief-F	NCA	PSO
Accuracy Obtained by using its own accuracy measurement (%)		94.60	93.80	96.75
Machine learning's Accuracy (%)	DT	60.32	65.00	51.20
	NB	59.67	64.83	50.50
	SVM	50.17	50.00	50.00
	Average Accuracy	56.72	59.94	50.57



	DT	NB	SVM	Average Accuracy
🟦 Relief-F	60.32	59.67	50.17	56.72
🟥 NCA	65	64.83	50	59.94
🟩 PSO	51.2	50.5	50	50.57

Fig. 6. Performance verification of feature selection method

Based on the performance obtain in Table 3, statistical features + NCA method (85.03%) is better compare to statistical features + PSO (76.50%). Therefore, the combination of statistical features (feature extraction method) and NCA method (feature selection method) are used for SMFS method. This result proves two stages of feature selection method is able to improve the accuracy compare to single stage of feature selection. The complete breast cancer detection framework with proposed SMFS after the experiment and analysis on feature extraction and selection method for breast cancer classification is as shown in Figure 7.

Table 3
Performance verification for Statistically Modelled Feature Selection (SMFS)

		Statistical Features + NCA	Statistical Features + PSO
Machine learning's Accuracy (%)	DT	85.64	78.74
	NB	84.41	74.25
	Average Accuracy	85.03	76.50

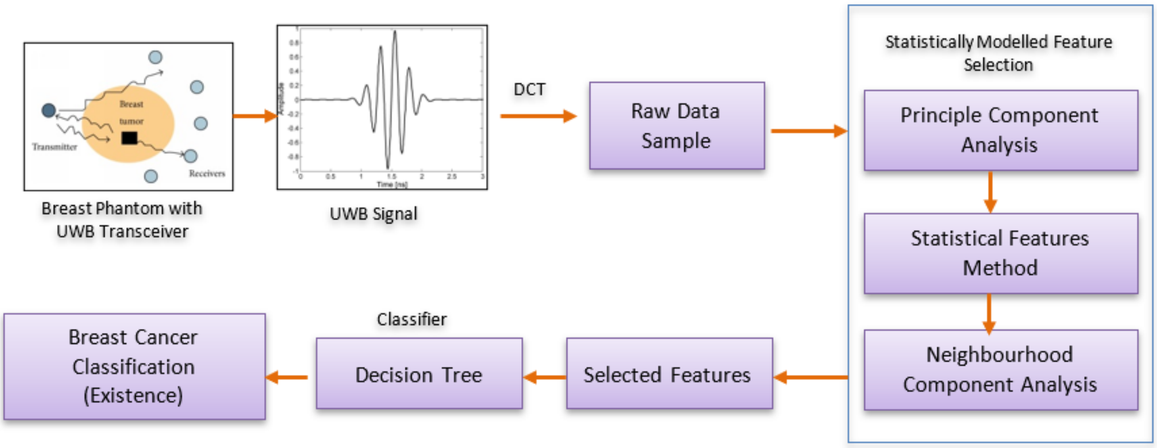

Breast tumor
Transmitter
Receivers
Breast Phantom with UWB Transceiver
Amplitude
Time [ns]
UWB Signal
DCT
Raw Data Sample
Statistically Modelled Feature Selection
Principle Component Analysis
Statistical Features Method
Neighbourhood Component Analysis
Selected Features
Classifier
Decision Tree
Breast Cancer Classification (Existence)

Fig. 7. The complete breast cancer detection framework by using proposed Statistically Modelled Feature Selection (SMFS)

4. Conclusions

SMFS framework is proposed to classify the existence of the breast cancer. Different types of feature extraction and feature selection methods are used to test its performance in classifying the breast cancer. The performance is measured based on Anova test and machine learning accuracy. NCA and statistical feature outperform compare to other methods. DT performs better compared to the other machine learnings. The final proposed SMFS is consisted of statistical features and NCA. The proposed SMFS is tested by using DT and obtained 86.64%. A complete framework is proposed and can be used for breast cancer classification by another researcher.

Acknowledgement

This research was funded by a grant from Ministry of Higher Education of Malaysia (FRGS/1/2021/TK0/UNIMAP/02/6).

References

[1] Kwon, Sollip, and Seungjun Lee. "Recent advances in microwave imaging for breast cancer detection." *International journal of biomedical imaging* 2016 (2016). https://doi.org/10.1155/2016/5054912

[2] Hang, Jo Ann, Lynn Sim, and Zulkarnay Zakaria. "Non-invasive breast cancer assessment using magnetic induction spectroscopy technique." *International Journal of Integrated Engineering* 9, no. 2 (2017).

[3] Abdul Halim, Ahmad Ashraf, Allan Melvin Andrew, Mohd Najib Mohd Yasin, Mohd Amiruddin Abd Rahman, Muzammil Jusoh, Vijayasarveswari Veeraperumal, Hasliza A. Rahim, Usman Illahi, Muhammad Khalis Abdul Karim, and Edgar Scavino. "Existing and emerging breast cancer detection technologies and its challenges: a review." *Applied Sciences* 11, no. 22 (2021): 10753. https://doi.org/10.3390/app112210753

[4] Aldhaeebi, Maged A., Khawla Alzoubi, Thamer S. Almoneef, Saeed M. Bamatraf, Hussein Attia, and Omar M. Ramahi. "Review of microwaves techniques for breast cancer detection." *Sensors* 20, no. 8 (2020): 2390. https://doi.org/10.3390/s20082390

[5] Martellosio, Andrea, Marco Pasian, Maurizio Bozzi, Luca Perregrini, Andrea Mazzanti, Francesco Svelto, P. E. Summers, G. Renne, and M. Bellomi. "0.5–50 GHz dielectric characterisation of breast cancer tissues." *Electronics Letters* 51, no. 13 (2015): 974-975. https://doi.org/10.1049/el.2015.1199

[6] Rahman, Ashiqur, Mohammad Tariqul Islam, Mandeep Jit Singh, Salehin Kibria, and Md Akhtaruzzaman. "Electromagnetic performances analysis of an ultra-wideband and flexible material antenna in microwave breast imaging: To implement a wearable medical bra." *Scientific reports* 6, no. 1 (2016): 38906. https://doi.org/10.1038/srep38906

[7] Halim, Ahmad Ashraf Abdul, Allan Melvin Andrew, Wan Azani Mustafa, Mohd Najib Mohd Yasin, Muzammil Jusoh, Vijayasarveswari Veeraperumal, Mohd Amiruddin Abd Rahman, Norshuhani Zamin, Mervin Retnadhas Mary, and

Sabira Khatun. "Optimized Intelligent Classifier for Early Breast Cancer Detection Using Ultra-Wide Band Transceiver." *Diagnostics* 12, no. 11 (2022): 2870. https://doi.org/10.3390/diagnostics12112870

[8] Vijayasarveswari, V., A. M. Andrew, M. Jusoh, R. B. Ahmad, T. Sabapathy, R. A. A. Raof, M. N. M. Yasin, S. Khatun, and H. A. Rahim. "Correction: Multi-stage feature selection (MSFS) algorithm for UWB-based early breast cancer size prediction." *Plos one* 16, no. 5 (2021): e0251679. https://doi.org/10.1371/journal.pone.0251679

[9] Wen, Tingxi, and Zhongnan Zhang. "Effective and extensible feature extraction method using genetic algorithm-based frequency-domain feature search for epileptic EEG multiclassification." *Medicine* 96, no. 19 (2017): e6879. https://doi.org/10.1097/MD.0000000000006879

[10] Yan, Jia, Xiuzhen Guo, Shukai Duan, Pengfei Jia, Lidan Wang, Chao Peng, and Songlin Zhang. "Electronic nose feature extraction methods: A review." *Sensors* 15, no. 11 (2015): 27804-27831. https://doi.org/10.3390/s151127804

[11] Agaal, Asma, and Mansour Essgaer. "Influence of feature selection methods on breast cancer early prediction phase using classification and regression tree." In *2022 International Conference on Engineering & MIS (ICEMIS)*, pp. 1-6. IEEE, 2022. https://doi.org/10.1109/ICEMIS56295.2022.9914078

[12] Elemam, Tarneem, and Mohamed Elshrkawey. "A highly discriminative hybrid feature selection algorithm for cancer diagnosis." *The Scientific World Journal* 2022 (2022). https://doi.org/10.1155/2022/1056490

[13] Alnowami, Majdi R., Fouad A. Abolaban, and Eslam Taha. "A wrapper-based feature selection approach to investigate potential biomarkers for early detection of breast cancer." *Journal of Radiation Research and Applied Sciences* 15, no. 1 (2022): 104-110. https://doi.org/10.1016/j.jrras.2022.01.003

[14] Sasikala, S., M. Ezhilarasi, and S. Arunkumar. "Feature selection algorithm based on binary BAT algorithm and optimum path forest classifier for breast cancer detection using both echographic and elastographic mode ultrasound images." *Journal of Cancer Research and Therapeutics* 19, no. 2 (2023): 191-197. https://doi.org/10.4103/jcrt.JCRT_324_19

[15] Han, Luyao, and Zhixiang Yin. "A hybrid breast cancer classification algorithm based on meta-learning and artificial neural networks." *Frontiers in Oncology* 12 (2022): 1042964. https://doi.org/10.3389/fonc.2022.1042964

[16] Hasan, Rakibul, and A. S. M. Shafi. "Feature selection based breast cancer prediction." *Int J Image Graph Signal Process* 15, no. 2 (2023): 13-23. https://doi.org/10.5815/ijigsp.2023.02.02

[17] Samee, Nagwan Abdel, Ghada Atteia, Souham Meshoul, Mugahed A. Al-antari, and Yasser M. Kadah. "Deep learning cascaded feature selection framework for breast cancer classification: Hybrid CNN with univariate-based approach." *Mathematics* 10, no. 19 (2022): 3631. https://doi.org/10.3390/math10193631

[18] da Silva, Amanda Lays Rodrigues, Maíra Araújo de Santana, Clarisse Lins de Lima, José Filipe Silva de Andrade, Thifany Ketuli Silva de Souza, Maria Beatriz Jacinto de Almeida, Washington Wagner Azevedo da Silva, Rita de Cássia Fernandes de Lima, and Wellington Pinheiro dos Santos. "Features selection study for breast cancer diagnosis using thermographic images, genetic algorithms, and particle swarm optimization." *International Journal of Artificial Intelligence and Machine Learning (IJAIML)* 11, no. 2 (2021): 1-18. https://doi.org/10.4018/IJAIML.20210701.oa1

[19] Haarika, Raye, Tina Babu, Rekha R. Nair, and T. M. Rajesh. "Breast cancer prediction using feature selection and classification with xgboost." In *2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC)*, pp. 1-6. IEEE, 2023. https://doi.org/10.1109/ICRTEC56977.2023.10111901

[20] Brancato, Valentina, Nadia Brancati, Giusy Esposito, Massimo La Rosa, Carlo Cavaliere, Ciro Allarà, Valeria Romeo *et al.,* "A two-step feature selection radiomic approach to predict molecular outcomes in breast cancer." *Sensors* 23, no. 3 (2023): 1552. https://doi.org/10.3390/s23031552

[21] Alzaqebah, Malek, Sana Jawarneh, Rami Mustafa A. Mohammad, Mutasem K. Alsmadi, Ibrahim Al-Marashdeh, Eman AE Ahmed, Nashat Alrefai, and Fahad A. Alghamdi. "Hybrid feature selection method based on particle swarm optimization and adaptive local search method." *International Journal of Electrical and Computer Engineering* 11, no. 3 (2021): 2414. https://doi.org/10.11591/ijece.v11i3.pp2414-2422

[22] Reza, Khondker Jahid, Sabira Khatun, Mohd Jamlos, Md Moslemuddin Fakir, and M. N. Morshed. "Performance enhancement of ultra-wideband breast cancer imaging system: proficient feature extraction and biomedical antenna approach." *Journal of Medical Imaging and Health Informatics* 5, no. 6 (2015): 1246-1250. https://doi.org/10.1166/jmihi.2015.1522

[23] Reza, Khondker Jahid, Sabira Khatun, Mohd F. Jamlos, Md Moslemuddin Fakir, and S. S. Mostafa. "Performance evaluation of diversified SVM kernel functions for breast tumor early prognosis." *ARPN Journal of Engineering and Applied Sciences* 9, no. 3 (2014): 329-335.

[24] Lever, Jake, Martin Krzywinski, and Naomi Altman. "Points of significance: Principal component analysis." *Nature methods* 14, no. 7 (2017): 641-643. https://doi.org/10.1038/nmeth.4346

[25] Parida, K., S. K. Mandal, S. S. Das, and A. R. Tripathy. "Feature extraction using K-means clustering: an approach & implementation." *International Journal of Computer Information Systems* 2, no. 2 (2011).

[26] Mishra, Sidharth Prasad, Uttam Sarkar, Subhash Taraphder, Sanjay Datta, D. Swain, Reshma Saikhom, Sasmita Panda, and Menalsh Laishram. "Multivariate statistical data analysis-principal component analysis

(PCA)." *International Journal of Livestock Research* 7, no. 5 (2017): 60-78. https://doi.org/10.5455/ijlr.20170415115235

[27] Piernik, Maciej, and Tadeusz Morzy. "A study on using data clustering for feature extraction to improve the quality of classification." *Knowledge and Information Systems* 63, no. 7 (2021): 1771-1805. https://doi.org/10.1007/s10115-021-01572-6

[28] Beltrán Segarra, Marc. "Study of reconstruction ICA for feature extraction in images and signals." (2017).

[29] Robnik-Šikonja, Marko, and Igor Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF." *Machine learning* 53 (2003): 23-69. https://doi.org/10.1023/A:1025667309714

[30] Stief, Anna, James R. Ottewill, and Jerzy Baranowski. "Relief F-based feature ranking and feature selection for monitoring induction motors." In *2018 23rd International Conference on Methods & Models in Automation & Robotics (MMAR)*, pp. 171-176. IEEE, 2018. https://doi.org/10.1109/MMAR.2018.8486097

[31] Djerioui, Mohamed, Youcef Brik, Mohamed Ladjal, and Bilal Attallah. "Neighborhood component analysis and support vector machines for heart disease prediction." *Ingénierie des Systèmes d Inf.* 24, no. 6 (2019): 591-595. https://doi.org/10.18280/isi.240605

[32] Yang, Wei, Kuanquan Wang, and Wangmeng Zuo. "Neighborhood component feature selection for high-dimensional data." *J. Comput.* 7, no. 1 (2012): 161-168. https://doi.org/10.4304/jcp.7.1.161-168

[33] Laghmati, Sara, Bouchaib Cherradi, Amal Tmiri, Othmane Daanouni, and Soufiane Hamida. "Classification of patients with breast cancer using neighbourhood component analysis and supervised machine learning techniques." In *2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet)*, pp. 1-6. IEEE, 2020. https://doi.org/10.1109/CommNet49926.2020.9199633

[34] Xie, Hailun, Li Zhang, Chee Peng Lim, Yonghong Yu, and Han Liu. "Feature selection using enhanced particle swarm optimisation for classification models." *Sensors* 21, no. 5 (2021): 1816. https://doi.org/10.3390/s21051816

[35] Halim, Ahmad Ashraf Abdul, Vijayasarveswari Veeraperumal, Allan Melvin Andrew, Mohd Najib Mohd Yasin, Mohd Zamri Zahir Ahmad, Kabir Hossain, Bifta Sama Bari, and Fatinnabila Kamal. "UWB-based early breast cancer existence prediction using artificial intelligence for large data set." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 29, no. 2 (2023): 81-90. https://doi.org/10.37934/araset.29.2.8190