IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Machine Learning Classification Techniques for Breast Cancer Diagnosis

To cite this article: David A. Omondiagbe *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **495** 012033

View the article online for updates and enhancements.

You may also like

- Simulated lesion, human observer performance comparison between thin-section dedicated breast CT images versus computed thick-section simulated projection images of the breast
L Chen, J M Boone, C K Abbey et al.
- Breast Cancer Detection Using Deep learning and IoT Technologies.
Siddhant Salvi and Ameya Kadam
- Use of risk projection models to estimate mortality and incidence from radiation-induced breast cancer in screening programs
M Ramos, S Ferrer, J I Villaescusa et al.

ECS Toyota Young Investigator Fellowship

For young professionals and scholars pursuing research in batteries, fuel cells and hydrogen, and future sustainable technologies.

At least one $50,000 fellowship is available annually. More than $1.4 million awarded since 2015!

Application deadline: January 31, 2023

Learn more. Apply today!

This content was downloaded from IP address 103.4.221.252 on 22/01/2023 at 07:39

Machine Learning Classification Techniques for Breast Cancer Diagnosis

David A. Omondiagbe [1], Shanmugam Veeramani [1*], Amandeep S. Sidhu [2]

1- Curtin University, Malaysia, CDT 250, Miri 98009, Sarawak, Malaysia
2- Curtin University, Kent St, Bentley WA 6102, Australia

*Corresponding author: s.veeramani@curtin.edu.my

Abstract— Breast cancer is one of the most widely spread disease and the second leading cause of cancer death among women. Breast cancer starts when malignant lumps which are cancerous begin to grow from the breast cells. Doctors may wrongly diagnose benign tumor (which is non-cancerous) as malignant tumor. There is need for a computer aided detection (CAD) systems which uses machine learning approach to provide accurate diagnosis of breast cancer. These CAD systems can aid in detecting breast cancer at an early stage. When, breast cancer is detected early enough, the survival rate increases because better treatment can be provided. This paper aims at investigating Support Vector Machine (using radial basis kernel), Artificial Neural Networks and Naïve Bayes using the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset. The focus of this paper is to integrate these machine learning techniques with feature selection/feature extraction methods and compare their performances to identify the most suitable approach. The goal is combining the advantages of dimensionality reduction and machine learning. This paper proposed a hybrid approach for breast cancer diagnosis by reducing the high dimensionality of features using linear discriminant analysis (LDA), and then applying the new reduced feature dataset to Support Vector Machine. The proposed approach obtained an accuracy of 98.82%, sensitivity of 98.41%, specificity of 99.07% and area under the receiver operating characteristic curve of 0.9994.

Keywords-Machine learning; Breast cancer; Support vector machine; Artificial neural network; Naïve Bayes.

1. Introduction

There is a growing interest in machine learning (ML) this last decade. This growing interest is accelerated by cheaper computing power and low-cost memory. Thus, large amount of data can be stored, processed and analyzed efficiently. Machine learning plays a vital role in a wide range of critical applications, such as data mining, natural language processing, image recognition, expert systems and prediction [1]. This paper focuses on breast cancer diagnosis. Breast cancer occurs majorly in women aged 40 years and above and it occurs when the cells in the glands that produce milk (called lobules) are abnormal and divide drastically. In Malaysia, breast cancer has been found to be the most common form of cancer affecting women [2].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Studies have shown [3] that breast cancer was the second among the most diagnosed cancers. Breast cancer being the most common cancer in women is known to affect about 10% of all women at some stages of their life [4]. According to studies [4], there is a rise in the incidence rate recently and data have shown that survival rate is 88% and 80 % after five and ten years respectively from diagnosis. Due to the severity of the disease, there is need for a computer aided detection (CAD) system using machine learning (ML) approach for breast cancer diagnosis

This paper proposes an automated method with a principled workflow for diagnosing breast cancer. The data used in this research work is the Wisconsin Diagnostic Breast Cancer Dataset (WDBC). The contribution of this research is to show that machine learning approaches which include Support Vector Machine (SVM), Artificial Neural Networks (ANN) and Naïve Bayes (NB) can solve pattern classification problems effectively. In addition, this provides the basis for performing a comparative analysis amongst these techniques. Another contribution was in feature selection/feature extraction techniques. It is important to note that feature selection/extraction techniques touch all disciplines that require knowledge discovery from big data. Finally, this study is important in selecting a suitable machine learning algorithm when building an integrated intelligent model.

The rest of the paper is organized as follows: Section 2 explains the background of the three machine learning classification techniques being investigated and the fundamental concepts of the dimensionality reduction methods used in this work. Section 3 presents literature review on related works. Section 4 presents the methodology and experimental setup. Section 5 shows the experimental results and discussion. Section 6 concludes and summarizes this research work. These sections are described in detail in the following paragraphs.

2. Background of machine learning & feature selection

Machine learning (ML) is a type of artificial intelligence which focuses on the development of computer programs which could change when exposed to a new data. It uses computer models and information obtained from past and previous data to aid classification, prediction and detection processes. This paper was designed to perform a review on some of the widely used classification algorithms and their application in breast cancer diagnosis.

Feature dimensions can be reduced using the appropriate feature selection or feature extraction method. There are several methods used to reduce the dimensions of features in a dataset. Feature selection techniques involve selecting a subset of features from the original set of features [5]. Feature extraction on the other hand aims at generating new features by merging the original features. This means that, they transform the features to artificial set and still retaining the information of the original dataset. Large number of features can affect the performance of a machine learning model. This work used four different methods to solve the high dimensionality problem.

2.1 Support vector machines

Support vector machine (SVM) is a supervised learning classification algorithm which builds hyper planes as decision surface thereby classifying input data to a high dimensional feature space. The hyper plane built distinguishes between the different classes examples and maximizes the separation margin. SVM model [6] represents instances as set of point, mapped so that the instances of the different classes are separated by a distinct line. The set of points are mapped and classified into classes according to the side of the line they belong to. That is how linear classification is performed by SVM. While, nonlinear classification is performed using kernel trick.

For a given training set $\{(x_i, y_i)\}_{i=1}^{N}$ with the input data as $x_i \in \mathrm{R}^{\mathrm{m}}$ and the corresponding class labels are given as $y_i \in \{-1,1\}$. The classifier model with hyper plane to split the two classes is given as [7]:

$$y = sign\,[w^T\,\phi(x) + b] \tag{1}$$

In Eqn. (1) through Eqn. (6), x is the feature vector, w denotes the weight vector perpendicular to the hyper planes, b is the bias value, $\emptyset$ is the nonlinear function and y is the class label. The nonlinear function $\emptyset(.)$ is the feature map which helps to transfer the input data to a high dimensional feature space. Maximizing the margin (M) gives the weight vector and the bias value. This is defined by Eqn. (2) [7].

$$M = \frac{2}{|w|} \tag{2}$$

The minimization is denoted in Eqn. (3) [7];

$$\frac{||w^2||}{2} = \frac{1}{2}w^T w \tag{3}$$

The features of the input data are split accordingly, and the features are mapped to the output class ($y_{i\,=}$ +1) if:

$$w^T \phi(x_i) + b \geq +1 \tag{4}$$

Also, the features correspond to output class ($y_{i\,=}$ -1) if:

$$w^T \phi(x_i) + b \leq -1 \tag{5}$$

These two sets of inequalities, Eqn. (4) and Eqn. (5) are combined into one giving Eqn. (6):

$$y(w^T \phi(x_i) + b) \geq +1 \tag{6}$$

Where i =1, 2…. N

The constrained optimization problem [7] which is also known as primal problem can be formulated when Eqn. (3) and Eqn. (6) are combined. This is shown in Eqn. (7) [7]:

$$\min_{w,b,\xi} J_{primal}(w, \xi) = \frac{1}{2}w^T w + C \textstyle\sum_{i=1}^{N} \xi_i \tag{7}$$

$$\text{Subject to: } y(w^T \phi(x_i) + b) \geq 1 - \xi_i \tag{8}$$

where, i =1, 2…. N, and $\xi_i \geq 0$ (ξ_i is slack variable for data point i and C is a regularization parameter)

There is need for slack variables ξ_i, so that misclassifications can be tolerated on the training data and this avoids overfitting the data [7]. The parameter C, is a regularization parameter and the choice of this parameter determines the balance between maximizing and minimizing the margin and classification error respectively. To solve the constrained optimization problem, the Lagrange function is constructed as shown in Eqn. (9) [7].

$$L\,(w, b, \xi, \alpha, \beta) = \frac{1}{2}w^T w + c\ \textstyle\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i \left\{y_i[w^T \phi(x_i) + b] - 1 + \xi_i\right\} + \sum_i^N \beta \xi_i \tag{9}$$

From Eqn. (9): α ≥ 0, β ≥ 0, where α and βare the Lagrange multipliers for $(i = 1,2,3 \ldots N)$. The solution of Eqn. (9) can be obtained by differentiating w.r.t w, b, ξ and assigning to zero value [7].

Thus, SVM can also be trained by solving the Lagrange dual of Eqn. (7) [8]. This is obtained when w is substituted by its expression in the Lagrange formed from the appropriate objective and constraints. The quadratic optimization problem or dual problem which corresponds to the primal problem is given in Eqn. (10) [8].

$$\max_{\alpha} J_{dual}\,(\alpha) = \textstyle\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i\, y_j K\left(x_i x_j\right) \tag{10}$$

Subject to: $0 \leq \alpha \leq c$ and $\sum_{i=1}^{N} \alpha_i y_i = 0$ Where, x = feature vector and α = Lagrange multiplier. K is the kernel function used and it maps the input features to a high dimensional feature space. The classifier for linear SVM can then be represented as Eqn. (11) [8]:

$$f(x) = w^T x + b \tag{11}$$

Where f is the linear function, x is the feature vector, b is the bias and w is the classifier weight which is computed using Eqn. (12) [8]

$$w = \textstyle\sum_{\alpha>0} \alpha_i y_i\, x_i \tag{12}$$

Therefore;

$$f(x) = \sum_{\alpha>0} \alpha_i y_i\,(x_i, x) + b \tag{13}$$

Furthermore, the nonlinear SVM can be derived by mapping the input data x into the feature space and the SVM is trained for the mapped features $\phi(x)$ [8]. SVM does this by using the kernel trick thereby replacing the inner product in Eqn. (13) with kernel $K\,(x_i, x_j)$ that corresponds to it. The kernel function maps the input reduced feature space and the goal is to find a separation between data. The classifier for nonlinear SVM is given in Eqn. (14) [8]:

$$f(x) = \sum_{\alpha>0} \alpha_i y_i\, K(x_i, x) + b \tag{14}$$

Figure 1 shows SVM solution for a binary classification problem. From the figure, the support vectors are the nearest datum to the optimal hyper plane, while the margin is the distance between the optimal hyper plane and the nearest datum to the optimal hyper plane.

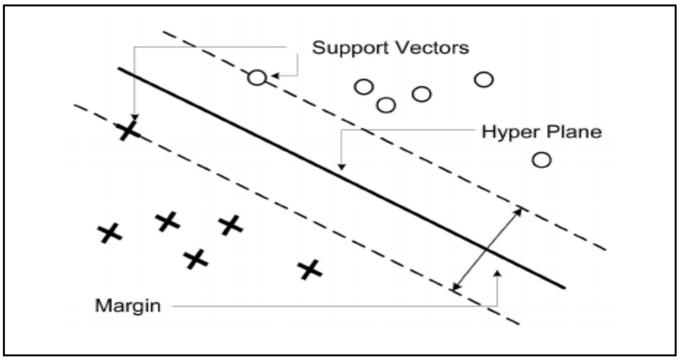


Figure 1. Working support vector machine for binary classification

2.2 Artificial neural networks

Artificial Neural Networks (ANN) or Neural Networks (NN) is composed of numerous processing elements that are highly connected and analogous to synapses [9]. These highly interconnected processing elements are called artificial neurons. The activation of artificial neuron is controlled by calculating inputs and weight using a mathematical equation. ANN contains intermediary layers between its input and output layer referred to as hidden layers and they have hidden nodes embedded in them. Thus, the resulting structure is known as multilayer neural network and the nodes in one layer are only connected to the nodes in the next layer [10].

The architecture of ANN is shown in figure 2 and this is a multilayer feed forward neural network. The multilayer feed forward neural networks have input layer, hidden layer and output layer. The output received from the input layer will be passed to the output layer after it must have been processed and computed in the hidden layer. The inherent structure of neural networks makes them a very powerful tool for processing complex dataset such as those in breast cancer studies that are characterized by highly nonlinear interactions between input data and target predictions [11]. Multilayer feed forward neural networks are trained with the standard back propagation (BP) algorithm and are broadly used for pattern classification because they learn to convert input data into results that are favorable [12]. The back propagation of errors from the output layer to the hidden layer is the main idea of this algorithm. A back propagation neural network represents a neuron and adjacent layers that are connected by weights

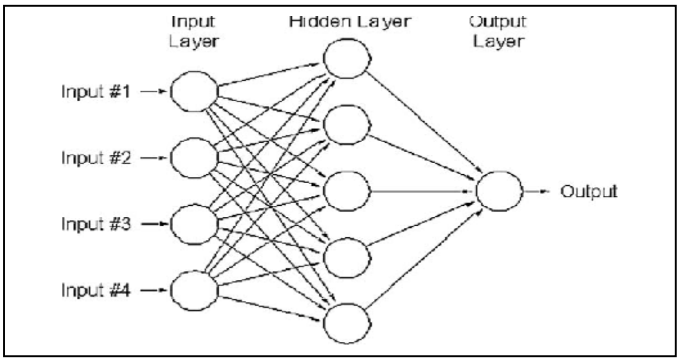


Figure 2. Artificial Neural Network Architecture.

The calculation can be explained using Eqn. (15) and Eqn. (16) to describe the activation process of hidden nodes.

$$I_j = \sum_i w_{ji} y_i + a_j \tag{15}$$

$$Y_j = f_j(I_j) \tag{16}$$

Where, f_j = activation function
w_{ji} = weight associated at the connection link between nodes in input layer i and nodes in hidden layer j
a_j = the bias associated at each connection link between the input layer and hidden layer.
y_i = input at nodes in input layer.
I_j = summation of weight inputs added with bias
Y_j = output of activation function at hidden layer.
The principle of output layer can be deduced using Eqn. (17) and Eqn. (18).

$$I_n = \sum_j w_{nj} y_j + b_n \tag{17}$$

$$Y_n = f_n(I_n) \tag{18}$$

Where, f_n = activation function
w_{nj} = weight associated at the connection link between nodes in hidden layer j and nodes in output layer n
b_n = the bias associated at each connection link between the hidden layer and output layer.
y_j = output at nodes in hidden layer
I_n = summation of weighted outputs at the output layer.
Y_n = final output at the output layer.
In Equations (15), (16), (17) and (18), it is important to note that i is the input layer, j is the hidden layer and n is the output layer.

2.3 Naïve Bayes classifier

The Naïve Bayes classifier (NBC) is a highly practical Bayesian learning technique. This classifier makes use of the Bayes rule which assumes independence among predictors. Simply put, NBC postulates that the presence of an attribute in a class is not related to the presence of any other attribute. The Bayes rule calculates conditional probability. Mathematically expressed, this is shown in Eqn. (19). The variables in Equations (19), (20) and (21) are defined as follows:

i.) P (Y|X) is the posterior probability of class Y given predictor (X).
ii.) P (Y) is the prior probability of class.
iii.) P (X|Y) is the probability of the predictor which is otherwise known as the likelihood.
iv.) P (X) is the prior probability of the predictor

$$P(Y|X) = \left(\frac{P(X|Y)P(Y)}{P(X)}\right) \tag{19}$$

By using this technique, all the features are presumed independent according to Bayes theorem which means there is no dependency among the attribute value on a given class and the other attributes [12]. The Bayes theorem enables us to express the posterior probability in terms of the prior probability P(Y), the class-conditional probability P (X|Y), and the evidence, P(X) as shown in Eqn. (19). The NBC works by estimating the class-conditional probability. By so doing, it assumes that the attributes are conditionally independent, given the class label y. The mathematical expression of the conditional independence assumption is given as:

$$P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y) \tag{20}$$

In Eqn. (20), each attribute set X: $\{X_1, X_2 \ldots. X_d\}$ consist of d attribute features. To classify a test dataset, NBC works by calculating the posterior probability of each class Y using Eqn. (21).

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)} \tag{21}$$

In Eqn. (21), P(X) is static for every Y, hence the class that maximizes the expression $P(Y)\prod_{i=1}^{d} P(X_i|Y)$ is chosen. NBC uses the conditional independence assumption to compute the conditional probability of each X_i Given Y, rather than calculating the class conditional probability of X_i.

2.4 *Correlation based feature selection*

Correlation based feature selection (CFS) is a feature selection technique that uses the filter approach. This feature selection technique does not depend on a ML algorithm to be applied to the selected features and the weights of feature attributes are assessed by looking only at the intrinsic properties of the data [13]. Often, the feature attributes in a dataset maybe highly correlated with each other. These features that highly correlate with other features give redundant information. Correlation based feature selection technique finds the correlation between features. Features that are highly correlated to other features are excluded by CFS. Similarly, features that highly interrelate with the class label are retained and selected.

In this paper, correlation between all features are computed and visualized. The correlation filter used is 0.7 and features with correlation greater than 0.7 are excluded from the training dataset. While the other features with lower mean are selected.

2.5 *Recursive feature elimination*

Recursive feature elimination (RFE) is a feature selection method that uses the wrapper approach. One of the limitations of this approach is that the literature available on wrapper techniques is not much compared to the filter algorithm [14]. RFE involves building a ML model with all the original features in the dataset and the features are ranked according to their quantitative importance to reducing the modeling error [15]. In this study, recursive feature elimination uses a random forest algorithm to test the combinations of features. Each feature subset is rated with an accuracy score. The subsets of features which have top ranking scores are chosen.

2.6 *Principal component analysis*

Principal component analysis (PCA) is a feature extraction method that transforms the original dataset into a reduced number of derived variables which do not correlate, and they are called principal components (PC) [16]. This work uses principal component analysis on neural networks. In performing PCA, cumulative variance is used as rule of thumb in reducing the feature dimension of the dataset. This rule of thumb chooses the number of principal components according to the Eigen value sizes or the proportion of variance each individual principal component explains [16].

In this study, PCA is applied to the WDBC dataset to identify the combination of attributes (principal components) that accounts for the most variance in the dataset.

2.7 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is one of the statistical feature extraction methods used to reduce high dimensions of data. It is mainly used for supervised dimension reduction. This means that it takes into consideration the different class labels. LDA is a feature extraction technique that computes transformation by maximizing the between class scatter and minimizing the within class scatter [17]. This is done simultaneously and the highest-class discrimination is achieved. To compute the optimal transformation in LDA, Eigen decomposition is used on the co-variance matrices.

LDA identifies attributes that account for the most variance between classes. Unlike PCA, LDA could produce better results. LDA training and LDA validation datasets are created.

3. Literature review

In this section, some of the related works previously done on breast cancer diagnosis by researchers using different machine learning approaches are discussed.

Ahmad et al. [4], compared the performance of decision tree (C4.5), SVM, and ANN. The dataset used was obtained from the Iranian center for breast cancer. Simulation results showed that SVM was the best classifier followed by ANN and decision tree.

Nematzadeh et al. [12], conducted a comparative study on decision tree, NB, NN and SVM with three different kernel functions as classifiers to classify WPBC and Wisconsin Breast Cancer (WBC). The experimental result showed that NN (10-fold) had the highest accuracy of 98.09% in WBC dataset, while SVM-RBF (10-fold) had the highest accuracy of 98.32% in WPBC dataset.

Hasan and Tahir [16], proposed an ANN classifier using PCA preprocessed data as optimal tool to improve differentiating between benign and malignant tumors on WBC dataset. They employed the three rules of thumb of PCA namely scree test, cumulative variance and Kaiser Guttman rule as feature selection. The result obtained showed that the method can distinguish between benign and malignant cases.

Ojha and Goel [18], use different ML algorithm to predict recurrent cases of breast cancer using the Wisconsin Prognostic Breast Cancer (WPBC) data set. The evaluation result produced SVM and decision tree (C 5.0) as the best predictors with 81% accuracy, while fuzzy c-means was found to have the lowest accuracy of 37%.

Ghosh et al. [19], diagnose and analyze breast cancer disease using two well-known classifiers which are Multilayer Perceptron using Back Propagation Neural Network (MLP BPN) and SVM. The experimental results of their work revealed SVM was the best classifier.

Osareh and Shadgar [20], investigated the issues of breast cancer diagnosis and prognostic risk evaluation of recrudescence and metastasis using SVM, K-nearest neighbor (KNN) and probabilistic neural network (PNN). These classifiers were combined with signal-to-noise ratio (SNR) feature ranking method, sequential forward selection-based (SFS) feature selection and PCA feature transformation. The SVM-RBF was found to obtain the best overall accuracies of 98.80%.

Bazazeh and Shubair [21], investigated SVM, random forest (RF) and Bayesian networks (BN) for breast cancer diagnosis and performed a comparative analysis on them. The WBC dataset was used as training set to evaluate the performance of the machine learning classifiers. The experimental results showed that SVM had the best performance in terms of accuracy, specificity and precision, while RF had the highest probability of correctly classifying tumors.

Azmi and Cob [22], built a system that can classify breast cancer tumor by employing neural network with feed –forward back propagation algorithm. The dataset used in their work was obtained from University of Wisconsin (UCI) machine learning repository. Experimental results revealed that neural network with hidden layer of 7 achieved the best accuracy of 96.63% when compared to others.

Gayathri and Sumathi [23], conducted a comparative study of relevance vector machine (RVM) with other ML algorithms used for detecting breast cancer. They used linear discriminant analysis (LDA) to reduce features. The data was classified by the RVM algorithm. The dataset used in this work is the WBC. The accuracy equals 96%. The sensitivity and specificity obtained from the simulation results are 98% and 94% respectively.

4. Experimental work

The proposed method used in this paper follows the concept of classification techniques. This concept generally follows two distinct steps. The first step involves building the classification model, in which the classification algorithm builds the model by learning from a training dataset which have class labels [19]. The second step is the prediction phase. The validation dataset is used to evaluate the classification accuracy of the model built. The experiment was conducted in RStudio using the R programming language for developing and testing the different ML models.

4.1 Data collection

Data collection involves gathering of data. The data used in this work was acquired from University of California- Irvine (UCI) machine learning repository. It is an open source. This dataset is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset which was donated on November 1st, 1995 and it consist of 569 instances of which 357 instances are benign and 212 are malignant cases [24]. It has 32 attributes which include two class attribute labels (diagnosis: B= benign, M= malignant), ID number and 30 real-value attributes. These attributes are computed from a digitized image of a fine needle aspiration (FNA) procedure of a breast mass and are used to describe the characteristics of the cell nuclei present in the image.

4.2 Data selection and preprocessing

Data preprocessing is performed to improve the quality of a dataset to get a clean data which can be useful for modeling [25]. There are several processes involved in data preprocessing. These processes include data cleaning, feature selection, feature extraction etc. Data cleaning involves removing noise and the inconsistencies which are present in the data, thereby improving the quality of the data. During the data preprocessing stage, the data is partitioned into the training dataset and validation dataset. The training dataset is used in training the machine learning model, while the validation dataset is used during the prediction stage. The training dataset consist of 399 observations of 31 variables, while the validation dataset has 170 observations of 31 variables. Other preprocessing operations performed on the dataset are centering and scaling.

Data selection has been an active research area in pattern recognition, statistics, and data mining. The goal of data selection is to reduce features by employing feature selection techniques and feature extraction methods on the training dataset. Feature selection involves selecting features combination which is important towards the target classification and ignores the less important features. Feature extraction on the other hand, reduces the number of dimensions by transforming features in high dimensional space to fewer dimensions. The feature selection techniques used in this work are CFS and RFE methods. While the feature extraction methods used are PCA and LDA

4.3 Apply machine learning techniques

After preprocessing the data, the next stage is to apply ML classification techniques on the processed data. During this stage, the processed data will be used to train and build the ML model. The ML classification algorithms considered in this work are SVM with RBF, ANN and NBC. These algorithms have been explained in section II. The training data which have the entire feature attributes is used to train the models to classify the data into benign and malignant tumors. Furthermore, the feature selection and feature extraction techniques discussed earlier reduces the dimensions of the data and the data with reduced features are used to train the models. This paper proposed an integrated method that combines ML classification based algorithm and feature selection/extraction technique. The experimental results showed that a model based on SVM with RBF kernel and LDA preprocessed data can produce promising results.

4.4 Performance evaluation

The evaluation of a ML algorithm performance involves testing the proposed model(s) built. In this work, the evaluation was done by comparing the model results with the real data value. This phase is the prediction phase whereby the test dataset is used to assess the performance of the models in classifying benign and malignant tumors.

The confusion matrix is built in comparing the predicted results with the actual values. The data in the matrix is used to compute the performance of the classifier [12]. Different performance metric that can be used to evaluate the performance of the ML model are accuracy, area under ROC curve, precision, recall, sensitivity, specificity, kappa statistic etc.

5. Results and discussion

Feature selection and feature extraction methods were performed on the data to reduce the dimension of features, thereby producing reduced versions of the original dataset. The methods considered are CFS, RFE, PCA and LDA. SVM, ANN and NBC were employed to train the datasets. Table 1 shows the classifier models built and their respective results in the different performance metrics.

The Receiver operating characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier. ROC graph plots the true positive rate (TPR) which is the benefit on y axis against the false positive rate (FPR) which is the cost on x axis [26]. The domains of both axes are stretched between 0 and 1 and graph is plotted by obtaining the TPR and FPR for every possible threshold value of the classifier [21]. ROC curve is used to visualize the classification models' performance by showing the tradeoff between the cost and benefit of that classifier [21]. The area under the ROC curve displays the performance of the ML model and high performance are denoted by values close to 1.

Accuracy is the ratio of the total number of correct predictions to the total number of samples. The Kappa or Cohen's kappa is just like the classification accuracy and it considers the expected rate of error. Precision shows the percent of the positive cases that were correctly predicted. Recall which is another performance metric is the percent of correctly identified positive cases. While sensitivity and specificity represent the true positive rate and true negative rate respectively.

5.1 Support vector machine performance

The simulation results revealed that reducing the high dimensionality of a dataset can improve the performance of a SVM model. Of all the dimensionality reduction method integrated with SVM, SVM and LDA combination tends to outperform the others. This LDA-SVM combination showed a classification accuracy of 98.82%. It also has a very good area under the ROC curve. The areas under the ROC curves are used to evaluate the performance of the algorithms.

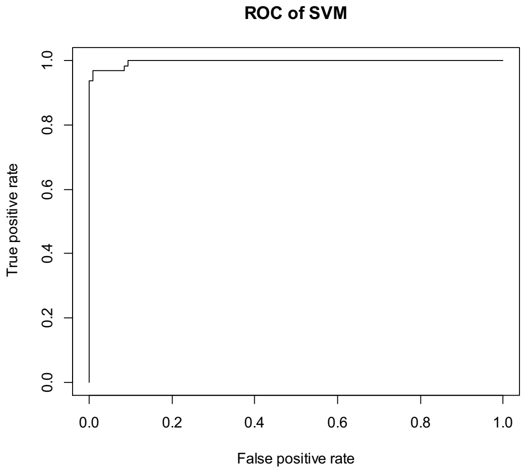


Figure 3. ROC plot of SVM

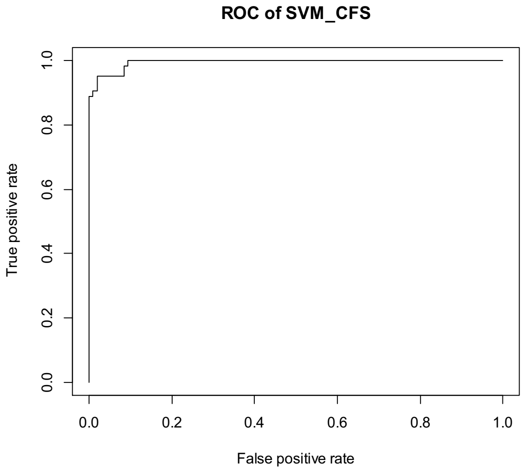


Figure 4. ROC plot of SVM and CFS combination

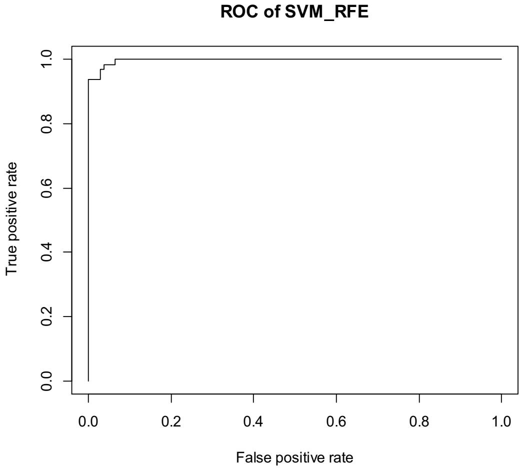


Figure 5. ROC plot of SVM and RFE combination

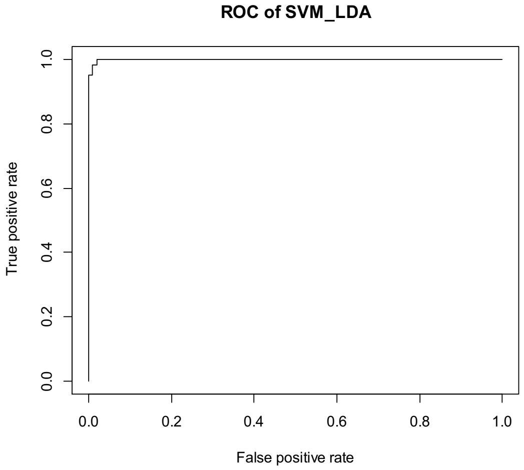


Figure 6. ROC plot of SVM and LDA combination

5.2 Artificial neural network performance

From the result, it can be observed that significant feature selection/feature extraction techniques have been found to improve the performance of ANN models. Using dataset with all the features to train the ANN model produced the poorest result as compared to the other ANN models built, while an ANN model built with LDA preprocessed data performs better than the other ANN models. The area under the ROC plots are used to evaluate the performance of the models.

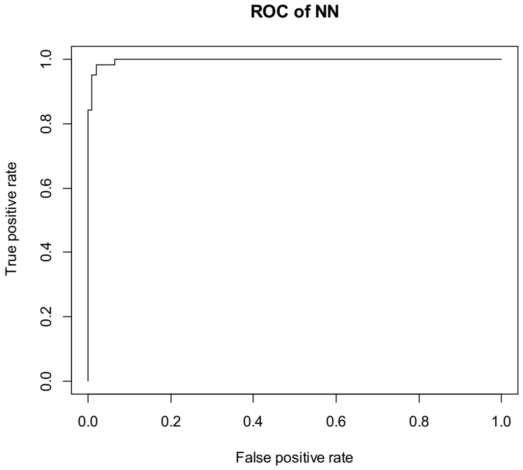


Figure 7. ROC plot of ANN

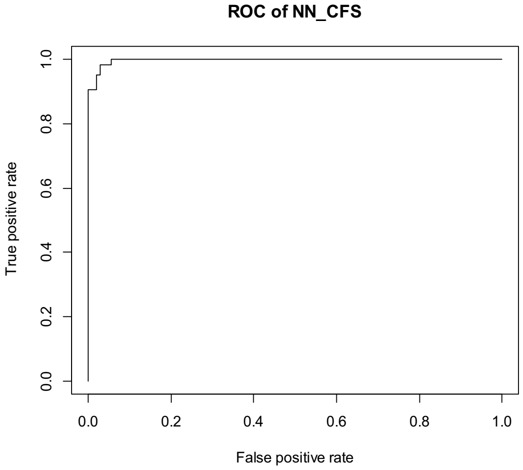
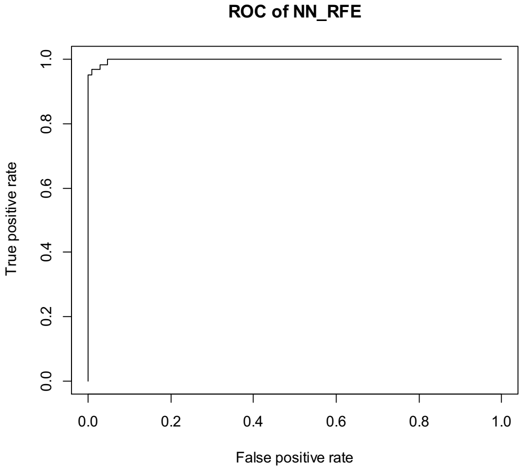


Figure 8. ROC plot of ANN and CFS combination



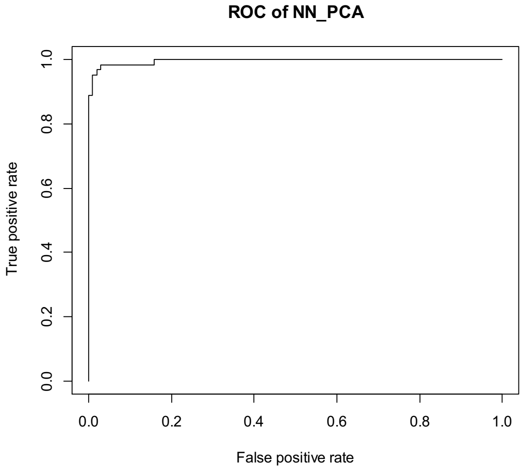


Figure 9. ROC plot of ANN and RFE combination

Figure 10. ROC plot of ANN and PCA combination

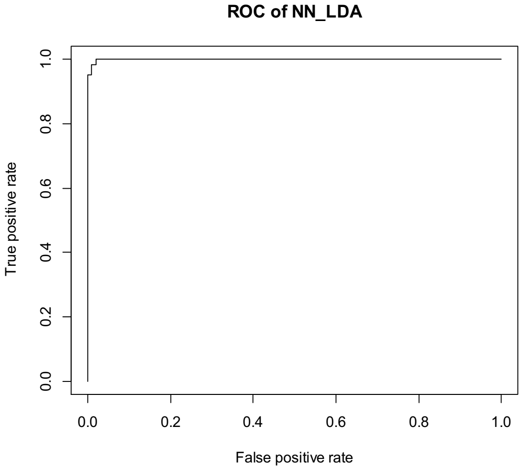


Figure 11. ROC plot of ANN and LDA combination

5.3 Naïve bayes classifier performance

From the experimental results, it was observed that CFS and LDA can affect the performance of NBC model. However, the combination of Naïve Bayes model and LDA outshines the other NBC models in terms of performance.

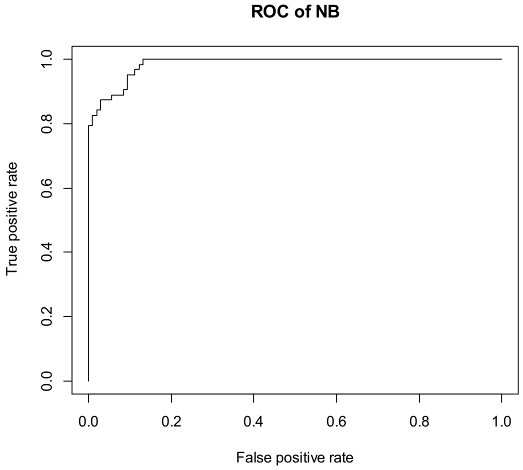


Figure 12. ROC plot of NB

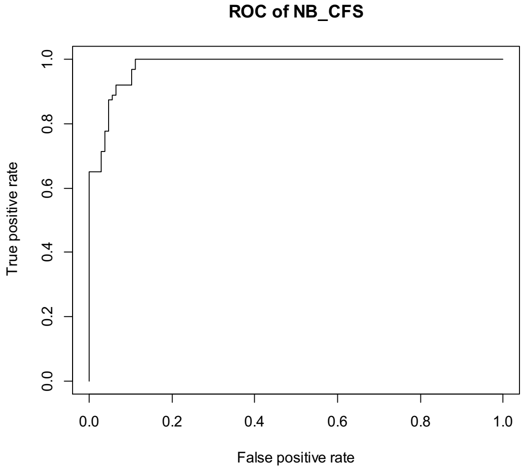


Figure 13. ROC plot of NB and CFS combination

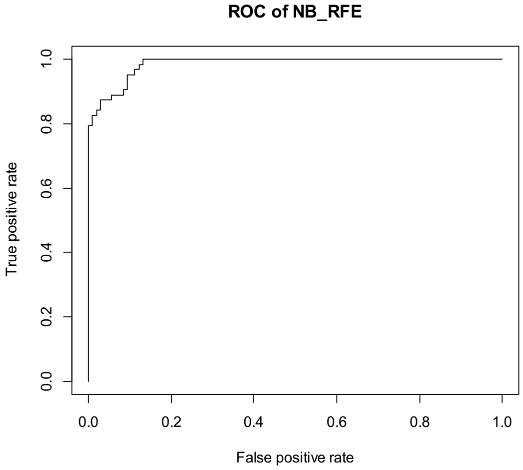


Figure 14. ROC plot of NB and RFE combination

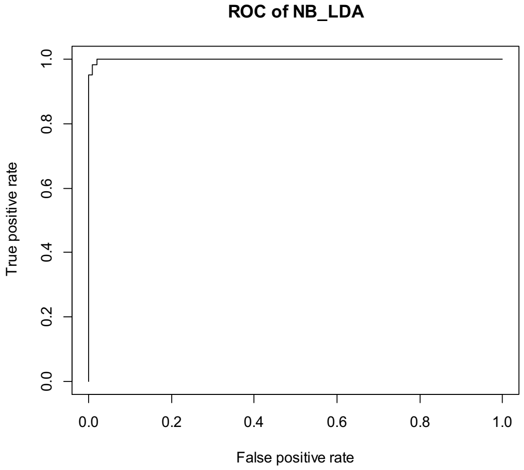


Figure 15. ROC plot of NB and LDA combination

5.4 Performance comparison of machine learning models

Table 1. Summary of machine learning models' results

Machine Learning Models	Accuracy	Area under ROC curve	Precision	Recall	Sensitivity	Specificity	Kappa
SVM	0.9647	0.9964	0.9385	0.9682	0.9682	0.9626	0.9248
SVM-CFS	0.9647	0.9954	0.9524	0.9524	0.9524	0.972	0.9243
SVM-RFE	0.9647	0.9976	0.9524	0.9524	0.9524	0.972	0.9243
SVM-LDA	0.9882	0.9994	0.9841	0.9841	0.9841	0.9907	0.9748
NN	0.9706	0.9985	0.9833	0.9365	0.9365	0.9907	0.9363
NN-CFS	0.9706	0.9973	0.9531	0.9683	0.9683	0.9719	0.9372
NN-RFE	0.9824	0.9989	0.9839	0.9683	0.9683	0.9907	0.962
NN-PCA	0.9765	0.9937	0.9836	0.9524	0.9524	0.9907	0.9492
NN-LDA	0.9882	0.9994	0.9841	0.9841	0.9841	0.9907	0.9748
NB	0.9118	0.9860	0.8750	0.8889	0.8889	0.9252	0.8115
NB-CFS	0.9176	0.9799	0.9152	0.8571	0.8571	0.9533	0.8211
NB-RFE	0.9118	0.9860	0.8750	0.8889	0.8889	0.9352	0.8115
NB-LDA	0.9824	0.9994	0.9839	0.9683	0.9683	0.9907	0.962

The results of the different classifier models are viewed and compared with each other. From the results displayed in table 1, it can be observed that SVM-LDA, NN-LDA, NN-RFE, NN-PCA, NB-LDA and NN-LDA combinations and NN alone obtained the best performance of 99.07% in terms of specificity. However, SVM-LDA combination and NN-LDA combination obtained the best performance in terms of sensitivity (98.41%), precision (98.41%), recall (98.41%), precision (98.41%), recall (98.41%) and accuracy (98.82%). The poorest performing ML model in terms of sensitivity and recall was NB-CFS combination, while in terms of precision, specificity and accuracy, NB and NB with RFE preprocessed data performed poorest.

In addition, dimensionality reduction affects the performance of a ML algorithm. Simulation results show that CFS and RFE methods increase the precision and specificity of SVM with radial kernel, while LDA increases both its accuracy and sensitivity (detection of malignant cases). In the case of ANN, CFS increases its sensitivity, while RFE, PCA and LDA increase its sensitivity and classification accuracy. The accuracy of NB is improved by CFS and LDA, while its sensitivity is improved by LDA. The machine learning classification algorithms are also compared using the area under their respective ROC plots and kappa values. SVM-LDA, NN-LDA and NB-LDA had the best area under their ROC curves. The value was 0.9994 and it displays their high performance. Moreover, the best kappa value of 0.9748 was obtained by SVM-LDA and NN-LDA.

Dimensionality reduction is very significant in the classification process. Feature extraction plays a vital role in the classification models. The main purposes of performing feature extraction are to improve the prediction performance and ensure faster prediction. Thus, the benefits of feature extraction cannot be over emphasized. When feature extraction is employed, data visualization and understanding are facilitated. Also, feature extraction decreases storage requirement and training times.

6. Conclusion

This paper analyzed WDBC dataset using dimensionality reduction techniques and three popular ML algorithms to classify malignant and benign tumors. The experimental work proves that classification performance is dependent on the ML classification technique chosen. Simulation results showed that SVM-LDA and NN-LDA outperforms the other ML classifier models. Nevertheless, SVM-LDA is chosen over NN-LDA because NN-LDA takes a longer computational time. Therefore, this paper proposes an intelligent approach which integrates linear discriminant analysis and support vector machine (with RBF kernel) for breast cancer diagnosis. This chosen approach showed good and promising results over the validation dataset. It obtained a classification accuracy of 98.82%, sensitivity of 98.41%, specificity of 99.07% and area under the receiver operating characteristic curve of 0.9994.

This research work reveals that feature selection and feature extraction can help improve the diagnosis of benign and malignant tumors using machine learning techniques. Thus, this work concludes that integrating significant dimensionality reduction methods with ML classification techniques produces better approaches for medical diagnoses (breast cancer diagnosis used as case study). The main idea is to combine the advantages of dimensionality reduction and ML algorithm.

Future work can be directed towards developing the chosen approach into a potential practical method for aiding and assisting doctors with quick second opinion in diagnosing breast cancer. Future work can also consider comparing more ML algorithms used for breast cancer diagnosis. More disease options can also be considered in future works.

References

[1] Niharika G. Maity and Dr. Sreerupa Das, "Machine learning for improved diagnosis and prognosis in healthcare", *2017 IEEE Aerospace Conf.*, IEEE, 2017.

[2] About Breast Cancer. [Online]. Available: https://www.breastcancerfoundation.org.my/about-breast-cancer/. Accessed [August] [2017].

[3] Purnami, Santi Wulan, S. P. Rahayu and Abdullah Embong, "Feature selection and classification of breast cancer diagnosis based on support vector machines.*" 2008 Int. Symp. on Information Technology* 1 (2008): 1-6.

[4] L.G. Ahmad, A.T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi and A.R. Razavi, "Using three machine learning techniques for predicting breast cancer recurrence," (2013), *J Health Med Inform* **4: 124**. doi:10.4172/2157-7420.1000124.

[5] Vimal Kumar Dubey and Amit Kumar Saxena, "Hybrid classification model of correlation-based feature selection and support vector machine," *Proc. in 2016 IEEE Int. Conf. on Current Trends in Advanced Computing (ICCTAC),* IEEE, 2016.

[6] Runjie Shen, Yuanyuan Yang and Fengfeng Shao, "Intelligent breast cancer prediction model using data mining techniques," *Proc. in 2014 6th Int. Conf. on Intelligent Human-Machine Systems and Cybernetics*, pp384-387, 2014.

[7] Johan A.K. Suykens, "Support vector machines and kernel-based learning for dynamical systems modelling", *15th IFAC Symp. on System Identification*, pp 1029-1037 July 6-8, 2009, Saint-Malo, France.

[8] Zhouyu Fu, Antonio Robles-Kelly and Jun Zhou. "Mixing linear SVMs for nonlinear classification," *IEEE Transactions on Neural Networks*, pp 1963-1975, IEEE, 2010.

[9] Vaibhav Narayan Chunekar and Hemant P. Ambulgekar, "Approach of neural network to diagnose breast cancer on three different data set," *Proc. in 2009 Int. Conf. on Advances in Recent Technologies in Communication and Computing*, IEEE, 2009.

[10] P.-N Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Addison-Wesley, 2006.

[11] Smita Jhajharia, Harish Kumar Varshney, Seema Verma and Rajesh Kumar, "A neural network-based breast cancer prognosis model with PCA processed features," *Proc. in 2016 Int. Conf. on Advances in Computing, Communications and Informatics (ICACCI),* pp 1896-1901, Sept. 21-24, 2016, Jaipur, India.

[12] Zahra Nematzadeh, Roliana Ibrahim and Ali Selamat, "Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques," *Proc. in 2015 10th Asian Control Conf. (ASCC),* pp 1-6, IEEE, 2015.

[13] K. Fathima Bibi and Dr. M. Nazreen Banu, "Feature subset selection based on filter technique," *Proc. in 2015 Int. Conf. on Computing and Communications Technologies (ICCCT'15),* pp 1-6, IEEE, 2015.

[14] Jozsef Suto, Stefan Oniga and Petrica Pop Sitar, "Comparison of wrapper and filter feature selection algorithms on human activity recognition," *Proc. in 2016 6th Int. Conf. on Computers Communications and Control (ICCCC),* pp 124-129, IEEE, 2016.

[15] H. Ravishankar, R. Madhavan, R. Mullick, T. Shetty, L. Marinelli, and Suresh E. Joel, "Recursive feature elimination for biomarker discovery in resting state functional connectivity," *Proc. in 2016 38th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society,* pp 4071-4074, IEEE, 2016.

[16] H. Hasan and N. M. Tahir, "Feature selection of breast cancer based on principal component analysis," in Signal Processing and Its Applications (CSPA), *2010 6th International Colloquium on, 2010*, pp. 1-4.

[17] H. Abbasian, B. Nasersharif, A. Akbari, M. Rahmani and M. S. Moin, "Optimized linear discriminant analysis for extracting robust speech features," *2008 3rd Int. Symp. on Communications, Control and Signal Processing,* pp 819-824, IEEE, 2008.

[18] Uma Ojha and Savita Goel, "A study on prediction of breast cancer recurrence using data mining techniques," *2017 7th Int. Conf. on Cloud Computing, Data Science & Engineering – Confluence,* pp 527-530, IEEE, 2017.

[19] Soumadip Ghosh, Sujoy Mondal and Bhaskar Ghosh, "A comparative study of breast cancer detection based on SVM and MLP BPN classifier," *2014 1st Int. Conf. on Automation, Control, Energy and Systems (ACES),* pp 1-4, IEEE, 2014.

[20] Alireza Osareh and Bita Shadgar. "Machine learning techniques to diagnose breast cancer," *2010 5th Int. Symp. on Health Informatics and Bioinformatics,* 20-23 April 2010, Antalya, Turkey.

[21] Dana Bazazeh and Raed Shubair. "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," *2016 5th Int. Conf. on Electronic Devices, Systems and Applications (ICEDSA),* 6-8 December 2016, Ras Al Khaimah, UAE.

[22] Muhammad Sufyian Bin Mohd Azmi and Z. C. Cob, "Breast cancer prediction based on backpropagation algorithm," *Proc. of 2010 IEEE Student Conf. on Research and Development (SCOReD 2010),* pp 164-168, 13 - 14 Dec 2010, Putrajaya, Malaysia.

[23] B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," *2016 IEEE Int. Conf. on Computational Intelligence and Computing Research (ICCIC),* pp 1-5, IEEE, 2016.

[24] UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names. Accessed [August] [2017].

[25] Poonam Pandey and Radhika Prabhakar, "An analysis of machine learning techniques (J48 & AdaBoost) -for classification," *2016 1st India Int. Conf. on Information Processing (IICIP)*, PP 1-6, IEEE, 2016, India.

[26] L. F. Carvalho, G. Fernandes, M. V. O. De Assis, J. J. P. C. Rodrigues, and M. Lemes Proenca, "Digital signature of network segment for healthcare environments support," *Irbm*, **vol. 35, no. 6**, pp. 299-309, 2014.