Jurnal Teknologi

Full paper

Cancer Detection Using Aritifical Neural Network and Support Vector Machine: A Comparative Study

Sharifah Hafizah Sy Ahmad Ubaidillah[a*], Roselina Sallehuddin[a], Nor Azizah Ali[a]

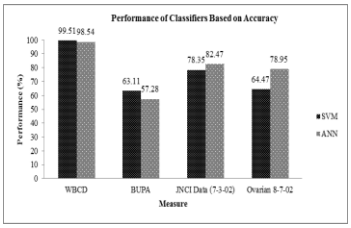
[a]*Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*

*Corresponding author: sh.hafizah1989@gmail.com

Article history

Received :14 June 2013
Received in revised form :
9 September 2013
Accepted :15 Octobe r2013

Graphical abstract



Abstract

Accurate diagnosis of cancer plays an importance role in order to save human life. The results of the diagnosis indicate by the medical experts are mostly differentiated based on the experience of different medical experts. This problem could risk the life of the cancer patients. From the literature, it has been found that Artificial Intelligence (AI) machine learning classifiers such as an Artificial Neural Network (ANN) and Support Vector Machine (SVM) can help doctors in diagnosing cancer more precisely. Both of them have been proven to produce good performance of cancer classification accuracy. The aim of this study is to compare the performance of the ANN and SVM classifiers on four different cancer datasets. For breast cancer and liver cancer dataset, the features of the data are based on the condition of the organs which is also called as standard data while for prostate cancer and ovarian cancer; both of these datasets are in the form of gene expression data. The datasets including benign and malignant tumours is specified to classify with proposed methods. The performance of both classifiers is evaluated using four different measuring tools which are accuracy, sensitivity, specificity and Area under Curve (AUC). This research has shown that the SVM classifier can obtain good performance in classifying cancer data compare to ANN classifier.

Keywords: Support vector machine; artificial neural network; classification; cancer; accuracy

Abstrak

Diagnosis kanser yang tepat memainkan peranan yang penting dalam usaha untuk menyelamatkan nyawa manusia. Keputusan diagnosis yang telah disahkan oleh pakar perubatan adalah berbeza mengikut pengalaman masing-masing. Masalah ini boleh membahayakan nyawa pesakit kanser. Dari kajian terdahulu, didapati bahawa penggunaan sistem pembelajaran Kepintaran Tiruan (AI) seperti Rangkaian Neural Buatan (ANN) dan Mesin Vektor Sokongan (SVM) boleh membantu doktor dalam mendiagnosis kanser dengan lebih tepat. Keupayaan kedua-dua algoritma ini telah terbukti dalam menghasilkan prestasi yang baik bagi pengelasan kanser. Tujuan kajian ini adalah untuk membandingkan prestasi ANN dan SVM dalam mengelaskan empat dataset kanser yang berbeza. Ciri-ciri data kanser payudara dan kanser hati adalah berdasarkan kepada keadaan organ-organ tersebut, manakala bagi kanser prostat dan kanser ovari, kedua-dua set data adalah dalam bentuk data ekspresi gen. Set data dikelaskan kepada dua jenis ketumbuhan iaitu ketumbuhan yang tidak berbahaya dan ketumbuhana yang merbahaya dengan menggunakan algoritma-algoritma yang dicadangkan. Prestasi kedua-dua algoritma dinilai menggunakan empat alat pengukur yang berbeza iaitu kejituan, kepekaan, keperincian dan nilai kawasan di bawah lengkungan (AUC). Kajian ini telah menunjukkan bahawa SVM mempunyai prestasi yang baik dalam mengklasifikasikan data kanser berbanding ANN.

Kata kunci: Mesin vektor sokongan; rangkaian neural buatan; pengelasan; kanser; kejituan

1.0 INTRODUCTION

Cancer is one of the leading causes of death in most countries around the world. The survival rate is strongly influenced by stage of the malignancy (malignant tumour) at the point of diagnosis [1]. Thus, an early diagnosis is needed in order to give the proper treatment to the patients and to help reduce the mortality and morbidity rate. Accurate diagnosis for different types of cancer plays an important role to the doctors to assist them in determining and choosing the proper treatment [2]. Undeniably, the decisions made by the doctors are the most important factors in diagnosis but lately, application of different AI classification techniques have been proven in helping doctors to facilitate their decision making process [3]. Recently, the use of AI classification techniques in the cancer classification in the medical field has increased gradually. Possible errors that might occur due to unskilled doctors can be minimized by using classification

techniques. This technique can also examine medical data in a shorter time and more precisely [3]. The aim of the classification is to develop a set of models that are able to correctly classify the class of different objects. There are three types of inputs to such models, which are; a set of objects or commonly described as training data, the dependent variables or classes which these objects belong to and the independent variables, which is a set of variables describing different characteristics of the objects. Once a classification model is built, it can be used to classify the class of the objects for which class information is unidentified [4].

There are many types of classification algorithm or commonly known as classifiers have been used for cancer diagnosis. Some of them are Artificial Neural Network (ANN), Support Vector Machine (SVM), Genetic Algorithm (GA), Fuzzy Set (FS) and Rough Set (RS). They are used to classify cancer dataset as malignant tumours (cancerous) and benign tumours (non-cancerous). However, ANN and SVM are the classifiers that received attention from most researchers. Both of them have been proven to produce good classification accuracy performance. Several comparative studies on ANN and SVM have been conducted by the researchers [3,5,6,7], however the result reported are inconsistent.

Due to the inconsistent result obtained, the aim of this study is to further validate the performance of both ANN and SVM in cancer classification. The performance of both classifiers will be tested and evaluate on four different cancer datasets which are divided into two type of cancer data namely; standard data and gene expression data. These four datasets are obtained from UCI Machine Learning Repository and National Cancer Institute (NCI). This paper is organized as follows. Section 2 provides the related studies carried out on cancer classification using ANN and SVM classifier and basic concept of ANN and SVM. In section 3, the explanation on dataset used will be explained. Section 4 described on the methodology of this study. In section 5, the results and discussion are summarized on tables to show the performance and the comparison of applied classifiers. Finally, the paper is concluded in section 6.

■2.0 LITERATURE REVIEW

2.1 Related Work

ANN and SVM classifiers have been the most useful AI techniques for the researchers' community to classify cancer. Both of them have obtained excellence performance in classifying cancer. For ANN, there are some studies that had proven the excellent performance of this classifier. In [8], a study on liver biopsy images using Probabilistic Neural Network (PNN) has been done. The result expresses high performance of PNN classifier with 92% of accuracy for testing set. Besides that, ANN classifier was also used for breast cancer classification for Wisconsin Breast Cancer Database (WBCD) dataset [9]. A neural network with feed-forward back propagation algorithm was used to classify the cancerous tumours from a symptom that causes the breast cancer disease. This model produces a correct classification rate of 96.63% for the testing set. In 2010, [10] applied ANN classifier to the lung cancer dataset. The dataset is in the type of CT images. They obtained 84.6% accuracy for the unknown samples of the dataset. [11] in 2006 focused on classifying the ovarian cancer dataset. A novel Radial Basis Function (RBF) neural network is used to classify the data. The results in the dataset show that the RBF neural network is able to achieve 100% accuracy.

SVM classifiers have also gained the attention of the researchers in classifying cancer. Recently, [12], proposed a cancer classification model using SVM for prostate magnetic resonance spectra dataset. The result stress that the SVM classifier can obtain high accuracy (95.85%). In the same year, [13] applied SVM classifier to the breast cancer dataset using digital ultrasound image database. They obtained 86.92% accuracy with 321 samples. In addition, the SVM classifier was also used for cancer classification of prostate cancer datasets by [14] in 2010. SVM classifier was used to classify the cancer dataset into two classes namely; normal and cancer samples. This model produces an accurate classification rate of 95.09%. In 2007, [15] successfully classified the breast cancer dataset by using LS-SVM classifier with an accuracy rate of 98.53%.

Several comparative studies have been done by the researchers in cancer classification in order to select the best techniques to classify cancer. However, the result obtains from the previous studies are inconsistent. Some studies state that ANN is better compare to SVM. In the study conducted by [5], which compare the performance of ANN and SVM on Dynamic Magnetic Resonance Imaging (MRI) of breast cancer data, had found that Probabilistic Neural Network (PNN) obtained the maximum accuracy among all classifiers. [6] also found that ANN outperforms SVM in the classification of Microclacification Clusters (MCCs) in mammogram imaging. [7] studied the performance of ANN and SVM classifier on Wisconsin Breast Cancer Database (WBCD). The research has demonstrated that Radial Basis Function Neural Network (RBFNN) outperformed the polynomial SVM for correctly classifying the tumours.

In contrast, some studies had found that SVM has better performance than ANN. In the study conducted by [16], the result obtained showed that Polynomial SVM gives better result than ANN in classifying the prostate cancer data. All of the results are determined based on the value of accuracy for each classifier. In 2011, [13] compared the performance of SVM and ANN in classifying breast cancer dataset. The experimental result demonstrate that the SVM classifier gives the best performance. Besides that, studied done by [14] also stressed that SVM results ineffectual and powerful classification of a prostate cancer dataset compare to ANN. Table 1 summarizes the result obtain for each comparative study and from this table it can be concluded that both of the classifiers can obtain good percentages performance in classifying cancer. However, all of the previous studies only compared the performance of both classifiers on one type of cancer data whether standard data or gene expression data. Thus, this study is conducted to further validate the performance of both classifiers in both type of cancer data in order to verify which classifier could performed better for both type of cancer data.

Table 1 S ummary of comparative studies

Author	ANN (%)	SVM (%)
[5]	94.00	88.00
[6]	78.00	72.00
[7]	96.57	92.13
[16]	79.30	81.10
[13]	86.60	86.92
[14]	94.11	95.09

2.2 Artificial Neural Network (ANN) Classifier

Artificial Neural Network (ANN) is a branch of computational intelligence that employs a variety of optimization tool to learn from past experiences and use that prior training to classify new

data, identify new patterns or predict. Artificial Neural Networks (ANNs) are gross simplifications of real (biological) networks of neurons. Inspired by the structure of the brain, a neural network consists of a set of highly interconnected entities, called nodes or units. Each unit is designed to mimic its biological counterpart, the neuron. Each accepts a weighted set of inputs and responds with an output [17]. Figure 1 shows the working of nodes in Artificial Neural Network (ANN).

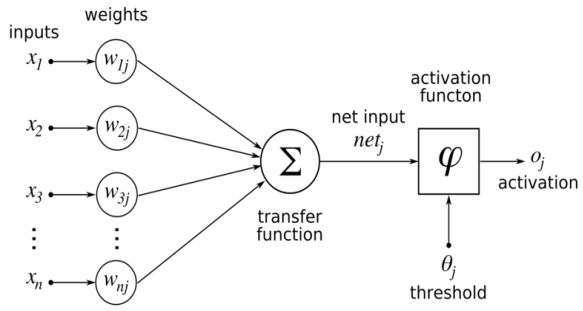


Figure 1 Working of node in ANN

The Back Propagation Neural Network (BPNN), also called multi-layer feed-forward neural network or multi-layer perceptron, is very popular and is most widely used [9,17]. The BPNN is based on the supervised procedure whereas the network constructs a model based on examples of data with known outputs [17]. Back propagation algorithm is a training algorithm where signals travel in one direction from input neuron to an output neuron without returning to its source [9]. Back propagation algorithm consists of at least three layers of units which are input layer, at least one hidden layer and output layer. The number of nodes in the input layer is corresponded to the number of input variables while for the number of nodes in the output layer is determined by the number of output variables [5]. In the context of cancer classification, the values of output variables are either zero for benign tumour or 1 for malignant tumours.

The term back propagation refers to the way the error computed at the output side is propagated backward from the output layer, to the hidden layer, and finally to the input layer. Each of the iteration in back propagation constitutes two sweeps: forward activation to produce a solution, and a backward propagation of the computed error to modify the weights. The forward and backward sweeps are performed repeatedly until the ANN solution agrees with the desired value within a pre-specified tolerance. The back propagation algorithm provides the needed weight adjustments in the backward sweep [9]. Table 2 shows the nine steps in BP algorithm.

2.3 Support Vector Machine (SVM) Classifier

Support Vector Machine (SVM) is a machine learning which has been extensively used as a classification tool and has found a great deal of success in many applications. Originally, SVM is developed based on the Vapnik-Chervonenkis (VC) theory and structural risk minimization (SRM) principle [18, 19] which is trying to find the trade-off between minimizing the training set error and maximizing the margin, in order to achieve the best generalization ability and remains resistant to over fitting. SVM is a method to estimate the function classifying the data into two classes [16]. For cancer classification, the classes will be divided into two which are benign and malignant tumours. A very brief review of SVM will be concentrated in this section. There are two types of SVM classifier which are linear SVM and non-linear SVM.

For linear SVM, consider *N* pairs of training samples:

$$(x_i, y_i),\ i = 1,2,.....n \tag{1}$$

Where $x_i \notin R^n$ is a *k*-dimensional feature vector and $y_i \notin \{+1,-1\}$ is the class label of x_i. A hyperplane in the feature space can be described as

$$w.x + b = 0 \tag{2}$$

where *w* is an orthogonal vector while *b* is a scalar. For linearly separable cases of training samples, SVM generate the optimal hyperplane that separates two classes with maximum margin and no training error [16, 20]. The hyper plane is placed midway between the two classes to maximize the margin [2]. Now maximizing the separating margin is equivalent to maximize the minimum value of signed distance *d(i)* from a point x_i to the hyperplane [20, 21]. The value of *d(i)* can be obtained by

$$d(i) = \frac{w.x_i + b}{\|w\|} \tag{3}$$

The parameter pairs of *w* and *b* that corresponding to the optimal hyperplane is the one that minimize

$$L(w) = \frac{1}{2}\|w\|^2 \tag{4}$$

subject to

$$y_i(w.x_i + b) \geq 1, i = 1,2,.....n \tag{5}$$

When the training samples are linearly non-separable, there is no such a hyperplane that is able to classify every training point correctly [20]. In order to solve the imperfect separation, the optimization idea can be generalized by introducing the concept of soft margin [21]. Thus, the new optimization problem becomes:

$$\text{Minimize } L(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \tag{6}$$

so that

$$y_i(w.x_i + b) \geq 1 - \xi_i, i = 1,2,.....n \tag{7}$$

where ξ_i are called as slack variables which are related to the soft margin, and *C* is the tuning parameter used to balance the margin and the training error. The optimization problem in (5) and (7) can be solved by using the Lagrange multipliers α_i that transform to quadratic optimization problem, for which there exist a unique solution. According to the KuhnTucker theorem of optimization theory [22], the optimal solution satisfy

$$\alpha_i\left|y_i.(w.x_i + b) - 1 = 0, i = 1,2,.....n\right| \tag{8}$$

(8) has non-zero Lagrange multipliers if and only if the points x_i satisfy

$$y_i.(w.x_i + b) = 1 \tag{9}$$

These points are called as Support Vector (SV) which lie either on or within the margin. Hence, if α_i is the non-zero optimal solution, the classification phase can be stated as

$$f(x) = \text{sgn}\left\{\sum_{i=1}^{n}\alpha_i y_i(x_i.x) + b\right\} \tag{10}$$

When a linear SVM does not gives good performance, non-linear SVM is used. The function of non-linear SVM is to map the feature vector, x by a non-linear mapping, $\phi(x)$ into a high dimensional feature space in which the optimal hyperplane is found [23]. The non-linear mapping can be perform into feature

space by using the kernel function, which computes the inner product of vectors $\phi(x_i)$ and $\phi(x_j)$. The kernel function can be explained as

$$k(x_i.x_j) = \{\phi(x_i).\phi(x_j)\} \quad (11)$$

The most commonly used kernel functions are the Radial Basis Function (RBF)

$$k(x_i.x_j) = \exp\left\{-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right\} \quad (12)$$

where σ is the parameter controlling the width of the kernel and the Polynomial Function

$$k(x_i.x_j) = (x_i.x_j + 1)^p \quad (13)$$

where the parameter, p is the polynomial order. In this study, the RBF kernel function is used.

Table 2 The summarized steps in BPNN algorithm

Step 1	Obtain a set of training patterns.
Step 2	Set up ANN model that consist of number of input neurons, hidden neurons and output neurons
Step 3	Set learning rate (h) and momentum rate (a)
Step 4	Initialize all connections (W_{ij} and W_{jk}) and bias weights (q_k and q_i) to random values.
Step 5	Set the minimum error, $E_{\min}$.
Step 6	Start training by applying input patterns one at a time and propagate through the layers then calculate total error.
Step 7	Back propagate error through output and hidden layer and adapt weights, W_{jk} and q_k .
Step 8	Back propagate error through hidden and input layer and adapt weights, W_{ij} and q_i .
Step 9	Check if Error < $E_{\min}$. If not, repeat steps 6-9. If yes, stop training

■3.0 EXPERIMENTAL DATA

The performance of the proposed method was tested and evaluated using four different types of cancers datasets which are breast cancer, liver cancer, prostate cancer and ovarian cancer. These dataset contains the samples of the benign and malignant tumours. The aim of this classification is to classify the benign and malignant tumours correctly using the ANN and SVM classifiers. Breast cancer and liver cancer dataset are obtained from the UCI Machine Library Database while ovarian and prostate cancers are obtained from the National Cancer Institute (NCI). The summary for all the datasets are shown in Table 3.

The breast cancer dataset which is Wisconsin Breast Cancer Database (WBCD) is given by W.Nick Street (1995) from University of Wisconsin. The dataset consist of 683 samples excluded missing values. These samples were divided into two classes: 444 benign tumours and 239 malignant tumours. There are 9 features in the dataset. For liver cancer, the BUPA Liver Disorders dataset which is created by BUPA Medical Research Limited is used. This dataset is given by Richard S.Forsyth in 1990. The total of data is 345 which 200 samples are benign tumours and 145 samples are malignant tumours. Each of the data has 6 features. The breast cancer and liver cancer dataset represent as standard data.

The prostate cancer dataset namely, JNCI Data (7-3-02) consists of 322 serum spectra composed of peak amplitude measurements at 15154 points stated by corresponding values in the range 0-20000 Da. There are 253 benign and 69 malignant samples in the dataset. The ovarian cancer dataset is labelled as "Ovarian 8-7-02", and consists of 253 dataset. An upgraded PBSII SELDI-TOF mass spectrometer was employed to generate the spectra, which include 91 benign samples and 162 malignant samples. Each spectrum includes peak amplitude measurements at 15154 points defined by corresponding m/z values in the range 0-20000 Da.

From the Table 3, we can see that each dataset is different in terms of number of features. For example, the breast cancer and liver cancer dataset have less number of features while prostate and ovarian cancer dataset have bigger number of features. The features of breast cancer is based on physical appearance of the tumours such as clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion and single epithelial cell size for breast cancer dataset. For liver cancer, the features that influence the tumours whether it is benign or malignant tumours is based on blood tests and number of half-pint equivalents of alcoholic beverages drunk per day. For prostate and ovarian cancer dataset, both of the datasets are in form of gene expression data which defined as the flow of genetic information from gene to protein. A data or commonly called as a mass spectrum in the context of gene expression data contain thousands of different mass/charge ratios. For both of the dataset, each data contain 15154 values of m/z in the range of 0-20000 Da. These values are then called as features in this study.

Table 3 The Summary of cancer datasets

Type of Cancer	Name of Dataset	Number of Samples	Number of Features	Benign Tumours	Malignant Tumours
Breast	Wisconsin Breast Cancer Dataset (WBCD)	683	9	444	239
Liver	BUPA Liver Disorders	345	6	200	145
Prostate	JNCI Data (7-3-02)	322	15154	253	69
Ovarian	Ovarian 8-7-02	253	15154	91	162

■4.0 METHODOLOGY

4.1 Development of SVM and ANN Classification Model

The development of ANN and SVM classification models consist of four steps which are input variable selection, data preprocessing and partitioning, setting of model parameter and model implementation. The difference between ANN and SVM classification models are lies in the setting of model parameter and model implementation. Figure 2 shows the summary of the steps involves in developing the classification model using ANN and SVM classifier. To facilitate the performance of the classifiers, Matlab R2012a Neural Network Toolbox is used to develop ANN classification model and LIBSVM package introduced by [24] is implemented in Matlab R2012a to develop the SVM classification model.

The first steps in developing the classification models are input variable selection. The network input variables are vary for each type of datasets. The second step is data preprocessing and

partitioning. Data is usually pre-processed before it can be used for training to accelerate convergence. Hence, data are normalized using a linear transformation. The actual data is transformed in the range of 0 to 1 using equation (20):

$$X_n = \frac{X_0 - X_{\min}}{X_{\max} - X_{\min}} \quad (14)$$

where X_n is the new value of X, X_0 is the initial value of X, $X_{\min}$ is the minimum value of X in the sample data and $X_{\max}$ is the maximum value of X in the sample data. Besides the data normalization, data preprocessing also involves the process of data conversion which requires that each data instance is represented as a vector of real numbers. Thus, data have to be converted into numeric data if they are categorical attributes. In the case of data conversion, [25] recommend using m numbers to represent a m-category attribute. For in tumours classification, it is usually classified to be whether benign or malignant, so it should be represented as (0,1) before it can be supplied into the classifiers.

Then the data are divided into two partitions which are training and testing set. There is no specific rule to determine the data division of training dataset and testing dataset [3]. In most cases, the researchers used different combinations of data division and it varies according to the problems. In this study, the datasets are split into training-test partitions namely, 70-30% respectively. The training set contains 70% of data from each tumour which are benign and malignant tumours while another 30% of the data used for testing set. For example WBCD dataset, 70% (311) of benign tumours data and 70% (133) of malignant tumours data are grouped as testing set. The other 30% of each tumours data in WBCD dataset are used for testing set. The division of data for all datasets are summarized in Table 4.

The third step is setting the model parameter and it is very important. The proper model parameters setting can improve the ANN and SVM classification accuracy performance. There are three types of parameters that should be considered for training the Back Propagation Neural Network (BPNN) that is network architecture, transfer function and learning parameters. The network architecture of the BPNN model consists of three layers; input, hidden and output. The number of hidden nodes in the hidden layer is different for each dataset; usually it depends on the number of input nodes used. The number of hidden nodes applied are important because it could effect the results of the experiments. Tangent sigmoid has been used in input and hidden layers as the transfer function. The scale-conjugate gradient (SCG) back propagation neural network was selected as learning parameter in this study. In SCG, the value of weight update is calculated as follows:

$$w(i+1) = w(i) + \eta(i)\rho(i) \quad (15)$$

$$\rho(i) = \frac{\partial \mathrm{E}(i)}{\partial w(i)} \quad (16)$$

Where i is the iteration count, η is the learning rate, and $\rho(i)$ is the step direction taken in the $i-th$ iteration step. For SVM classification model, there are two parameters that should be considered in RBF kernel function, namely regularization parameter, C and gamma parameter, γ. C determines the trade-off cost between minimizing the training error and the complexity of the model, while γ defines the non-linear mapping from the input space to some high dimensional space [21]. For this study, a parameter search is conducted in order to identify the best values of parameters (C,γ) using trial and error approach.

The last steps in developing the classification models are model implementation. For ANN classification model, the best classification model is chosen based on the smallest value of Mean Square Error (MSE) obtained during the training phase and used to classify the testing dataset while in SVM classification model, the SVM model is trained until the best pairs of parameters (C,γ) are obtained. This process involved cross validation techniques. In this study, 3-fold cross validation is used. Meaning that, for each of 3 subsets acts as an independent holdout test set for the model trained with the rest of 2 subsets. The advantage of k-fold cross validation are the impact of data dependency is minimized and the reliability of result can be improved. The best parameter pairs (C,γ) are used to create the classification model. The selected classification model is then tested on the testing dataset.

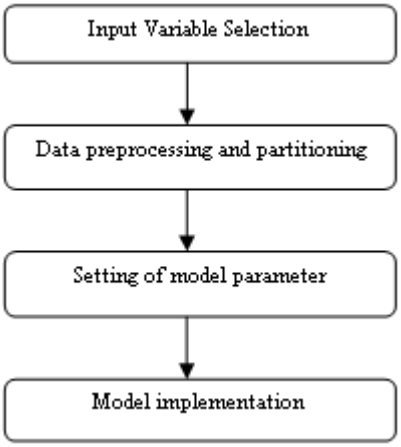


Figure 2 Steps involve in developing the classification model using ANN and SVM classifier

Table 4 Division of datasets

Name of Dataset	Training Set (70%)	Testing Set (30%)
Wisconsin Breast Cancer Dataset (WBCD)	478	205
BUPA Liver Disorders	242	103
JNCI Data (7-3-02)	225	97
Ovarian 8-7-02	177	76

4.2 Performance Measure

The performance of the classifiers was evaluated by the percentage of accurately assigned new samples of cancer data to its correct class such as benign and malignant. There are several measuring tools to evaluate the performance of the classifiers that have been proposed. They are sensitivity, specificity, accuracy and area under receiving operating characteristic curve (AUC). Each of them are used to measure different aspects of performance for example, or in other word, the performance of ANN is quantified based on how accurate the ANN could classify the benign and malignant tumours correctly on the dataset that never been used in training.

Sensitivity which is also defined as True Positive Rate (TPR) is the percentage of benign tumours data classified as benign by the classifier. The classifier that can correctly classify benign tumours will have a higher result in sensitivity. Sensitivity is defined as follows [5,7]:

$$\text{Sensitivity (\%)} = \text{TPR} = \frac{TP}{(FN+TP)} \times 100 \quad (17)$$

Specificity is the percentage of malignant tumours data classified as malignant by the classifiers. The classifier that can correctly classify malignant tumours will have a better result in specificity. Specificity was also known as the True Negative Rate (TNR) and was calculated as follows [6,16]:

$$\text{Specificity (\%)} = \text{TNR} = \frac{TN}{(TN + FP)} \times 100 \quad (18)$$

Accuracy evaluates the performance of the classifier that can correctly classify both types of tumours. The higher value of accuracy indicates better performance of the classifier. It is given by [9,11,13,14]:

$$\text{Accuracy (\%)} = \frac{(TP + TN)}{(TP + FN + TN + FP)} \times 100 \quad (19)$$

AUC represents a common measure of sensitivity and specificity over all possible thresholds. The AUC value of 100% represents perfect discrimination (the classifier can classify the tumours correctly), whereas an AUC value of 50% is equivalent to random model. AUC was calculated as follows [5]:

$$\text{AUC (\%)} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \times 100 \quad (20)$$

Even though there are several measuring tools used to evaluate the performance of the classifiers but the best classifiers is chosen based on the classification accuracy [5,6,7].

5.0 RESULTS AND DISCUSSION

In this study, 8 different classification models have been built by using two different classifiers (i.e., ANN and SVM). The best classifiers for each dataset are determined based on four measuring tools such as accuracy, sensitivity, specificity and AUC. The result obtains is summarized in Table 5. Based on the result obtained, the best classification techniques vary for each dataset.

Table 5 Summary of the results obtain on four datasets

	SVM				ANN			
	accuracy	**sensitivity**	**specificity**	**AUC**	**accuracy**	**sensitivity**	**specificity**	**AUC**
WBCD	99.51%	99.25%	100.00%	99.63%	98.54%	99.25%	97.22%	98.24%
BUPA	63.11%	36.67%	100.00%	68.34%	57.28%	75.00%	32.56%	53.78%
JNCI Data	78.35%	100.00%	0.00%	50.00%	82.47%	100.00%	19.05%	59.52%
Ovarian	64.47%	0.00%	100.00%	50.00%	78.95%	40.74%	100.00%	70.37%

In terms of accuracy (Figure 3), SVM classifier outperforms ANN classifier for WBCD and BUPA dataset. On the other hand, ANN classifier obtains higher performance in accuracy for JNCI (7-3-02) and Ovarian 8-7-02 dataset. This indicates that, SVM has the highest capability in classifying dataset with a smaller number of input features while ANN has better performance of accuracy in classifying dataset with larger number of input features.

Figure 4 and Figure 5 showed the performance of ANN and SVM classifiers in terms of sensitivity and specificity. Based on Figure 4, ANN classifier has better result for BUPA and Ovarian 8-7-02 dataset. Both of the classifier obtains the same result in sensitivity for WBCD and JNCI (7-3-02). For specificity (Figure 5), SVM classifier outperforms ANN classifier for WBCD and BUPA dataset while ANN classifier gives better results in specificity for JNCI (7-3-02) dataset compare to SVM. Both of the classifiers obtains similar results for Ovarian 8-7-02 dataset.

From this result, it can be seen that SVM classifier is better than ANN in classifying the dataset that represent malignant tumours (specificity). This is because the SVM classifier has obtained 100% in specificity performance for three dataset which is WBCD, BUPA and Ovarian 8-7-02 dataset. Even though, ANN classifier has higher performance of sensitivity in two dataset which are BUPA and Ovarian 8-7-02 dataset compared to SVM classifier, but both of the classifier obtains similar results in the other two dataset. Thus, it can be said that the SVM classifier can also obtain good result in sensitivity or in other word; SVM can correctly classify the dataset which belongs to benign tumours.

For JNCI (7-3-02) and Ovarian 8-7-02 dataset, which have bigger number of input features, the imbalanced distribution of data for benign and malignant tumours has a big effect on the performance of the SVM classifier in sensitivity and specificity. This can be proven in Figure 4 and Figure 5, where the percentage of sensitivity (correct classification of benign tumours) for Ovarian 8-7-02 dataset and the percentage of specificity (correct classification of malignant tumours) for JNCI (7-3-02) dataset obtained by SVM classifier is zero. However, for the datasets which have less number of input features such as WBCD and BUPA dataset, the imbalanced distribution of data for both tumours did not affect the performance of the SVM classifier in sensitivity and specificity. Unlike SVM classifier, the number of input features and the imbalanced distribution of data for both tumours affected the performance of ANN classifiers in sensitivity and specificity. The sensitivity and specificity value obtain by ANN classifier are higher in the class of tumours which contain more data.

Lastly, as it can be seen in Figure 6, ANN obtains higher performance in terms of AUC value for JNCI (7-3-02) and Ovarian 8-7-02 dataset while SVM classifier outperforms ANN classifiers in terms of AUC for the other two datasets. Similar to accuracy, SVM has higher value of AUC for classifying dataset with a smaller number of input features while ANN has better performance of AUC in classifying dataset with bigger number of input features.

As the conclusions, it can be seen that both of the AI classification techniques can do well in classifying cancer dataset. Both of the classifiers obtained good performances in accuracy based on the datasets used. ANN classifier can obtain good classification performance in the dataset with bigger amount of input features (prostate and ovarian cancer dataset) while SVM classifier can have better performance in the dataset with smaller amount of input features (breast cancer and liver cancer dataset). Although both of the classifiers have good result in accuracy and AUC but the SVM classifier is better in classifying data which belongs to each tumours (sensitivity and specificity).

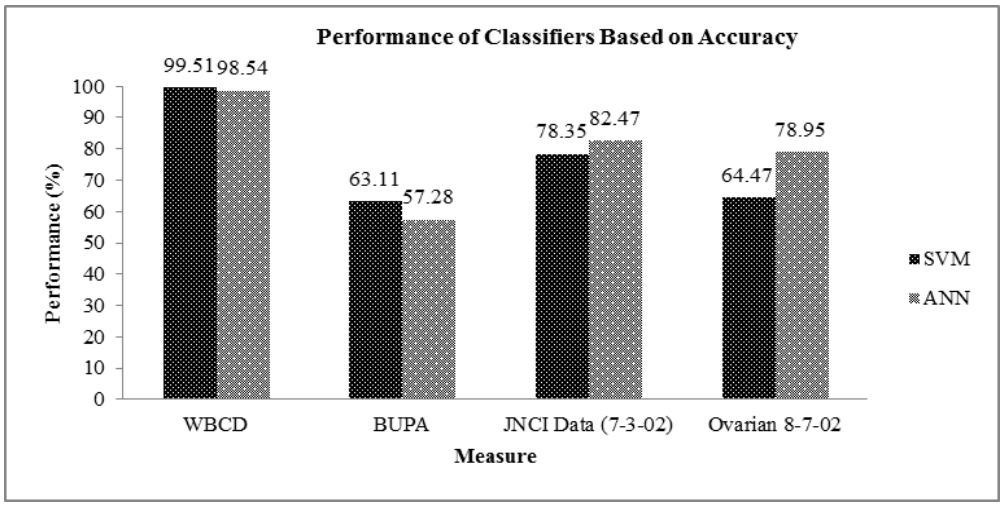


Figure 3 Performance of classifiers based on accuracy

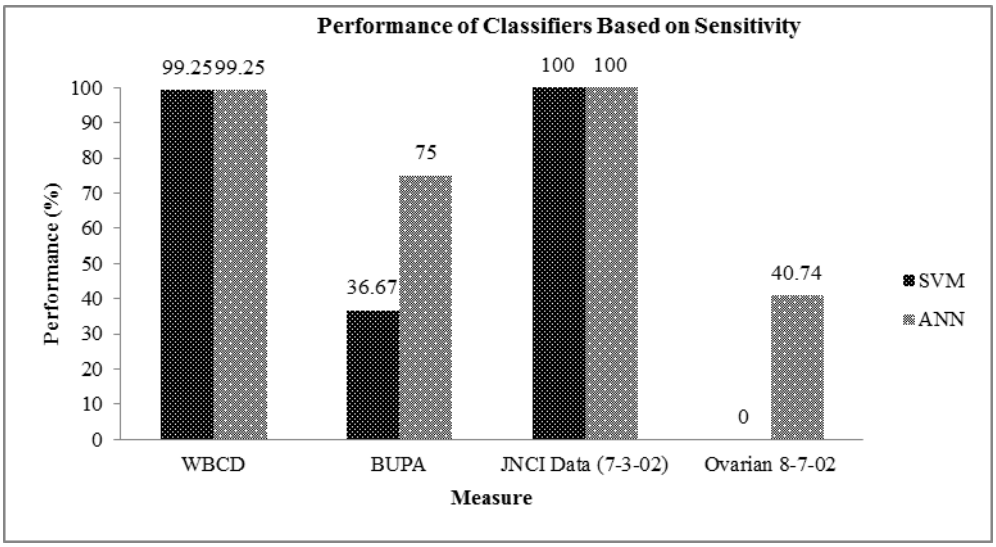


Figure 4 Performance of classifiers based on sensitivity

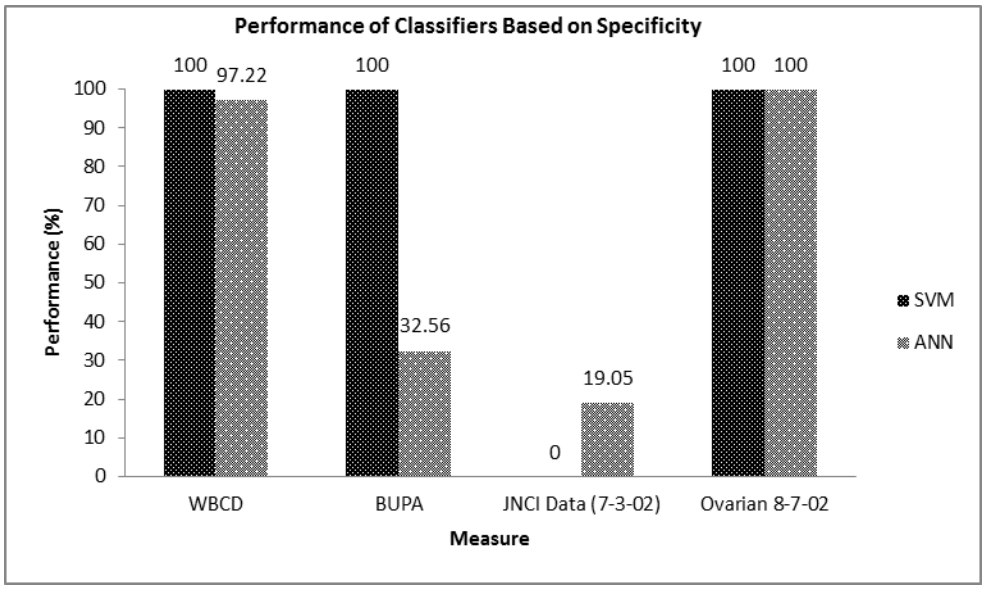


Figure 5 Performance of classifiers based on specificity

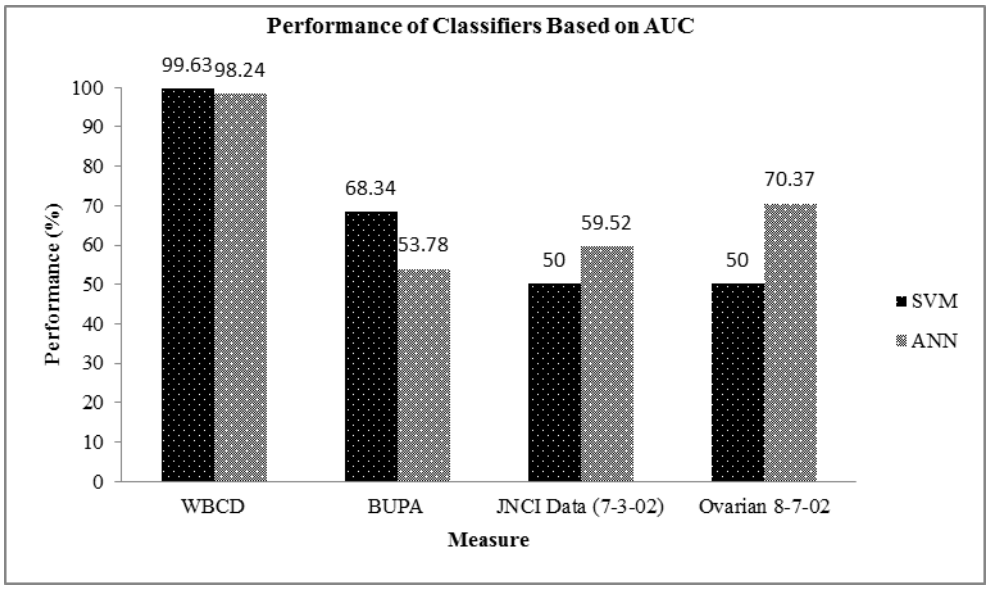


Figure 6 Performance of classifiers based on AUC

6.0 CONCLUSIONS

Accurate cancer classification is important in order to save human's life. Despite using common diagnosis tools, most of the researchers nowadays are interested in using AI classification techniques to classify cancer. This study is conducted in order to compare the performance of two AI classification techniques which are SVM and ANN in classifying cancer data. Both of the techniques can be effective tools in order to classify cancer data. In the future study, different training rules can be used for training ANN while SVM classifier can also be train by using different kernel functions in order to improve the performance of the classifiers.

Acknowledgement

This study is supported by a Fundamental Research Grants Scheme (vot : 4F086) that sponsored by Ministry of Higher Education (MOHE) and GUP grant (vot:Q.J130000.2628.08J02). Authors would like to thank Research Management Centre (RMC) Universiti Teknologi Malaysia, for the research activities and Soft Computing Research Group (SCRG) for the support and motivation in making this study a success.

References

[1] Sattlecker M. 2011. Optimisation of Machine Learning Methods for Cancer Diagnostics using Vibrational Spectroscopy. PhD Thesis. Cranfield University, United Kingdom.
[2] Chu F., W. Xie, and L. Wang. 2004. Gene Selection and Cancer Classification using A Fuzzy Neural Network. *IEEE*.
[3] Polat K., S. Sahan, H. Kodaz, and S. Gunes. 2007. Breast Cancer and Liver Disorders Classification using Artificial Immune Recognition System (AIRS) with Performance Evaluation by Fuzzy Resource Allocation mechanism. *Expert System with Applications*. 32: 172–183.
[4] Saravanan V., and R. Mallika. 2009. An Effective Classification Model For Cancer Diagnosis using Micro Array Gene Expression Data. *38th International Conference on Computer Engineering & Technology*. 1: 137–141.
[5] Keyvanfard F., M. A. Shoorehdeli, and M. Teshnehlab. 2011. Feature Selection and Classification of Breast Cancer on Dynamic Magnetic Resonance Imaging using ANN and SVM. *American Journal of Biomedical Engineering*. 1: 20–25.
[6] Ren J. 2012. ANN vs. SVM: Which One Performs Better in Classification of MCCs in Mammogram Imaging. *Knowledge-Based Systems*. 26: 144–153.
[7] Subashini T. S., V. Ramalingam, and S. Palanivel. 2009. Breast Mass Classification Based on Cytological Patterns using RBFNN and SVM. *Expert Systems with Applications*. 36: 5284–5290.
[8] Pan S. M., and C. H. Lin. 2010. Fractal Features Classification for Liver Biopsy Images Using Neural Network-Based Classifier, *International Symposium on Computer, Communication, Control and Automation*. 2: 227–230.
[9] Azmi M. S., and Z. C. Cob. 2010. Breast Cancer Prediction Based on Backpropagation Algorithm. *Student Conference on Research and Development (SCOReD)*. 164–168.
[10] Wu Y., N. Wang, H. Zhang, L. Qin, Z. Yan, and Y. Wu. 2010. Application of Artificial Neural Networks in the Diagnosis of Lung Cancer by Computed Tomography. *Sixth International Conference on Natural Computation (ICNC)*. 1: 147–153.
[11] Chu F., and L. Wang. 2006. Applying RBF Neural Networks to Cancer Classification Based on Gene Expressions, International Joint Conference on Neural Network. 1930–1934.
[12] Parfait S., P. M. Walkera, G. Créhangea, X. Tizond, and J. Mitérana. 2011. Classification of Prostate Magnetic Resonance Spectra using Support Vector Machine. *Biomedical Signal Processing and Control*. 7: 499–508.
[13] Liao R., T. Wan, and Z. Qin. 2011. Classification of Benign and Malignant Breast Tumors in Ultrasound Images Based on Multiple Sonographic and Textural Features. *Third International Conference on Intelligent Human-Machine Systems and Cybernetic*. 1: 71–74.
[14] Chen A. H., and C. H. Lin. 2011. A Novel Support Vector Sampling Technique to Improve Classification Accuracy and To Identify Key Genes of Leukemia and Prostate Cancers. *Expert Systems with Applications*. 38: 3209–3219.
[15] Polat K., and S. Güneş. 2007. Breast Cancer Diagnosis using Least Square Support Vector Machine. *Digital Signal Processing*. 17(4): 694–701.
[16] Murat C., E. Mehmet, Z. B. Erkan, and Y. A. Ziya. 2009. Early Prostate Cancer Diagnosis by using Artificial Neural Networks and Support Vector Machines. *Expert Systems with Applications*. 36: 6357–6361.
[17] Mumtaz K. 2009. Evaluation of Three Neural Network Models using Wisconsin Breast Cancer Database. *International Conference on Control, Automation, Communication and Energy Conservation*. 1–7.
[18] Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.
[19] Vapnik, V. 1998. *Statistical Learning Theory*. New York: Wiley.

[20] Liu Y., and Y. F. Zheng. 2004. FS_SFS: A Novel Feature Selection Method for Support Vector Machines. *IEEE International Conference on Acoustic, Speech, and Signal Processing*. 5: 797–800.
[21] Chen H. L. 2011. A Support Vector Machine Classifier with Rough Set-Based Feature Selection, *Expert System Appl*. 38(7): 9014–9022.
[22] Bertsekas D. P. 1995. *Nonlinear Programming*. Belmont: Athena Scientific.
[23]] Akay M. F. 2009. Support Vector Machines Combined with Feature Selection for Breast Cancer Diagnosis, *Expert System Appl*. 36: 3240–3247.
[24] Chang C. C., and C. J. Lin C. J. *LIBSVM: A Library for Support Vector Machine*. S oftware available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[25] Hsu H. H., C. W. Hsieh, and M-D. Lu. 2011. Hybrid Feature Selection By Combining Filters and Wrappers. *Expert System Appl*. 38(7): 8144–8150.