Original Research

Machine Learning Study of Several Classifiers Trained With Texture Analysis Features to Differentiate Benign from Malignant Soft-Tissue Tumors in T1-MRI Images

Jaber Juntu, MSc,* Jan Sijbers, PhD, Steve De Backer, PhD, Jeny Rajan, MTech, and Dirk Van Dyck, PhD

Purpose: To study, from a machine learning perspective, the performance of several machine learning classifiers that use texture analysis features extracted from soft-tissue tumors in nonenhanced T1-MRI images to discriminate between malignant and benign tumors.

Materials and Methods: Texture analysis features were extracted from the tumor regions from T1-MRI images of clinically proven cases of 49 malignant and 86 benign soft-tissue tumors. Three conventional machine learning classifiers were trained and tested. The best classifier was compared to the radiologists by means of the McNemar's statistical test.

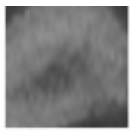
Results: The SVM classifier performs better than the neural network and the C4.5 decision tree based on the analysis of their receiver operating curves (ROC) and cost curves. The classification accuracy of the SVM, which was 93% (91% specificity; 94% sensitivity), was better than the radiologist classification accuracy of 90% (92% specificity; 81% sensitivity).

Conclusion: Machine learning classifiers trained with texture analysis features are potentially valuable for detecting malignant tumors in T1-MRI images. Analysis of the learning curves of the classifiers showed that a training data size smaller than 100 T1-MRI images is sufficient to train a machine learning classifier that performs as well as expert radiologists.

Key Words: texture analysis; soft-tissue tumors; T1-MRI tumors classification; machine learning; ROC curves analysis; classifiers comparison
J. Magn. Reson. Imaging 2010;31:680–689.

MAGNETIC RESONANCE IMAGING (MRI) is currently regarded as the standard diagnostic tool for detection and grading of soft-tissue tumors (1). Radiologists often look for certain features in MRI images to decide whether a given tumor is benign or malignant. For example, radiologists differentiate tumors based on features that describe the biological activity of the tumor such as the tumor size, the shape of the boundaries of the tumor, the presence of necrosis, edema, and invasion of the surrounding tissue. Some important characteristic features of a benign tumor include small size, well-defined margins, and homogeneous T1-MRI signal intensity. Malignant tumors typically have larger size, are poorly marginated, heterogeneous in signal intensity in T1-MRI, may exhibit peritumoral edema, and invasion or encasement of adjacent structures (2). Among the mentioned features, the MRI signal homogeneity in T1-MRI is the most correlated feature with tumor pathology. It has been reported that if the tumor region in T1-MRI is inhomogeneous, then it is 90% of the time a malignant tumor (1). Malignant tumors show inhomogeneous (heterogeneous) signal in T1-MRI signal because they have an increased vascularity and have large extracellular spaces compared to benign tumors. However, there are some exceptions where some types of benign tumors show an inhomogeneous T1-MRI signal while some types of malignant tumors show homogeneous T1-MRI signal. Examples of benign tumors that are likely to show inhomogeneous T1-MRI intensities are: hemangioma (cavernous, spindle cell), giant cell tumor (tenosynovial diffuse), and schwannoma. Examples of malignant tumors that may show homogeneous MRI intensities are: leiomyosarcoma, fibrosarcoma (adult type myxoid-myxofibrosarcoma), and synovial sarcoma. Due to the signal intensity overlap, the discrimination between benign and malignant tumors based on visual assessment of signal homogeneity or heterogeneity cannot always be conclusive. As human visual abilities to differentiate within a wide range of texture is very limited (3,4), analyzing tumors by automated texture analysis algorithms might have

Vision Laboratory (VisieLab), Department of Physics, University of Antwerp, Belgium.
Contract grant sponsor: IWT SBO Project Quantiviam.
*Address reprint requests to: J.J., Vision Laboratory (VisieLab), Department of Physics, University of Antwerp (CDE) Universiteitsplein 1 (N Building) B-2610, Antwerp, Belgium.
E-mail: Jaber.Juntu@ua.ac.be
Received January 30, 2008; Accepted December 29, 2009.
DOI 10.1002/jmri.22095
Published online in Wiley InterScience (www.interscience.wiley.com).



a ROI from a bengin tumor

b ROI from a malignant tumor

Features	Skewness	Perc.1%	Per.10%	Per.50%	s(0,2) SumVarnc	s(0,4) Contrast	Horzl_fraction	dgr_LngREmph45	GrSkewness	Teta2	Sigma	WavEnHH_s1
Benign ROI	0.26	106	112	119	9.62	2.94	0.49	4.38	0.90	0.71	0.18	0.10
Malignant ROI	0.21	39	45	51	5.78	2.11	0.58	3.41	0.73	0.54	0.25	0.16

Figure 1. Two regions cut from soft-tissue tumors; one is cut from a benign tumor and the other from a malignant tumor. It is very difficult to distinguish between the tumors visually. The two rows at the bottom are a subset of the extracted texture analysis features that clearly show the difference quantitatively.

the potential to discriminate between some cases of malignancy that humans cannot recognize easily by visual inspection.

The signal homogeneity or heterogeneity of the tumor areas in T1-MRI images can be quantified by texture analysis algorithms (5–8). Applying texture analysis algorithms to the tumor areas results in numerical values that quantify the degree of homogeneity or heterogeneity of the tumor. Figure 1 shows an example of the value of some texture features extracted from malignant and benign soft-tissue tumors. The figure clearly shows that benign and malignant tumors have very distinctive numerical values.

Characterizing texture in MRI images is still assessed visually by radiologists. Such subjective methods can produce a high risk of misinterpretation and can also contribute to inter- and intraobserver variation in making the correct classification. A further limitation is associated with the visual coarse categorization of texture variation into two categories (homogeneous versus heterogeneous) that makes it difficult to distinguish small differences in texture changes. The ability to measure small differences in texture is particularly important to reduce the diagnosis errors caused by the overlap of the textures between the benign and the malignant tumors.

Even though there are several biological features that can be used to differentiate benign from malignant tumors, from an image analysis perspective texture features are preferred over other tumor features such as the mean MRI signal intensity or the shape of the boundaries of the tumor for several reasons. First, texture features are very easy to calculate in a reasonable amount of time. Second, texture features are largely correlated to tumor pathology (9,10). Third, some texture features are very robust to changes in MRI acquisition settings, invariant to changes in image resolution, and unaffected by the corruption of the MRI image by magnetic field inhomogeneity. Two recent research articles (8,11) extensively studied the value of texture analysis features for classification of soft-tissue tumors. In Ref. 8 it is shown that texture analysis features extracted from T1-weighted MRI images acquired by different MRI machines are statistically very similar for the same tissue type. The work reported in Ref. 11 extended the previous study and showed that texture analysis features are useful for discrimination between benign and malignant soft-tissue masses in MRI images.

The main objective of the current study was to extract a large number of texture features from tumor areas and to comprehensively evaluate several machine learning classifiers that are trained with these texture analysis features to discriminate malignant from benign soft-tissue tumors. From the tumor regions in the T1-MRI images with pathologically proven soft-tissue tumors, 300 texture features were extracted. The texture features were reduced from 300 to 13 texture features by an optimal feature selection procedure. Three classifiers were trained with the selected 13 texture features. Several comparative studies were performed to find the best classifier that fits the problem domain well.

For estimating the classification errors and comparing the classifier models several sources of variation were taken into account (12,13). The first source of variation is the influence of using a specific limited training dataset on the classifiers' performance. We used the learning curves plots to study how the classifiers performance might be affected by changing the size of the training data. From the learning curves we estimated the minimum size of the training dataset needed to train the classifiers with an optimal performance.

The second source of variation is the internal randomness of the training algorithms. Classifier performances are largely affected by the training set characteristics as well as the parameter setting of the classifiers. Therefore, it is dangerous to draw conclusions from training a single classifier or running a single testing experiment. A single training and testing experiment is not a reliable estimator of the true error rate of a classification scheme. A random

Table 1
Patients Database*

Type of tumor	Benign tumors ($n = 86$)	Malignant tumors ($n = 49$)
Fibrous tissue tumors	Fibroma (3)	Fibrosarcoma, myxoid (8)
	Fibromatosis (10)	Fibrosarcoma infantile (2)
Lipomatous tumors	Lipoma deep (intramuscular, Perineural) (16)	Dedifferentiated liposarcoma (7)
	Lipoma cutaneous (3)	Myxoid liposarcoma (5)
	Lipomabalstoma (2)	Liposarcoma (4)
Cartilage and bone tumors	Myositis ossificans (3)	Extraskeletal chondrosarcoma (well differentiated) (5)
Neural tumors	Degenerated schwannoma (4)	Malignant peripheral nerve sheath tumor (MPNST) (4)
	Plexiform schwannoma (3)	
	Usual schwannoma (neurilemoma) (12)	
Endothelial tumors of blood and lymph vessels	Cavernous hemangioma (14)	Angiosarcoma (1)
Smooth muscle tumors	Leiomyoma (3)	Leiomyosarcoma (8)
	Angiomyoma (2)	
Miscellaneous tumors	Granular cell tumor (5)	Synovial sarcoma (4)
	Myxoma (intramuscular) (6)	Ewing's sarcoma (1)

The second and third columns show the pathology of the tumors (number of patients).

subsampling of the training dataset by a cross-validation procedure should be used to minimize the classification error estimation bias. We trained the classifiers several times by randomly sampling small subsets from the full dataset using the 10-fold cross-validation procedure. For consistency, exactly the same data were used to train and test all classifiers; this is often called a paired experimental design. The results of testing the classifiers were validated by applying several statistical tests such as computing the confidence intervals of the receiver operating curves (ROC).

Finally, the way the classifiers are tested and evaluated could bias the final results if not done properly. Different methods were applied to test and compare the classifiers' performance. For example, we plotted the ROC curves and cost curves of the classifiers. We computed the confidence intervals of the area under the curves (AUC) and applied a pairwise statistical test to compare the classifiers. We applied the nonparametric McNemar's statistical test to compare the performance of the best trained classifier with the radiologists.

MATERIALS AND METHODS

Patient Dataset and MRI Images

The datasets used in this study were collected from the database of the Belgian Soft Tissue Neoplasm Registry (BSTNR) which was collected by the University Hospital of Antwerpen (UZA) over several years. The BSTNR is a multi-institutional database project involving nearly all MRI centers in Belgium with the cooperation of some European MRI centers. This initiative, which started before the year 2001, had two main goals. First, it provides a second opinion report for diagnosing difficult cases of soft-tissue tumors as a benefit toward all cooperating radiologists. Second, it serves as a scientific databank of soft-tissue tumors that are rare lesions in the daily radiological practice. Currently, the center database contains more than 1500 histologically confirmed cases of soft-tissue tumors. For this study we obtained a collection of non-enhanced T1-weighted MRI images of 135 patients. Among the 135 cases there were 49 cases with malignant tumors and 86 cases with benign tumors. All tumor cases were pathologically confirmed, which we used as the gold standard for training and testing the classifiers. This collection of T1-MR images were acquired using various scanner types (Siemens, Philips, and GE). The pulse sequence for all the MRI images was a T1-weighted imaging sequence, however, with different repetition time (TR) and echo time. On average, the MRI images were acquired using a TR (average = 580 msec) and TE (average = 15 msec). The age range of the patients with benign tumors was 18–73 years old, while for the patients with malignant tumors the age range was 15–85 years old. Some more details about the dataset are listed in Table 1. For constructing Table 1 we adopted the classification system proposed by the World Health Organization (WHO) (14) and followed the guidelines given in Ref (15).

Most texture features are commonly measured in small square or rectangular areas (area of interest [AOI]) from the MRI images. For such purposes we selected non-overlapping areas of sizes 50×50 pixels from the tumor regions in the T1-MRI images. Since training machine learning classifiers require large datasets and given the fact that we had only 135 different cases of soft-tissue tumors, we increased the size of the training dataset by computing texture features from several regions for each patient. However, we made sure that such regions were not selected from the same MRI image but selected from several MRI images. In total, we selected 681 tumor regions (253 benign tumors regions and 428 malignant tumors regions). Increasing the size of the training dataset by adding virtual data samples is a common practice in machine learning (16,17). First, virtual examples prevent classifiers from overfitting the training data, which is equivalent to adding a regularization term to the cost function that has to be minimized (18). Second, adding virtual examples is equivalent to incorporating a sort of prior knowledge

Table 2
Texture Analysis Methods and the Corresponding Texture Parameters

Methods	Calculated parameters
First order:	
Histogram	Mean, minimum, variance, skewness, kurtosis, 1%, 10%, 50%, 90%, 99% percentiles.
Second order:	
Coocurrence matrix	
$\{\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ and $\{d = 1,2,3,4,5\}$	Angular second moment, contrast, sum of squares, inverse difference moment, sum average, correlation, entropy, difference variance, difference entropy.
Absolute gradient distribution	Mean of absolute gradient, variance of absolute gradient, skewness of absolute gradient, kurtosis of absolute gradient.
Higher order:	
Runlength graylevel matrix	Short run emphasis moment, long run emphasis moment, run length nonuniformity, fraction of image in run.
Autoregressive texture model	θ_1, θ_2, θ_3, θ_4, θ_5, σ.
Filtering techniques:	
Wavelets	Energies of wavelet coefficients of sub bands at successive scales.

which makes the trained classifiers more robust to variation in texture features caused by a change of imaging acquisition settings or differences in MRI machine models (16–18).

Texture Features Computation

For feature computation, the 681 tumor subimages with sizes of 50 × 50 pixels were imported to the software package MaZda 3.20 (Institute of Electronics, Technical University of Lodz, Poland) (19). The MaZda software program allows computation of 300 texture features based on statistical, wavelet filtering, and model-based methods. To ensure the consistency of the calculated texture feature across all the tumor subimages, we wrote some macros for the MaZda program that read tumor subimages and calculate the tumors' texture features with the same texture analysis parameters setting. All the 300-texture features offered by the MaZda program were calculated for the 681 tumor subimages. Table 2 shows the texture analysis methods and the calculated texture features used in this study.

Feature Selection

Training machine learning classifiers with large numbers of texture features can lead to classifier overfitting, reduces the generalization capabilities of the classifiers and slows down the training process. Before training the classifiers the number of texture features were reduced by a feature selection procedure to remove the unimportant and uninformative texture features. Since MaZda program has a limited number of feature selection methods, we exported the computed texture features to the machine learning package Weka 3.5.8 (20) to experiment with different feature selection methods. To keep the loss of texture information by the feature selection procedure to a minimum we tested eight feature selection methods that belong to three feature selection families as follows:

1) Subset feature search selection methods: Such methods literally search the set of possible features for the optimal subset. We used four different techniques, namely: the forward search, the backward search, the bidirectional search, and the greedy search.
2) Feature ranking methods: Rank the texture features by a numeric value and eliminate all features that do not achieve an adequate score. We used two ranking methods, one that ranks features by the chi-square statistics and the other method ranks features by the information gain criteria.
3) Embedded methods: The feature selection is combined with a classifier to evaluate the discrimination power of a selected subset. We used two embedded methods, one that uses the C4.5 decision trees classifier and one based on the Vapnik's SVM classifier.

In total, eight feature selection methods were tested to select eight optimal features subsets out of the full 300 texture features. We used simple Bayes classifier to evaluate the power of the eight features subsets for discrimination between the benign and the malignant tumors. The results of applying the Bayes classifier trained with the full 300 texture features were considered a baseline for the effectiveness of the selected texture features subsets. Based on the Bayes classifier results, the texture features subset that was selected by the forward search method was the best subset with a minimum loss of information (Table 3).

Trained Classifiers

We trained three classifiers that represent typical implementations of three machine learning algorithms. The first classifier is a multilayer perceptron neural network classifier with two hidden layers, an input layer of 13 nodes that correspond to the selected texture feature subset by the forward search method, and two nodes at the output layer corresponding to the benign and malignant classes. The neural network trained with a backpropagation algorithm with a learning rate = 0.3, momentum = 0.2, and a training time of 1000 epochs. The second

Table 3
Eight Texture Feature Subsets Selected by the Three Feature Selection Methods*

	Best features selection				Feature ranking		Wrappers	
Feature name	Forward search	Backward search	Bidirectional search	Greedy stepwise	Chi-squares statistics	Information gain	C4.5 decision trees	Vapnik's SVM
Skewness	•	•	•	•				
Kurtosis								•
Per 0.01%	•	•	•	•			•	•
Per 0.1%	•	•	•	•	•	•		
Per 0.5%	•	•	•	•				
S(1,0),Contrast	•							•
S(1,0), InvDfMom					•	•		
S(1,0), DifVarnc					•	•	•	
S(1,1), Contrast					•	•		
S(1,1), Correl							•	
S(1,1), DifVarnc					•	•		
S(1,1), DifEntrp					•	•		
S(1,−1), Contrast							•	
S(1,−1), InvDfMom					•	•	•	
S(1,−1), Entropy								
S(1,−1), DifVarnc							•	
S(2,0), SumVarnc	•		•	•				
S(2,−2), Correlat								•
S(3,0), SumVarnc		•						
S(3,0), DiffEntrp							•	
S(4,0), Entropy							•	
S(4,0), Contrast	•		•					
S(0,4), InvDfMom								•
S(0,4), DifEntrp								•
S(0,5), AngScMom								•
S(0,5), Contrast								•
Horzl_RlNonUni					•	•		
Horzl_LngREmph		•		•	•	•	•	
Horzl_ShrtREmp								
Horz_Fraction	•		•		•			
Vertl_RLNNonUni								•
dgr_LngREmph45	•	•	•	•	•			
dgr_Fraction45					•			
dr_LngREmph135					•	•		
GrSkewness	•	•	•	•				
Teta2	•	•	•	•				
Teta4		•						
Sigma	•	•	•	•		•		
WavEnHH_s-1	•	•	•	•		•		
ACC%	76.8	77.7	77.1	78	67.99	65.34	70.77	78
TP	0.8	0.8	0.79	0.83	0.65	0.56	0.7	0.86
TN	0.74	0.74	0.73	0.69	0.73	0.81	0.73	0.64
AUC	0.87	0.85	0.86	0.83	0.72	0.75	0.78	0.84

The shaded columns show the best texture features subset which was used to train the classifiers. The last four rows are the results of evaluation of each texture subset using the Bayes classifier.

classifier was Quinlan's C4.5 decision tree classifier built with a minimum of two instances per leaf. After training, each node in the tree was guaranteed to have at least two objects. The tree was pruned with a 3-fold pruning procedure such that 2-fold were used for growing the tree and one for the pruning process. The confidence factor that controls the pruning process was set to a small value so that there was a balance between growing the tree and the pruning process. Finally, the third classifier was the Vapnik's nonlinear SVM classifier that used the RBF kernel with a large bandwidth ($\sigma = 1000$) and a cost coefficient ($C = 1.0$). The bandwidth and the cost coefficient were selected empirically by a grid search method. We preferred to keep the default parameter settings of the trained classifiers as originally defined by the PRTOOLs 4.0, the machine learning Matlab (MathWorks, Natick, MA) toolbox unless we found that the default parameters were not set properly for certain classifier. In such cases we estimated the classifier parameters using the training data. We sampled small random subsets from the training data, estimated the classifier parameters, and took the average of the estimated parameters. For the purpose of classifiers

comparison we preferred to use the default classifiers parameter setting to avoid tuning the classifiers to that specific tumor training dataset. For example, for the neural network classifier we kept the learning rate, the momentum, and the initial conditions to their default values. Such simple approach might result in lower estimates of the true error rate; however, it minimizes the effect of expert bias that tunes the classifiers to the specific problem domain. Furthermore, the bias introduced by such a parameter selection scheme affects all the learning algorithms equally.

A single training and testing partition are not reliable estimators of the true error rate of a classification method on a limited dataset. A leave-one-out classification method is too computationally expensive to be used. According to previous recommendations (12,21), a 10-fold cross-validation procedure was chosen to train the classifiers. The 10-fold cross-validation method has been found to provide an adequate and accurate estimate of the true error rate. In each iteration the 10-fold cross-validation splits the data into two random parts, 90% training part and 10% testing part. However, it ensures that the proportions of examples from each class remains fixed throughout all iterations.

Sample Size and the Classifier Performance

Classifiers become unstable when trained by small size training dataset (22). The performance of most classifiers can be improved by increasing the size of the training dataset. Learning curves are commonly used to study the generalization properties of the trained classifiers as a function of the number of training data examples.

To plot the learning curves the selected classifiers were trained with training subsets of different sizes that were randomly sampled from the full training data. On a typical learning curve the horizontal axis (x-axis) represents the number of examples used for training, while the vertical axis (y-axis) represents the error rate of the classifiers tested against a set of examples unseen during the training process. We randomly sampled with replacement small training subsets that contained 5, 10, 15, 20, 25, 30, 35…, 200 examples from the full data. The three classifiers, the neural network, the C4.5 decision trees, and the SVM were trained using the training subsets and tested on the rest of the data that were not included in the training subsets. This procedure was repeated 10 times and the averages of the classification errors were calculated and plotted.

ROC and the Cost Curves

Analyzing the ROCs of classifiers is a well-established method in the machine learning literature for evaluating and comparing classifier performances. The ROC encapsulates all information contained in the confusion matrices of several training and testing experiments. With the ROC we can visualize classifier performances over a wide range of the training data distribution. The ROC is a valuable tool to find an optimal operating point or a decision threshold for a given classifier trained by certain training dataset.

Table 4
Table Constructed From the SVM Classifier and the Radiologists Diagnosis for Performing McNemar's Test

$n_{00} = 8$	$n_{01} = 2$	$n_{00} + n_{01} = 10$
$n_{10} = 10$	$n_{11} = 115$	$n_{10} + n_{11} = 125$
$n_{00} + n_{10} = 18$	$n_{01} + n_{11} = 117$	$n = 135$

To summarize the information contained in the ROC curve, the AUC of an ROC curve is recommended as a single number evaluation method of a machine learning algorithm. The AUC is a global summary measure of the classifier performance that is independent of the training dataset probability distribution, independent of the classification error costs and the decision threshold (23). The cost curves, which have been proposed recently (24), complement the ROC curves analysis by displaying the same information contained in the ROC in a different way. However, they are much better for comparison between the classifiers, especially in the case when the ROC curves cross.

To compare the selected classifiers we trained them using the 10-fold cross-validation method and plotted their ROC and cost curves. We performed a pairwise statistical comparison and computed the confidence intervals and *P*-values of the AUC of the ROC. We also plotted the cost curves to find out why the performances of the classifiers overlap in some areas of the ROC curves.

Comparison Between the Radiologists and the Best Classifier

Once the classifiers are trained, evaluated, and compared, the next logical step is to apply some statistical tests such as McNemar's test (12,13) to compare the performance of the best classifier to expert radiologists.

The first step to apply McNemar's test is to construct a table as shown in Table 4. The table cells summarize the number of agreements and disagreements between the trained classifier and the radiologists for classifying benign and malignant tumors. The entries in Table 4 are as follows: The diagonal elements: $n_{00} = 8$ is the number of tumor examples misclassified by both the radiologists and the SVM classifier, $n_{11} = 115$ is the number of tumor examples that are correctly classified by both the radiologists and the SVM classifier. The off-diagonal elements: $n_{01} = 2$ is the number of tumor examples that are correctly classified by the radiologists but incorrectly classified by the SVM classifier, and $n_{10} = 10$ is the number of tumor examples that are correctly classified by the SVM classifier but incorrectly classified by the radiologists. Notice that the off-diagonal elements in Table 4 are the only numbers used in McNemar's test. The second step is to test the null hypothesis: H_0: $n_{01} = n_{10}$ that the classifier and the radiologists have the same error rate against the alternative hypothesis: H_1: $n_{01} \neq n_{10}$. We computed the chi-square value using the equation shown below and tested the

results against the theoretical chi-square with one degree of freedom.

$$\tilde{\chi}^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \quad [1]$$

RESULTS

Feature Selection Analysis

In the texture feature selection step, eight different texture subsets were selected as shown in the Table 3. The first column of the table shows the texture features' names as specified by the MaZda package. Each selected subset is displayed in a single column in Table 3 where the specific selected texture features are indicated by black bullets. Each one of the eight feature selection methods selected slightly different texture subsets. However, feature selection methods that belong to the same family selected very similar texture subsets. The Bayes classifier was used to identify a single optimal texture subset out of the eight subsets. The Bayes classifier results of evaluation of the eight texture subsets are listed in the last bottom rows of Table 3. We listed the accuracy of classification (ACC %), the true positive rate (TP), the true negative rate (TN), and finally the AUC. The result of training the Bayes classifier with the full 300 texture features was: ACC = 73.7%, TP = 0.74, TN = 0.69, and AUC = 0.71, which we used as a baseline to test the effectiveness of the selected texture subset. The texture subset that was selected by the forward search method has the highest AUC value (AUC = 0.78) compared to the AUC values of the other feature selection methods; hence, it was chosen for training the classifiers.

Analysis of the Learning Curves and Some Observations

Figure 2 shows the learning curves of the three trained classifiers: the neural network, the C4.5 decision tree, and the support vector machine. As expected, the error rates of the trained classifiers decrease as the number of training examples increase. The learning curves have a fast decreasing portion early in the curve, followed with a relatively slow decreasing portion, and finally a plateau portion when the learning error rates no longer decrease with more data. A reasonable small classification error rate is achievable across the three classifiers with a training sample size of around 60 training samples. As the size of training data increases over 60 samples, the performance of the classifiers improve very slowly. This observation points to the fact that training the classifiers with a very large training dataset will not improve the classifier performances very significantly. The training curves do not cross in the plateau area and are relatively smooth, which indicates that the classifiers were nearly stable and robust to random variation of the training data samples. The average error rates of the classifiers after training with more than 60 data samples are: 0.14 for the support vector machine, 0.17 for the C4.5 decision tree, and 0.22 for the backpropagation neural network. The C4.5 decision tree and the support vector machine are relatively stable after training with a training set of size of around 60 samples. However, the neural network classifier became stable after training with around 100 samples. That might be explained by the fact that the neural network classifier is very flexible and can easily overfit the training data when the size of the data is small. Neural networks do not have a built-in mechanism to control overfitting the training data. However, there are some heuristics that can be followed to reduce the overfitting process such as reducing the training time to keep the node firing in the linear regions of the sigmoid functions. The support vector machine classifier is similarly very flexible, like the neural network; however, it became stable very early during the training process since it has a built-in mechanism to control overfitting the training data.

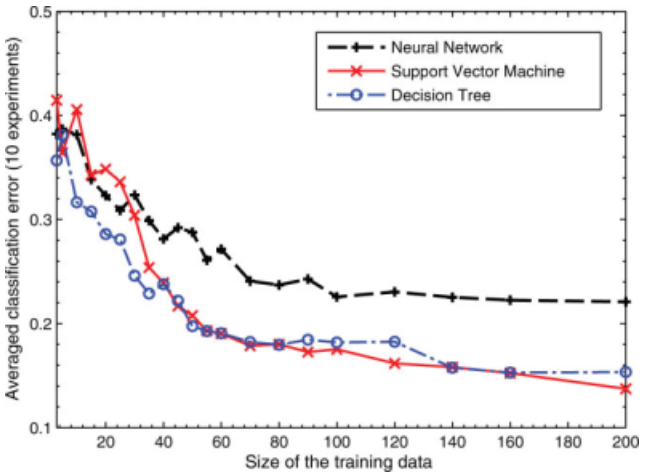


Figure 2. The learning curves were produced by training the three classifiers with different training data sizes. Each point in each learning curve is the averaged classification error of 10 random cross-validation experiments. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Analysis of the ROC and the Cost Curves

Figure 3 shows the ROC of the support vector machine, the neural network, and the decision trees classifier. In general, the three classifiers perform much better than the trivial random classifier since the three curves lie above the diagonal line. The best classifier is the support vector machines since its ROC curve is located at the top of Fig. 3. The AUC, as a general summary measure of the classifiers performance, was AUC = 0.91 for the support vector machine, AUC = 0.88 for the C4.5 decision tree, and AUC = 0.85 for the backpropagation neural network. The optimal operating points of the three classifiers are indicated by the solid circles at (FP = 0.15, TP = 0.91) for the SVM, (FP = 0.20, TP = 0.89) for the decision trees, and (FP = 0.49,TP = 0.93) for the neural network, where the numbers inside the brackets indicate the false positive (FP) and the true positive (TP). These three points are the maximum optimal (FP, NP)

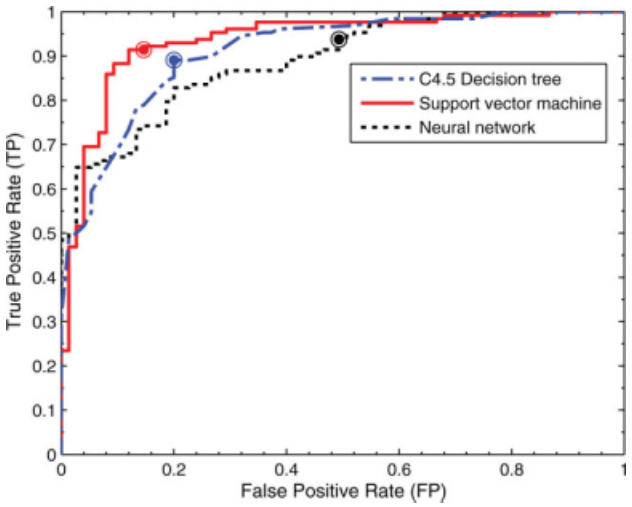


Figure 3. The ROCs of the neural network, the decision tree, and the support vector machine. The curves cross, which indicate that the classifiers behavior change based on the distribution of the malignant and benign cases in the training dataset. The closed circles on the top of the ROCs represent the optimal operating points of the three classifiers based on the current training dataset. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

points when the three classifiers are trained with the full training dataset.

From the ROC we calculated the AUC and some other related statistics and obtained the following: the SVM (AUC = 0.91, SE = 0.0246, CI = 0.827–0.921), the neural networks (AUC = 0.85, SE = 0.0331, CI = 0.707–0.827), and the C4.5 decision tree (AUC = 0.88, SE = 0.0251, CI = 0.821–0.917) where SE is the standard error and CI represents the 95% confidence interval. Apparently the SVM has the highest AUC value, which indicates that it is the best classifier. The standard errors are reasonably small, resulting in relatively narrow confidence intervals. The lower limits of the confidence intervals of the AUC of the three classifiers exceed the 0.5 value (the AUC of a random classifier) and hence the three classifiers perform much better than a random classifier.

To compare the three classifiers, we applied pairwise statistical tests based on the difference in the AUC of the three classifiers. The null hypothesis is that the difference in the AUC between two classifiers is equal to zero. The results of the pairwise comparison of the AUC of the three ROC curves are shown in Table 5. The pairwise comparison between the SVM versus the neural networks and the SVM versus the C4.5 decision trees show that their confidence intervals do not include the 0 value with a < 0.001, which means that in both cases the difference in AUC of the ROC are statistically significant. On the other hand, the pairwise comparison between the neural network versus the C4.5 decision tree shows no significant difference between their AUC values since the confidence interval contains the 0 value with nonsignificant *P*-value = 0.683.

The former analysis of the pairwise comparison test can be supported by plotting the cost curves of the three classifiers. Figure 4 clearly shows that the cost curve of the SVM vector machine is the lower one since it has the lowest normalized expected cost of about 0.1 for a wide range of probability cost function (PCF = 0.20–0.75). In that range the support vector machine is insensitive to the distribution of the benign and malignant tumors in the training dataset or to change of the cost of misclassification. On the other hand, the cost curves of the C4.5 decision tree and the neural network overlap at a probability cost function of about PCF = 0.42. Such overlap indicates that their performances are sensitive to the tumors class distribution in the training dataset and also sensitive to the cost of misclassification. In Fig. 4 the cost curves of the three classifiers are completely located within the triangular area that is defined by the long-dashed lines. All classifiers for which their cost curves completely fall inside this triangular area always perform much better than a random trivial classifier. The short-dotted lines touching each cost curve are the operating lines of the three classifiers. The operating lines are equivalent to the black circles that represent the operating points of the classifiers in the ROCs. As a matter of fact, the cost curves and the ROCs are dual representations of the same information content where every point in the ROC maps into a line above the cost curve and vice versa (see Ref. 23 for more details).

Statistical Comparison Between the SVM Classifier and the Radiologists

For each tumor case we calculated the texture features from one tumor subimage. We trained with a leave-one-out cross-validation procedure the Vapnik's support vector classifier (which was the best classifier). We constructed Table 4 by comparing the radiologist's diagnosis of the tumor images and the classification results of the support vector machine classifier. We obtained the results of the radiologists' diagnosis

Table 5
Pairwise Comparison Between the AUC of the Three Classifiers*

	SVM vs. NN	SVM vs. C4.5	NN vs. C4.5
Difference between AUC areas	0.06	0.03	0.03
Standard error (SE)	0.00898	0.0154	0.0121
95% confidence interval (CI)	0.0911–0.126	0.0735–0.134	−0.0188–0.0287
Significance level	$P < 0.001$	$P < 0.001$	$P = 0.683$
z statistics	12.07	6.731	0.408

The SVM is better than the NN and C4.5 since the *P*-value in the first and second columns are less than 0.001 and the confidence intervals do not include the 0 value. See text for more details.

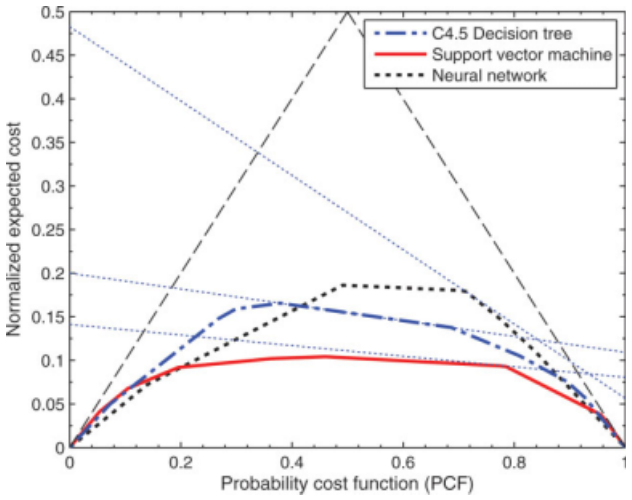


Figure 4. The cost curves of the neural network, the C4.5 decision tree, and the support vector machine. All cost curves fall inside the triangular area. The cost curve of the support vector machine classifier is the lowest curve and does not cross the other two curves for a wide range of probability cost function (the x-axis of the plot), which shows that it is the best classifier. The C4.5 decision tree classifier and the neural networks classifier have overlapping performance depending on the characteristics of the training dataset and the misclassification cost. The optimal operating lines are the thin lines on top of the cost curves. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

of the training data from the database of soft-tissue tumors (BSNR) of the University Hospital of Antwerpen where the MRI images of each patient were assessed by at least two radiologists.

The with 95% CI was calculated using McNemar's test. Using Eq. [1] and the data in Table 4 we obtained the estimated chi-square value, which is larger than the tabulated theoretical value. Hence, we rejected the null hypothesis that both the radiologists and the support vector machines perform equally. To validate the statistical testing result we calculated the kappa statistics ($\kappa = 0.526$) which shows that the strength of agreement is considered "moderate." Since the SVM classifier has a higher classification accuracy than the radiologists, we conclude that the SVM is better. Finally, Table 6 lists some performance parameters of the support vector machines and the radiologist.

Overall, the SVM classifier performs better than the radiologists since it has better classification accuracy. The specificity of the radiologists is slightly better than the SVM classifier since the SVM classifier mistakenly classify few benign tumors cases as malignant tumors. In Table 7 we list the tumors that are missed by the support vector machine classifier.

Table 6
Several Performance Parameters for the Radiologists and the SVM Classifier

	Radiologists	SVM
Accuracy %	90	93
Sensitivity %	81	94
Specificity %	92	91
AUC %	85	92

Table 7
Errors in Benign vs. Malignant Classification of the Full Training Set by SVM Classifier

Pathology diagnosis	SVM classification	Histology
Benign	Malignant	Fibroma
Benign	Malignant	Deep lipoma
Benign	Malignant	Leiomyoma
Benign	Malignant	Fibromotosis
Benign	Malignant	Lipoma cutaneous
Malignant	Benign	Fibrosarcoma, Myxoid
Malignant	Benign	Leiomyosarcoma
Malignant	Benign	Synovail sarcoma
Malignant	Benign	Myxoid liposarcoma

DISCUSSION

Machine learning classifiers trained with texture analysis features extracted from the tumor areas in T1-weighted MR images are potentially valuable tools for the differentiation between malignant and benign tumors. We demonstrated that the classification results correlated very well with the clinical status of the patients. Even though we used a large training dataset compared to what we have seen in other published studies, the size of the training MRI dataset still represents a small random sample from the whole population of soft-tissue tumors. For practical applications we need much larger training samples to cover all different pathological types of malignant and benign tumors. Several published studies have shown the importance and value of texture features for discriminating benign from malignant soft-tissue tumors by training some machine learning classifiers such as neural networks and the k-NN classifier (8,9,11). However, these studies did not evaluate the trained classifiers comprehensively, which we tried to cover in this study. Comparison of the SVM with the radiologist based on McNemar's nonparametric statistical test showed that the SVM performed as good as or better than the radiologists. The SVM classifier was trained with texture analysis features that quantify the signal homogeneity and heterogeneity of the T1-MRI. The radiologists diagnosed the tumors using different MR images modalities (T1-MRI, PD-MRI, T2-MRI, and chemically enhanced images) and other nonimage information such as the laboratory tests and the medical history of the patients.

In conclusion, the results were not surprising given the fact that texture analysis by computer algorithms can extract more texture information from the tumor regions compared to what humans can do based only on visual assessment. We mention a similar example from the analysis of digitized mammograms field where a computerized method for calculating a breast density index (BDI) can quantitatively model the radiologists' perception of the breast density (25). We think that an index derived from texture analysis can play a similar

role for discrimination between malignant and benign soft-tissue tumors since it can subjectively model the human perception of signal homogeneity and heterogeneity of the tumor areas in T1-MRI.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their hints and comments for improvement of an earlier version of this article. We also thank the editor of the journal for his patience and continued feedback during the revision process.

REFERENCES

1. Schepper AMD, Vanhoenacker F, Parizel PM, Gielen J (eds.). Imaging of soft tissue tumors, 2nd ed. Berlin: Springer; 2005.
2. Weatherall PT. Benign and malignant masses: MR imaging differentiation. Magn Reson Imaging Clin N Am 1995;3:669–694.
3. Julesz B, Gilbert EN, Shepp LA, Frisch HL. Inability of humans to discriminate between visual textures that agree in second-order statistics. Perception 1973;2:391–405.
4. Julesz B. Experiments in visual perception of texture. Sci Am 1975;232:34–43.
5. Materka A, Strzelectky M. Texture analysis methods: a review. Technical University of Lodz, COST B11-techincal report. 1998; 11:873–887.
6. Tuceryan M, Jain AK. Texture analysis. The handbook of pattern recognition and computer vision, 2nd ed. Dartmouth, MA: University of Massachusetts, Word Scientific; 1998. p 207–248.
7. Wagner T. Texture analysis. The handbook of computer vision and applications, Vol. 2. Signal processing and pattern recognition, New York: Academic Press; 1999;275–308.
8. Mayerhoefer ME, Breitenseher MJ, Kramer J, Aigner N, Hofmann S, Materka A. Texture analysis for tissue discrimination on T1-weighted MR images of knee joint in a multicenter study: transferability of texture features and comparison of feature selection methods and classifiers. J Magn Reson Imaging 2005;22:674–680.
9. Castellano G, Bonilha L, Li LM, Cendes F. Texture analysis of medical images. Clin Radiol 2004;59:1061–1069.
10. Huang Y-L, Wang K-L, Chen D-R. Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines. Neural Comput Appl 2006;15:164–169.
11. Mayerhoefer ME, Breitenseher M, Amann G, Dominkus M. Are signal intensity and homogeneity useful parameters for distinguishing between benign and malignant soft tissue masses on MR images? Objective evaluation by means of texture analysis. Magn Reson Imaging 2008;26:1316–1322
12. Alpaydin E. Introduction to machine learning, 5th ed. Cambridge, MA: MIT Press; 2004.
13. Salzberg S. On comparing classifiers: pitfalls to avoid and a recommended Approach. Data Mining Knowl Disc 1997;1:317–328.
14. Fletcher CDM, Unni KK, Mertens F (eds.). World Health Organization classification of tumours: pathology and genetics of tumours of soft tissue and bone. Vol. 5. Lyon, France: IARC Press; 2002.
15. Rosai J. Rosai and Ackerman's surgical pathology, 2 volume set, 9th ed. St. Louis, MO: Mosby; 2004.
16. Girosi F, Chan NT. Prior knowledge and the creation of virtual examples for RBF. Neural Networks for Signal Processing, V. Proc 1995 IEEE Workshop, Cambridge, MA; 1995:201–210.
17. Niyogi P, Girosi F, Poggios T. Incorporating prior information in machine learning by creating virtual examples. Proc IEEE 1998; 86:2196–2209.
18. Bishop CM. Training with noise is equivalent to Tikhonov regularization. Neural Comput 1995;7:108–116.
19. The freeware MaZda 3.20 software program for texture analysis made by the Institute of Electronics, Technical University of Lodz, Poland. Available at: http://www.eletel.p.lodz.pl/programy/cost/software.html Accessed March 10, 2009.
20. The open source machine learning package Weka 3.5.8 (Waikato Environment for Knowledge Analysis) made by the University of Waikato, New Zealand. Available at: http://www.cs.waikato.ac.nz/ml/weka/ Accessed: March 10, 2009.
21. Witten IH, Frank E. Data mining-practical machine learning tools and techniques, 2nd ed. San Francisco: Morgan Kaufmann. 2005.
22. Sarunas JR, Anil KJ. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Machine Intell 1991;13:252–264.
23. Jin Huang, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17: 299–310.
24. Drummond C, Holte RC. Cost curves: an improved method for visualizing classifier performance. Machine Learn 2006;26: 95–130.
25. Boone JM, Lindfors KK, Beatty CS JA. A breast density index for digital mammograms based on radiologists' ranking. J Digit Imaging 1998;11:101–115.