See SunSCAN 3D in action:
Request your demo today

Introducing SunSCAN™ 3D

The Next-Generation Cylindrical Water Scanning System

SunSCAN 3D simplifies beam scanning with SRS-class accuracy and user-centered design.

It enables faster, easier workflows, and hyper-accurate dosimetry for today's busy clinics.

Learn more:
sunnuclear.com

SUN NUCLEAR
A MIRION MEDICAL COMPANY

SunSCAN™ 3D is not available for sale in all markets. CE Mark pending.

The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process

M. Elter[a)]
Fraunhofer Institute for Integrated Circuits (IIS), Am Wolfsmantel 33, 91058 Erlangen, Germany

R. Schulz-Wendtland
Institute of Radiology, Gynaecological Radiology, University Erlangen-Nuremberg, Universitätsstraße 21-23, 91054 Erlangen, Germany

T. Wittenberg
Fraunhofer Institute for Integrated Circuits (IIS), Am Wolfsmantel 33, 91058 Erlangen, Germany

(Received 12 April 2007; revised 24 August 2007; accepted for publication 28 August 2007; published 15 October 2007)

Mammography is the most effective method for breast cancer screening available today. However, the low positive predictive value of breast biopsy resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes. To reduce the high number of unnecessary breast biopsies, several computer-aided diagnosis (CAD) systems have been proposed in the last several years. These systems help physicians in their decision to perform a breast biopsy on a suspicious lesion seen in a mammogram or to perform a short term follow-up examination instead. We present two novel CAD approaches that both emphasize an intelligible decision process to predict breast biopsy outcomes from BI-RADS™ findings. An intelligible reasoning process is an important requirement for the acceptance of CAD systems by physicians. The first approach induces a global model based on decison-tree learning. The second approach is based on case-based reasoning and applies an entropic similarity measure. We have evaluated the performance of both CAD approaches on two large publicly available mammography reference databases using receiver operating characteristic (ROC) analysis, bootstrap sampling, and the ANOVA statistical significance test. Both approaches outperform the diagnosis decisions of the physicians. Hence, both systems have the potential to reduce the number of unnecessary breast biopsies in clinical practice. A comparison of the performance of the proposed decision tree and CBR approaches with a state of the art approach based on artificial neural networks (ANN) shows that the CBR approach performs slightly better than the ANN approach, which in turn results in slightly better performance than the decision-tree approach. The differences are statistically significant (p value <0.001). On 2100 masses extracted from the DDSM database, the CRB approach for example resulted in an area under the ROC curve of $A(z)=0.89\pm0.01$, the decision-tree approach in $A(z)=0.87\pm0.01$, and the ANN approach in $A(z)=0.88\pm0.01$. © *2007 American Association of Physicists in Medicine*. [DOI: 10.1118/1.2786864]

Key words: mammography, computer-aided diagnosis, knowledge-based, receiver-operating characteristic (ROC)

I. INTRODUCTION

According to the American Cancer Society, breast cancer is the most common cancer among women, accounting for nearly one in three cancers diagnosed in US women. With more than 40 000 women dying from breast cancer in the United States each year, only lung cancer accounts for more women dying from cancer.[1] Breast cancers in early stages unfortunately produce no symptoms when the tumor is still small and hence well treatable. Therefore, it is difficult but at the same time important to detect breast cancers at an early stage. Both randomized trials and population-based evaluations of screening mammography have shown that early detection of breast cancer through mammography greatly improves the chances of survival.[2–5] Mammography can identify cancer several years before physical symptoms are produced and therefore is recognized as the most effective breast cancer screening method available today. However, about 5%–10% of the mammography results are interpreted as abnormal or inconclusive until further examinations like ultrasound imaging or breast biopsy lead to a final interpretation of normal or benign breast tissue. It is reported that only 10%–30% of all breast biopsies actually show a malignant pathology.[6] The high number of unnecessary breast biopsies causes major mental and physical discomfort for the patients as well as unnecessary expenses spent for examinations. In the last several years computer aided diagnosis systems have been proposed that use lesion descriptions based on the BI-RADS™[7] standard lexicon as input attributes to support the physician's decision to perform a breast biopsy or a short follow-up diagnosis on a suspicious region seen in a mammogram. Baker *et al.* and Markey *et al.*[8,9] have proposed an artificial neural network (ANN) approach to deduce

diagnosis proposals from BI-RADS descriptions. Alternative approaches based on case-based reasoning (CBR) and Bayesian networks were later proposed by Floyd *et al.*,[10] Bilska-Wolak *et al.*[11–13] and Fischer *et al.*[14] The prime advantage of CBR over the earlier proposed approaches is the intelligible reasoning process that leads to the systems diagnosis suggestion. A CBR-based CAD system reasons based on stored knowledge (prior cases with associated ground truth) and its final diagnosis suggestion is based on the ground truth of the stored cases that are most similar to the query case. Hence, its reasoning process is much easier to comprehend for the physician than an ANN system that acts like a black box. In this paper we propose a novel approach, based on decision-tree learning (DT), for building a CAD system that predicts breast cancer biopsy outcomes based on BI-RADS compliant lesion descriptions. Similar to the cited CAD systems based on case-based reasoning, a CAD system based on decision trees features a very transparent reasoning process. However, in contrast to a CBR system, a decision-tree learner abstracts a global model of the decision process from the prior cases with associated ground truth instead of directly using these cases in the decision process. This global model is even easier to understand and predictions based on it are even easier to comprehend for the physician than those of a CBR system. In a second approach, we propose an extension of the state of the art CBR approaches that features an entropic distance measure. It provides a solid mathematical basis for measuring the similarity of cases that have attributes of different types (e.g., nominal and numeric). It furthermore provides a clean mathematical foundation for handling missing attributes, in contrast to previous approaches.[15] We have evaluated the proposed CAD approaches on two large, publicly available mammography databases to compare both their performance and features.

II. MATERIALS AND METHODS

In the following we specify the mammography case databases that we have used to develop and evaluate our CAD systems on. We furthermore present our novel CAD system based on decision-tree learning, and a novel extension of the state of the art CBR approaches. We furthermore give a short overview of a CAD approach based on artificial neural networks that we have implemented to compare our novel approaches to a state of the art approach. We conclude Sec. II by describing the methods that we have used to evaluate and compare the performance of all three approaches.

II.A. The mammography case databases

A meaningful and reproducible evaluation of a CAD system requires a sufficiently large and preferably public case database. We have chosen the digital database for screening mammography (DDSM) which is publicly available from the University of South Florida[16] to evaluate and compare our CAD approaches on. The DDSM is intended as a benchmark database for mammography CAD algorithms. It contains 2620 cases acquired in the early 1990s. Each case includes two images of each breast, along with the patient's age, BI-RADS breast density rating, subtlety rating for abnormalities, and BI-RADS standard lexicon compliant descriptions of abnormalities. The database contains normal, benign, and malignant cases. The images are conventional screen film mammograms that have been digitized using scanners from three different manufacturers. We have extracted all benign and malignant abnormalities that contain masses or calcifications. We have discarded those abnormalities that contain *both* masses and calcifications, architectural distortions, or tissue asymmetries. The case database therefore contains 2100 abnormalities (regions) containing masses and 1359 regions containing calcifications. In the following we will refer to the set of mass regions as M_{DDSM} and to the set of calcification regions as C_{DDSM}. Both sets are balanced with respect to the cases being classified (ground truth) as benign or malignant. 1055 (50.2%) of the mass regions are benign and 1045 (49.8%) are malignant. 749 (55.2%) of the calcification regions are benign and 610 (44.8%) are malignant. We use the patient's age and the available BI-RADS descriptions (mass shape, mass margins, calcification type, calcification subtype, and calcification distribution) as input attributes for the CAD systems.

In addition to the DDSM data we have collected our own mammography database containing 961 mass regions at a large radiological center, which we will make publicly available at the well-known UCI machine learning repository[17] with the publication of this paper. The database is based on modern full-field digital mammograms and contains cases that have been acquired from 2003 to 2006. We will refer to this set of mass regions as M_{UCI}. 515 (53.6%) of these mass regions are benign and 446 (46.4%) are malignant. Similar to the DDSM data we use the patient's age and BI-RADS descriptions as input attributes for the CBR system. In addition to the BI-RADS descriptions mentioned earlier, the mass density was available as a BI-RADS compliant attribute for these cases and we therefore use it as an additional input attribute.

II.B. Decision trees

II.B.1. Concept

It is both natural and intuitive to classify a suspicious lesion seen in a mammogram using a sequence of questions, in which the next question depends on the answer to the previous question. A typical example would be to classify a suspicious mass in a mammogram by starting with the question "What is the shape of the mass?" and if the answer is "irregular" to proceed with the question "How does the margin of the mass look?" The answer to this second question might be "spiculated," in which case the classification process might already be finished with the final classification "malignant." Such sequences of questions can be represented by a directed tree structure called a decision tree. In a decision tree, nodes represent questions and directed links between the nodes represent the possible answers to the questions. Terminal nodes are called leaf nodes and represent

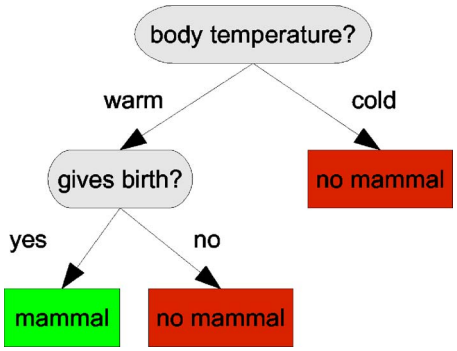

FIG. 1. A sample decision tree for the classification of animals as mammals.

classifications like "benign" or "malignancy." Figure 1 shows a very simple example of a decision tree for the classification of animals.

II.B.2. Decision-tree learning

Many different algorithms have been proposed to automatically induce a decision tree from instances (in our case mammography cases) that are represented as attribute-value pairs. However, most of these algorithms share a common core algorithm, which is a greedy search through the space of all possible decision trees. A well-known representative of such decision-tree learning algorithms is the ID3 algorithm proposed by Quinlan.[18] In the ID3 algorithm, a decision tree is induced from sample instances by constructing it in a top-down approach. It starts by choosing the attribute as root node that best classifies the training examples alone. A child node is then created for each possible value of this attribute and the training samples are assigned to the appropriate child nodes. This process is repeated recursively by considering each of the new child nodes as a root node of a subtree and by selecting the best attribute for this point of the tree. The algorithm stops when for each leaf node of the tree, all samples that are assigned to that leaf share the same classification.

II.B.3. Attribute selection

The core of the above described ID3 algorithm is the choice of the best attribute for a given node in the decision tree. The natural choice is the attribute that is most useful for classifying the examples assigned to that node. The ID3 algorithm uses a property called information gain as the quantitative measure of the worth of an attribute for the classification of a set of examples. The definition of the information gain is based on the definition of the entropy, a measure that is well known from information theory. Considering a set S of examples, the entropy $H(S)$ characterizes the (im)purity of the set S with respect to the target attribute. Let the target attribute be a nominal value that can have n different values (e.g., the attribute "mass shape" can have one of the $n=4$ values "round," "oval," "lobular," and "irregular"). The entropy of the set of examples S is then defined as

$$H(S) = \sum_{i=1}^{n} -p_i \log_2 p_i,$$

with p_i denoting the fraction of S that belongs to class i.

With this definition of the entropy as a measure of the impurity of a set of examples S, the information gain $G(S,a)$ can now be defined as the expected reduction in entropy caused by partitioning the example set S according to the attribute a:

$$G(S,a) = H(S) - H(S,a),$$

with $H(S,a)$ denoting the expected value of the entropy of the set S after partitioning S using the attribute a. $H(S,a)$ is the sum of the entropies of each subset S_p, weighted by the fraction of examples $|S_p|/|S|$ that belong to S_p. More precisely, it is defined as

$$H(S,a) = \sum_{p} \frac{|S_p|}{|S|} H(S_p).$$

II.B.4. Extensions

We use an extension of the ID3 algorithm called C4.5, which was also proposed by Quinlan.[19] C4.5 extends the ID3 algorithm by supporting numeric in addition to nominal attributes. It furthermore employs a pruning technique that prevents overfitting by pruning (cleaning) the decision tree. It also contains extensions that allow samples to have missing attribute values, which in our case is important for samples where not all BI-RADS compliant attributes are given by the physician (e.g., if the mass margin is not provided by the physician).

II.B.5. Classification

Once a decision tree has been constructed, classifying a query case is straightforward. Starting at the root node, the node attribute tests are applied to the query case and the appropriate links are followed. This ultimately leads to a leaf node, and the class label associated with that leaf node is assigned to the query case. Besides the class label, a confidence value for the assigned classification can be defined based on the ratios of the classifications of training cases that were assigned to the leaf node during tree construction.

II.C. Case-based reasoning

A case-based reasoning (CBR) system stores expertise as a library of cases with known outcomes. For our mammography CBR system, each case is a region of interest containing a mammographic abnormality together with a set of input attributes as described in the following. The known outcome is the biopsy result (benign or malignant) for the abnormality. To generate a diagnosis proposal for a new region, the BI-RADS attributes of the query region are matched against the BI-RADS attributes of all the regions in the case library using a similarity metric. The most similar regions are retrieved from the database. The known classifications of re-

trieved regions are then used to suggest a solution for the query region (benign or malignant) based on a decision rule.

II.C.1. Similarity metric

The similarity metric which is used to find the most similar regions in the case database for a given query region is the key component of a CBR system. State of the art CBR systems that deduce diagnosis proposals from BI-RADS attributes use either the Hamming distance or the Euclidean distance as a similarity metric. While the Hamming distance works well for categorical attributes (like the mass shape or the calcification distribution), it does not work well for numeric attributes like the patient's age. In contrast to the Hamming distance, the Euclidean distance works well for numeric attributes but is not well defined for categorical attributes. Furthermore, both approaches have the drawback that they do not provide a mathematically well founded approach to handle missing attribute values. To solve these problems we propose using an entropy-based distance measure for a CBR-based mammography CAD system. It provides a consistent approach to handling both categorical and numeric attributes, and it additionally provides a mathematically well founded approach to handle missing attribute values. This distance measure was originally proposed by Cleary and Trigg[20] as distance metric for an instance-based machine learning algorithm called κ^*, which is closely related to the classic k-nearest-neighbor classifier. The key idea of this entropy-based distance metric is to define the distance between two instances as the complexity of transforming one instance into the other. Cleary and Trigg define a "program" that transforms one instance into the other as a finite sequence of atomic transformations. The usual definition of the (Kolmogorov) complexity of such a program is the length of the shortest string representing the program. A more robust approach is to sum the length over all possible programs that transform an instance into the other. This approach is known from computational biology where it is used to measure the distance between two sequences of DNA. The mathematical details of defining this metric for the different types of attributes is rather comprehensive and hence we refer the reader to the work of Cleary and Trigg for the mathematical details. However, we include the rather short definition of the metric for categorical attributes in the following to give the reader at least a first impression of how the theoretical definition based on the length of the shortest string representing a transformation program is realized in practice.

The set of transformations that are defined for categorical attributes in Ref. 20 are the transformations of any (of the discrete set of categorical) attribute values to any other. Given n possible attribute values, let p_i be the probability for a certain value i, with $1 \leq i \leq n$, to occur. Let s be the probability of a value staying the same and $(1-s)p_j$ the probability of a value transformed to the value j. As mentioned earlier, the transformation probability is defined by summing over all possible transformations:

$$P^*(j|i) = \begin{cases} (1-s)p_j & \text{if } i \neq j \\ s + (1-s)p_j & \text{if } i = j. \end{cases}$$

The distance K^* from instance a to instance b is then defined as $K^*(b|a) = -\log_2(P^*(b|a))$. For the probability s a reasonable value must be chosen depending on the data being modeled. For the details please refer to Ref. 20.

II.C.2. Decision rule

A diagnosis (benign or malignant) is suggested for a new (query) case q using a decision rule that is based on the similarity values of the query region and all regions d_l stored in the database and their associated ground truth. The similarity value s_{q,d_l}, which denotes the similarity between the query region q and the stored region d_l, is calculated for all N regions d_l, $0 < l < N-1$ stored in the database. Then the k most similar regions $d_m, 0 < m < k-1$ are selected and the fraction fm of malignant regions is calculated. Given a malignancy fraction cut-off C_f, *with* $0 \leq C_f \leq 1$, the diagnosis *malignant* is suggested if $f_m \geq C_f$ and the diagnosis *benign* otherwise.

II.D. Artificial neural network

As mentioned in Sec. I, a state of the art approach to deduce diagnosis proposals from BI-RADS attributes is to use an artificial neural network (ANN). To be able to compare the performance and the properties of the CBR and the decision-tree approach that we propose in this work with the state of the art, we have implemented an ANN that is trained using the same set of attributes and the same case databases that we use for the proposed approaches. We have implemented an ANN with a network layout as close as possible to the one described in Ref. 8. The layout is not exactly the same, as our case databases contain a different set of BI-RADS attributes than those used in Ref. 8. The ANN is a three-layer, feed-forward network and it is trained using backpropagation. The layers consist of an input layer with one input node per attribute, a hidden layer with four nodes, and an output layer with a single output node.

II.E. Performance evaluation

The performances of the CAD approaches have been evaluated using receiver operating characteristic (ROC) analysis and bootstrap sampling. They have been compared using Fisher's analysis of variance (ANOVA) approach.[21]

II.E.1. ROC analysis

By computing the sensitivity (true positive fraction) and the specificity (1 - false positive fraction) of the CBR system for malignancy fraction cut-offs C_f varying over the range from zero to one, a ROC curve[22,23] is generated. For the decision-tree approach, the ROC curve is generated by varying a confidence threshold for cases being classified as malignant over the range from zero to one. We use the area under the ROC curve $A(z)$ as a metric for the performance of the system. We calculate $A(z)$ using numeric integration. For

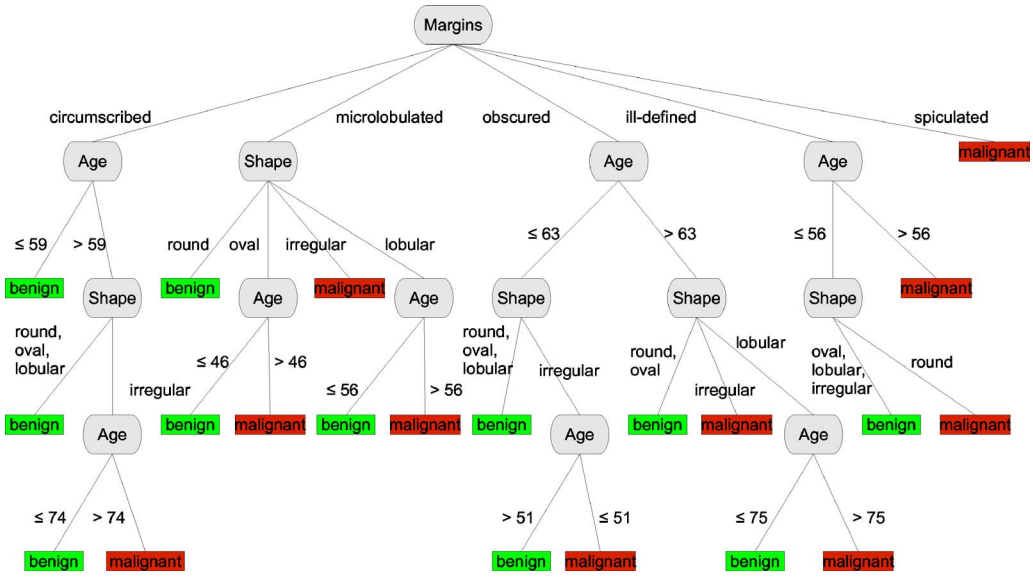

FIG. 2. The decision-tree model for the mass regions of the DDSM database (M_{DDSM}). Each path from the root to a leaf node represents a decision rule. Masses with microlobulated margin and oval shape for example are classified as malignant, if the patient is older than 46.

breast cancer prediction high sensitivity is usually considered as more important than high specificity, i.e., it is better to falsely classify a benign region as malignant rather than to miss a breast cancer by classifying a malignant region as benign. Hence, we also use the area under the partial ROC curve for a sensitivity of 0.9 or greater $A(z)_{0.9}$ and the specificity $\mathrm{Spec}_{0.95}$ at the given sensitivity 0.95 as performance measures.

II.E.2. Bootstrap sampling

Bootstrap sampling[24–26] is a resampling technique that can be used to estimate a statistic Θ (e.g., the standard deviation of the area under a ROC curve). Consider a set S containing n samples. A bootstrap set $\hat{S}$ is created by randomly selecting n samples from S, with replacement. Because of sampling with replacement and because of both S and $\hat{S}$ containing n samples, $\hat{S}$ basically always contains duplicated samples. To estimate a statistic Θ using bootstrap sampling, L independent bootstrap samples $\hat{S}_i$ are sampled from S and the bootstrap estimate of Θ is simply the mean of the L individual values of Θ. We use a generalization of the bootstrap sampling concept to estimate the three performance measures described earlier for both CAD systems that we have presented. We generate L bootstrap samples from the input set of samples and train L instances of each of the two CAD systems on these sets. To make sure that training and testing sets are always disjoint, each instance of the CAD systems is tested on those samples of the input set that have not been selected for the bootstrap set the instance was trained on. The probability of a sample being chosen for a bootstrap set is $1-(1-1/n)^n$. For a sufficiently large n, this probability approaches $1-e^{-1}=0.632$. On average a bootstrap sample of size n, therefore, contains 63.2% and a test set on average contains 36.8% of the samples in the input set. Using this approach, we estimate the mean and the standard deviations of the performance measures $A(z)$, $A(z)_{0.9}$, and $\mathrm{Spec}_{0.95}$ introduced earlier.

III. RESULTS

All performance values given in the following are the mean values of the bootstrap samples ± the standard deviations from the mean. The ROC performance of the decision-tree approach on the mass regions extracted from the DDSM database (M_{DDSM}) evaluated using 1000 bootstrap samples as described in Sec. II E results in an area under the ROC curve of $A(z)=0.872\pm0.011$, a partial area under the high sensitive region of the ROC curve of $A(z)_{0.9}=0.650\pm0.031$, and a specificity of $\mathrm{Spec}_{0.95}=0.554\pm0.057$ for a fixed sensitivity of 0.95. The same kind of evaluation of the case-based reasoning approach resulted in an area under the ROC curve of $A(z)=0.885\pm0.010$, a partial area under the high sensitive region of the ROC curve of $A(z)_{0.9}=0.672\pm0.03$, and a specificity of $\mathrm{Spec}_{0.95}=0.593\pm0.04$ for a fixed sensitivity of 0.95. The ANN approach resulted in an area under the ROC curve of $A(z)=0.882\pm0.010$, a partial area under the high sensitive region of the ROC curve of $A(z)_{0.9}=0.672\pm0.025$, and a specificity of $\mathrm{Spec}_{0.95}=0.587\pm0.035$ for a fixed sensitivity of 0.95. The differences of all three performance metrics are statistically significant (p value <0.001). The decision-tree model resulting from using the full sets of DDSM masses is shown in Fig. 2. Each path from the root to a leaf node represents a decision rule. Masses with microlobulated margin and oval shape, for example, are classified as malignant, if the patient is older than 46.

On the set of mass regions extracted from the UCI database (M_{UCI}), the performance measures for the decision-tree approach are $A(z)=0.838\pm0.017$, $A(z)_{0.9}=0.477\pm0.06$, and $\mathrm{Spec}_{0.95}=0.298\pm0.076$. For the case-based reasoning approach they are $A(z)=0.857\pm0.016$, $A(z)_{0.9}=0.505\pm0.063$,

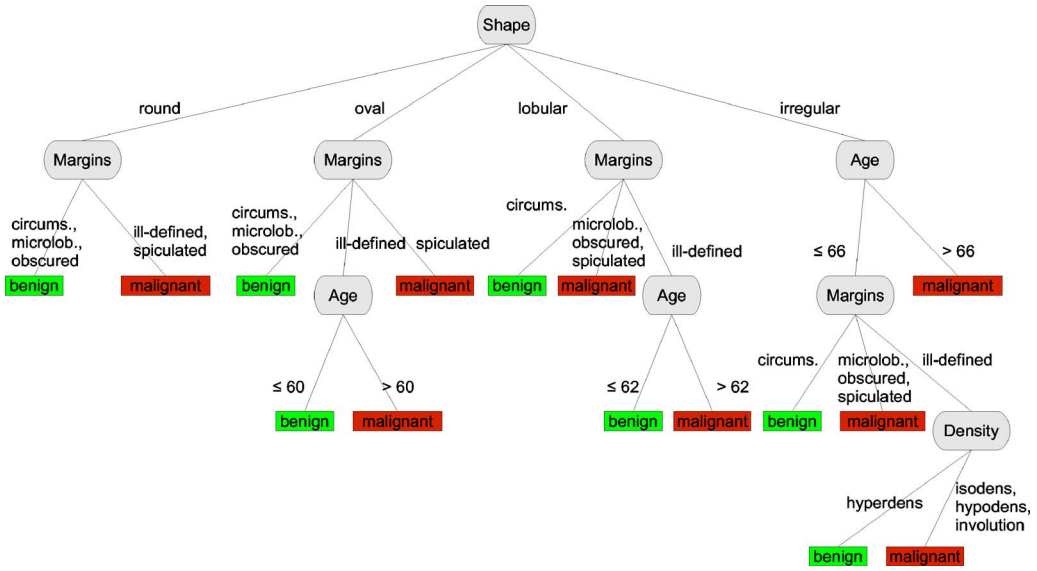

FIG. 3. The decision-tree model for the mass regions of the UCI database (M_{UCI}).

and $Spec_{0.95}=0.313\pm0.054$. The ANN approach resulted in an area under the ROC curve of $A(z)=0.847\pm0.017$, a partial area under the high sensitive region of the ROC curve of $A(z)_{0.9}=0.521\pm0.055$ and a specificity of $Spec_{0.95}=0.338\pm0.07$ for a fixed sensitivity of 0.95. The decision-tree model resulting from using the full sets of UCI masses is shown in Fig. 3. The differences of all three performance metrics are statistically significant (p value <0.001 for $A(z)$ and $A(z)_{0.9}$, p value <0.02 for $Spec_{0.95}$).

On the set of calcifications regions extracted from the DDSM database (C_{DDSM}) the performance measures for the decision-tree approach are $A(z)=0.737\pm0.021$, $A(z)_{0.9}=0.292\pm0.045$, and $Spec_{0.95}=0.125\pm0.037$. For the case-based reasoning approach they are $A(z)=0.761\pm0.018$, $A(z)_{0.9}=0.341\pm0.040$, and $Spec_{0.95}=0.166\pm0.050$. The ANN approach resulted in an area under the ROC curve of $A(z)=0.755\pm0.019$, a partial area under the high sensitive region of the ROC curve of $A(z)_{0.9}=0.332\pm0.05$, and a specificity of $Spec_{0.95}=0.151\pm0.06$ for a fixed sensitivity of 0.95. The decision-tree model resulting from using the full set of DDSM calcifications is shown in Fig. 4. The differences of all three performance metrics are statistically significant (p value <0.001).

Table I summarizes the ROC performance values that have been achieved using the three different classifiers (decision tree, case-based reasoning, and neural network) on the different region of interest sets (M_{DDSM}, M_{UCI}, and C_{DDSM}).

Each case in the DDSM and the UCI database has an associated BI-RADS number ranging from 2 (definitely benign) to 5 (highly suggestive of malignancy) assigned in a double-review process by physicians. Assuming that all cases with BI-RADS numbers greater than or equal to a given value (varying from 2 to 5) are malignant and the other cases benign, sensitivities and associated specificities can be calculated. Tables II–IV show these values for the case sets M_{DDSM}, M_{UCI}, and C_{DDSM} in comparison to the symmetric 95% confidence intervals of the specificities achieved by our CAD systems for the same sensitivities. Typical sensitivities

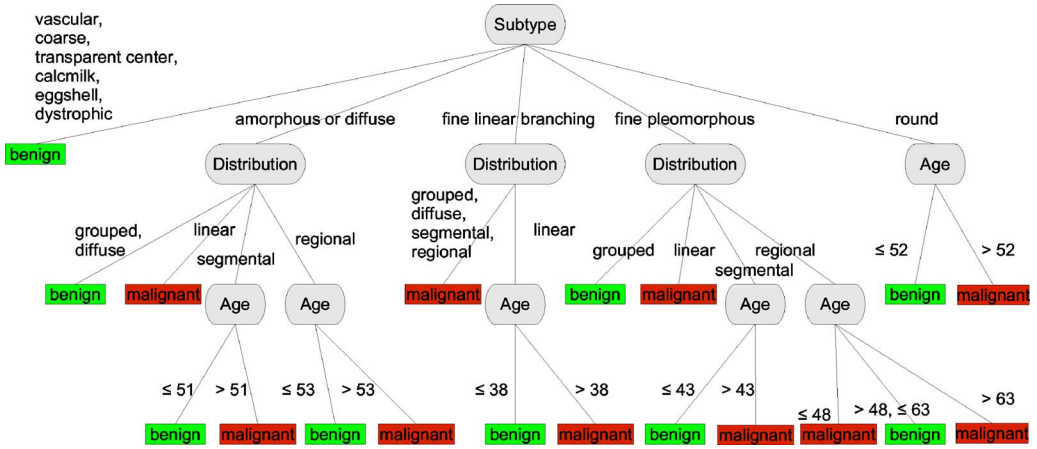

FIG. 4. The decision-tree model for the calcifications regions of the DDSM database (C_{DDSM}).

TABLE I. Summary of the ROC performance values achieved using the different classifiers on the different region of interest sets.

		M_{DDSM}	M_{UCI}	C_{DDSM}
DT	$A(z)$	0.872±0.011	0.838±0.017	0.737±0.021
	$A(z)_{0.9}$	0.650±0.031	0.477±0.060	0.292±0.045
	$Spec_{0.95}$	0.55.4±0.057	0.298±0.076	0.125±0.037
CBR	$A(z)$	0.885±0.010	0.857±0.016	0.761±0.018
	$A(z)_{0.9}$	0.672±0.030	0.505±0.063	0.341±0.040
	$Spec_{0.95}$	0.593±0.040	0.313±0.054	0.166±0.050
ANN	$A(z)$	0.882±0.010	0.847±0.017	0.755±0.019
	$A(z)_{0.9}$	0.672±0.025	0.521±0.055	0.332±0.050
	$Spec_{0.95}$	0.587±0.035	0.338±0.070	0.151±0.060

TABLE II. Comparison of the specificity achieved for case set M_{DDSM} (masses) by the physicians and the symmetric 95% confidence intervals of the specificities achieved by the CAD systems for given sensitivities.

BI-RADS	Sens.	Spec.	CBR spec.	DT spec.	ANN spec.
≥2	0.99	0.05	[0.284,0.296]	[0.185,0.195]	[0.204,0.216]
≥3	0.98	0.09	[0.438,0.442]	[0.316,0.324]	[0.375,0.385]
≥4	0.91	0.36	[0.668,0.672]	[0.648,0.652]	[0.658,0.662]
≥5	0.46	0.98	[0.959,0.961]	[0.949,0.951]	[0.949,0.951]

TABLE III. Comparison of the specificity achieved for case set M_{UCI} (masses) by the physicians and the symmetric 95% confidence intervals of the specificities achieved by the CAD systems for given sensitivities.

BI-RADS	Sens.	Spec.	CBR spec.	DT spec.	ANN spec.
≥2	1.00	0.00	[0.038,0.042]	[0.009,0.010]	[0.038,0.042]
≥3	0.99	0.03	[0.107,0.113]	[0.086,0.092]	[0.096,0.104]
≥4	0.98	0.09	[0.187,0.193]	[0.148,0.152]	[0.166,0.174]
≥5	0.70	0.92	[0.878,0.882]	[0.858,0.862]	[0.868,0.872]

TABLE IV. Comparison of the specificity achieved for case set C_{DDSM} (calcifications) by the physicians and the symmetric 95% confidence intervals of the specificities achieved by the CAD systems for given sensitivities.

BI-RADS	Sens.	Spec.	CBR spec.	DT spec.	ANN spec.
≥2	1.00	0.00	[0.076,0.082]	[0.029,0.031]	[0.048,0.052]
≥3	0.99	0.07	[0.118,0.122]	[0.079,0.081]	[0.089,0.091]
≥4	0.93	0.13	[0.298,0.302]	[0.208,0.212]	[0.258,0.262]
≥5	0.22	0.99	[0.979,0.981]	[0.959,0.961]	[0.969,0.971]

TABLE V. Comparison of the information gain of the attributes for the case set M_{DDSM}. Note that the information gain values are relative values normalized to a range of zero to one. The fact that the age attribute has zero information gain therefore does not mean that it is useless for the classification but simply that in relation to the other attributes it is the least useful one.

Mass margins	Mass shape	Age
1.00	0.38	0.00

TABLE VI. Information gain of the attributes for the case set M_{UCI}.

Mass margins	Mass shape	Age	Mass density
1.00	0.91	0.52	0.00

considered for mammography diagnosis are greater than or equal to 0.9, and both CAD systems outperform the physicians for high sensitivities. Note that for the physicians' performance values no standard deviations can be calculated using this approach. Hence no test on the statistical significance of the differences of the physicians' and the CAD performances was possible.

As stated in Sec. II B 3 the information gain is a quantitative measure of the worth of an attribute for the classification of a set of examples. A comparison of the information gains of the input attributes that have been used for classification hence gives an impression of the relative usefulness of the individual attributes. Tables V–VII show the information gain values for the attributes used for the classification of the case sets M_{DDSM}, M_{UCI}, and C_{DDSM}, respectively. Note that the information gain values are relative values normalized to a range of zero to one.

While a comparison of the information gain gives a quantitative impression of the usefulness of the individual attributes, a visual impression can be obtained by plotting histograms of the values that occur for a given attribute in a case set. Figures 5 and 6 show the histograms of the age attribute for case sets M_{DDSM} and C_{DDSM}, which have been found to be especially interesting. Each figure contains two histograms, one for benign regions (blue) and one for malignant regions (red). While the age is obviously a useful attribute to distinguish between the benign and malignant masses of case set M_{DDSM}, it is almost useless to distinguish between the benign and malignant calcifications of case set C_{DDSM}.

While an evaluation of the performance of a CAD approach using bootstrap sampling on one of the two case databases (DDSM or UCI) allows one to estimate the generalization error for new cases to be collected at the same radiologic institutions and under similar conditions, it does not give an insight into how well a model that is built using one of the two case databases performs on a case database collected at different institutions and under different conditions. Therefore for all three CAD approaches a model was built using the DDSM data and tested on the UCI data and vice versa.

The ROC performance of the decision-tree approach trained on M_{DDSM}, and evaluated on M_{UCI} resulted in an area under the ROC curve of $A(z)=0.791$, a partial area under the high sensitive region of the ROC curve of $A(z)_{0.9}=0.31$, and

TABLE VII. Information gain of the attributes for case set C_{UCI}.

Calc subtype	Calc distribution	Calc type	Age
1.00	0.67	0.47	0.00

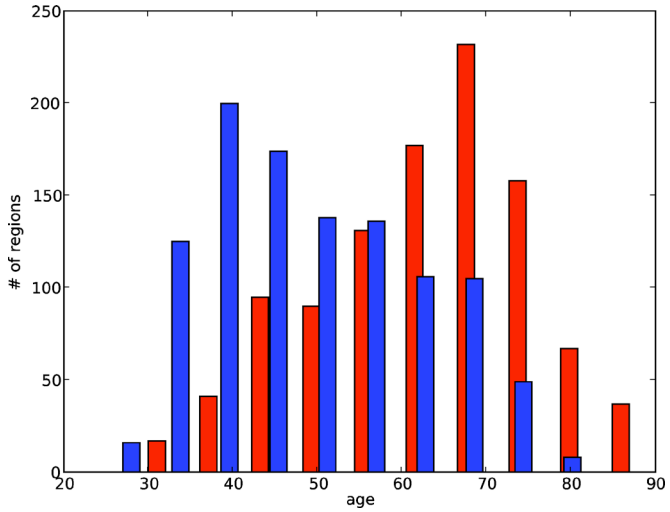

FIG. 5. Histogram (with 10 bins) of the age attribute for case set M_{DDSM}. The figure contains two histograms. One for benign regions (dark gray) and one for malignant regions (light gray).

a specificity of $Spec_{0.95}=0.16$ for a fixed sensitivity of 0.95. The case-based reasoning approach resulted in an area under the ROC curve of $A(z)=0.823$, a partial area under the high sensitive region of the ROC curve of $A(z)_{0.9}=0.445$, and a specificity of $Spec_{0.95}=0.233$ for a fixed sensitivity of 0.95. The ANN approach resulted in an area under the ROC curve of $A(z)=0.791$, a partial area under the high sensitive region of the ROC curve of $A(z)_{0.9}=0.24$, and a specificity of $Spec_{0.95}=0.14$ for a fixed sensitivity of 0.95. The differences of all three performance metrics are statistically significant (p value <0.001).

The ROC performance of the decision-tree approach trained on M_{UCI}, and evaluated on M_{DDSM} resulted in an area under the ROC curve of $A(z)=0.767$, a partial area under the high sensitive region of the ROC curve of $A(z)_{0.9}=0.316$, and a specificity of $Spec_{0.95}=0.188$ for a fixed sensitivity of 0.95.

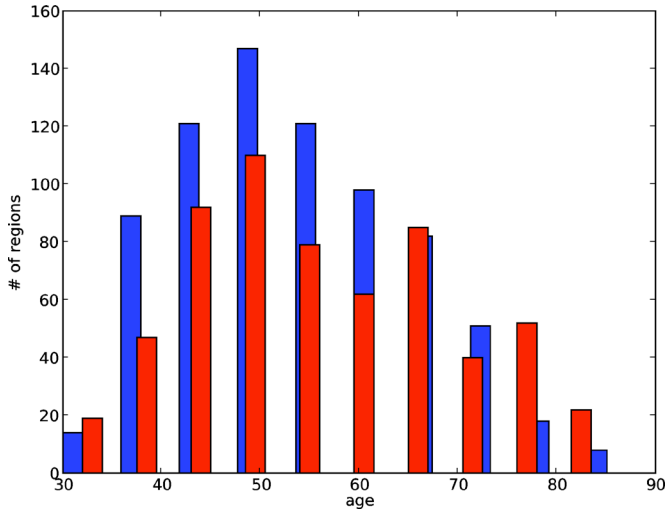

FIG. 6. Histogram (with 10 bins) of the age attribute for case set C_{DDSM}. The figure contains two histograms. One for benign regions (dark gray) and one for malignant regions (light gray).

The case-based reasoning approach resulted in an area under the ROC curve of $A(z)=0.815$, a partial area under the high sensitive region of the ROC curve of $A(Z)_{0.9}=0.563$, and a specificity of $Spec_{0.95}=0.425$ for a fixed sensitivity of 0.95. The ANN approach resulted in an area under the ROC curve of $A(z)=0.753$, a partial area under the high sensitive region of the ROC curve of $A(z)_{0.9}=0.309$, and a specificity of $Spec_{0.95}=0.161$ for a fixed sensitivity of 0.95. The differences of all three performance metrics are statistically significant (p value <0.001). Furthermore, the decreases in performance compared to the ROC performances listed in Table I are statistically significant (p value <0.001), too.

IV. DISCUSSION AND CONCLUSIONS

We have presented two CAD systems that both emphasize an intelligible decision process for the prediction of breast cancer biopsy outcomes from BI-RADS input attributes. The first approach is based on decision-tree learning and the second on case-based reasoning using an entropic distance measure. We have evaluated and compared the performance of both systems using ROC analysis and bootstrap sampling. Both systems outperform the diagnosis decisions of the physicians. Hence, both systems have the potential to reduce the number of unnecessary breast biopsies. However, these results have been obtained using a retrospective approach using relatively small case databases containing only a few thousand cases. Therefore further investigations are needed, and these results probably need to be confirmed in a larger prospective clinical study.

A comparison of the ROC performances of the two proposed CAD approaches with a state of the art ANN approach shows that the CBR approach significantly outperforms the ANN approach, which in turn significantly outperforms the decision-tree approach.

While there are significant but only small performance differences between the two proposed CAD approaches, there are fundamental differences regarding their properties. Hence, a decision to choose one over the other should be mainly based on evaluating these properties with regard to the intended clinical application. One fundamental difference is that while decision-tree systems are so-called eager learners, which induce a global model from the training data, their CBR counterparts are lazy learners, which directly use the training data to deduce a diagnosis suggestion. The advantage of a global decision-tree model is that it is a compact representation of the decision process, which can be understood simply by looking at the decision tree. Once the model is induced from the training data it can even be printed on a paper and applied to new cases without the need of a computer system. In contrast, the decision process of a CBR system is transparent to the physician in a different way. While it does not induce a global model, it induces local models for each new query case. These local models, which are simply based on the attributes and classification of the most similar database cases for a query case are also intelligible for the physician and can potentially model more complex decision processes than a global model.

The proposed CAD approaches both feature intelligible reasoning processes, which is an important requirement for the acceptance of CAD systems by physicians. One of the authors of this work is a physician and the head of a large screening mammography center. In his opinion, the reasoning processes of both the CBR and decision-tree-based approaches are easy to comprehend and hence probably useful in future clinical practice.

The CAD approaches have been evaluated on two different mammography databases (DDSM and UCI) and all three show better ROC performance on the DDSM than on the UCI database. While there are some obvious possible explanations for this finding (full field digital versus screen film mammograms and the fact that the databases have been collected at different institutions), the exact reason needs further investigations.

BI-RADS attributes are subjective to some degree, and different physicians trained at different radiology centers probably have slightly different styles of assigning a value for a given BI-RADS attribute (like the mass shape) to a suspicious region seen in a mammogram. Therefore we have investigated the performance of the CAD approaches trained on the DDSM set of masses M_{DDSM} and tested on the UCI set of masses M_{UCI} as well as vice versa. While the DDSM cases have been collected at radiological centers in the United States, the UCI data cases have been collected at a radiological center in Europe. The results of this "cross-center" validation show that both the two proposed as well as the state of the art CAD approach generalize less well to cases from different case databases than to cases from the same case database. While we assume that this effect is due to the fact that BI-RADS attributes are subjective to a certain degree, this needs further investigation. The proposed CBR approach shows the best generalization performance, which we assume is due to its local, implicit model of the data. However, this, too, is a point for further research.

We have evaluated the CAD approaches on public databases that contain only mammography BI-RADS input attributes. However, physicians base their diagnosis decision in many cases on the results of additional examinations using ultrasound or MRI systems. In the last few years, BI-RADS standard lexicon descriptions for both ultrasound and MRI breast examinations have been defined. Hence, a promising future extension for our CAD systems is to collect associated BI-RADS descriptions for mammography, ultrasound, and MRI examinations and to apply both CAD approaches to these extended attribute sets.

ACKNOWLEDGMENT

This work was supported by the Bayerische Forschungsstiftung within the scope of the project Mammo-iCAD.

[a]Electronic mail: matthias.elter@iis.fraunhofer.de

[1]A. Jemal, T. Murray, and E. Ward, "Cancer statstics, 2005," CA-Cancer J. Clin. **55**, 10–30 (2005).
[2]L. L. Humphrey, M. Helfand, B. K. Chan, and S. H. Woolf, "Breast cancer screening: A summary of the evidence for the U.S. Preventive Services Task Force," Ann. Intern Med. **137**, 347–360 (2002).
[3]L. Tabar, M. F. Yen, B. Vitak, H. H. Tony Chan, R. A. Smith, and S. W. Duffy, "Mammography service screening and mortality in breast cancer patients: 20-years follow-up before and after introduction of screening," Lancet **361**, 1405–1410 (2003).
[4]S. W. Duffy, L. Tabar, and H. H. Chen, "The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish counties," Cancer **95**, 458–469 (2002).
[5]L. Tabar, B. Vitak, H. H. Chen, M. F. Yen, S. W. Duffy, and R. A. Smith, "Beyond randomized controlled trials: Oraganized Mammographic screening substantially reduces breast carcinoma mortality," Cancer **91**, 1724–1731 (2001).
[6]D. B. Kopans, "the positive predictive value of mammography," AJR, Am. J. Roentgenol. **158**, 521–526 (1992).
[7]American College of Radiology, *Breast Imaging Reporting and Data System* ((BI-RADS©), Altas 2006).
[8]J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd, "Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon," Radiology **196**, 817–822 (1995).
[9]M. K. Markey, J. Y. Lo, R. Vargas-Voracek, G. D. Tourassi, and C. E. Floyd, "Perception error surface analysis: A case study in breast cancer diagnosis," Comput. Biol. Med. **32**, 99–109 (2002).
[10]C. E. Floyd, J. Y. Lo, and G. D. Tourassi, "Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions," AJR, Am. J. Roentgenol. **175**, 1347–1352 (2000).
[11]A. O. Bilska-Wolak and C. E. Floyd, "Investigating different similarity measures for a case-based reasoning classifier to predict breast cancer," Proc. SPIE **4322**, 1862–1866 (2001).
[12]A. O. Bilska-Wolak and C. E. Floyd, "Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with BI-RADS lexicon," Med. Phys. **29**, 2090–2100 (2002).
[13]A. O. Bilska-Wolak, C. E. Floyd, J. Y. Lo, and J. A. Baker, "Computer aid for decision to biopsy breast masses on mammography: Validation on New Cases," Acad. Radiol. **12**, 671–680 (2005).
[14]M. K. Markey, E. A. Fischer, and J. Y. Lo, "Bayesian networks of BI-RADS descriptors for breast lesion classifications," in *International Conference of the IEEE Engineering in Medicine and Biology Society* (San Francisco, California, 2004), pp. 3031–3034.
[15]M. K. Markeya, G. D. Tourassi, M. Margolis, and D. M. Delong, "Impact of missing data in evaluating artificial neural networks trained on complete data," Comput. Biol. Med. **36**, 516–525 (2006).
[16]M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer Jr., "The digital database for screening mammography," in *The Proceedings of the 5th International Workshop on Digital Mammogrphy*, Madison, WI (Medical Physics Publishing, Toronto, Canada, 2000).
[17]D. J. Newmann, S. Hettich, C. L. Blake, and C. J. Merz, "UCI repository of machine learning databases," Univeristy of California, Irvine, Dept. of Information and Computer Sciences, http://www.ics.edu/simmlearn/MLRepository.html, 1998.
[18]J. R. Quinlan, "Induction of decision trees," Mach. Learn. **1**, 81–106 (1986).
[19]J. R. Quinlan, *C4.5: Programs for Machine Learning* (Kaufmann, San Francisco, 1993).
[20]J. G. Cleary and L. E. Trig, "K*: An instance-based learner using an entropic distance measure," in *Proceedings of the 12th International Conference on Machine Learning* (Tahoe City, California, 1995), pp. 108–114.
[21]H. R. Lindman, *Analysis of Variance in Complex Experimental Designs* (Freeman, San Francisco, 1974).
[22]L. B. Lusted, "Signal detectability and medical decision-making," Science, **171**, 1217–1219 (1971).
[23]J. P. Egan, *Signal Detection Theory and ROC Analysis* (Academic, New York, 1975).
[24]A. K. Jain, R. C. Dubes, and C. C. Chen, "Bootstrap techniques for error estimation," IEEE Trans. Pattern Anal. Mach. Intell. **9**, 628–633 (1987).
[25]B. Efron and R. Tibshirani, *An Introduction to the Bootstrap* (Chapman and Hall, London, 1993).
[26]B. Efron and R. C. Tibshirani, "Improvements on cross-validation: The 632+ bootstrap method," J. Am. Stat. Assoc. **92**, 548–560 (1997).