Image processing and machine learning for fully automated probabilistic evaluation of medical images

Luka Šajn, Matjaž Kukar*

University of Ljubljana, Faculty of Computer and Information Science,
Tržaška 25, SI-1001 Ljubljana, Slovenia,
{luka.sajn,matjaz.kukar}@fri.uni-lj.si

Abstract

The paper presents results of our long-term study on using image processing and data mining methods in a medical imaging. Since evaluation of modern medical images is becoming increasingly complex, advanced analytical and decision support tools are involved in integration of partial diagnostic results. Such partial results, frequently obtained from tests with substantial imperfections, are integrated into ultimate diagnostic conclusion about the probability of disease for a given patient. We study various topics such as improving the predictive power of clinical tests by utilizing pre-test and post-test probabilities, texture representation, multi-resolution feature extraction, feature construction and data mining algorithms that significantly outperform medical practice. Our long-term study reveals three significant milestones. The first improvement was achieved by significantly increasing post-test diagnostic probabilities with respect to expert physicians. The second, even more significant improvement utilizes multi-resolution image parametrization. Machine learning methods in conjunction with the feature subset selection on these parameters significantly improve diagnostic performance. However, further feature construction with the principle component analysis on these features elevates results to an even higher accuracy level that represents the third milestone. With the proposed approach clinical results are significantly improved throughout the study. The most significant result of our study is improvement in the diagnostic power of the whole diagnostic process. Our compound approach aids, but does not replace, the physician's judgment and may assist in decisions on cost effectiveness of tests.

Keywords: machine learning, coronary artery disease, medical diagnostics, multi-resolution image parameterization, association rules, principal component analysis.

*Corresponding author: Matjaž Kukar, University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, SI-1001 Ljubljana, Slovenia, matjaz.kukar@fri.uni-lj.si

1 Introduction

Internal medicine and, in particular, cardiovascular medicine have a plethora of diagnostic imaging tests available to physicians to help identify various problems and abnormalities. Diagnostic imaging uses a variety of methods to create pictures of structures and function inside the human body. The choice of imaging technology depends on exhibited symptoms, the part of the body being examined, and its cost and availability. X-rays, computer tomography (CT) scans, nuclear medicine scans (including scintigraphy), magnetic resonance imaging (MRI) scans and ultrasound are all types of diagnostic imaging.

Many imaging tests are painless and easy. Some are slightly uncomfortable, as they require the patient to stay still for a long time inside a machine. Certain tests involve radiation, but these are generally considered safe because the dosage is very low. In some imaging tests, an implement (a tiny camera or other sensing device) is attached to a long, thin tube and inserted in the body. These procedures are quite unpleasant and often require anaesthesia. If possible, such procedures should preferably be replaced by less invasive ones.

Cardiovascular diseases, specifically coronary artery disease (CAD), are among the world's premier causes of mortality. Currently, cardiovascular disease diagnosis relies on diagnostic imaging tests that require expensive, specialized equipment and trained personnel (both technicians and physicians) for efficient operation. The goal of our research is to improve the diagnostics of CAD from a computational perspective. Our early research on this topic, conducted between 1995 and 1998 [1] , showed that machine learning methods may enable objective interpretation of available diagnostic images and, as a result, increase the accuracy and reliability of the diagnostic process. Experiments conducted with various learning algorithms showed that these were able to achieve performance levels comparable to those of clinicians. The algorithms were also extended to deal with non-uniform misclassification costs in order to perform ROC analysis to control the trade-off between sensitivity and specificity. The ROC analysis showed significant improvements of sensitivity and specificity of machine learning algorithms compared to the performance of clinicians. The predictive power of standard tests can thus be significantly improved using machine learning techniques.

The main problem with this study was that all data, including evaluation of diagnostic images, had to be provided by expert physicians, this causing a major data acquisition bottleneck and a certain reluctance to accept the procedure in everyday practice. In the current study, we aim to alleviate the data acquisition problem by introducing algorithms for completely automatic evaluation (parameterization) of diagnostic images and for suggesting the most useful (informative) resolutions [2]. In this process we also introduce a feature extraction method (principal component analysis) that helps in achieving excellent results [3]. Our paper briefly describes the methodology used and relates the results of both studies. It also introduces assessment of the diagnostic power and the value of ROC analysis.

1.1 Coronary artery disease

Coronary artery disease (CAD) is the result of the accumulation of atheromatous plaques within the walls of the coronary arteries that supply the myocardium with oxygen. While the symptoms and signs of coronary artery disease are noted in the advanced state of disease, most individuals with coronary artery disease show no evidence of disease for decades as the disease progresses before the first onset of symptoms finally arises (a sudden heart attack). As the coronary artery disease progresses, there may be near-complete obstruction of the coronary artery, severely restricting the flow of oxygen-carrying blood to the myocardium. Individuals with this degree of coronary artery disease typically have suffered from one or more myocardial infarctions, and may have signs and symptoms of chronic coronary ischemia, including symptoms of angina at rest and flash pulmonary edema.

The usual clinical process of coronary artery disease diagnostics consists of four steps:

1. evaluation of signs and symptoms of the disease and electrocardiogram (ECG) at rest
2. ECG testing during controlled exercise
3. myocardial scintigraphy
4. coronary angiography

In this process, the fourth diagnostic level (coronary angiography) is considered to be the best reference method by physicians. Given that this diagnostic procedure is invasive and unpleasant for patients, as well as relatively expensive, there is an incentive to improve diagnostic performance of earlier diagnostic levels, especially of myocardial scintigraphy [1, 4]. Approaches used for this purpose include applications of neural networks [5–7], expert systems [8], subgroup mining [9], statistical techniques [10], and rule-based approaches [11].

In our study, we focus on various aspects of improving the diagnostic performance of the third diagnostic level. Myocardial perfusion scintigraphy consists of acquiring a series of medical images using an inexpensive and non-invasive procedure when the patient is at rest and during a controlled physical exercise. Subsequently, expert physicians use their medical knowledge and experience to manually describe (parameterize) and evaluate the images, often with the help of image processing capabilities provided by various imaging software.

Our first study was based on patients' data compiled entirely by physicians – either by extracting data from medical records, or from test results (ECGs and scintigraphic images). Using these data, our machine learning algorithms showed excellent diagnostic accuracy and reliability in the diagnosis of coronary artery occlusions [4].

The current study presents an innovative alternative method to manual image evaluation. The method consists of automatic multi-resolution image parameterization, based on texture description with specialized association rules, coupled with

image evaluation with machine learning methods. Since this approach yields a large number of relatively low-level features (though much more informative than simple pixel intensity values), we have used additional dimensionality reduction techniques, either by throwing away some features (feature selection), or combining them into more informative, high-level features (feature construction). Our results show that multi-resolution image parameterization equals or even outperforms physicians in terms of the quality of image parameters. By using automatic image description parameters, diagnostic performance can be significantly improved with respect current clinical practice.

2 Methods

2.1 Image parameterization

Image parameterization is a technique for describing bitmapped images with numerical parameters - features or attributes. Traditionally, popular image features are first- and second-order statistics, structural and spectral properties, and several others. Image parameterization is used in quality control, identification, image grouping, surveillance, image storage and retrieval, and image querying. Over the past few decades, image parameterization has been extensively applied to medical domains where texture classification is closely related to diagnostic process [12]. This complements medical practice, where manual image parameterization (evaluation of medical images by expert physicians) frequently plays an important role in diagnostic process.

Images in digital form are normally described with spatially complex data matrices. Such data, however, are insufficient to distinguish between the predefined image classes. Determining image features that can discriminate between observed image classes is a difficult task for which several algorithms exist [13]. They transform the image from the matrix form into a set of numeric or discrete features (parameters) that convey useful high-level (compared to simple pixel intensities) information for discriminating between classes.

For the purposes of diagnosis from medical images, structural description seem most appropriate [14]. Structural representations have several good properties like invariance to global brightness and invariance to rotation. To obtain structural descriptions, we applied spatial association rules to textures using the ArTex algorithm (described in Sec. 2.3) Association rules algorithms can be used for describing textures if an appropriate texture representation formalism is used. Association rules capture structural and statistical information and are very convenient to identify the structures that occur frequently and have most discriminative characteristics.

2.2 Image classification with machine learning methods

Provided that medical images are described with informative numerical attributes, various machine learning algorithms can be used [15] to generate a classification system (classifier) for patient diagnosis. Although many machine learning methods are available, we decided to use decision trees, naive Bayesian classifiers, Bayesian networks, K-nearest neighbors, and support vector machines based on our previous experience with medical diagnostics [1]and their use in other studies (e.g. [16, 17]).

Our early work in the problem of diagnosing the coronary artery disease from myocardial scintigraphy images [2] indicates that the naive Bayesian classifier gives the best results. Our results are consistent with several other studies [18, 19] that also find that the naive Bayesian classifier frequently outperforms other, often much more complex, classifiers in medical diagnoses. In addition, many authors have established feature subset selection as a necessary step before decision tree induction [20, 21], therefore this must be taken into account when classifying images.

The performance of a diagnostic test is described with diagnostic accuracy, sensitivity, and specificity:

$$
\begin{aligned}
accuracy &= \frac{\#true\ positives\ +\ \#true\ negatives}{\#all\ patients} \\
sensitivity &= \frac{\#true\ positives}{\#all\ patients\ with\ the\ disease} \\
specificity &= \frac{\#true\ negatives}{\#all\ patients\ without\ the\ disease}
\end{aligned}
$$

The *true positives* are all patients with the disease and a positive test result, whereas the true negatives are all patients without the disease and negative test result.

2.3 ArTex and ARes algorithms

For efficient heart-scintigraphy image classification we need an appropriate image parameterization tool. The second study introduces the ArTex (Association rules for Textures - ArTex) [22] algorithm for parameterizing textures with association rules belonging to structural parameterization algorithms. The ArTex algorithm gives a texture representation, which is an appropriate formalism that allows straightforward application of association rules algorithms. This representation has several good properties like invariance to global lightness and invariance to rotation. Association rules capture structural and statistical information and are very convenient to identify the structures that occur most frequently and have the highest discriminative power.

Initially, the ArTex algorithm was used for texture classification. We have later discovered that it also performs well in heart-scintigraphy despite the fact that scintigraphy images do not exhibit a pattern.

The obtained high quality image parameters can be used to describe images with a relatively small number of features, which allows their use in machine learning process. Images of patients with known confirmed diagnosis can be used as learning data that, in conjunction with the applied machine learning methods, produces reliable decision support tools (classifiers) for the diagnostic problem at hand. In order to justify the use of the ArTex algorithm, its performance was compared to the performance of three other image parameterization algorithms (Haar wavelets [23], Laws filters [24] and Gabor filters [25]).

The current study also introduces the improvement of the parameterization with a multi-resolution approach. From our experiments with synthetic data, we have observed that using parameterization-produced features at several different resolutions usually improves the classification accuracy of machine learning classifiers [26]. This parameterization approach is very effective in analyzing myocardial scintigraphy. The algorithm ARes (ArTex with resolutions - ARes) for selecting the resolution set yields more informative parameterization attributes when combining the parameters from the proposed resolutions. The idea of the ARes algorithm derives from the SIFT (Scale Invariant Feature Transform) algorithm [27]. ARes was designed especially for structural image parameterization algorithms, specifically for the ArTex algorithm. ArTex and ARes are independent of the used machine learning algorithm.

A detailed presentation of both ArTex and ARes algorithms is given in [26]. The obtained texture parameters are subsequently used for image classification with machine learning methods [15].

2.4 Dimensionality reduction with principal component analysis

Dimensionality reduction is a mapping from a multidimensional space into a space of fewer dimensions. It is often the case that data analysis can be carried out in the reduced space more accurately than in the original space. More formally, the dimensionality reduction problem can be stated as follows: given the a-dimensional random variable $\mathbf{x} = (x_1, \ldots, x_a)$ find a lower dimensional representation of it, $\mathbf{s} = (s_1, \ldots, s_k)$ with $k < a$, that captures the content in the original data, according to some criterion.

Principal components analysis (PCA) is a linear transformation that chooses a new coordinate system for the data such that the greatest variance by any projection of the data set lies on the first axis (called the first principal component), the second greatest variance on the second axis, and so on [28]. PCA can be used for reducing dimensionality in a dataset while retaining those characteristics of the dataset that contribute most to its variance by eliminating the lesser principal components (by a more or less heuristic decision).

PCA is sometimes used to extract features directly from images in matrix form, where pixel intensity values are used as primary features. Our experiments with using such a feature extraction on CAD images produced such dismal results of machine learning (on par with a simple majority classifier) that we were discour-

aged to further pursue in this direction. So in the case of CAD diagnostics from scintigraphic images, several thousands of ArTex/ARes-generated image features are used as an input for PCA.

In our current study, we use PCA to reduce the high number of ArTex/ARes-generated image features (several thousands), to more manageable levels (a few tens of compound attributes that explain most of data variance).

3 Materials

In our early (1994) study, we used a dataset of 327 patients (250 males, 77 females) selected from a population of approximately 4000 patients examined at the Nuclear Medicine Department between 1991 and 1994. All selected patients had complete diagnostic procedures (all four levels) [29], consisting of clinical and laboratory examinations, exercise ECG, myocardial scintigraphy, and coronary angiograph. The features from the ECG and scintigraphy data were extracted manually by clinicians. Angiography confirmed the disease in 229 cases and excluded it in 98 cases. 162 patients had suffered a recent myocardial infarction. Our experiments were conducted on four problems. They differ in the amount of clinical and laboratory data (attributes) available for learning, corresponding to different diagnostic levels (Table 1).

Table 1: Old (1994) and new (2006) CAD data for different diagnostic levels. Of the attributes belonging to the coronary angiography diagnostic level, in the new study only the final diagnosis – the two-valued class – was used in experiments.

Diagnostic level	study	Number of attributes		
		Nominal	Numeric	Total
1. Signs and symptoms	1994	23	7	30
	2006	22	5	27
2. Exercise ECG	1994	7	9	16
	2006	11	7	18
3. Myocardial scintigraphy	1994	22	9	31
(+9 image series)	2006	8	2	10
4. Coronary angiography	1994	1	6	7
	2006	1	6	7
Class distribution	1994	98 (29.97%) CAD negative 229 (70.03%) CAD positive		
	2006	129 (46.40%) CAD negative 149 (53.60%) CAD positive		

In our recent (2006) CAD study we use a newer dataset of 288 patients who completed clinical and laboratory examinations, exercise ECG, myocardial scintigraphy (including complete image sets) and coronary angiography because of sus-

pected CAD. The features from the ECG an scintigraphy data were extracted manually by the clinicians. Ten patients were later excluded for data pre-processing and calibration required by ArTex/ARes, so only 278 patients (66 females, 212 males, average age 60 years) were used in actual experiments. In 149 cases the disease was angiographically confirmed and in 129 cases it was excluded. The patients were selected from a population of several thousands patients who were examined at the Nuclear Medicine Department between 2001 and 2006. Again we selected only the patients with complete diagnostic procedures (all four levels), and for whom the imaging data was readily accessible. Some characteristics of the dataset are shown in Table 1.

Although both data sets were collected in exactly the same way, there is a significant difference in class distributions between the two sets (see Table 1). Due to improved diagnostic clinical capabilities as well as an ageing population, the patients included in our recent (2006) study are much more difficult to diagnose. Although the total number of patients examined for CAD is increasing, an increasing number of patients is being reliably diagnosed with less invasive diagnostic tests. The population in our recent (2006) study therefore consists of patients that defy reliable diagnostics on lower diagnostic levels, and thus represents a challenge even for expert physicians.

Several patients in our 2006 dataset had already undergone cardiac surgery or dilatation of coronary vessels. This clearly reflects the situation in Central Europe with its ageing population. Our results are therefore not applicable to the general population, and vice versa, general findings only partially apply to our population.

The myocardial scintigraphy group of attributes consists of evaluation of myocardial defects (no defect, mild defect, well defined defect, serious defect) that could be observed in images either while resting or during a controlled exercise. They are assessed for four different myocardial regions: LAD, LCx, and RCA vascular territories, as well as ventricular apex. Additional two attributes concern effective blood flow and volumes in myocardium: left ventricular ejection fraction (LVEF) and end-diastolic volume (EDV).

In our clinical practice, four expert physicians regularly assess myocardial scintigraphy images. They estimate the level of coronary artery congestion and produce attribute values for different myocardial regions. The final diagnosis summarizes the obtained attribute values. It is difficult to precisely describe how the attribute values are determined, as it is based on years of experience and medical knowledge of the myocardium. As an important step in data pre-processing, and to insure reliability, an additional expert physician re-evaluated all images. Only images whose original and retrospective diagnoses were in accord were retained for the experiments.

3.1 Scintigraphic images

Scintigraphic images were obtained using the General Electric eNTE- GRA SPECT camera, producing grayscale images with a 64 x 64 8-bit pixel resolution. Images

were obtained while the patient was at rest and following a controlled exercise, producing a total of 64 images. Due to patient's movements and partial obscuring of the heart muscle by other internal organs, these images were not suitable for further use without heavy pre-processing. For this purpose, the ECToolbox workstation software [30] was used to generate a series of 9 polar map images for each patient. Polar maps were chosen because previous work in this field [31] had shown that they have useful diagnostic value. Polar images, usually referred to as a bull's-eye plot, present the short axis section as rings of increasing diameter, with the apex at the center and the base of the heart at the periphery. This allows quick assessment of the number and the area of any defects at stress and rest. Comparison with a database of normal images highlights areas of reduced activity which meet a predefined criterion for significance, and subtraction images are produced to illustrate the extent of reversibility. The 9 polar map images for each patient consist of the following images [30]:

- three raw images (the stress image, and the rest image, and the reversibility image, calculated as a difference between normalized rest and stress images);
- three blackout (defect extent) images (the stress image and the rest image, both compared to the respective database of normal images, and suitably processed). Again the reversibility image is calculated as a difference between normalized rest and stress blackout images;
- three standard deviation images that show relative perfusion variance when compared to the respective database of normal images.

An example of polar map images for three patients is shown in Figure 1. First patient (A) exhibits very clear manifestation of CAD. The second patient (B) represents a moderate manifestation of CAD whereas the third patient (C) exhibits atypical signs of the disease.

Unfortunately, in most cases (and especially in our specific population) the differences between images taken during exercise and at rest are not as clear-cut as shown in Figure 1. Interpretation and evaluation of scintigraphic images therefore requires considerable knowledge and experience of expert physicians. Although specialized tools such as the ECToolbox software can aid in this process, they still require a lot of training and in-depth medical knowledge for evaluation of results.

4 Results

4.1 Experimental methodology

To objectively evaluate the proposed methodology, experiments were performed on CAD images as well as synthetic data in the following manner. First, ten learning examples (images or sets of nine [1] images for CAD) were excluded for data pre-

[1] Physicians observe typical polar maps taken after exercise, at rest, and their difference. For each type the raw image, the blackout image, and the standard deviation image is used. ($3 \times 3 = 9$ images)

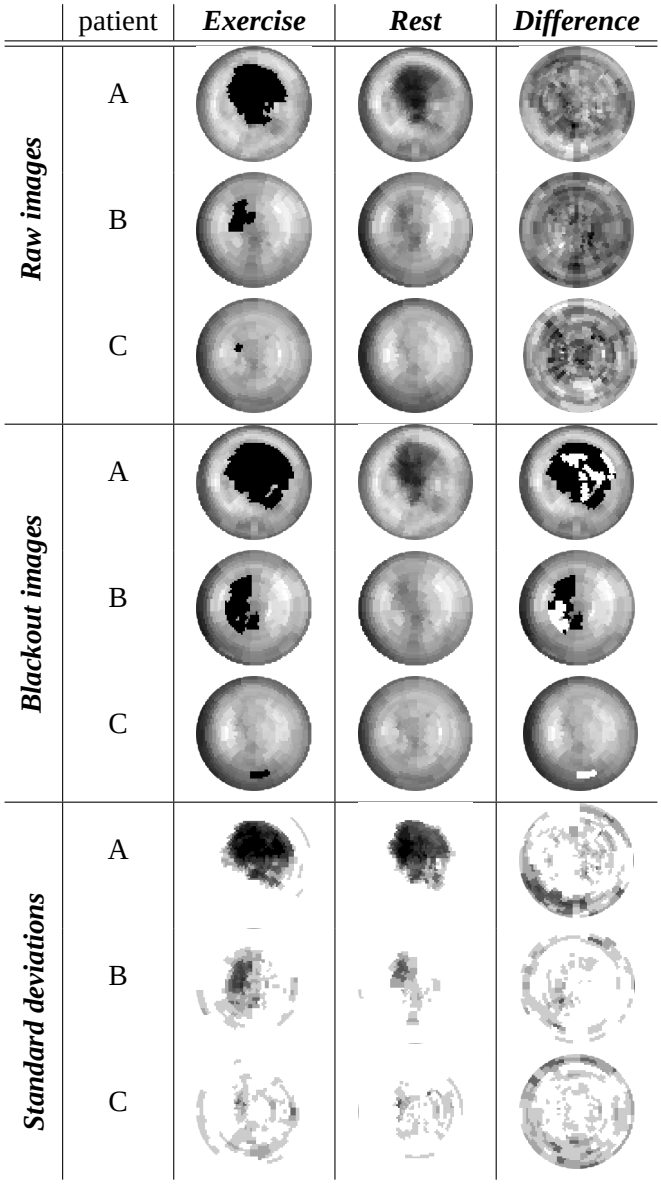


Figure 1: Typical polar maps taken after exercise, at rest, and their difference for three patients exhibiting severe (A), moderate (B) and atypical (C) manifestation of CAD. The first three rows show raw images, the second three show blackout images, and the last three show standard deviation images. Black regions indicate insufficiently perfused cardiac tissue (a potential defect).

processing and calibration of ArTex/ARes. Images from the remaining examples were parameterized and only the obtained parameters were subsequently used for evaluation. Further testing was performed in the ten-fold cross-validation setting: at each step 90% of examples were used for building a classifier and the remaining

10% of examples were used for testing.

For CAD data diagnostics, the set of parameters generated from the set of nine polar map images for each patient were reduced by extracting the ten most informative parameters using either feature extraction or feature selection methods. Feature extraction consisted of applying PCA to the full set of parameters and retaining the 10 best principal components that together accounted for not less than 70% of data variance. Feature selection consisted of applying ReliefF [2] [32] attribute quality estimation and retaining only 10 best ArTex/ARes generated attributes. In addition, ten of the best attributes provided by physicians were also used, as estimated by ReliefF.

In each cross-validation step the real-valued attributes were discretized in advance using the Fayyad-Irani [33] algorithm if the applied method (such as the naive Bayesian classifier) required only discrete attributes.

We applied four popular machine learning algorithms: naive Bayesian classifier, tree-augmented Bayesian network, support vector machine (SMO using RBF kernel), and J4.8 (C4.5) decision tree. We performed experiments with both Weka [34] and Orange [35] machine learning toolkits. For CAD diagnostics, aggregated results of the coronary angiography (CAD negative/CAD positive) were used as the class variable. The results of clinical practice were validated by careful blind evaluation of images by an independent expert physician. Differences between physician and machine learning results were evaluated for statistical significance by using McNemar's test [36].

4.2 Validation of the proposed approach

A thorough performance overview of the ArTex/ARes combination and comparison with the SIFT [27] algorithm and geometrical resolutions on synthetic data was published in [26]. Tests were performed on eight synthetic datasets that are publicly available [37, 38] and commonly used in testing the parameterization and classification algorithms. Performance was assessed with Friedman's rank tests [36]. In seven of eight datasets ArTex/ARes significantly outperformed other methods. In the remaining case the performance difference was not statistically significant $(p < 0.05)$.

The proposed ArTex/ARes combination was also used in the bone scintigraphy diagnostics, where significant diagnostic improvements were achieved [14].

4.3 Results in CAD diagnostics

As described in Secs. 3 and 4.1, out of the 288 patients, 10 were excluded for data preprocessing and calibration required by ArTex/ARes. These patients were not used in further experiments. The remaining 278 patients with 9 images each

[2] ReliefF is an advanced algorithm for feature selection, as it is not "near-sighted" and can be used in real-world domains (improves certainty of estimates, deals with noisy and missing data and solves multi-class problems).

were parameterized for three resolutions in advance. ARes proposed three[3] resolutions: $0.95\times$, $0.80\times$, and $0.30\times$ of the original resolution, producing together 2944 additional attributes (features, parameters). Since this number is too large for most practical purposes, it was reduced either by applying feature selection (with ReliefF) or by feature extraction (with PCA).

The ReliefF algorithm [32] was used to evaluate all 2944 features by assigning each a numerical value. Features were ranked according to their relevance, and only the topmost (most relevant) 10 features were used in subsequent experiments.

We also performed some experiments with other image parameterization approaches such as wavelet and DFT transform (Haar and Laws), Gabor filters and SIFT features; they, however, invariably performed significantly worse than ArTex/ARes [26].

Experimental results are compared with diagnostic accuracy, specificity and sensitivity of expert physicians after evaluation of scintigraphic images (Table 2). The results of clinical practice were validated by careful blind evaluation of images by the expert physician.

For machine learning experiments we considered five different settings that are described in more detail in subsequent sections:

1. evaluation of machine learning methods only on physician-provided attributes (Section 4.3.1, Table 2),

2. evaluation of all ArTex/ARres-generated attributes (Section 4.3.2, the first half of Table 3),

3. evaluation of all ArTex/ARres-generated attributes together with all attributes provided by physicians, (Section 4.3.2, the second half of Table 3),

4. evaluation of 10 best attributes (accounting for 70% of data variance) extracted by either ReliefF (the first half of Table 4) or PCA (the first half of Table 5) from ArTex/ARres-generated attributes (Section 4.3.3),

5. evaluation of the same 10 best attributes extracted by either ReliefF (the second half of Table 4) or PCA (the second half of Table 5) in conjunction with 10 best attributes provided by physicians, as estimated by the ReliefF algorithm (Section 4.3.3).

4.3.1 Contribution of machine learning methods on original attributes provided by physicians.

In this setup, only attributes provided by physicians were used for learning classifiers and subsequent classification in 10-fold cross validation setting. Thus we can evaluate the contribution of machine learning methods alone. From Table 2 we can see that machine learning algorithms are approximately on the level of expert physicians when evaluating the original data, as collected by physicians. The

[3] A resolution of $0.30\times$ means $0.30 \cdot 64 \times 0.30 \cdot 64$ pixels instead of 64×64 pixels

Table 2: Diagnostic results of the physicians compared with results of machine learning classifiers obtained from the original attributes, as extracted by physicians. Results (classification accuracies) that are significantly $(p < 0.05)$ different (better or worse) from clinical results are emphasized.

	All attributes provided by physicians		
	Accuracy	Specificity	Sensitivity
Clinical	64.00	71.10	55.80
Naive Bayes	**68.34**	69.80	67.10
Bayes Net	**67.14**	68.20	66.70
SMO (RBF)	65.10	62.80	67.10
J4.8	**57.19**	53.50	60.40

Naive Bayesian classifier achieves significantly higher classification ($(p < 0.05)$) accuracy and slightly (insignificantly) lower sensitivity than physicians, while the J4.8 decision tree achieves significantly lower classification accuracy. However, for physicians, improvements of specificity are more important than improvements of sensitivity or overall classification accuracy, since increased specificity decreases the number of unnecessarily performed higher-level diagnostic tests, and consequently shorter waiting times for truly ill patients. Unfortunately, no applied machine learning method attained this goal at this stage.

4.3.2 Evaluation of all attributes generated by ArTex/ARres

To establish the adequacy of the multi-resolution approach, we first examine results obtained by using all the parameters provided by ArTex/ARes (Table 3). Results show only a slight improvement of classification accuracy with respect to machine learning on clinical data in both experimental setups (image attributes only, and both image and physicians' attributes).

In case of using only image attributes (upper half of the Table 3) we have a truly automated approach where diagnosis is proposed without any physician involvement. While the experimental result by itself look very nice, as they significantly improve diagnostic accuracy with respect to physicians, they do not improve in all three criteria (diagnostic accuracy, specificity, and sensitivity).

From Table 3 we can also note that some machine learning algorithms — especially decision trees and surprisingly SMO[4] have some trouble handling a huge number (2944) of additional attributes with only 278 learning examples. This leads to overfitting the learning data and reduction of diagnostic performance. When using all 2944 attributes, the naive Bayesian classifier produced best results – significantly better than physicians as well as other tested machine learning algorithms. Using these 2944 attributes together with the physician-provided attributes again

[4]Support vector machines are supposed to perform well on high-dimensional data.

Table 3: Experimental results of machine learning classifiers on parameterized images obtained by using all available ArTex/ARes attributes as well those provided by physicians. Results in diagnostic accuracy that are significantly $(p < 0.05)$ different from clinical results are emphasized.

	All image attributes		
	Accuracy	Specificity	Sensitivity
Naive Bayes	**70.14**	68.50	72.10
Bayes Net	**69.20**	68.10	70.30
SMO (RBF)	61.15	58.10	63.80
J4.8	59.71	63.80	55.00
Clinical	64.00	71.10	55.80
	All image and physicians' attributes		
	Accuracy	Specificity	Sensitivity
Naive Bayes	**70.50**	69.10	72.10
Bayes Net	**69.80**	68.30	71.40
SMO (RBF)	**69.40**	69.80	69.10
J4.8	65.10	60.50	69.10

slightly improves the classification results. This improvement was significant in two of three cases. Especially notable is the improvement in the previously low-performing J4.8 (decision trees) and SMO, as they clearly cannot extract all the available knowledge from a large set of individual, possibly correlated attributes.

It is reasonable to assume that physicians' attributes are considerably more complex and much more informative than simple numerical features provided by ArTex/ARes. From machine learning and data mining theory [15] we know that machine learning algorithms benefit considerably by using dimensionality reduction techniques. Specifically, extraction of new, less numerous possibly uncorrelated and more informative composite features usually contribute to more successful machine learning (in terms of higher classification accuracy and larger area under ROC curve – AUC).

4.3.3 Evaluation of best attributes extracted by PCA from ArTex/ARres-generated attributes

In this setting, we either extracted 10 best principal components (linear combinations of original ArTex/ARes attributes) by PCA, or selected 10 best original attributes with ReliefF from the set of 2944 ArTex/ARes attributes. We also enriched the data representation by using the same number (10) of best physicians' attributes as evaluated by ReliefF and compared with the results of machine learning.

Tables 4 and 5 and Figures 2 and 3 depict the results. It is gratifying to see that without any special tuning of learning parameters, the results are in all cases significantly better than the results of physicians in terms of classification (diagnostic)

accuracy. Especially good results are that of the naive Bayesian classifier (Table 5), that improve in all three criteria: diagnostic accuracy (by 17.3%), sensitivity (by 23.4%) and specificity (by 12.6%). Another interesting issue is that including the best physician-provided attributes does not necessarily improve diagnostic performance (SMO, J4.8 in Table 5). It seems that there is some level of redundancy between physicians' and principal components generated from ArTex/ARes attributes that bothers some methods more than the others. Consequently, it seems that some of automatically generated attributes are (from the diagnostic performance point of view) at least as good as the physician-provided ones, and may therefore represent new knowledge about CAD diagnostics.

Table 4: Experimental results of machine learning classifiers on parameterized images obtained by selecting only the best 10 attributes from either ArTex/ARes (also combined with 10 best attributes provided by physicians). Classification accuracy results that are significantly better $(p < 0.05)$ than clinical results are emphasized.

	ArTex/ARes		
	Accuracy	Specificity	Sensitivity
Naive Bayes	**69.4%**	58.9%	78.5%
Bayes Net	**69.4%**	58.9%	78.5%
SMO (RBF)	**71.9%**	65.1%	77.9%
J4.8	**70.9%**	61.2%	79.2%
Clinical	64.0%	71.1%	55.8%
	ArTex/ARes+physicians		
	Accuracy	Specificity	Sensitivity
Naive Bayes	**74.8%**	70.5%	78.5%
Bayes Net	**74.4%**	69.8%	78.5%
SMO (RBF)	**73.4%**	65.9%	79.9%
J4.8	**68.0%**	63.6%	71.8%

4.4 Explaining the meaning of the new attributes.

Since the diagnostic performance of machine learning methods turned out to be significantly better than that of expert physicians, a question whether some new knowledge had been induced from images is imminent. To gain some insight into the new attributes, we performed an analysis of associatons between best physician-provided and Artex/ARes-generated attributes.

– When reviewing associations between in ARES-generated and physicians' attributes, several highly confident (more than 99%) rules of shape *"IF sex=Male AND value of ARES attribute is high THEN lbbb is absent"* surfaced. Left bundle branch block (lbbb) is a cardiac conduction abnormality seen on the electrocardiogram (ECG) and if present, may cause false readings of scintigrams. It seems that some generated attributes describe the absence of this anomaly in

Table 5: Experimental results of machine learning classifiers on parameterized images obtained by selecting only the best 10 attributes from PCA on ArTex/ARes (also combined with 10 best attributes provided by physicians). Classification accuracy results that are significantly better $(p < 0.05)$ than clinical results are emphasized.

	PCA on ArTex/ARes		
	Accuracy	Specificity	Sensitivity
Naive Bayes	**81.3%**	83.7%	79.2%
Bayes Net	**71.9%**	69.0%	74.5%
SMO (RBF)	**78.4%**	76.0%	80.1%
J4.8	**75.2%**	78.3%	72.5%
Clinical	64.0%	71.1%	55.8%
	PCA on ArTex/ARes+physicians		
	Accuracy	Specificity	Sensitivity
Naive Bayes	**79.1%**	82.9%	75.8%
Bayes Net	**79.1%**	83.7%	75.2%
SMO (RBF)	**76.6%**	77.5%	75.8%
J4.8	**74.1%**	73.6%	74.5%

male patients. Another interesting type of rules associates (although with lower confidence about 70%) values of some ARES attributes with results of scintigraphy during rest and controlled exercise, and thus supports (or even improves on) physicians' findings. An example of such a rule is *"IF no anomalous reading at rest AND value of ARES attribute is high THEN no anomalous reading during exercise"*.

– When reviewing associations between principal components and physicians' attributes, we found two rules associating two PCA components with low HDL level and diabetes with confidence of about 90%. There were also a few rules relating PCA attributes with scintigraphic results in LCx territory (both during rest and stress), also with confidence over 90%.

In the analysis of graphical representation of causal networks (Figure 4, results shown for PCA+physicians only), causal relations are indicated by edges in the graph between scintigraphic attributes describing test results in RCA and LCx territories during rest and stress, for both ARES- and PCA-on-ARES- generated attributes. Although there are some similarities between physicians and association rules or causal networks, it seems that the new attributes convey considerably different diagnostic information and may therefore contribute new medical knowledge.

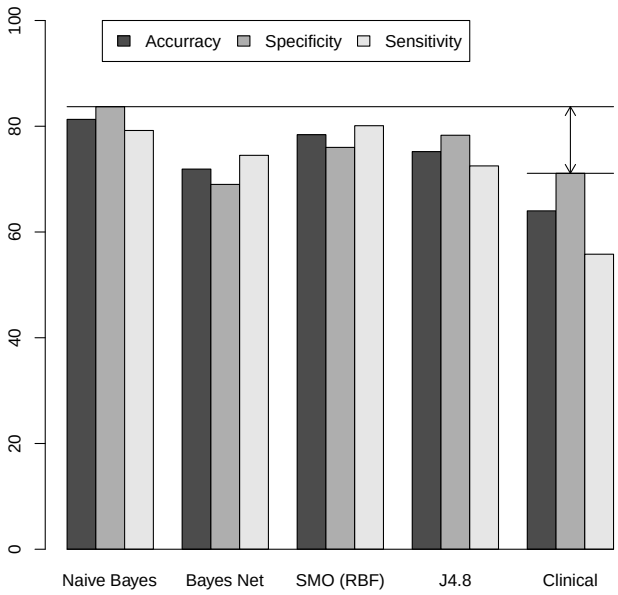


Figure 2: Comparison of clinical results and results of machine learning classifiers on parameterized images from Table 5.

4.5 Assessing the diagnostic power

In order to assess the diagnostic power of our compound approach, we applied the post-test probability calculation method as described in [40] for assessing reliability (probability of a correct diagnosis) of machine learning classifications in stepwise diagnostic process, To determine the pre-test probability we applied tabulated values (Table 6) as given by [41]. For each patient, the table was indexed by a subset of "signs and symptoms" attributes (age, sex, type of chest pain).

For both physicians and machine learning methods we calculated the post-test probabilities in the stepwise manner, starting from the pre-test probability and proceeding with evaluation of signs and symptoms, exercise ECG, and myocardial scintigraphy. For myocardial scintigraphy, physicians achieved 64% diagnostic accuracy, 71.1% specificity, and 55.8% sensitivity. For the reliability threshold of 90%, 52% of diagnoses could be considered as reliable (their post-test probability was higher than 90% for positive, or lower than 10% for negative diagnoses). On the other hand, naive Bayesian classifier achieved for myocardial scintigraphy 81.3% diagnostic accuracy, 83.7% specificity, and 79.2% sensitivity. For the reliability threshold of 90%, 69% of diagnoses could be considered as reliable. Improvement in 17% of reliable diagnostic accuracy is a result of 19% improvement for reliable positive diagnoses, and 16% for reliable negative diagnoses.

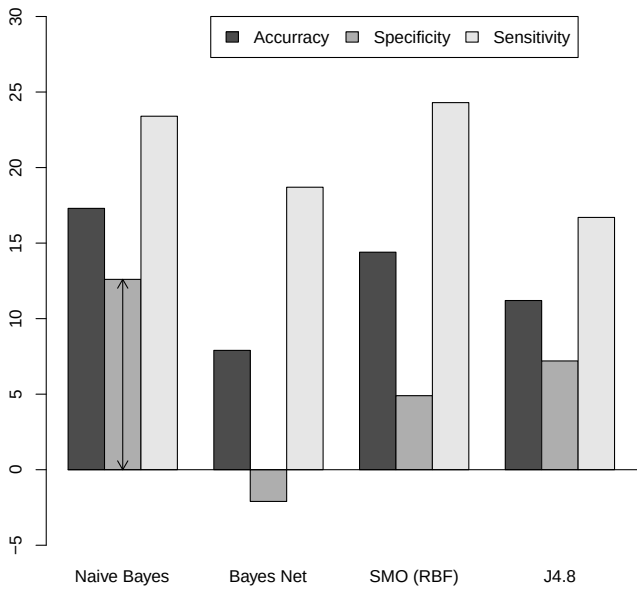


Figure 3: Improvements of machine learning classifiers on parameterized images from Table 5 relative to clinical results (baseline 0%).

Table 6: Pre-test probabilities for the presence of CAD.

Sex	Age	Asymptomatic patients	Nonang. chest pain	Atypical angina	Typical angina
Female	35-44	0.007	0.027	0.155	0.454
	45-54	0.021	0.069	0.317	0.677
	55-64	0.054	0.127	0.465	0.839
	65-74	0.115	0.171	0.541	0.947
Male	35-44	0.037	0.105	0.428	0.809
	45-54	0.077	0.206	0.601	0.907
	55-64	0.111	0.282	0.690	0.939
	65-74	0.113	0.282	0.700	0.943

We also depict results of both physicians' and automatic approach in ROC curves, obtained by varying reliability threshold between 0 and 1 (Figures 5(a) and 5(b)). A fully automatic approach (Naive Bayes on parameterized images) has considerably higher ROC curve than physicians, both for reliable positive (AUC=0.90 vs. 0.82) and reliable negative patients (AUC=0.91 vs. 0.83). Of improvements

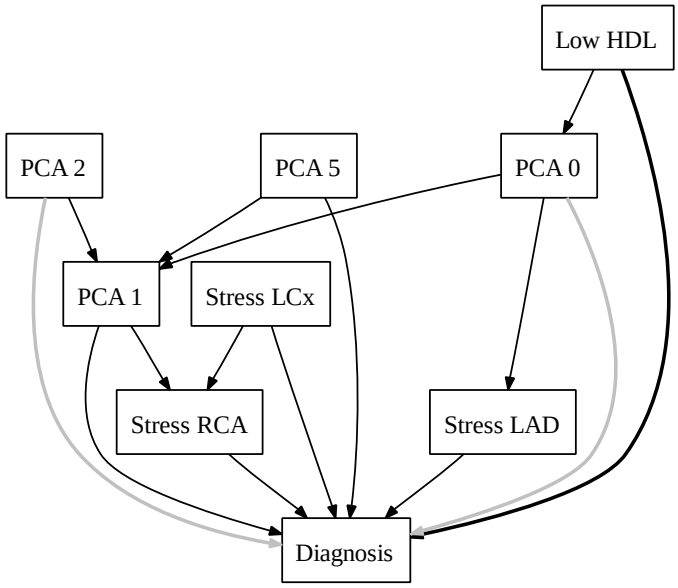


Figure 4: Causal network of best physician-provided and best four PCA-generated attributes (linear combination of ArTex generated attributes). Arrows represent cause-effect relations, as induced from learning data by the TAN algorithm [39]. Bold black line represent strong positive causality, whereas bold gray line represents negative causality.
Explanation of abbreviations: *LCx – left circumflex, LAD – left anterior descending and RCA – right coronary artery* $\rightarrow$ these abbreviations denote the location of the ischemia (restriction in blood supply); *HDL – the level of HDL cholesterol*.

in positive and negative reliable diagnoses, by far the more important is the 16% improvement for reliable negative diagnoses. The reason for this is that positive patients undergo further pre-operative diagnostic tests in any case, while for negative patients diagnostic process can reliably be finished on the myocardial scintigraphy level.

4.6 Summary of the results achieved through the study

Table 7 summarizes experimental results of our earlier (1994) study [4]. Compared to our current study, both diagnostic accuracy and sensitivity were considerably higher, whereas specificity was about the same. According to expert physicians' explanation this is a direct consequence of aging population and improved early diagnostic tests. Therefore, our current study comprises a population that is much more difficult to diagnose reliably.

In Table 8 we summarize results on the new (2006) study and compare them with results from clinical practice (Table 8, first row). It is obvious that machine learning algorithms have some trouble when handling a huge number (2944) of attributes, with only 278 learning examples (Table 8, second and third row). This can lead to overfitting the learning data and thus lower their diagnostic performance.

Only naive Bayesian classifier is significantly better than physicians when using all 2944 attributes. However, using these 2944 attributes together with the original attributes invariably improves the physicians' results, in two of three cases even significantly.

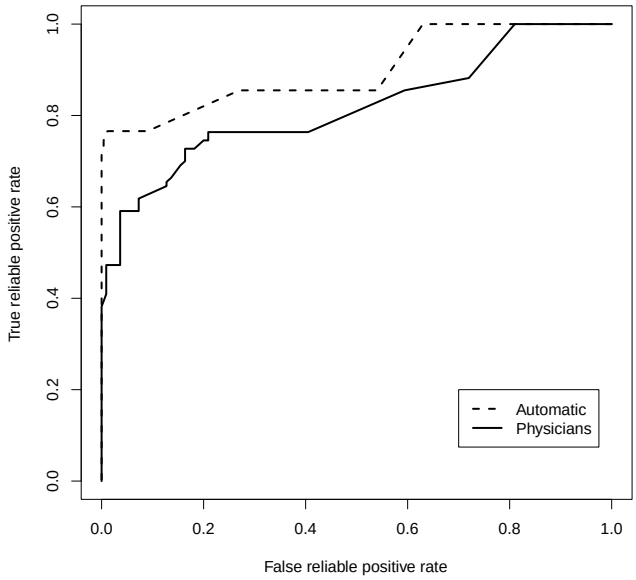
Table 7: Experimental results of our earlier (1994) CAD study compared to respective expert physicians' results.

Results of earlier (1994) study	Accuracy	Specificity	Sensitivity
Physicians	83%	85%	83%
Naive Bayes (scintigraphy)	90%	81%	94%
Naive Bayes (all attributes)	91%	81%	96%

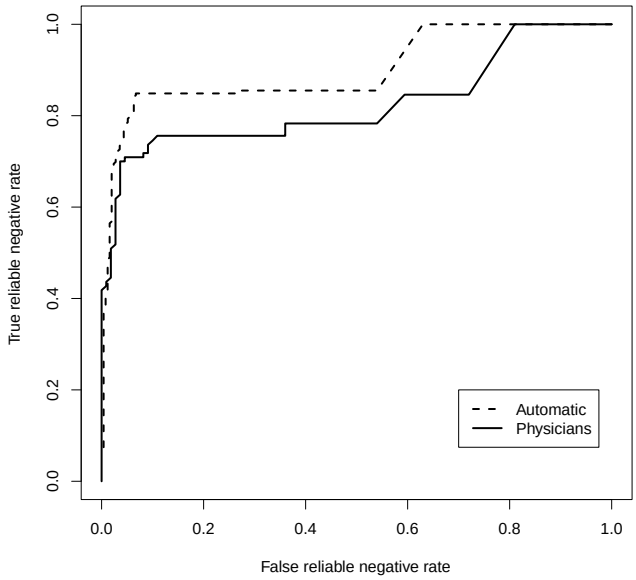
Most utilized learning algorithms benefit considerably from attribute filtering. In all cases the results are significantly better than the results of physicians. Especially good results are that of naive Bayesian classifier, which improves diagnostic accuracy, sensitivity and specificity (Table 8, fourth row). Attributes were filtered with the ReliefF algorithm [32, 42]. However, even better results are achieved by extracting higher-level, compound attributes with principal component analysis (Table 8, fifth row).

Table 8: Diagnostic performance of machine learning classifiers achieved over the study. Results (classification accuracies) that are significantly $(p < 0.05)$ better than clinical results are emphasized.

Study description	*Accuracy*	*Specificity*	*Sensitivity*
1. Clinical results	64.0%	71.1%	55.8%
2. Machine learning results on the original attributes as extracted by physicians.	All attributes provided by physicians		
	68.34	69.80	67.10
3. Results of machine learning on parameterized images obtained by using all available attributes.	All image (2944) and basic attributes		
	70.50	69.10	72.10
	All image attributes (2944)		
	70.14	68.50	72.10
4. Results of machine learning on parameterized images obtained by selecting only the best 200 attributes.	200 best image and basic attributes		
	74.10	79.80	69.10
	200 best image attributes		
	72.30	79.80	65.80
5. Results on parameterized images obtained by selecting the best 10 attributes from PCA on ArTex/ARes parameters	PCA on ArTex/ARes		
	81.3%	83.7%	79.2%
	PCA on ArTex/ARes + 10 best basic attributes		
	79.1%	82.9%	75.8%



(a) ROC curves for reliable positive diagnoses. AUCs are respectively 0.82 (for physicians) and 0.90 (for automated diagnostics).



(b) ROC curves for reliable negative diagnoses. AUCs are respectively 0.83 (for physicians) and 0.91 (for automated diagnostics).

Figure 5: Comparison of ROC curves, obtained by varying reliability threshold between 0 and 1, for reliable positive and negative diagnoses.

5 Discussion

A major bottleneck in clinical evaluation of medical imaging test results is that expert physicians need to be involved – by using their medical knowledge and experience as well as image processing capabilities provided by various imaging software – to manually describe (parameterize) and evaluate the images. We describe an innovative alternative to manual image evaluation - automatic multi-resolution image parameterization based on spatial association rules (ArTex/ARes) supplemented with feature selection or (preferably) feature extraction. Our results show that multi-resolution image parameterization equals or even betters the physicians in terms of diagnostic quality of image parameters. By using these parameters for building machine learning classifiers, diagnostic performance can be significantly improved with respect to the results of clinical practice. We also explore relations between newly generated image attributes and physicians' description of images. Our findings indicate that ArTex/ARes with PCA is likely to extract more useful information from images than the physicians do, as it significantly outperforms them in terms of diagnostic accuracy, specificity and sensitivity.

Utilizing machine learning methods can help interns or inexperienced physicians to reliably evaluate medical images and thus improve their diagnostic accuracy, sensitivity and specificity. From the practical use of described approaches two-fold improvements of the diagnostic procedure can be expected. Higher diagnostic accuracy (up to 17.3%) and sensitivity (up to 23.4%) represent a very considerable gain. Due to higher specificity of tests (up to 12.6%), fewer patients without the disease would have to be examined with the invasive and possibly dangerous coronary angiography. Together with higher sensitivity this would save money and shorten the waiting times of the truly ill patients. Also, new attributes generated by ArTex/ARes with PCA are invoking considerable interest from expert physicians, since they significantly contribute to increased diagnostic performance and may therefore convey some novel medical knowledge of the CAD diagnostics problem.

Finally, we need to emphasize again the caveat that the results of our current study are based on data from a significantly restricted population and therefore may not be generally applicable to the normal population or to all the patients coming to the Nuclear Medicine Department, University Clinical Centre Ljubljana, Slovenia.

5.1 Future work

The utilized combination of machine learning and image parameterization algorithms opens a new research area of multi-resolution image parameterization and could be utilized in several medical, industrial and other domains where textures or texture-like surfaces are classified. The resolution selection algorithm ARes can be improved with additional domain-specific resolution search refinements, and with heuristic methods for controlling selection of resolutions.

In our case study – the CAD diagnostics problem – we intend to concentrate

even more on improving the diagnostic performance of the myocardial perfusion scintigraphy and assess problem-dependent criteria for resolution quality. We will study in-depth relations between automatically generated and physician-provided attributes and try to establish the possible correspondence between them. Potential improvements of the parameterization and classification scheme will be used in the post-test probability estimation setting [40] for evaluating the reliability of machine-generated diagnoses.

Acknowledgements

This work was supported by the Slovenian Ministry of Higher Education, Science, and Technology. Special thanks to nuclear medicine specialist Ciril Grošelj at the University Medical Centre in Ljubljana for his help and support.

References

[1] M. Kukar, I. Kononenko, C. Grošelj, K. Kralj, and J. Fettich. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial Intelligence in Medicine*, 16 (1):25–50, 1999.

[2] M. Kukar, L. Šajn, C. Grošelj, and J. Grošelj. Multi-resolution image parametrization in sequential diagnostics of coronary artery disease. In R. Bellazzi, A. Abu-Hanna, and J. Hunter, editors, *Artificial intelligence in medicine*, pages 119–129, Berlin, Heidelberg, 2007. Springer.

[3] M. Kukar and L. Šajn. Improving probabilistic interpretation of medical diagnoses with multi-resolution image parameterization: a case study. In C. Combi, Y. Shahar, and A. Abu-Hanna, editors, *12th Conference on artificial intelligence in medicine*, pages 136–145. Springer, 2009.

[4] M. Kukar and C. Grošelj. Machine learning in stepwise diagnostic process. In W. Horn, Y. Shahar, G. Lindberg, S. Andreassen, and J. Wyatt, editors, *Proc. Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making*, pages 315–325, Aalborg, Denmark, 1999. Springer.

[5] J. S. Allison, J. Heo, and A. E. Iskandrian. Artificial neural network modeling of stress single-photon emission computed tomographic imaging for detecting extensive coronary artery disease. *The American Journal of Cardiology*, 95(2):178–81, 2005.

[6] B. A. Mobley, E. Schechter, and W. E. Moore. Predictions of coronary artery stenosis by artificial neural network. *Artificial Intelligence in Medicine*, 18(3):187–203, 2000.

[7] M. Ohlsson. WeAidU–a decision support system for myocardial perfusion images using artificial neural networks. *Artificial Intelligence in Medicine*, 30:49–60, 2004.

[8] E. V. Garcia, C. D. Cooke, R. D. Folks, C. A. Santana, E. G. Krawczynska, L. De Braal, and N. F. Ezquerra. Diagnostic performance of an expert system for the interpretation of myocardial perfusion spect studies. *The Journal of Nuclear Medicine*, 42(8):1185–91, 2001.

[9] D. Gamberger, N. Lavrač, and G. Krstačič. Active subgroup mining: a case study in coronary heart disease risk group detection. *Artificial Intelligence in Medicine*, 28 (1):27–57, 2003.

[10] P. J. Slomka, H. Nishina, D. S. Berman, C. Akincioglu, A. Abidov, J. D. Friedman, S. W. Hayes, and G. Germano. Automated quantification of myocardial perfusion spect using simplified normal limits. *Journal of Nuclear Cardiology*, 12(1):66–77, 2005.

[11] L. A. Kurgan, K. J. Cios, and R. Tadeusiewicz. Knowledge discovery approach to automated cardiac spect diagnosis. *Artificial Intelligence in Medicine*, 23(2):149–169, 2001.

[12] J. Fitzpatrick and M. Sonka. *Handbook of Medical Imaging, Medical Image Processing and Analysis*, volume 2. SPIE, Bellingham, 2000. ISBN 0-8194-3622-4.

[13] M. Nixon and A.S. Aguado. *Feature Extraction and Image Processing*. Academic Press, Elsevier, Amsterdam, 2^{nd} edition, 2008. ISBN 978-0-12372-538-7.

[14] L. Šajn and I. Kononenko. *Computational Intelligence in Medical Imaging: Techniques & Applications*, chapter Image segmentation and parametrization for automatic diagnostics of whole-body scintigrams, pages 347–377. CRC Press; Taylor & Francis Group, cop., Boca Raton, London, New York, 2009.

[15] I. Kononenko and M. Kukar. *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing, cop., Chichester, Great Britain, 2007.

[16] Y. Peng, B. Yao, and J. Jiang. Knowledge-discovery incorporated evolutionary search for microcalcification detection in breast cancer diagnosis. *Artificial Intelligence in Medicine*, 37(1):43–53, 2006.

[17] C.D. Katsis, Y. Goletsis, A. Likas, D.I. Fotiadis, and I. Sarmas. A novel method for automated EMG decomposition and MUAP classification. *Artificial Intelligence in Medicine*, 37(1):55–64, 2006.

[18] I. Kononenko. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7:317–337, 1993.

[19] I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 3:89–109, 2001.

[20] J. Jelonek and J. Stefanowski. Feature subset selection for classification of histological images. *Artificial Intelligence in Medicine*, 9(3):227–239, 1997.

[21] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. In *IEEE Intelligent Systems*, pages 380–385, 1998.

[22] M. Bevk and I. Kononenko. Towards symbolic mining of images with association rules : preliminary results on textures. *Intelligent data analysis*, 10(4):379–393, 2006.

[23] C.K. Chui. *An Introduction to Wavelets*. Academic Press, San Diego, 1992.

[24] K. I. Laws. *Textured image segmentation*. PhD thesis, Dept. Electrical Engineering, University of Southern California, Los Angeles, Calif, USA, 1980.

[25] S.E. Grigorescu, N. Petkov, and P. Kruizinga. Comparison of texture features based on gabor filters. *IEEE Trans. on Image Processing*, 11(10):1160–1167, 2002.

[26] L. Šajn and I. Kononenko. Multiresolution image parametrization for improving texture classification. *EURASIP Journal on Advances in Signal Processing*, 2008(1): 1–12, 2008.

[27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. ISSN 0920-5691.

[28] K. Pearson. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, page 559, 1901.

[29] M. Kukar, C. Grošelj, I. Kononenko, and J. Fettich. An application of machine learning in the diagnosis of ischaemic heart disease. In *Proc. Sixth European Conference of AI in Medicine Europe AIME'97*, pages 461–464, Grenoble, France, 1997.

[30] General Electric. ECToolbox Protocol Operator's Guide, 2001.

[31] D. Lindahl, J. Palmer, J. Pettersson, T. White, A. Lundin, and L. Edenbrandt. Scintigraphic diagnosis of coronary artery disease: myocardial bull's-eye images contain the important information. *Clinical Physiology*, 6(18), 1998.

[32] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53:23–69, 2003.

[33] U. M. Fayyad. Multi-interval discretization of continuous-valued attributes for classification learning. In R. Bajcsy, editor, *Proc. International Joint Conferences on Artificial Intelligence*, pages 1022–1027, New York, San Mateo, 1993. Morgan Kaufmann.

[34] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2^{nd} edition, 2005.

[35] J. Demšar, B. Zupan B, and G. Leban. Orange: From experimental machine learning to interactive data mining, white paper, http://www.ailab.si/orange (Accessed: 16 august 2009), 2004.

[36] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[37] P. Brodatz. *Textures - A Photographic Album for Artists and Designers*. Reinhold, Dover, New York, 1966.

[38] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllönen, and S. Huovinen. Outex - new framework for empirical evaluation of texture analysis algorithms. In *International Conference on Pattern Recognition*, volume 1, page 10701, Los Alamitos, CA, USA, 2002. IEEE Computer Society.

[39] M. Sahami. Learning limited dependence Bayesian classifiers. In *KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 335–338. AAAI Press, 1996.

[40] M. Olona-Cabases. The probability of a correct diagnosis. In J. Candell-Riera and D. Ortega-Alcalde, editors, *Nuclear Cardiology in Everyday Practice*, pages 348–357. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

[41] B. H. Pollock. Computer-assisted interpretation of noninvasive tests for diagnosis of coronary artery disease. *Cardiovascular Rev. Rep. 4*, pages 367–375, 1983.

[42] K. Kira and L. Rendell. A practical approach to feature selection. In D. Sleeman and P. Edwards, editors, *Proc. Intern. Conf. on Machine Learning*, pages 249–256, Aberdeen, UK, 1992. Morgan Kaufmann.