AN INTELLIGENT SYSTEM FOR THE EARLY DIAGNOSIS OF CORONARY ARTERY DISEASE

D. Frossyniotis, S. Vrettos, D. Vogiatzis & A. Stafylopatis,
C. Constantinopoulos, D.I. Fotiadis, & A. Likas
G. Potamias
A. Naka, Th. Tzimas & L.K. Michalis

13 – 2001

Preprint, no 13 – 01 / 2001

Department of Computer Science
University of Ioannina
45110 Ioannina, Greece

4th International Conference "Neural Networks and Expert Systems in Medicine and Healthcare" NNESMED 2001
20-22 June 2001, Milos Island, Greece

An Intelligent System for the Early Diagnosis of Coronary Artery Disease

D. Frossyniotis, S. Vrettos, D. Vogiatzis & A. Stafylopatis
Dept. of Electrical and Computer Engineering, National Technical University of Athens, GR 157 73 Zographou, Athens, Greece

C. Constantinopoulos, D.I. Fotiadis & A. Likas
Dept. of Computer Science, University of Ioannina, GR 451 10 Ioannina, Greece

G. Potamias
ICS/FORTH, Herakleion, Greece

A. Naka, Th. Tzimas & L.K. Michalis
Dept. of Cardiology, Medical School, University of Ioannina, GR 451 10 Ioannina, Greece

BSTRACT: In this paper we present and compare several approaches for the design of an intelligent system for the early diagnosis of coronary artery disease. The aim of our work is the development of an innovative non-invasive methodology for the early diagnosis and monitoring of coronary artery disease based on the evaluation of risk factors and vascular status indices. A database was set up consisting of 139 patients without medical history of coronary artery disease, each patient record containing 24 attributes. The data are preprocessed using Principal Component Analysis and several methods are used for classification, such as artificial neural networks (MLP, RBF), support vector machines (SVM), fuzzy neural networks (fuzzy-ARTMAP), probabilistic belief networks and decision tree learning algorithms. The problem of missing values in the input data is addressed and the diagnostic value of the proposed methods is discussed.

INTRODUCTION

Atherosclerosis, the principal cause of coronary, cerebrovascular and peripheral vascular disease, is responsible for more than 50% of all mortality and morbidity in westernised countries. The understanding of the pathophysiology of the atherosclerotic disease, the diagnosis in its early and reversible stages, the prevention of its development, the identification and effective modification of the risk factors and the treatment of e disease are of crucial importance. It is becoming increasingly accepted that the vascular endothelium, and its dysfunction, plays a pivotal role in the pathogenesis and development of atherosclerosis through its effects on vasoregulation, platelet and mocyte adhesion, vascular smooth muscle growth, and coagulation [1]. The endothelial dysfunction appears to have detrimental functional consequences as well as adverse long-term effects [2,3] and the relation between coronary risk factors and endothelial function is complex [4]. However, the association between risk factors and endothelial dysfunction is strengthened by observations that risk factor modification such as cholesterol lowering, smoking cessation, exercise, and oestrogen replacement improves endothelial function [5,6] even in the presence of atherosclerosis [7].

Risk factor assessment by the doctor during history taking is still the crude measure used to estimate the overall risk of a patient for coronary artery disease. Angiography remains the gold standard for the diagnosis of overt atherosclerosis. However, it is an invasive and costly procedure that cannot be used for the early diagnosis of atherosclerosis, screening of large population or the close follow-up of treatment. Other recently developed non-invasive methods suffering from similar problems are the computerised tomography for the detection of coronary calcification and magnetic resonance imaging of atheroma for the assessment of the function of the aorta [8,9].

The assessment of the endothelial function in both coronary and peripheral arteries has expanded as a research tool to identify potential factors and mechanisms related to the atherogenic process and investigate methods of treatment [6]. In addition, evidence of increased arterial stiffness (reduced aortic distensibility, increased aortic pulse wave velocity – which can be measured non-invasively using for example applanation tonometry) has been proposed as a convenient surrogate marker of atherosclerosis [10], but its use for early detection of endothelial dysfunction has been questioned [11]. On the other hand, the effects and impact of physical exercise on the patient with atherosclerotic disease has not yet been fully determined.

The development of an automated diagnosis system that will take all potential risk factors into account would benefit from the introduction of arterial stiffness index and probably endothelial health (change of arterial stiffness following exercise stimulated NO release). The diagnostic power of such a system will increase in the early stages of atherosclerotic disease when the damage is still reversible and will throw some light into the complex mechanism of the atherogenic process. A diagnostic system can be used for the evaluation of the importance and contribution of all risk factors in the diagnosis of the various stages of atherosclerosis in different disease

states and potentially help in the identification of the appropriate treatment.

THE PROBLEM AND THE DATASET

In our study, we present and compare several approaches for the design of an intelligent system for the early diagnosis of coronary artery disease (CAD). The aim of our work is the development of an innovative non-invasive methodology for the early diagnosis and monitoring of coronary artery disease based on the evaluation of risk factors and vascular status indices. The importance of the proposed approach lies in the data-driven method used for the design of the intelligent diagnostic system and the evaluation of different factors that can reduce the development of the disease.

A database was set up collecting data for patients of the Cardiology Department of the University Hospital. The database consists of 139 patients (35 normal / 104 pathological) without medical history of coronary artery disease. Each patient record is composed of 24 input attributes including exercise testing results, several risk factors for the coronary artery disease (such as age, sex, family history for coronary artery sease, smoking history, diabetes history, hypertension history, distribution of body fat, body mass index, physical condition, diet) and several measures such as lipid profile, plasma glucose, oxidative stress markers, lipid peroxidation products and indices of arterial stiffness (pulse wave velocity) [14-16].

Several classification methods are applied and compared for the design of an intelligent diagnostic system based on the association of the above set of symptoms and measurements. These approaches include computational intelligence techniques, such as artificial neural networks (MLP, RBF), support vector machines (SVM), fuzzy neural networks (fuzzy-ARTMAP), probabilistic belief networks and decision tree learning algorithms.

The focus of our study is on the prediction of the presence of CAD with the results of coronary angiography used as the golden standard. In our experiments, we preprocess the data set by applying principal component analysis (PCA). In addition, e retain only those components which contribute more than a specified fraction (defined 0.01) of the total variation in the dataset. The dimension of the final input patterns is 22. The problem of missing input data is also addressed. As a first approach, missing values were replaced by 0 (data set: Value_0). Another approach to the problem of incomplete data is to substitute the missing value with an estimated value, in particular we used the averaged value of the attribute to replace the missing value (data set: Value_av).

The performance of the proposed diagnostic methods is described using the classification accuracy, sensitivity and specificity defined as:

$$accuracy = \frac{\text{number of true positive + true negative test results}}{\text{number of all patients}}$$

$$sensitivity = \frac{\text{number of true positive test results}}{\text{number of all patients with disease}}$$

$$specificity = \frac{\text{number of true negative test results}}{\text{number of all patients without disease}}$$

The goal of our study is to improve the diagnostic performance using non-invasive diagnostic methods by evaluating all the available diagnostic information using Machine Learning (ML) techniques which are presented below.

ARTIFICIAL NEURAL NETWORKS

Neural networks offer several advantages over conventional computing architectures. In this paper we present an extensive comparison among several feedforward neural network models in the context of the detection of coronary artery disease. We present results from the application of multilayer perceptrons (MLP) and radial basis function networks (RBF).

Multilayer Perceptrons

Multilayer perceptrons are the most popular feedforward neural network models consisting of fully interconnected layers of neurons. The conventional training algorithm for training MLPs is the backpropagation algorithm which constitutes an implementation of the gradient descent method and is relatively slow. To improve training speed and effectiveness, we have employed the *BFGS quasi-Newton* optimization algorithm [12]. Two termination criteria have been considered: (1) maximum number of iterations (in our case 100) or (2) achieved mean square training error less than 0.05. To assess the generalisation performance, a separate test is presented to the MLP after the completion of training.

We considered MLP architectures consisting of 22 input units (set of symptoms and measurements after applying PCA), 2 hidden layers (5 to 20 sigmoid hidden units) and 2 output units (one corresponding to each of the two classes). Weights were randomly initialised in the range [-1,1].

To compare the different network architectures, several series of experiments have been conducted. For each type of MLP, ten experiments were performed with random splits of data into training and test sets of fixed size (119 patterns for training and 20 for testing). We study both Value_0 and Value_av data sets. The average results were calculated from these ten trials and the best results are summarized in Table 1. The outputs of the MLP provide values in the range between 0.0 and 1.0 and the input sample is attributed to the class whose output neuron has the highest value.

Table 1: Average results using BFGS quasi-Newton algorithm for training MLP, for both Value_0 and Value_av data sets.

Data set	MLP*	Average Accuracy (%)	Average Sensitivity (%)	Average Specificity (%)
Value_0	15 + 5	74	84	41
Value_av	10 +10	76	86	54

*Network architecture: Number of units in the first hidden layer + number of units in the second hidden layer.

Moreover, in order to increase the classification reliability of the method [22], we have used a cut-off point of 0.8 for the output value of the winning class, below which the result was considered as unclassified. The average results from ten trials using a cut-off point are summarized in Table 2.

Table 2: Average results using a cut-off point for both Value_0 and Value_av data sets.

Data set	MLP*	Average Accuracy (%)	Average Sensitivity (%)	Average Specificity (%)
Value_0	15 + 5 (19)	76	86	46
Value_av	10+10 (18.8)	78	88	46

*Network architecture: Number of units in the first hidden layer + number of units in the second hidden layer. The average number of test patterns classified from ten experiments is indicated in parentheses.

Radial Basis Function Networks

The Radial Basis Function (RBF) neural network is a simple, yet powerful, machine learning algorithm for nonlinear regression and classification. Consider the approximation of a function $f:R^d \to R$ based on samples $z_i = (\vec{x}_i, y_i), i = 1 \ldots n$ in the form of a linear combination of m basis functions:

$$f_m(\vec{x}, \vec{w}) = \sum_{j=1}^{m} w_j g_j(\vec{x}, \vec{v}_j) + w_0$$

or in normalized form

$$f_m(\vec{x}, \vec{w}) = \frac{\sum_{j=1}^{m} w_j g_j(\vec{x}, \vec{v}_j)}{\sum_{k=1}^{m} g_k(\vec{x}, \vec{v}_j)}$$

In the case of RBF, the basis functions are usally radially symmetric gaussians of the form:

$$g_j(\vec{x}, \vec{v}_j) = \exp\left(-\frac{\| \vec{x} - \vec{v}_j \|^2}{2a_j}\right)$$

The different layers of an RBF network (hidden layer of basis functions and output layer of their combination) perform different tasks, so we are able to separate the optimization of the hidden and output layers using different techniques and operating in different time scales. That is, the parametric vector $\vec{w}$ of the output layer is estimated from data via linear least uares, while the positions of the centers $\vec{v}_j$ and their widths a_j can be estimated using several learning methods [25,30].

The unsupervised training procedure used to estimate the parameters of the gaussian can be summarized by the following algorithm [29]:

1. Choose the number of basis functions in the hidden layer (m).
2. Estimate centers $\vec{v}_j$ using the $\vec{x}$ values of training data by unsupervised training such as k-means clustering or self-organizing maps (SOM).
3. Determine width parameter a_j using the following heuristic rule: For a given center $\vec{v}_j$, find the distance to the closest center r_j and set the width parameter $a_j = \gamma r_j$, where γ is a parameter controlling the overlap between basis functions (in practice $1 \le \gamma \le 3$).
4. Estimate $\vec{w}$ via linear least squares.

The Radial Basis Function classifier uses multi-output nonlinear regression to built a decision boundary and training is performed in the same way as in the case of RBF regression.

In this dataset, the standard SOM method, as well as the following SOM-based variant were used to estimate parameters $\vec{v}_j$ and a_j. Suppose that a $m_1 \times m_2$ self-organizing map was trained using the $\vec{x}$ values of training data. Instead of using the $m = m_1 \times m_2$ units and their corresponding weight vectors as RBF centers, the following two-step operation (pruning + splitting) has been performed:

1. Prune all units that are favored by less than τ training patterns (where τ is a threshold value).
2. Split each unit (cluster) into so many units as the categories of the training patterns that are in favor of the particular unit. Set the weight vector of each of unit to be the centroid of the patterns that belong to the same category.
3. Estimate parameters a_j and $\vec{w}$ as described previously.

The two algorithms were used to estimate the parameters of a two – class normalized RBF neural network and the results are illustrated in Table 3. The data were normalized so that their mean is zero and their standard deviation is one. PCA was found to deteriorate the results under these methods. Twenty experiments were performed with random splits of data into training and test sets of fixed size (119 patterns for training and 20 patterns for testing). The 20 patterns of the test set are restricted to correspond to 15 positive and 5 negative patients.

Table 3: Average results using RBF-SOM and RBF-SOM Pruning + Splitting, for both Value_0 and Value_av data sets.

(Dataset) - Method	Average Accuracy (%)	Average Sensitivity (%)	Average Specificity (%)
(Value_0)RBF - SOM	76	86	47
(Value_0)RBF –SOM Pruning + Splitting	77	85	54
(Value_av)RBF - SOM	77	89	41
(Value_av)RBF –SOM Pruning + Splitting	78	78	75

In the experiments presented in Table 3, we set parameter γ = 1.5 and parameter τ = 2. For this particular dataset, pruning without splitting has improved accuracy but has resulted in worse specificity while the application of splitting has resulted in more balanced classification results.

SUPPORT VECTOR MACHINES

The Support Vector Machine (SVM) is a constructive learning procedure that applies the Structural Risk Minimization principle to construct rules that exhibit good generalization abilities [13]. In doing so, it extracts a small subset of the training data called the "*Support Vectors*".

A support vector machine is based on the following two operations:

- Nonlinear mapping of an input into a high-dimensional feature space.

- Construction of an optimal hyperplane for separating the above features.

The applied SVM approach uses radial basis functions as the nonlinear transformations from the input space to the feature space:

$$f(\vec{x}) = sign\left[\sum_{i=1}^{N} a_i \exp\left(-\frac{\left|\vec{x}-\vec{x}^i\right|^2}{\sigma^2}\right)\right]$$

Here σ is a width parameter defined a priori and α_i , i=1,...,N (N: number of training patterns) are parameters (Lagrange multipliers) determined optimally by the SVM algorithm. The SVM was trained several times with different values for σ and the regularization parameter *C* of the SVM algorithm. For each SVM, ten experiments were performed with random splits of data into training and test sets of fixed size (119 patterns for training and 20 for testing). Both Value_0 and Value_av data sets have been studied. The average results were calculated from these ten trials and the best results are summarized in able 4.

Table 4: Average results using SVM for both Value_0 and Value_av data sets.

Data set	SVM*	Average Accuracy (%)	Average Sensitivity (%)	Average Specificity (%)
Value_0	(14.2 , 20)	67	70	63
Value_av	(14.2 , 1000)	75	83	54

*SVM parameters: (σ , C).

FUZZY ARTMAP NETWORKS

ARTMAP networks [21] are incrementally trained classification networks. An ARTMAP model consists of two adaptive resonance modules (ARTa and ARTb) that create stable recognition categories in response to sequences of input patterns. During supervised learning the ARTa module ceives a stream {α(p)} of input patterns, and ARTb receives a stream {b(p)} of input patterns, where {b(p)} is the correct prediction given {α(p)}. These modules are linked through an associative network and a controller.

Table 5: Results using ARTMAP networks for both Value_0 and Value_av data sets.

Data set	#ARTa nodes*	Average Accuracy (%)	Average Sensitivity (%)	Average Specificity (%)
Value_0	4 voters 24.35 nodes (16-37)	76%	81%	47%
Value_av	8voters, 25.24nodes (18-40)	78%	83%	53%
Training Time: 44sec (user time), Ultra Spark 4.				

*Network architecture: average number/range of ART_a nodes.

The current application was based on an ARTMAP simulator (see http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/fuzzy/systems/artmap/0.html), which implements a voting strategy. The ARTMAP network was trained several times on input sets with different orderings of patterns. Each of these orderings constitutes the training set of a voter (recall that the training order in ARTMAP is important because of the incremental learning which takes place). The final prediction for a given test set is the one made by the largest number of voters. The data (139 vectors, of 24 dimensions, 2 categories) are split into the training set (119 vectors) and test set (20 vectors). The results, which are reported in Table 5, are averages over 10 experiments. The ARTMAP parameters were set as follows : α=1.0, ε=0.1, Learning Rate=1, Min ARTa Vigilance=0.1, ARTb Vigilance=1.0.

PROBABILISTIC NEURAL NETWORK MODELS

Consider a classification problem with K classes and a training set $X=\{(x^n,k^n), n=1,...,N)\}$, where x^n is a d-dimensional pattern and k^n is an integer in the range $(1,K)$ indicating the class of the pattern x^n. The original set X can be easily partitioned into K independent subsets X_k so that each subset contains only the data of the corresponding class. Let N_k denote the number of patterns of class C_k, ie. $N_k=|X_k|$.

The application of probabilistic models to the above classification problem presupposes the estimation (using the available training set X) of the conditional density $p(x|C_k)$ of each class C_k [23]. Then, in order to classify a new pattern x, the conditional densities are combined with prior probabilities $P(C_k)$ through Bayes' theorem, and provide the posterior probabilities $P(C_k|x)$:

$$P(C_k \mid x) = \frac{P(C_k)\,p(x \mid C_k)}{\sum_{k'=1}^{K} P(C_{k'})\,p(x \mid C_{k'})}$$

The final classification decision is performed by selecting the class with maximum posterior probability.

The computation of each conditional density $p(x|C_k)$ is performed though the training of probabilistic neural networks, ie. neural network models whose outputs correspond to probability density functions. The most widely used probabilistic neural network model is the *Gaussian mixture model* [24]:
A Gaussian mixture model with M spherical kernels is described as follows:

$$p(x) = \sum_{j=1}^{M} \pi_j\, p(x \mid j)$$

where

$$p(x \mid j) = \frac{1}{(2\pi\sigma_j^2)^{d/2}} \exp\left[-\frac{\|x-\mu_j\|^2}{2\sigma_j^2}\right]$$

with μ_j, σ_j representing the center and variance of kernel j, while for the priors π_j $(0 \le \pi_j \le 1)$ it holds that

$$\sum_{j=1}^{M} \pi_j = 1$$

Two approaches have been proposed for training of probabilistic neural networks: the separate mixtures model [23,25] and the Probabilistic RBF model [26,27].

The separate mixtures model (SM)

In the first approach [23, 25], a separate Gaussian mixture model (with M_k kernels) is employed to estimate each class conditional density $p(x|C_k)$ using only the data of the corresponding class C_k. For such models efficient unsupervised density estimation training procedures have been developed that are based on the EM algorithm [25, 28].

The Probabilistic RBF model (PRBF)

The second approach is based on the Probablistic RBF model [26, 27] which comprises a special case of the RBF neural network where the basis functions are taken to be probability densities and the second layer weights are constrained to present prior probabilities. In this way, the outputs of the RBF represent class conditional densities. This interpretation of the outputs has given the opportunity to treat RBF training as a maximum likelihood problem and derive an one-stage EM algorithm for adjusting the model parameters.

Assume that we have a number of M spherical Gaussian kernel functions $p(x|j)$:

$$p(x \mid j) = \frac{1}{(2\pi\sigma_j^2)^{d/2}} \exp\left[-\frac{\|x-\mu_j\|^2}{2\sigma_j^2}\right]$$

We would like to utilize $p(x|j)$ for estimating the conditional densities of all classes by considering the kernels as a common pool. Thus, each class conditional density function $p(x|C_k)$ is modeled as

$$p(x \mid C_k) = \sum_{j=1}^{M} \pi_{jk}\, p(x \mid j)$$

here the mixing coefficient π_{jk} represents the prior probability of the pattern x having been generated from kernel j, given that it belongs to class C_k. The priors take positive values and satisfy the following constraint for each k:

$$\sum_{j=1}^{M} \pi_{jk} = 1$$

Training of the PRBF network can be performed using the EM algorithm for likelihood maximization [26,27]. The parameter update equations at $t+1$ iteration of the algorithm are as follows:

$$P^{(t)}(j \mid C_k, x) = \frac{\pi_{jk}^{(t)} p^{(t)}(x \mid j)}{\sum_{j'} \pi_{j'k}^{(t)} p^{(t)}(x \mid j')}$$

$$\mu_j^{(t+1)} = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N_k} P^{(t)}(j \mid C_k, x^n)\, x^n}{\sum_{k=1}^{K}\sum_{n=1}^{N_k} P^{(t)}(j \mid C_k, x^n)}$$

$$\sigma_j^{2\,(t+1)} = \frac{1}{d} \frac{\sum_{k=1}^{K}\sum_{n=1}^{N_k} P^{(t)}(j \mid C_k, x^n) \left\|x^n - \mu_j^{(t+1)}\right\|^2}{\sum_{k=1}^{K}\sum_{n=1}^{N_k} P^{(t)}(j \mid C_k, x^n)}$$

$$\pi_{jk}^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N_k} P^{(t)}(j \mid C_k, x^n)$$

The training algorithm for the PRBF network can be summarized as follows:

1. Specify the number of kernels M and the initial parameter values.
2. E-step: For each training point $(x^n, k^n) \in X$ compute the posterior probabilities $P^{(t)}$ using the current parameters $\theta^{(t)} = \left(\mu_j^{(t)}, \sigma_j^{(t)}, \pi_{jk}^{(t)}\right)$.
3. M-step: Find the new parameter vector $\theta^{(t+1)} = \left(\mu_j^{(t+1)}, \sigma_j^{(t+1)}, \pi_{jk}^{(t+1)}\right)$.
4. Iterate going to step 2 until convergence is achieved.

Experimental results

We applied the two probabilistic models (SM and PRBF) to the PCA preprocessed two-class data. The leave-one-out cross validation method was employed in order to estimate the classification performance. The results are illustrated in Table 6 for both methods. In the SM case M_k denotes the number of kernels of each separate mixture, while in the PRBF case, M denotes the total number of utilized kernels.

Table 6: Classification performance for different SM and PRBF architectures for Value_0 data set.

	Average Accuracy (%)	**Average Sensitivity (%)**	**Average Specificity (%)**
SM (M_k = 1)	68	68	68
SM (M_k = 2)	67	67	68
SM (M_k = 3)	66	66	65
PRBF (M = 4)	58	59	57
PRBF (M = 6)	63	60	68
PRBF (M = 8)	67	64	74

DECISION-TREES

Decision Trees [17-19] constitute one of the most effective and successfully utilized methods of inductive learning [20] that is particularly suited for classification and prediction tasks. A decision tree structure is either,

- a *leaf*, indicating a class or,
- a *decision node* that specifies some test to be carried out on a single feature value, with on branch and sub-tree of the node for each possible outcome of the test (feature's values).

Starting at the root and moving down the tree until a leaf is encountered, we are able to identify all the relevant and informative feature-values for concluding the class recorded at the leaf.

The way that the tree is constructed (induced) is based on a "*divide and conquer*" strategy. Within this approach, the given set of data is partitioned into subsets and for each of the subsets the 'most informative' feature (based on the Information Gain metric) is identified. Following a "*hill-climbing*" approach and iteratively dividing the data set into subsets by selecting the

most informative feature, the whole tree is constructed. An outline of the decision tree construction algorithm is shown in Figure 1 and constitutes a variant of Quinlan's C4.5 algorithm [18].

Main
Step1: E = All set of observations (examples)
Step2: Call **Best_Feature(E,BestFeature)** sub-routine
Step3: Divide E into sub-sets E[BestFeature(Value)]
Step4 -**Iterate**
Set E = E[(BestFeature(Value)]
If
Call **Single_Class(E[Best_Feature(Value)])** sub-routine
Then Stop for current branch and conclude to a leaf labeled with the single class labeling all examples in E[BestFeature(Value)]
Else Goto Step2

Best_Feature(E,BestFeature)
Step1: Compute Entropy(E)
Compute Entropy(E/Feature$_i$)
Compute InfoGain(Feature$_i$)
Step2: Compute BestFeature = max{Feature$_i$/Feature$_i$ >Feature$_j$, 1≤i, j≤ #Features}

Single_Class(E[Best_Feature(Value)])
Check if all examples in data set E[BestFeature(Value)], i.e., all examples have value Value for BestFeature, are assigned to the same class.

Figure 1: Outline of the Decision Tree construction algorithm

Experimental results

The testing-course with the coronary artery disease dataset include the set of experiments summarized in Table 7. The available dataset was carefully cleaned for cases with too-many unknown values in order to exclude as-much-as-possible the potential noise in the data. The resulted dataset consists of 124 cases.

- All the experiments were conducted with the use of the See-C5 decision-tree construction algorithm, a commercially available offspring of C4.5.
- For all experiments a 10-fold cross-validation procedure was followed (using the 100% of the data) in order to assess the predictive accuracy of the final output.
- For each of the experiments, the rules induced using the 100% of the respective dataset are to be considered as the knowledge-base to be incorporated in the intelligent system for the primary diagnosis of the coronary artery disease.

Table 7: The specifics of the experiments conducted with the decision-tree induction approach

Experiment name	Explanation	#cases
All_Normal_vs_Abnormal	The whole dataset is used (both male and female patients included); cases are assigned into two- (2) classes: patients with 'normal' status of vessels (i.e., no infraction) and patients with 'abnormal' status of vessels (i.e., vessels with less50% and more50% infraction).	124
Males_Normal_vs_Abnormal	Focus on restricted dataset of male patients; cases are assigned into two- (2) classes: patients with 'normal' status of vessels (i.e., no infraction) and patients with 'abnormal' status of vessels (i.e., vessels with less50% and more50% infraction).	96
Males_Less50_vs_More50	Focus on restricted dataset of male patients with abnormal vessels status; cases are assigned into two- (2) classes: patients with 'less50%' infraction, and patients with 'more50%' infraction.	81
Males_More50	Focus on restricted dataset of male patients with more50% infraction vessels status; cases are assigned into three- (3) classes: 'in_1', 'in_2', and 'in_3' for more50% infraction in only one vessel, in two vessels and in three vessels, respectively.	66
Females_Normal_vs_Abnormal	Focus on restricted dataset of female patients; cases are assigned into two- (2) classes: patients with 'normal' status of vessels (i.e., no infraction) and patients with 'abnormal' status of vessels (i.e., vessels with less50% and more50% infraction).	27

The results of all experiments are summarized in Table 8.

Table 8: Results on the set of experiments conducted with the decision-tree induction approach.

Experimet name	Predictive Accuracy V-fold$_{(V=10)}$/ Standard Error	# Rules	Complexity of Rules Min/Mean/Max* number of conditions
All_Normal_vs_Abnormal	**79.8** / 3.6	9	1 / 2.6 / 4
Males_Normal_vs_Abnormal	**83.3** / 3.8	14	2 / 3.4 / 5
Males_Less50_vs_More50	**77.8** / 4.6	7	1 / 2.4 / 6
Males_More50	**49.0** / 6.6	12	1 / 2.4 / 6
Females_Normal_vs_Abnormal	**79.8** / 3.6	9	1 / 2.6 / 4

* Min, Mean, and Max are the minimum, mean (=total_#_of_conditions_in_all_rules / #_of_rules) and maximum number of conditions used in the respective induced set of rules. These measures are used as an indication of the complexity of the final learning result.

CONCLUSIONS

In this paper we present and compare several classification methods, applied to the problem of the early diagnosis of coronary artery disease. The classification method along with the data selection-storage system (knowledge base) composes the diagnostic system for coronary artery disease, which accepts as input the data of a patient and provides a decision for the patient. The proposed intelligent system architecture is

quite general and applicable to a wide class of models and classification techniques, allowing the training and evaluation of different approaches and the combination of their decisions.

The results of our study, using the available patient data, give us very promising classification performance. Due to the high specificity obtained, only a small number of persons not suffering from the disease have to be examined with angiongraphy which is invasive and painful. The proposed approach results in shorter treatment time for the patients and is more cost effective. However, it should be taken into account that the results are obtained on a significantly restricted population and therefore the proposed approach may not be generally applicable to the normal population.

In the near future, the proposed algorithms have to be adapted to design an intelligent system for the classification of patients in five categories: patients with normal coronary arteries, patients not featuring alterations in angiography but with endothelial disfunction, patients having atheromatic coronary vessels but without significant stenosis (<50 %) and patients with significant coronary arterial stenosis (>50 %).

CKNOWLEDGEMENTS

The present work is partially supported by the Greek General Secretariat of Research and Technology as part of the project "PENED99-318: Development of an Intelligent System for the Early Diagnosis of Coronary Artery Disease".

REFERENCES

1. R.A. Vogel, Coronary Risk Factors, Entothelial Function and Atherosclerosis: A Review, Clin. Cardiol. 20, 426-432 (1997).
2. A.H. Henderson, St Cyre's Lecture: Entdothelium in Control, Br. Heart J. 65, 116-125 (1991).
3. H. Drexler and B. Hornig, Endothelial Dysfunction in Human Disease, J. Mol. Cell Cardiol. 31(1), 51-60 (1999).
4. J.M. McLenachan, J.K. Williams, R.D. Fish, P. Ganz, A.P. Selwyn, Loss of Flow – Mediated Endothelium – Dependent Dilation Occurs early in the Development of Atherosclerosis, Circulation 84, 1273-1278 (1991).
5. D.S. Celermajer, K.E. Sorensen, C. Bull, J. Robinson and J.E. Deanfield, Endothelium – Dependent Dilation in the Systemic Arteries of Asymptomatic Subjects Relates to Coronary Risk Factors and Their Interaction, J. Am. Coll. Cardiol. 24, 1468-1474 (1994).
6. T.J. Anderson, A. Uehata, M.D. Gerhard, I.T. Meredith, S. Knab, D. Delagrange, E.H. Lieberman, P. Ganz, M.A. Greager, A.C. Yeung, A.P. Selwyn, Close Relationship of Endothelial Function in the Human Coronary and Peripheral Circulations, J. Am. Coll. Cardiol. 26, 1235-1241 (1995).
7. K.H. Benzuly, R.C. Radgett, S. Kaul, D.J. Piegors, M.L. Armstrong and D.D. Heistad, Functional Improvement Precedes Structural Regression of Atherosclerosis, Circulation 89, 1810-1818 (1994).
8. A. Schmermund, D. Baumgart, G. Gorge, R. Seibel, D. Gronemeyer and R. Erbel, Non-Invasive Visualisation of Coronary Arteries with and without Calcification be Electron Beam Computed Tomography, Herz 21(2), 118-126 (1996).
9. R.S. Underwood and R.H. Mohiaddin, Magnetic Resonance Imaging of Atherosclerotic Vascular Disease, Am. J. Hypertens. 6, 335S-339S (1993).
10. E.D. Lehmann, K.D. Hopkins, R.M. Marsden, I. Brown, R.L. Jones, R.C. Turay, M.G. Taylor, and R.G. Gosling, Aortic Compliance Measured by Non-Invasive Doppler Ultrasound: Application of a Personal Computer MkII System and its Repeatability, Med. Engng. Phys. 16(3), 213-221 (1994).
11. J.L. Megnien, A. Simon, N. Denarie, M. Del-Pino, J. Gariepy, P. Segond and J. Levenson, Aortic Stiffening does not Predict Coronary and Extracoronary Atherosclerosis in Asymptomatic Men at Risk for Cardiovascular Disease, Am. J. Hypertens. 11, 293-301 (1998).
12. J.E. Dennis, R.B. Schnabel, Numerical methods for unconstrained optimization and nonlinear equations. Englewood Cliffs, NJ: Prentice-Hall, (1983).
13. V. Vapnik, The nature of statistical learning theory, Springer-Velag, New York (1995).
14. AP. Michaelides et. al. Improved detection of coronary artery disease by exercise electrocardiography with the use of right precordial leads. N Engl J Med;340:340-345(1999).
15. W. Nichols, M. O' Rourke. McDonald's Blood Flow in arteries. In: Edward Arnold (eds.), (1990).
16. LJ Shaw, ED Peterson, LK Shaw, KL Kesler, ER DeLong, FE Jr Harrell, LH Muhlbaier, DB Mark. Use of a prognostic treadmill score in identifying diagnostic coronary disease subgroups. Circulation 1998;98:1622-1630.
17. R.J. Quinlan, Generating production rules from decision trees. Procs. 10th International Joint Conference on Artificial Intelligence (IJCAI 87), 304-307 (1987).
18. R.J. Quinlan, C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA, (1993).
19. T. M. Mitchell, Machine Learning, McGraw-Hill, (1997).
20. V. Moustakis, G. Potamias, L. Gaga, M. Blazadonakis, P. Vassilakis and G. Charissis, Bias identification using inductive learning techniques - a medical case study. In B.G. Silverman (Ed.) Expert Judgment, Human Error, and Intelligent Systems Workshop, European Conference on Artificial Intelligence - ECAI 92, (1992).
21. A.G. Carpenter, S. Grossberg, N. Markuzon, J. Reynolds. , D. Rosen, Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps, IEEE Transactions on Neural Networks, 3(5), (1992).
22. L.P. Cordella, P. Foggia, C. Sansone, F. Tortorella and M. Vento, Reliability parameters to improve combination strategies in multi-expert systems, Pattern Analysis and Applications, 2:205-214, Springer-Verlag, (1999).
23. R.O. Duda and P. E. Hart, Pattern classification and Scene Analysis. Wiley, New York, (1973).

24. G.J. McLachlan and D. Peel, Finite Mixture Models. Wiley, (2000).

25. C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, (1995).

26. M. Titsias, A. Likas. A Probabilistic RBF network for Classification, Proc. of International Joint Conference on Neural Networks, Komo, Italy, (July 2000).

27. M. Titsias, A. Likas. Shared Kernel Models for Class Conditional Density Estimation, IEEE Trans. on Neural Networks, accepted with revision.

28. A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society B, vol. 39, pp. 1-38, (1977).

29. V. Cherkassky, F. Mulier, Learning from Data, Concepts Theory , and Methods. John Wiley & Sons Inc, (1998).

30. S. Haykin, Neural Networks: A comprehensive foundation, Second Edition, Prentice Hall, (1999).