CMS 1: 245–274 (2004)

DOI: 10.1007/s10287-004-0015-8

A multiple decision trees architecture for medical diagnosis: The differentiation of opening snap, second heart sound split and third heart sound

A. Ch. Stasis[1]**, E.N. Loukis**[2]**, S.A. Pavlopoulos**[1]**, D. Koutsouris**[1,⋆]

[1] National Technical University of Athens, Biomedical Engineering Laboratory, Iroon Polytexniou Zografou, Athens, Greece, (e-mail: astasis@biomed.ntua.gr)
[2] Dept. of Information and Communication Systems Engineering, University of Aegean, Samos, Greece, (e-mail: eloukis@aegean.gr)

Abstract. In this paper a Decision Support System Architecture is proposed for the heart sound diagnosis problem, and in general for complex medical diagnosis problems. It is based on the division of a complex diagnostic problem into simpler sub-problems; each of them is handled by a specialized decision tree. This Multiple Decision Trees Architecture in general consists of a network of detection decision trees and arbitration decision trees, and can also incorporate other classification methods as well (e.g. patterns recognition, neural networks, etc.). The initial motivation for developing this Multiple Decision Trees Architecture has been the problem of differentiation among Opening Snap (OS), 2nd Heart Sound Split (A2_P2), and 3rd Heart Sound (S3), which is a crucial and at the same time difficult and complicated part of the heart sound diagnosis problem. The Multiple Decision Tree Architecture developed for the above diagnosis/differentiation problem has been tested with real heart sound signals, and its performance and generalisation capabilities were found to be higher than the previous 'traditional' architectures.

Keywords: Heart sound diagnosis, Medical diagnosis, Decision trees, Decision support systems, Opening snap, 3rd heart sound, 2nd heart sound split

Mathematics Subject Classification (2000): 68U35

[⋆] The authors would like to thank the clinician Dr D.E. Skarpalezos for his clinical support, Dr G. Koundourakis, and Neurosoft S.A. for their support and provision of Envisioner, a data-mining tool that was used to execute algorithms related to the Decision Trees.

CMS Computational Management Science

1 Introduction

The heart sound diagnosis is one of the most useful and at the same time lowest cost screening methods for diagnosing heart pathologic conditions. The only equipment it requires is the stethoscope, which is available even in the smallest primary healthcare units of all urban and rural areas; it is of very low cost, and also it is portable, so it can be easily transported and used in any place (e.g. even for homecare). New technologies, like Echocardiography, CT, and MRI, do provide more accurate evidence, but they require expensive equipment, specialized technicians, high cost consumables, a permanent place to be installed (i.e. they are not portable), and generally they require many resources in order to function properly. These requirements are usually met only in the hospitals, so these new technologies are not suitable for use in homecare, in rural areas, and generally in primary healthcare (DeGroff et al., 2001). Therefore in numerous cases the heart sound diagnosis is the only feasible alternative, and the health, or even the life, of numerous patients depend critically on it.

However, the internal medicine and the cardiology training programs focus mainly on the above new technologies and underestimate the value of cardiac auscultation, so the young clinicians are not well trained in this field (Criley et al., 2000). Also the medical personnel in the community clinics are usually young, inexperienced, and with limited cardiac auscultation skills. Therefore efficient Decision Support Systems can be very useful for supporting clinicians in making reliable heart sound diagnosis, and especially in making correct differential diagnosis (i.e. discriminate among heart diseases resulting in similar heart sound signals), especially in rural areas, in homecare and in primary healthcare. These systems are usually based on digital electronic stethoscopes (for heart sound signals acquisition) and computers (for the analysis of the acquired signals based on appropriate algorithms) (Myint and Dillard, 2001; Lukkarinen et al., 1997).

As concluded from the relevant literature, which is reviewed in the following paragraphs, the development of efficient algorithms for supporting the computer assisted heart sound diagnosis is a wide, challenging and open research subject. In general, heart sound diagnosis is a very difficult and complex diagnostic problem: from highly noisy heart sound signals, it should be diagnosed whether the heart is healthy or not, and if not, the exact heart disease should be determined, among a big number of possible heart diseases. Also, there are heart diseases resulting in similar heart sound signals, and the discrimination among them is quite difficult. It should also be noted that there are many other medical diagnosis problems with similar difficulties and complexities. For handling them highly sophisticated decision support algorithms and architectures are required, involving complex signal processing, classification and in general decision support methods.

In the last decade extensive research work has been conducted, concerning heart sound diagnosis, with promising results. The detection of the first and the second heart sound, i.e. the detection of the boundaries of each heart sound cycle, has been a vital subject for the computer assisted heart sound diagnosis, therefore many researchers have worked on it (Haghighi-Mood and Torry, 1995; Liang et al., 1997a; Liang et al., 1997b; Hebden and Torry, 1996). The signal processing research in this area has been focused on the analysis of the first, the second heart sound and the heart murmurs (Tovar-Corona and Torry, 1998; White et al. 1996; Wu et al. 1995; Wang et al., 2001). The pattern recognition research has been focused on features extraction and classification of heart sounds and murmurs (Liang and Hartimo, 1998; Sharif et al., 2000). The algorithms, which have been utilized for the analysis of the heart sound signals and for the features extraction, were based on:

1) Auto Regressive and Auto Regressive Moving Average Spectral Methods (Haghighi-Mood and Torry, 1995; Wu et al., 1995),
2) Power Spectral Density (Haghighi-Mood and Torry, 1995),
3) Trimmed Mean Spectrograms (Leung et al., 2000),
4) Sort Time Fourier Transform (Wu et al., 1995),
5) Wavelet Transform (Liang et al., 1997a; Tovar-Corona and Torry, 1998; Wu et al., 1995), and
6) Wigner-Ville distribution, and generally the ambiguous function (White et al., 1996).

The classification algorithms were mainly based on:

i) Discriminant analysis (Leung et al., 1998),
ii) Nearest Neighbour (Durand et al., 1993),
iii) Bayesian networks (Durand et al., 1993; Wu, 1997),
iv) Neural Networks (DeGroff et al., 2001; Hebden and Torry, 1996; Leung et al., 2000) (backpropagation, radial basis function, multilayer perceptron, self organizing map, probabilistic neural networks etc), and
v) rule-based methods (Sharif et al., 2000).

It is worth emphasizing that the clinicians, for making heart sound diagnosis, take into account not only the heart sound signal but also the way that the examination is done, the patient's physical examination variables, the patient's history, etc. The heart sound signal information alone is not adequate in most cases for a heart disease diagnosis, so very few research works focus on the exploitation of heart sound signals for the direct diagnosis of cardiac diseases. One of the most important research works on direct diagnosis is the research of M. Akay and co-workers in Coronary Artery Disease (Akay et al., 1994, and many other publications), and also the research of Hedben and Torry in the identification of Aortic Stenosis and Mitral Regurgitation (Hebden and Torry, 1997).

The basic motivation for conducting the research work described in this paper has been the problem of differentiation among the Opening Snap (OS), the Second Heart Sound Split (A2_P2) and the Third Heart Sound (S3). This is a very significant problem in cardiology; very often there is confusion between these three morphological characteristics, leading to uncertain conclusions. The correct differentiation among them is of critical importance for making the correct diagnosis of the heart disease and recommending the appropriate treatment.

Concerning the (S3), most of the previous research is mainly trying either to explain the mechanism that generates S3, or to find a mathematical model that describes S3 genesis and its physiologic disappearance with aging (Longhini et al., 1996). Additionally some research works try to associate S3 with some heart phenomena, such as ischemic heart disease, transmitral flow etc (Aggio et al., 1990). Concerning the OS, from the relevant literature review it is concluded that very limited research has been made about it so far. The S-transform can be used for the differentiation of OS from S3 as described in Livanos et al. (2000).

The research work presented in this paper treats the problem of differentiation among OS, A2_P2 and S3 as a classification problem and aims at examining whether Decision Trees-based classifier algorithms and architectures can give reliable solutions for the above problem. This is the first study of the performance of classification algorithms in such difficult and complex medical problem. We chose to examine first the decision trees as classification algorithm for this problem, because their knowledge representation model, i.e. decision trees, can offer to the user-clinician not only a recommended diagnosis for the examined heart sound signal, but also a justification of it as well, which is quite necessary. In other words this method does not work as a ‘black box’ for the clinicians (i.e. in medical terms). On the contrary other classification methods, e.g. neural networks, genetic algorithms, etc are working as ‘black box’ for the clinicians: they can offer only a recommended diagnosis, but they do not offer any justification of it.

By using decision trees clinicians can trace back the model, examine the justification of the recommended diagnosis, and either accept or reject it. This capability increases the confidence of the clinician in the recommended diagnosis. As described in Sect. 3, for this purpose we also chose appropriate heart sound features, which are similar to the ones that doctors usually use in heart sound diagnosis (i.e. shape features and frequency features).

In particular, we evaluate and compare three different Decision Tree Architectures for the above differential diagnosis problem. The first Decision Tree Architecture is based on one ‘Differentiation Decision Tree’, which aims at differentiating among the above three morphological characteristics (OS, A2_P2, S3) of the heart sound, and the problem is treated as one differentiation problem. The second Architecture is based on three ‘Detection Decision Trees’, each of them corresponding to one of the above three morphological characteristics. Each of these Detection Decision Trees decides whether the corresponding morphological characteristic exists or not in the examined heart sound; so the initial problem is treated as three separate Detection sub-problems. Finally we develop and evaluate a new ‘Multiple Decision Trees Architecture’, consisting of a network of Detection Decision Trees and Differentiation Decision Trees. In all the above evaluations, the generalization capabilities of the above Decision Tree architectures were examined, which is a very important issue, due to the high levels of noise in the heart sound signals, the different signal acquisition methods, etc. The structure of this paper is as follows: in Sect. 2, the heart sound diagnosis problem and the problem of differentiation of OS, A2_P2 and S3 are presented. In Sect. 3, is described a method for preprocessing the collected heart sound signals and calculating the corresponding patterns (feature vectors). In the same section is also described the Decision Trees construction method. In Sect. 4 are presented the 'Differentiation Decision Tree Architecture' and the 'Detection Decision Trees Architecture', their application on the above problem, and also the evaluation of the obtained results. In Sect. 5, a new 'Multiple-Decision Tree Architecture' is developed and evaluated. Also a generalization of this Architecture is presented for the whole heart sound diagnosis problem, which has a wide applicability to many other medical diagnosis problems of high complexity. Finally, the conclusions derived from this research work are given in Sect. 6.

2 The problem of heart sound diagnosis and the problem of differentiation of OS, A2_P2 and S3

A typical normal heart sound signal that corresponds to a heart cycle consists of four structural components:

- The first heart sound, which is produced from the closing of the mitral and the tricuspid valves (S1).
- The systolic phase.
- The second heart sound, which is produced from the closing of the aortic and pulmonary valves (S2).
- The diastolic phase.

Heart sound signals with various additional sounds are observed in patients with heart diseases. The tone of these sounds can be either like murmur or click-like. The murmurs are produced from the turbulent blood flow and are named after the phase of the heart cycle where they are best heard, for instance systolic murmur (SM), diastolic murmur (DM), pro-systolic murmur (PSM) etc. The heart sound diagnosis problem consists in the diagnosis from noisy heart sound signals:

a) whether the heart is healthy or not,
b) if it is not healthy, which is the exact heart disease.

Sometimes the aortic (A2) and pulmonary (P2) components of the S2 are not heard simultaneously, giving the impression of two different click-like sounds with approximate time difference 100ms (Fig. 1). This phenomenon is referred as Second Heart Sound Split (A2_P2). This happens due to the delayed extraction of the blood from the right ventricle and consequently the late closing of the pulmonary valve during the cardiac contraction, either because there is plenty of blood quantity (for instance during inspiration), or because there is an obstacle to the exit towards the lungs (for instance pulmonary stenosis).

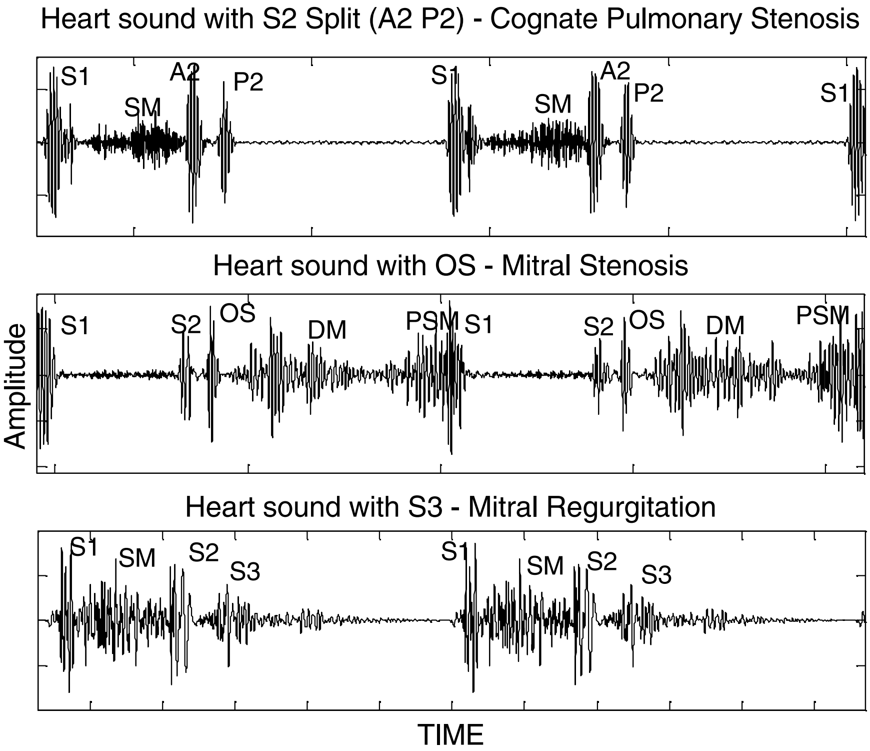

Fig. 1. Heart Sound Signals with A2_P2, OS and S3, corresponding to two Heart Cycles

The opening snap (OS) is another click-like sound and may occur 60-110ms after the A2 component of S2, while the mitral valve is opening (Fig. 1). It is a sort duration sound with high frequency. If an A2_P2 precedes it, the A2-P2-OS combination may be heard as a trilling sound (Livanos et al., 2000). Another click sound that may be heard shortly after S2 is the third heart sound (S3). S3 lasts longer and has lower frequency than OS (Fig. 1). It occurs around 150ms after A2 and is a manifestation of sudden intrinsic limitation in the expansion of the left ventricle during the early diastole.

It is obvious from the above descriptions (Fig. 1) that shortly after the aortic component of the S2 there is confusion as to what sounds are heard. The OS and S3 differ in the frequency content and in the duration. Additionally the OS is mostly related with the mitral stenosis, usually having diastolic murmur. On the contrary the S3 is mostly related with mitral regurgitation, usually having systolic murmur. The shape of the heart sound signals, which are produced from patients with pulmonary stenosis, is depended on the phase of the respiration cycle. The delay of the P2 component as regards the A2 component, during S2 split, is also dependent on the respiration cycle. The respiration cycle affects the blood quantity that flows from the right part of the heart to the lungs and vice-versa. All these facts describe the problem of differentiation of OS, A2_P2 and S3. In the following sections we propose a method that uses time-frequency features and decision tree classifiers for solving this problem.

3 The decision tree method

3.1 Collecting heart sound data

The characteristics of the recorded heart sound signals are significantly affected by many factors related to the acquisition method, such as:

- *Factors related to the digital electronic stethoscope used.*
  The heart sound signal is recorded, using a digital electronic stethoscope; therefore the kind of stethoscope used and the kind of sensor that the stethoscope has (e.g. microphone, piezoelectric film) affects the recorded heart sound signal.
- *Factors related to the patient's condition during auscultation.*
  Such factors are, the condition of patient's heart (e.g. healthy, weak), the medicine that the patient is using during the auscultation (e.g. vasodilators), the exercise that he is taking before and during auscultation (e.g. handgrip, valsalva), the phase of his respiration cycle (inspiration, expiration).
- *Factors related to the way that the auscultation is done.*
  The mode that the stethoscope is used (i.e. bell, diaphragm, extended) and the way the stethoscope is pressed on the patient's skin (firmly or loosely) affects the frequency content of the acquired signal. The shape of the recorded heart sound signal is also affected by the patient's position during auscultation (e.g. supine position, standing, squatting) and by the auscultation areas (i.e. apex, lower left sternal border, pulmonic area, aortic area)
- *Factors related to the filters, which are applied while recording the heart sound signals.*
  The digital filters, which are applied while recording (e.g. anti-tremor filter, respiratory sound reduction, noise reduction) also affect the characteristics of the heart sound signal.

The above factors cannot be controlled, therefore they are a source of high levels of noise in the heart sound signals, making the whole diagnostic problem much more difficult. Therefore a heart sound diagnosis algorithm should be capable of coping with these high levels of noise, and also be based on and tested with heart sounds signals from many different sources and recorded with many different acquisition methods.

For this purpose we tried to collect heart sound signals from different heart sound sources (the heart sound sources are referred at the end of the paper) and create a representative heart sound database. These heart sounds were recorded from different educational audiocassettes, audio CDs, CD ROMs and files from heart sound databases; each of these heart sounds had already been diagnosed and related to a specific heart disease. Then we chose, based on advice from clinicians, from the available heart sounds that were stored in the above database, all the heart sounds containing S2 split (A2_P2), OS and S3, that is 46 heart sounds containing A2_P2, 36 containing OS, and 53 containing S3. In this way we created a dataset of 135 heart sounds in total, which was used in the next phases of this research work, as described in the next sections. It is worth emphasizing that these heart sounds were both from normal subjects, and from patients with a variety of diseases (such as atrial septal defect, ventricular septal defect, congestive heart failure, Ebstein's anomaly, hypertrophic cardiomyopathy, ischemic cardiomyopathy, mitral stenosis, mitral regurgitation, aortic stenosis, aortic regurgitation, mitral valve prolapse, tricuspid regurgitation, left or right bundle branch block, pulmonary stenosis, myxoma (tumor plop), pericardial disease, etc). Therefore the above heart sounds data set that we created for this medical problem was highly representative.

3.2 Pre-processing of Heart Sound Data

In order to preprocess the heart sound signals and calculate the appropriate features, a preprocessing method has been developed and applied. It consists of two phases, which are shown in Fig. 2. Initially in the first phase each of the above data set of 46+36+53=135 heart sounds signals was pre-processed in six steps in order to detect the cardiac cycles, i.e. detect the S1 and S2 (Stasis et al., 2003):

1. The wavelet decomposition described in Liang et al. (1997a) (the only difference being that we kept the 4th and 5th level detail, i.e. frequencies from 34 to 138Hz).
2. The calculation of the normalized average Shannon Energy (Liang et al., 1997b).
3. A morphological transform action that amplifies the sharp peaks and attenuates the broad ones (Haghighi-Mood and Torry, 1995).
4. A method, similar to the one described in Liang et al. (1997b) that selects and recovers the peaks corresponding to S1 and S2 and rejects the others.
5. An algorithm that determines the boundaries of S1 and S2 in each heart cycle.
6. A method that distinguishes S1 from S2 similar to the one described in Hebden and Torry (1996).

Then in a second phase, from each transformed heart sound signal (that has undergone the above 6 steps of processing in the first phase) we extracted the heart sound features. For this purpose we calculated for each heart sound signal two mean signals for each of the four structural components of the heart cycle, namely two for the S1, two for the systolic phase, two for the S2 and two for the diastolic phase. The first of these mean signals focused on the frequency characteristics of the heart sound; it was calculated as the mean value of each component, after segmenting and extracting the heart cycle components, time warping them, and aligning them. The second mean signal focused on the morphological time characteristics of heart sound; it was calculated as the mean value of the normalized average Shannon Energy Envelope of each component, after segmenting and extracting the heart cycle components, time warping them, and aligning them.

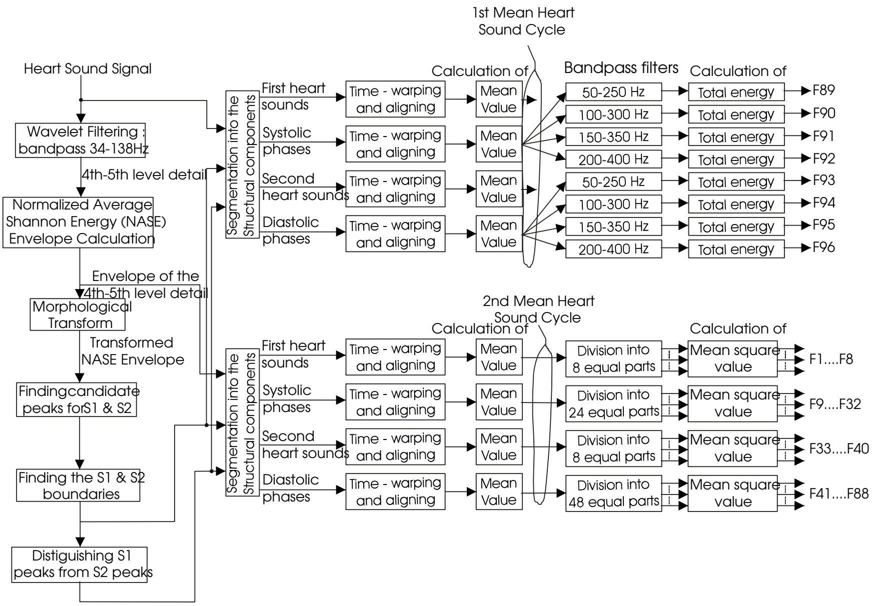
1st Mean Heart Sound Cycle
Heart Sound Signal
Wavelet Filtering : bandpass 34-138Hz
4th-5th level detail
Normalized Average Shannon Energy (NASE) Envelope Calculation
Envelope of the 4th-5th level detail
Morphological Transform
Transformed NASE Envelope
Findingcandidate peaks forS1 & S2
Finding the S1 & S2 boundaries
Distiguishing S1 peaks from S2 peaks
Segmentation into the Structural components
First heart sounds
Systolic phases
Second heart sounds
Diastolic phases
Time - warping and aligning
Calculation of
Mean Value
Bandpass filters
50-250 Hz
100-300 Hz
150-350 Hz
200-400 Hz
Calculation of
Total energy
F89
F90
F91
F92
F93
F94
F95
F96
2nd Mean Heart Sound Cycle
Division into 8 equal parts
Division into 24 equal parts
Division into 8 equal parts
Division into 48 equal parts
Mean square value
F1....F8
F9....F32
F33....F40
F41....F88

Fig. 2. The preprocess method for conversion of the heart sound signal into a heart sound pattern

Then the second S1 mean signal was divided into 8 equal parts. For each part we calculated the mean square value and this value was used as a feature in the corresponding heart sound feature vector. In this way we calculated 8 scalar features for S1 (F1–F8); similarly we calculated 24 scalar features for the systolic period (F9–F32), 8 scalar features for S2 (F33–F40) and 48 scalar features for the diastolic period (F41–F88).

The systolic and diastolic phase components of the above first mean signal were also passed through four bandpass filters: a) a 50–250Hz filter, giving the low frequency content, b) a 100–300Hz filter, giving the medium frequency content, c) a 150–350Hz filter, giving the medium-high frequency content and d) a 200–400Hz filter, giving the high frequency content. For each of these 8 outputs, the total energy was calculated and was used as a feature in the heart sound vector (F89–F96). With the above two phases of preprocessing every heart sound signal was transformed into a heart sound feature vector (pattern) with dimension 1×96.

These preprocessed data feature vectors were stored in a database table with 135 records; each record describes the feature vector (pattern) of a heart sound signal and has 98 attributes-fields: one attribute named ID, for the pattern identification code, one attribute named OS_AP_S3 for the pre-existing characterization (diagnosis) of the corresponding heart sound signal as containing OS or A2_P2 or S3, and also 96 attributes for the above 96 heart sound features (F1–F96).

Before constructing and utilizing the Decision Tree Classifiers we tried, using Relevance Analysis (Kamber et al., 1997), to improve the classification efficiency by eliminating the less useful (for the classification) features and reducing the amount of input data to the classification stage. For example, from the previous description of the database table attributes it is obvious that the pattern identification code (ID) is irrelevant to the classification, therefore the classification algorithm should not use this attribute.

For this purpose, we used the value of the *Uncertainty Coefficient* (Koundourakis, 2001; Kamber et al., 1997) of each of the above 96 features, in order to rank these features according to their relevance to the classifying attribute (dependent variable), which in our case is the OS_AP_S3 attribute. The Uncertainty Coefficient of each feature-attribute is calculated according to the following steps (Fig. 3), which are described in more detail in Koundourakis (2001):

i) Initially each numeric attribute is transformed into a categorical attribure (CA). For this purpose, as described in more detail in Koundourakis (2001), a 'temporary' decision tree is created, by using as its training set the above numeric attribute and the classifying attribute (dependent variable) of all the 135 patterns of our dataset. In this way a series of classification rules are generated: each of them constitutes a range of values of the numeric attribute that correspond to a specific value of the classifying attribute. By assigning each of these ranges to a new category, we transform the initial numeric attribute into a categorical one, with the highest possible relevance to the classifying attribute.
ii) The Information Entropy (IE) of the initial dataset is calculated. The initial dataset has 135 records, which according to the classifying attribute belong to one out of the three discrete classes (OS class, A2_P2 class and S3 class).
iii) The CA is then used to partition the initial dataset into subsets. The new IE is calculated, as the weighted average of the Information Entropies of these new subsets.
iv) The IE gained by the above partitioning according to this CA is calculated (as the difference of the initial IE from this new IE).
v) The Uncertainty Coefficient U(C*A*) for a categorical attribute is obtained by normalizing this Information Entropy Gain of C*A*, via dividing it with the initial IE so that U(C*A*) ranges from 0% to 100%.

A low value of U(CA) near 0% means that there is no increase in homogeneity (and therefore in classification accuracy) if we partition the initial set according to this CA, therefore there is low dependence-relevance between the categorical attribute C*A* and the classifying attribute-dependent variable. On the contrary a high value of U(CA) near 100% means that there is strong relevance between the two attributes.

3.3 Construction of Decision Tree Classifier Structures

A Decision Tree is a class discrimination tree structure, consisting of non-leaf nodes (internal nodes) and leaf nodes (final-without children nodes), (Koundourakis, 2001; Stasis, 2003; Han and Kamber, 2001). For constructing a Decision Tree

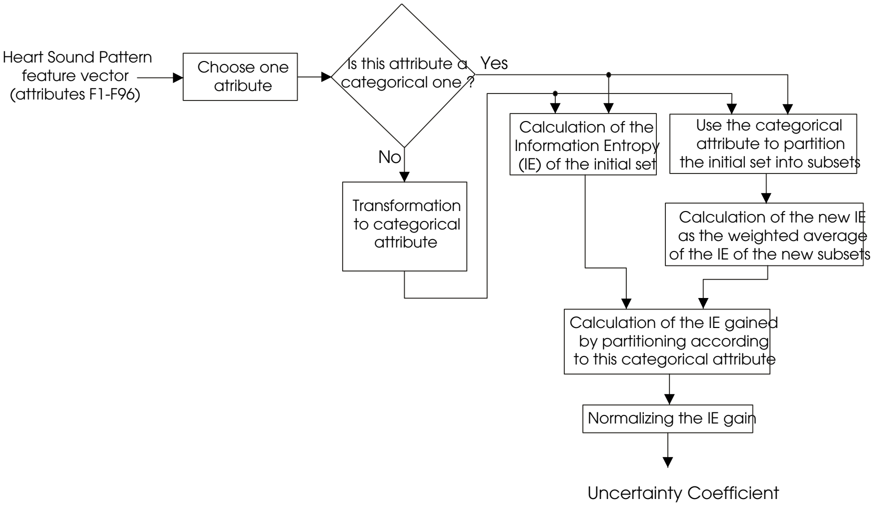
Heart Sound Pattern feature vector (attributes F1-F96)
Choose one atribute
Is this attribute a categorical one?
Yes
No
Transformation to categorical attribute
Calculation of the Information Entropy (IE) of the initial set
Use the categorical attribute to partition the initial set into subsets
Calculation of the new IE as the weighted average of the IE of the new subsets
Calculation of the IE gained by partitioning according to this categorical attribute
Normalizing the IE gain
Uncertainty Coefficient

Fig. 3. The calculation of the Uncertainty Coefficient

we use a training data set, i.e. a set of records, for which we know the values of all the feature-attributes (independent variables) and the classifying attribute (dependent variable). Starting from the root node we determine the best test (= attribute + splitting condition) for splitting the training data set, which creates the most homogeneous subsets concerning the classifying attribute and therefore gives the highest classification accuracy. Each of these subsets can be further split in the same way etc. Such a splitting test at a Decision Tree non-leaf node usually has exactly two possible outcomes (binary Decision Tree), leading to two corresponding child nodes. The left child node inherits from the parent node the data that satisfies the splitting test of the parent node, while the right child node inherits the data that does not satisfy the splitting test. As the depth of the tree increases, the size of the data in the nodes is decreasing and also the probability that the data of each node belongs to only one class increases (leading to nodes of lower impurity concerning the classifying attribute).

During the construction of the tree, the goal at each node is to determine the best splitting test, which best divides the training records belonging to that leaf into the most homogeneous subsets concerning the classifying attribute. For the evaluation of alternative splitting tests in a node, various splitting indexes can be used. In this work the splitting index that has been used is the E*ntropy Index* (Han and Kamber, 2001). In order to find the best splitting test for a node, we examine, for each attribute of the data set, all the possible splitting tests that can be formed based on the values of this attribute. For each of these splitting tests, we calculate the Entropy Index. We finally select the splitting test with the lowest value of Entropy Index in order to split the node. The expansion of a Decision Tree can continue, dividing the training set into subsets (corresponding to new child nodes), until we have subsets-nodes with “homogeneous” records having all the same value of the classification attribute.

Finally the constructed Decision Tree structure is used as a classifier in order to classify new data sets; in our work we used this classifier to classify both the training data set and the test data set.

3.4 Pruning of Decision Tree Classifier Structures

The expansion of a Decision Tree structure, as described in Sect. 3.3, usually leads to a Decision Tree structure that is over-fitted to the training data set and to the noise this data set contains. Consequently such a Decision Tree structure does not have good generalization capabilities (i.e. high classification accuracy for other test data sets). Stopping rules can therefore be adopted in order to prevent such an over-fitting; this is usually referred to as Decision Tree Pruning (Weiss and Indurkhya, 1994). A simple stopping rule that has been examined in this research work was a restriction concerning the minimum node size. The minimum node size sets a minimum number of records required per node (pre-pruning approach); in our work for the pruned Decision Tree structure described in Sect. 4.3 a node was not split if it had less records than a predefined percentage of the initial training data set records (percentages 5%, 10%, 15%, and 20% were tried).

3.5 Selection of Training and Test Pattern Sets

In practical situations, where the acquisition of heart sound signals for all probable cases is time consuming and almost impossible, we are very much interested in the capability of the constructed-trained Decision Tree structure to generalize successfully. As a first step in order to examine the generalization capabilities of the constructed Decision Tree structures and architectures, the complete pattern set was divided in two subsets. The first subset included 54 patterns, 18 containing A2_P2, 14 containing OS and 22 containing S3 (40% of each class of the heart sound patterns dataset) randomly selected out of each pattern class (A2_P2, OS and S3) of the patterns set. This subset was used as the training set. The other subset included the remaining 81 patterns (28 belonging to the A2_P2 class, 22 belonging to the OS class and 31 belonging to the S3 class) and was used as the test set. In this way the first training-test set scheme was developed. The division of the pattern set was repeated, keeping the same proportions (40% training set – 60% test set), but using different patterns, in this way giving a second scheme. In the same way were created 8 more schemes with different proportions, in order to examine the impact of the training and test data set size on the performance of the Decision Tree classifier. We can see these 10 schemes in Table 1.

Table 1. The training-test set data schemes that were tried

Schemes	Number of records of Training data set	Number of records of Test data set
40% a	54	81
40% b	54	81
50% a	67	68
50% b	68	67
60% a	81	54
60% b	81	54
70% a	95	40
70% b	95	40
80% a	108	27
80% b	108	27

Next, as a second step in this direction, in order to examine more systematically the generalization capabilities of the constructed Decision Tree structures and architectures, the k-fold cross-validation method was used (Han et al., 2001). According to this method the initial dataset is partitioned into k mutually exclusive subsets (or 'folds') of approximately equal size $S_1, S_2, \ldots, S_K$. Then training and testing is performed k times. In the j^{th} iteration, the subset S_j is used as the test set, while the remaining k−1 subsets are used as the training set. Finally the accuracy estimate is calculated as the average of the accuracies (percentages of correct classifications) achieved in the above k iterations. In particular, according to the recommendations of the literature, we performed 5-fold validation and 10-fold validation.

4 Results of differentiation of A2_P2, OS and S3, with Decision Tree Classifier

4.1 Relevance analysis

In order to examine the relevance and the contribution to the differentiation of A2_P2, OS, and S3 for each of the above mentioned 96 heart sound features, the Uncertainty Coefficients were calculated for each of them, considering the OS_AP_S3 field as the classifying attribute. The calculation was made separately for the training data set of each of the 10 schemes outlined in Table 1 and for each heart sound feature. Then we calculated the average value and the standard deviation of the Uncertainty Coefficients, taking into account the 10 values that were calculated from these 10 schemes. The average values and the standard deviations of the Uncertainty Coefficient for the most important features are presented in Fig. 4. We remark that the most relevant features are the frequency features (i.e. E_dias_hf = High Frequency Energy in diastolic phase, E_dias_mf = Medium Frequency Energy in diastolic phase, E_sys_hf = High Frequency Energy in systolic phase, E_dias_mh = Medium High Frequency Energy in diastolic phase, etc) and the morphological features that describe the S2 (i.e. s2_1 ... s2_7). These results are compatible with our physical understanding of the problem; in particular the A2_P2, OS and S3

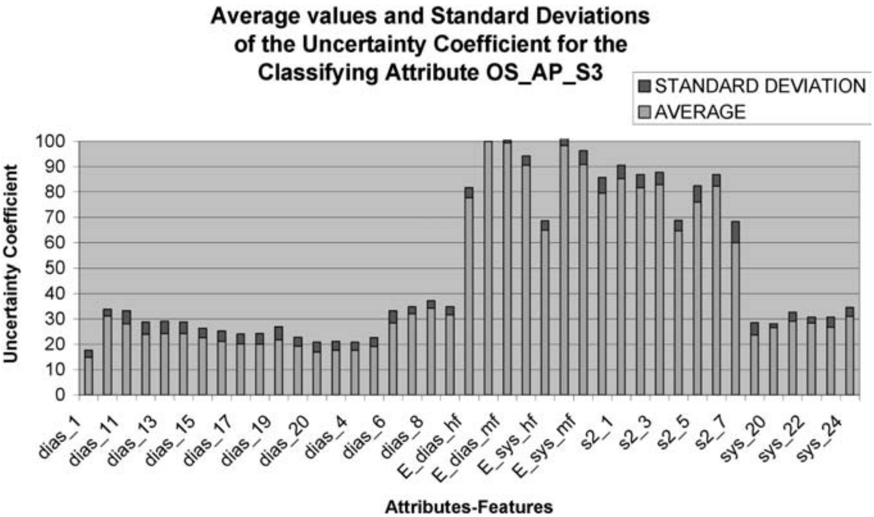

Fig. 4. Average values and Standard Deviations of the Uncertainty Coefficient for the most important features regarding the OS_AP_S3 as classifying attribute

click-like sounds are almost appearing simultaneously with S2 and usually each of these click-like sounds is related to specific heart diseases, which have heart sound murmurs in the systolic and the diastolic phase. The heart sound murmurs related to these heart diseases differ mostly in the frequency content and less in the envelope shape. The standard deviation values are generally smaller that 5%, showing that the Uncertainty Coefficients calculated from each scheme separately, especially the ones of the most relevant features, are similar and consistent.

As we have already mentioned among the evaluated architectures is the one based on three Detection Decision Trees. Each Detection Decision Tree corresponds to one out of the three morphological characteristics, i.e. one Detection Decision Tree for OS, one Detection Decision Tree for A2_P2 and one Detection Decision Tree for S3. Each of these Detection Decision Trees determines whether the corresponding morphological characteristic exists or not in the examined heart sound signal. In order to study the relevancies for this architecture we repeated the Relevance Analysis three times: the first time using the existence of OS as the classifying attribute, the second time using the existence of A2_P2 as the classifying attribute, and the third time using the existence of S3 as the classifying attribute. The results concerning the average values of the Uncertainty Coefficient are shown in Fig. 5; they are similar to the ones of Fig. 4, confirming the above relevant conclusions.

4.2 Fully Expanded Decision Tree

According to the methodology described in Sect. 3.3, initially we constructed the Differentiation Decision Tree (DiDT) structure with no restriction to the nodes

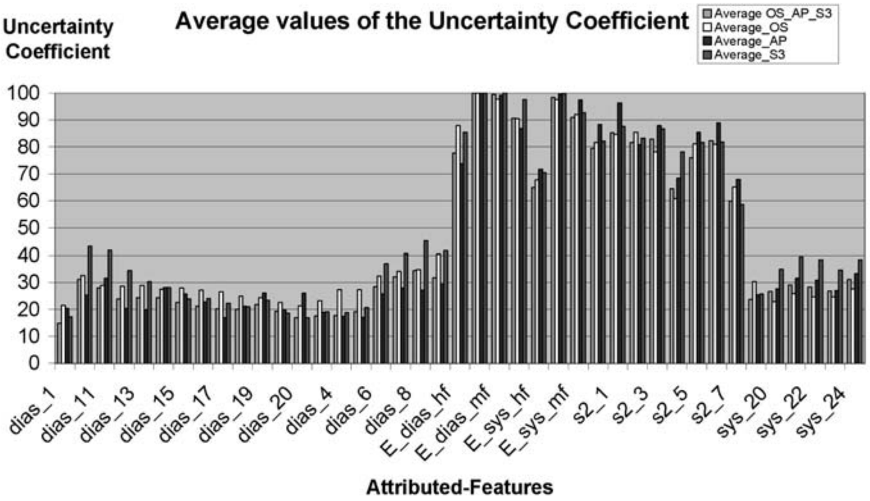

Fig. 5. Average values of the Uncertainty Coefficient for the most important features regarding the OS_AP_S3, the OS, the AP and the S3 as classifying attributes

(without pruning). The training data set we initially used was from the scheme 40%a (Table 1) and had 54 records (heart sound patterns). The remaining 81 patterns were used as a test set. The DiDT was constructed based on the training data set and regarding the OS_AP_S3 attribute as the classifying attribute (dependent variable). Then using this DiDT the patterns of the test data set were classified in order to examine the generalization capabilities of this Decision Tree. Finally the percentage of the correctly classified patterns of the test data set (according to the OS_AP_S3 attribute) was calculated. The classification performance for the 40%a scheme was: 56 patterns were classified correctly (14 OS, 18 AP and 24 S3), while for 25 the classification was wrong (8 OS, 10 AP, 7 S3); therefore the percentage of classification accuracy defined as:

$$Accuracy = \frac{\text{correctly\_classified(OS sounds} + \text{A2\_P2 sounds} + \text{S3 sounds)}}{\text{tested(OS sounds} + \text{A2\_P2 sounds} + \text{S3 sounds)}} \tag{1}$$

is 56/81*100 = 69.14%.

It should be noted that the structure of this DiDT (which has been constructed using all the attributes and features) uses 13 attributes out of the 96.

In order to exploit the results of the Relevance Analysis, described in Sect. 4.1, we constructed five more DiDTs using only the heart sound features that have Uncertainty Coefficient above i) 50%, ii) 60%, iii) 70%, iv) 80% and v) 90%. For example in the case v of 90% the features that were used were the following five: 1) E_dias_lf, 2) E_dias_mf, 3) E_dias_mhf, 4) E_sys_lf, and 5) E_sys_mf. These five DiDTs were used for classifying the patterns of the test data set; the results were almost identical to the ones achieved by using all the available features. Therefore using only 5 of the 96 features we can get almost identical levels of classification accuracy with much less computational effort.

Then for each of the data schemes of Table1 we repeated the above calculation and finally we calculated the average classification accuracy and its standard deviation for the above five cases i to v (= using different numbers of features, according to their Uncertainty Coefficients). The results are shown in Fig. 6. We remark from Fig. 6 that the classification accuracy results for cases i to v are identical, and the standard deviation is very low (zero in most of the cases) confirming the above-mentioned relevant conclusions. These results confirm once more the conclusion from relevance analysis results i.e. the A2_P2, OS and S3 click-like sounds are usually related to specific heart diseases, which have specific heart sound murmurs in the systolic and the diastolic phase and these murmurs differ mostly in the frequency content.

The striped columns in Fig. 6 represent the mean values of the Average Classification Accuracies (for the above five cases i to v) of the schemes, having the same proportion between the number of records of the training data set and the number of records of the test data set; this is referred in Fig. 6 as mean/proportion (e.g. mean value between average classification accuracy of 40%a and of 40%b schemes, ... mean value between average classification accuracy of 80%a and of 80%b schemes). The mean/proportion values are rated between 66,5% and 68,5% with the only exception being the mean/proportion value of schemes 60%a and 60%b, which is considered as outlier. Also 5-fold cross-validation and 10-fold cross-validation was performed, resulting in average accuracies of 62,96% and 63,51% respectively. Therefore it is concluded that Differentiation Decision Trees if properly trained can give a low to satisfactory level of classification accuracy concerning the differentiation of OS, A2_P2 and S3. Also it is worth mentioning that the Classification Accuracy for the training data set was 100% for all the examined cases.

In order to study whether the classification accuracy can be improved, we examined the Detection Decision Tree architecture. For each data scheme outlined in Table 1, we constructed three Detection Decision Trees (DeDT), one for each of the three morphological characteristics, i.e. one Detection Decision Tree for OS, one Detection Decision Tree for A2_P2 and one Detection Decision Tree for S3. The first DeDT (OS DeDT), was constructed according to the methodology described in Sect. 3.3, for each data scheme, using the training data set of the corresponding data scheme, and regarding the existence of OS in the heart sound signal as the classification attribute (the existence_OS attribute was the dependent variable of the OS DeDT). The existence_OS attribute was related to the OS_AP_S3 attribute with the following generalization rule {if OS_AP_S3= “OS” then existence_OS=”yes” else existence_OS=”no”}. The second and the third DeDTs (A2_P2 DeDT and S3 DeDT), were constructed in the same way for each data scheme, but considering as classifying attributes the existence of A2_P2 (existence_A2_P2) and the existence of S3 (existence_S3) attributes respectively. Then using these DeDTs we classified

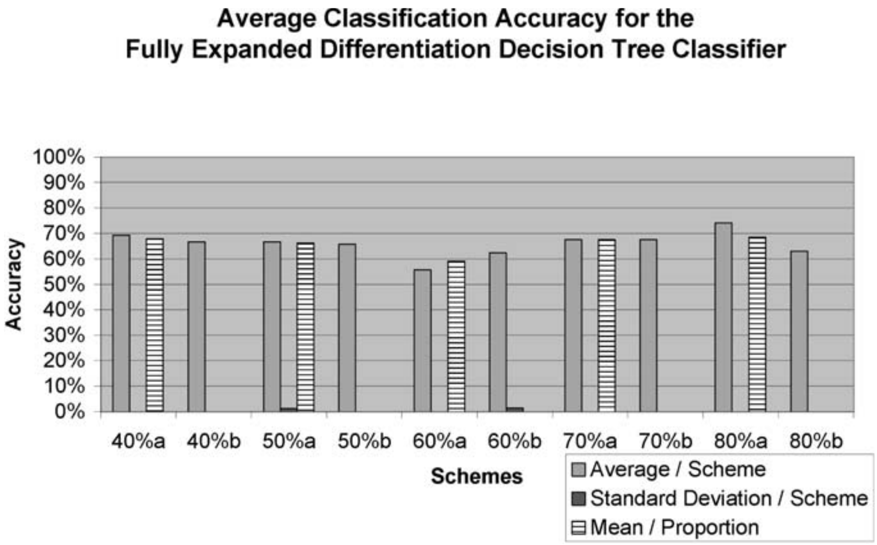

Fig. 6. Average Classification Accuracy for all data schemes and cases for the Fully Expanded Differentiation Decision Tree

the test data sets of the corresponding data schemes outlined in Table 1. Afterwards we calculated, for each data scheme, the classification accuracy for the corresponding OS DeDT, based on the results from the classifications of the corresponding test data sets. The OS DeDT classification accuracy is defined as:

$$Accuracy\_OS = \frac{\text{correctly classified(OS sounds + non OS sounds)}}{\text{tested(OS sounds + non OS sounds)}} \tag{2}$$

Using similar definition we also calculated the classification accuracies for the A2_P2 DeDTs and for the S3 DeDTs. The classification accuracy of A2_P2 DeDT is given by:

$$\text{Accuracy\_A2P2} = \frac{\text{correctly classified(A2P2 sounds + non A2P2sounds)}}{\text{tested(A2P2 sounds + non A2P2 sounds)}} \tag{3}$$

Respectively the classification accuracy of S3 DeDT is given by:

$$\text{Accuracy\_S3} = \frac{\text{correctly classified(}S3\text{ sounds + non }S3\text{ sounds)}}{\text{tested(}S3\text{ sounds + non }S3\text{ sounds)}} \tag{4}$$

Considering the Relevance Analysis described in Sect. 4.1, we also constructed, for each data scheme outlined in Table 1, five more DeDT architectures using only the heart sound features that have Uncertainty Coefficient above i) 50%, ii) 60%, iii) 70%, iv) 80% and v) 90%. Each of these architectures has also three Detection Decision Trees i.e. the OS DeDT, the A2_P2 DeDT and the S3 DeDT. All the DeDTs of each of the five architectures were used to classify the corresponding test data set according to the corresponding data scheme. Then the classification accuracies of these DeDTs were calculated. Finally the Average Classification Accuracy of the OS DeDTs, the A2_P2 DeDTs and the S3 DeDTs and also the Standard Deviation of them were calculated taking into account the values of all accuracies for cases i to v.

In order to simplify the presentation and the interpretation of the results we also calculated the Mean Value of the Average Classification Accuracies of the above five cases i to v (for each of the three kinds of DeDTs separately) of the schemes, having the same proportion between the number of records of the training data set and the number of records of the test data set. For instance we calculated the mean value between the OS DeDT Average Classification Accuracy of 40%a and of 40%b schemes, etc. These results are shown in Fig. 7. The standard deviation of all the examined cases was zero, proving once more that the 5 most relevant features out of the 96 features contain all the information we need for the differentiation of OS, A2_P2, and S3; for this reason the results of the calculations related to the standard deviation are not presented in Fig. 7. It should also be noted that the structures of the DeDTs (which has been constructed using all the attributes and features) used about 9 attributes out of the 96, so the computational effort was significantly reduced when the relevance analysis results were considered during the construction of the DeDTs.

We can see from Fig. 7 that for the schemes with the largest training data sets (= 80%a and 80%b schemes, with the training data set of 108 records-patterns) we have almost the same level of classification accuracy, about 80%, for all three kinds of DeDTs. For the other schemes of the classification accuracy is around 75,5%. This is probably due to the smaller size of the corresponding training sets.

Also it should be noted that for some of these test patterns more than one of the above tree DeDTs detect the existence of the corresponding morphological characteristics. For those patterns an arbitration mechanism is required, as described in the next Sect. 5.

Furthermore, a 5-fold cross-validation was performed, which resulted in average accuracies of 77,78%, 79,63% and 79,26% for the OS DeDT, the A2_P2 DeDT and the S3 DeDT respectively, and also a 10-fold cross-validation, which resulted in average accuracies of 77,53%, 76,98% and 78,46% respectively.

Therefore it is concluded that Detection Decision Trees if properly trained can give a higher level of classification accuracy than the Differentiation Decision Tree concerning the differentiation of OS, A2_P2 and S3. This is due to the higher simplicity and specialisation of each Detection Decision Tree (their task consists in a simple binary (Yes/No) decision concerning the existence or not of a specific morphological characteristic) in comparison with the Differentiation Decision Tree. Therefore the division of a complex diagnostic problem into simpler sub-problems, each of them being handled by a separate specialised decision tree, can result in significant increase of classification accuracy. This conclusion opens up a huge research area aiming at the development of effective 'architectures' of decision trees

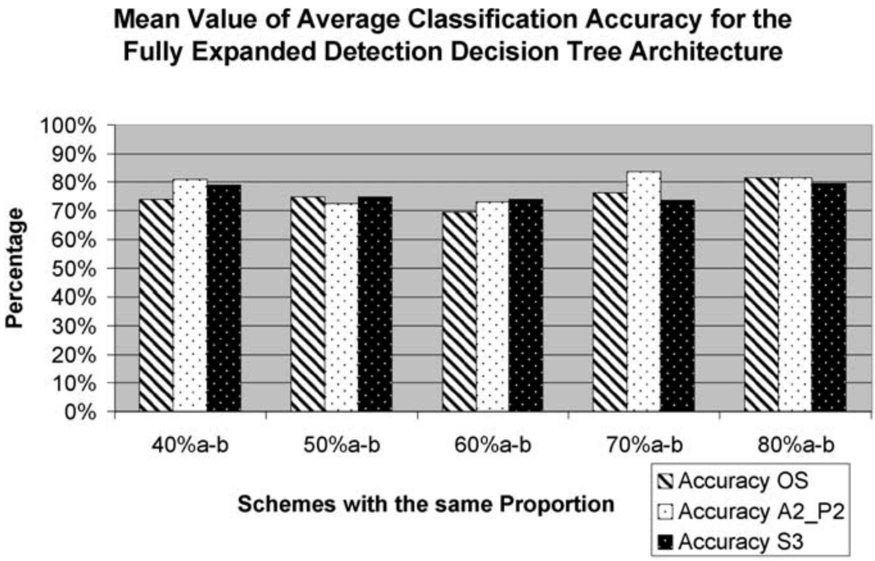

Fig. 7. Mean Value of Average Classification Accuracy for all cases and data schemes with the same proportion for the Fully Expanded Detection Decision Tree Architecture

networks (i.e. with multiple detection decision trees, arbitration decision trees, etc.) for important categories of diagnostic problems. A first step towards this direction is described in Sect. 5.

From Fig. 7 we can see that the Classification Accuracies of the above three kinds of DeDTs is not the same. In order to study more systematically this observation we calculated the Average Classification Accuracies separately for the OS DeDT (= 75,22%), the A2_P2 DeDT (= 78,37%) and the S3 DeDT (= 76,26%).

Then we tried to test statistically the above conclusions by performing three t-tests. The hypothesis statements (H0 and H1) of these tests are described below:

- Hypothesis statements for the first t-test:
  H0:OS_Accur=A2_P2_Accur, H1:OS_Accur≠A2_P2_Accur
- Hypothesis statements for the second t-test:
  H0:S3_Accur=A2_P2_Accur, H1:S3_Accur≠A2_P2_Accur
- Hypothesis statements for the third t-test:
  H0:S3_Accur=OS_Accur, H1: S3_Accur≠OS_Accur

Assuming a level of significance (probability of rejecting H_0 while it is true) a=1%, we found that the null hypothesis H_0 should be rejected in the first and the second t-test and accepted in the third t-test. Therefore it is statistically confirmed that in general the above specialised DeDTs have different classification accuracies.

Also it is worth mentioning that the DeDT Classification Accuracies for the training data set was once more 100% for all the examined cases.

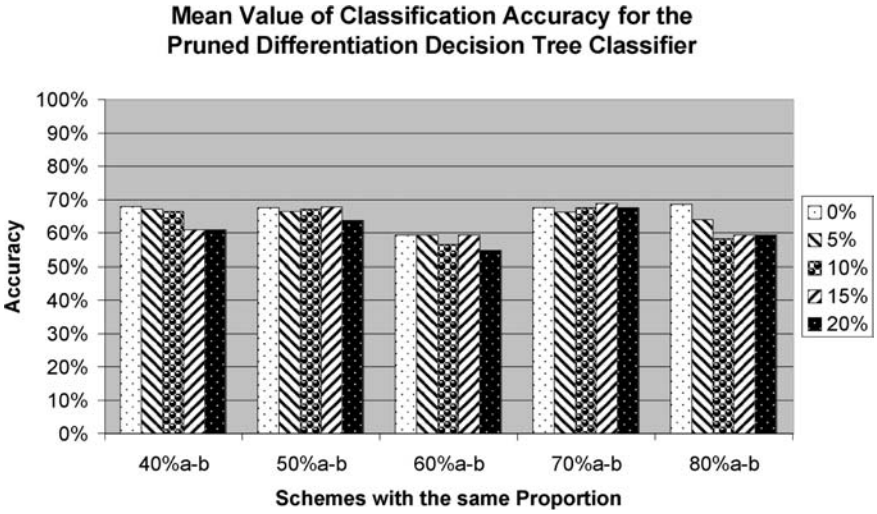

Fig. 8. Mean Value of Classification Accuracy for data schemes with the same proportion for the Pruned Differentiation Decision Tree

4.3 Pruned Decision Tree

In this part of our work we tried to examine if we could improve the generalization performance of the Decision Trees by placing a restriction during the training phase concerning the creation of new nodes of the Decision Tree and not allowing it to be fully expanded, as described in 3.4. The restriction we placed was that the number of records-patterns in a node, as a percentage of the total number of records of the initial training set, should be higher than a predefined percentage, usually referred to as the 'minimum support' of a node. The predefined percentages (minimum support levels) we examined were 5%, 10%, 15% and 20%. In particular, in the construction of a Decision Tree structure during the training phase, for each pattern node was examined the size (number of records-patterns) of its children nodes. If the size of one of these children nodes was less than the above predefined percentage of the initial training set, then the parent node was not further split, was converted to a leaf node and the Decision Tree was not further expanded from that path.

The same calculations, which are described in Sect. 4.2, were repeated with pruning (namely the only difference being that the leaf nodes of the Decision Tree had at least the predefined minimum support level, using all the available features).

In Fig. 8 are shown the mean values of the Classification Accuracies, which were achieved with the Pruned Differentiation Decision Tree, for the data schemes, having the same proportion between the number of records of the training data set and the number of records of the test data set. The cases of Pruned Differentiation Decision Tree that are shown are the ones with minimum leaf node support at least 5%, 10%, 15%, and 20%, compared with the previous case of the Fully expanded Decision Tree (0% minimum support at leaf nodes).

From Fig. 8 we can see that in 2 out of the 5 cases (50%a–b, 60%a-b) Pruned Differentiation Decision Tree gives almost the same mean value of Classification Accuracy with the corresponding Fully Expanded Differentiation Decision Tree, in one case (70%a–b) pruning slightly increases the mean value of Classification Accuracy, while in the remaining 2 cases (40%a–b, 80%a–b) pruning slightly decreases the mean value of Classification Accuracy. Therefore it is concluded that for our specific differentiation problem pruning does not significantly change the Classification Accuracy of the Differentiation Decision Trees. We performed four individual t-tests, to test statistically this conclusion. The hypothesis statements ($\mathrm{H}_{0\mathrm{X}\%}$ and $\mathrm{H}_{1\mathrm{X}\%}$) were defined as:

$$H_{0X\%} : support\_0\%\_Accur = support\_X\%\_Accur$$
$$H_{1X\%} : support\_0\%\_Accur \neq support\_X\%\_Accur$$

The X% stands for the minimum support levels i.e. 5% or 10% or 15% or 20%, used separately in each one of the four t-tests. Assuming a level of significance a=1%, we found that the null hypothesis $\mathrm{H}_{0\mathrm{x}\%}$ should be accepted. This statistical result confirms once more the above conclusion, namely the pruning of the Differentiation Decision Tree does not significantly change the classification accuracy.

We also examined the impact of pruning to the Classification Accuracy of the Detection Decision Tree Architecture. The Detection Decision Tree Architecture, as we have already mentioned in 4.2, has three Decision Trees, namely the OS DeDT, the A2_P2 DeDT and the S3 DeDT. Each of these Detection Decision Trees was studied independently from the other two, in a way similar to the one used above for the Differentiation Decision Tree, i.e. performing the same calculations, with the same restrictions to the leaf nodes (5%, 10%, 15% and 20%), the same attributes (as independent variables), the same data schemes and generally the same cases. In Fig. 9 are shown the mean values of classification accuracies related to the OS DeDT, while the ones for A2_P2 DeDT and S3 DeDT are shown in Fig. 10 and Fig. 11 respectively.

From Fig. 9 we can see that in 2 out of the 5 cases (40%a–b, 70%a–b) Pruned OS DeDT gives almost the same mean value of Classification Accuracy with the corresponding Fully Expanded OS DeDT, in 2 cases (50%a–b, 60%a–b,) pruning slightly increases the mean value of Classification Accuracy, while in the remaining one case (80%a–b) pruning slightly decreases the mean value of Classification Accuracy.

From Fig. 10 we can see that in 3 out of the 5 cases (50%a–b, 70%a–b, 80%a–b) Pruned A2_P2 DeDT gives almost the same mean value of Classification Accuracy with the corresponding Fully Expanded A2_P2 DeDT, while in the remaining 2 cases (40%a–b, 60%a–b) pruning slightly decreases the mean value of Classification Accuracy.

From Fig. 11 we can see that in one out of the 5 cases (50%a–b) Pruned S3 DeDT gives almost the same mean value of Classification Accuracy with the corresponding

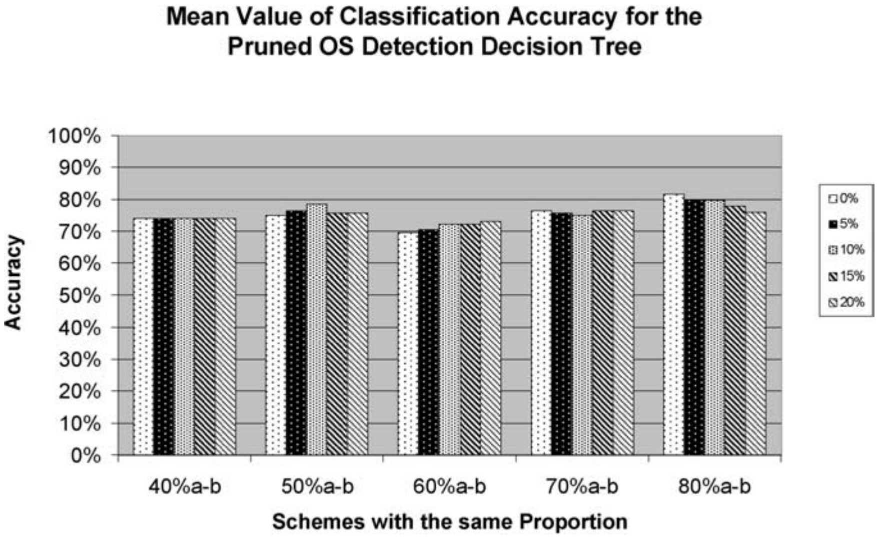

Fig. 9. Mean Value of Classification Accuracy for data schemes with the same proportion for the Pruned OS Detection Decision Tree

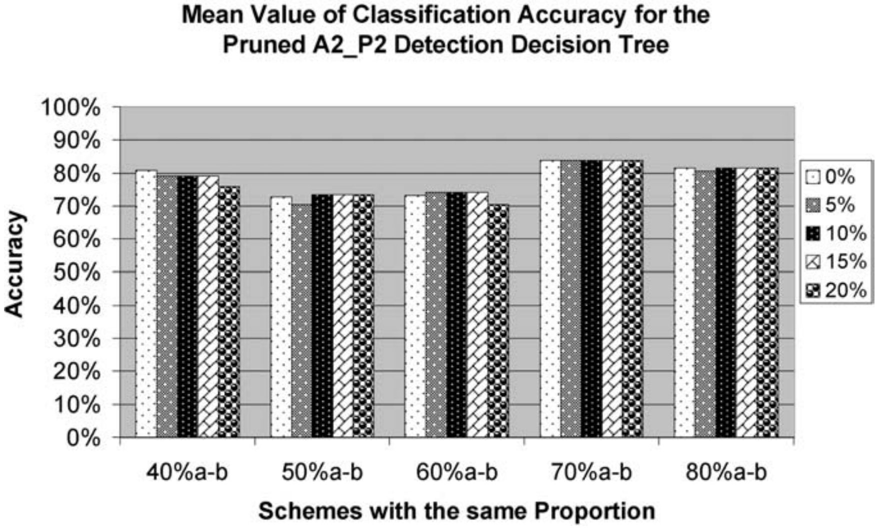

Fig. 10. Mean Value of Classification Accuracy for data schemes with the same proportion for the Pruned A2_P2 Detection Decision Tree

Fully Expanded S3 DeDT, in one case (70%a–b) pruning slightly increases the mean value of Classification Accuracy, while in the remaining 3 cases (40%a–b, 60%a–b, 80%a-b) pruning slightly decreases the mean value of Classification Accuracy.

Considering the results from Figures 9, 10, and 11, and after performing statistical hypothesis testing (using the appropriate t-tests), we concluded that the pruning

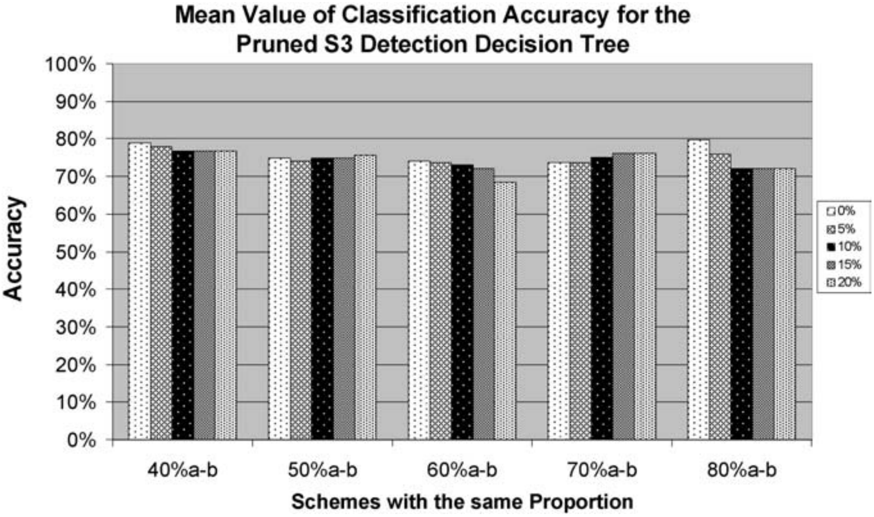

Fig. 11. Mean Value of Classification Accuracy for data schemes with the same proportion for the Pruned S3 Detection Decision Tree

does not significantly change the Classification Accuracy of the Detection Decision Tree Architecture, i.e. neither improves nor worsens the Classification Accuracy.

It should be noted that in most of the examined cases pruning changes the structure of the corresponding Decision Trees, and either improves or worsens the classification accuracy of the Decision Trees (DiDTs and DeDTs). However, the average impact of pruning in classification accuracy is not statistically significant, probably due to the relatively small training and test data set. This shows that the training data set is not consistent, despite the extensive preprocessing they undergo (that reduces to some extent the differences among the patterns of this data set) according to Sect. 3.2.

In general we believe that in the ‘real life’ the problem of differentiation among OS, A2_P2 and S3 needs a mechanism to handle the over-fitting problems, which we will definitely have, given the numerous factors affecting the characteristics of the acquired heart sound signals, that have been described in 3.1. For these reasons in our future research work we are going to use bigger training and test data sets, and also investigate various post-pruning approaches and combined approaches.

5 Development of Multiple Decision Tree Architecture

An effective medical diagnosis Decision Support System, in order to assist the clinicians to make better diagnoses, should provide reliable and consistent diagnostic suggestion and advice, especially concerning new and unknown cases. In particular, considering the diagnostic problem studied in this paper, an effective Decision

Support System for the differentiation of OS, A2_P2 and S3 should have high classification accuracy in new test data sets i.e. high generalization capabilities.

For this purpose, based on the conclusions drawn from the simpler architectures examined in the previous sections, a Multiple Decision Trees Architecture has been developed. This proposed Multiple Decision Tree Architecture (Stasis et al., 2003) consists of a first stage with various Suggestion modules and a second stage with an Arbitration module. The Suggestion modules can be either Detection Decision Trees, or Differentiation Decision Trees, or even be based on any other classification method (e.g. patterns recognition, neural network, etc.). The Arbitration module decides, whenever the above Suggestion modules produce inconsistent suggestions, which of these suggestions should be accepted and which of them should be rejected. The arbitration rules are defined considering the specific physical problem and the type of the available suggestions proposed by the Suggestion modules.

For instance in Fig. 12A, is shown a Multiple Decision Tree Architecture for the Differentiation among OS, A2_P2 and S3. This architecture combines the Differentiation Decision Tree Architecture (DiDT) and the Detection Decision Tree Architecture (DeDT). The DiDT Architecture provides one final suggestion for each heart sound pattern, but the classification accuracy of the test data set is not very high, according to 4.2. The Detection Decision Tree Architecture has better classification accuracy, but provides three suggestions for each heart sound pattern. If these three suggestions are not consistent, they may be confusing and probably not useful to the clinician. Totally the available suggestions, using the Multiple Decision Tree Architecture shown in Fig. 12A, are the three suggestions from the DeDTs and the one suggestion from the DiDT. So the DeDTs suggestions are consistent only when one of them detects the corresponding morphological characteristic, and the other two do not detect their corresponding characteristic. The final suggestion is consistent when the suggestion of DeDT architecture is the same with the suggestion of DiDT architecture.

Therefore, given that the DeDT suggestion is preferable (because the DeDT architecture has better classification accuracy for this problem than the DiDT architecture), the arbitration rule for this case can be: “If only one DeDT detects its corresponding characteristic then the final suggestion is this DeDT suggestion, otherwise the final suggestion is the DiDT Suggestion”.

It is obvious that we may have suggestions from various additional Suggestion modules, such as one Decision Tree specialized in the differentiation between OS and A2_P2, one specialized in the differentiation between OS and S3, and another one specialized in the differentiation between A2_P2 and S3 etc (Fig. 12B). Then the arbitration rules should be modified in order to take into account the new suggestions. Also, in order to save computer processing power, the Differentiation Decision Trees could be activated only if it is required, i.e. if the corresponding Detection Decision Trees give inconsistent suggestions. For this purpose a ‘multilayer architecture’ should be adopted: the Detection Decision Trees would constitute the first ‘layer’ of the architecture, the Differentiation Decision Trees would constitute

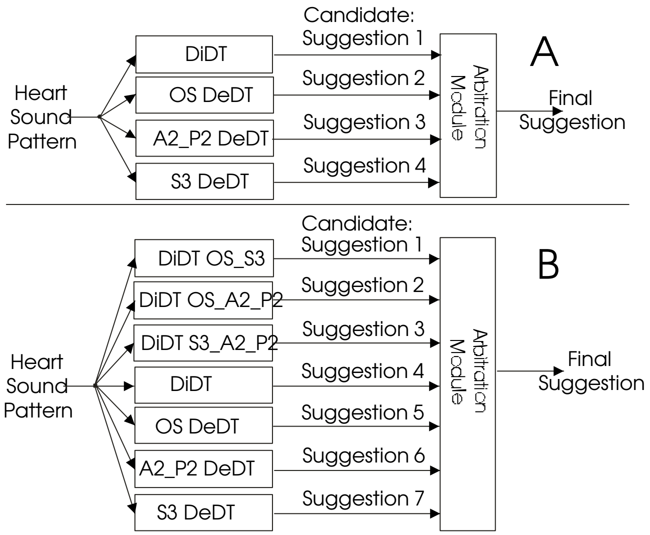

Fig. 12. Multiple Decision Tree Architecture

the second 'layer' activated by the first, and the arbitration module would be the third 'layer'.

We studied in detail the Multiple Decision Tree Architecture that is shown in Fig. 12A, using the corresponding arbitration rule described above. We repeated the same calculations as the ones described in Sect. 4.2, considering also the Relevance Analysis described in Sect. 4.1, in order to have comparable results with the DiDT Architecture. The results of the Multiple Decision Tree Architecture (of Fig. 12A) and the DiDT Architecture are shown in Fig. 13.

We remark that this new Multiple Decision Tree Architecture improves the classification accuracy in comparison with the DiDT Architecture. The classification accuracy of the 80%b data scheme was greater than 80%, indicating that the new Multiple Decision Tree Architecture if properly trained can achieve very good performance. This new architecture is statistically superior to the DiDT architecture, as concluded from the t-test we performed. In particular, the hypothesis statements (H0 and H1) for this test were:

$$\text{H0} : \text{Mul\_Accur} = \text{DiDT\_Accur}, \text{H1} : \text{Mul\_Accur} > \text{DiDT\_Accur}$$

Assuming level of significance a=1%, we found that the null hypothesis H_0 should be rejected.

We also performed a 5-fold cross-validation of the new Multiple Decision Tree Architecture, which resulted in an average accuracy of 69,84%, and a 10-fold cross-validation of it, which resulted in an average accuracy of 68,35%. These results confirm the above conclusions.

It should also be noted that this Multiple Decision Trees Architecture offers a critical advantage over the other classification methods (e.g. neural networks,

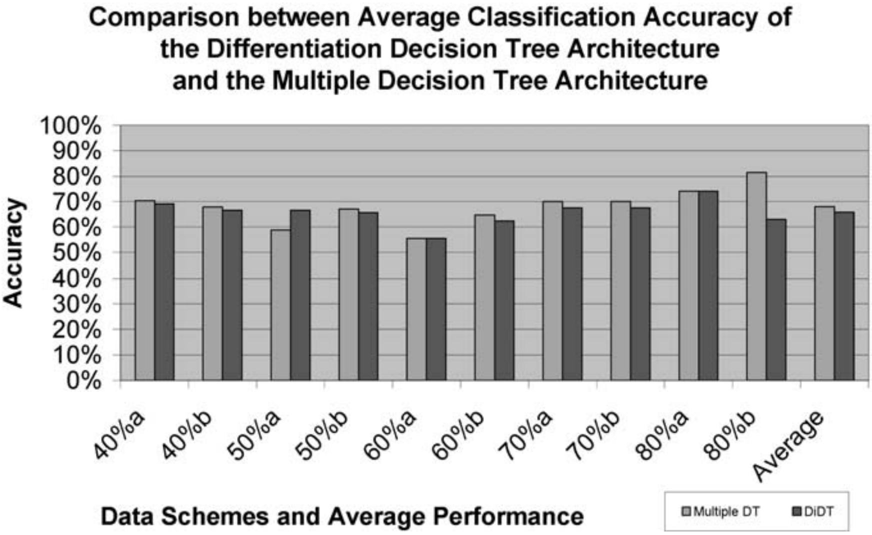

Fig. 13. Comparison, between the DiDT Architecture and the Multiple Decision Tree Architecture

pattern recognition, etc.); it provides the user-clinician not only a recommended diagnosis for the examined heart sound signal, but also a full justification of it as well, using a model that is similar to the differential diagnosis that clinicians are familiar with.

This Multiple Decision Tree Architecture can be generalized in order to offer decision support for the whole heart sound diagnosis problem. For this purpose it has to be extended towards a 'multilayer' network of nodes: each of these nodes will be specialised for handling a part of the initial highly complex problem and can follow any of the existing classification methods (e.g. neural network, pattern recognition, etc.); however decision trees are preferable, because they can provide fully justified suggestions. The first layers will consist of detection nodes; each of them will be specialised in the binary detection (Yes/No) of a specific morphological characteristic or disease. The second group of layers will consist of differentiation nodes; each of them will be specialised in the differentiation among a set of morphological characteristics or diseases with similar heart sound signals. Each of these differentiation nodes will be connected with the corresponding detection nodes of the first layers, which detect the corresponding morphological characteristics or diseases, and will be activated by them only if their suggestions are inconsistent. Finally the third group of layers will consist of arbitration nodes; their input will be the suggestions produced by the nodes of the above first and second layer, and based on them they will determine the final suggestion.

Given that there are many other medical diagnosis problems with similar characteristics, difficulties and complexities with the ones studied in the previous sections, the above 'extended' Multiple Decision Trees Architecture has a very wide applicability to all these medical diagnosis problems. Therefore a huge research area is opened up, aiming at the development of effective ‘architectures’ of such decision support networks for important categories of diagnostic problems.

6 Conclusions

The research work that has been described in this paper concerns the development of a Decision Support System Architecture for the heart sound diagnosis problem, and in general for complex medical diagnosis problems. The initial motivation for developing this Multiple Decision Trees Architecture has been the problem of Differentiation among Opening Snap (OS), 2nd Heart Sound Split (A2_P2), and 3rd Heart Sound (S3), which is a crucial and at the same time difficult and complicated part of the heart sound diagnosis problem. Heart sound diagnosis in general is very important especially in homecare, in rural areas and generally in the primary healthcare, where more sophisticated equipment is not available.

A number of Decision Tree architectures have been evaluated for the differentiation of OS, A2_P2 and S3. The main criterion for this evaluation was the classification accuracy of the training and test data set. The most important conclusions are the following:

- The Decision Tree algorithms can be used successfully as a basis for a Decision Support Systems, which will assist and advice the young and inexperienced clinicians in order to make better heart sound diagnosis. In this way the unnecessary patient transfers to Hospitals and the redundant examinations can be reduced. It should also be noted that such decision support systems should not be considered as a substitute for the clinician, due to the high complexity and life-criticality of the heart sound diagnosis.
- An extensive and sophisticated pre-processing of the initial heart sound signals is initially required, because of the high complexity and the high levels of noise of these signals
- Using Relevance Analysis we can determine a small subset of the initial set of features, that contains most of the information required for the differentiation. In this way the required computational effort for constructing, training and using these Decision Trees can significantly decrease.
- The DiDT architecture provides consistent suggestions but with rather low classification accuracy. The DeDT architecture provides higher classification accuracy, but its suggestions sometimes may be inconsistent. The Multiple Decision Tree Architecture is the preferable architecture, which achieved good classification accuracy and always provides consistent suggestions (Fig. 12A).
- In general the Decision Trees based methods in appropriate architectures can give satisfactory levels of generalization capabilities. This is very important, due to the difficulty and the high cost of having enough training data for every possible case.

- Increasing the size of the training data sets results in improvements of the Classification Accuracy achieved afterwards, using other test data sets, and consequently the general reliability of the whole Decision Support System.
- The proposed Multiple Decision Trees Architecture, and in general all the Decision Trees based architectures, offer a critical advantage over the other classification methods (e.g. neural networks, pattern recognition, etc.): they provide the user-clinician not only a recommended diagnosis for the examined heart sound signal, but also a full justification of it.
- This Multiple Decision Trees Architecture can be generalized in order to offer decision support for the whole heart sound diagnosis problem. The general heart sound diagnosis problem should be divided into a number of simpler subproblems, such as: detection of systolic murmur, detection of diastolic murmur, determination of the type of the murmurs (crescendo, decrescendo), determination of the frequency content (low, high, medium), detection of Mid-systolic click, arrhythmia, of premature ventricular contraction, differential diagnosis between heart diseases with similar heart sound signals, differentiation of OS, A2_P2, and S3, differentiation of the 4rth heart sound, ejection clicks and the split of S1 etc. All these simpler subproblems can be handled by separate specialized subsystems, which can be based on different methods, algorithms and features. The partial diagnosis given by these subsystems can then be combined to give a total diagnosis by an arbitration/combination subsystem. The combination of all these subsystems can lead to an integrated Decision Support System architecture for Heart Sound Diagnosis (Stasis et al., 2003).
- The above Multiple Decision Trees Architecture has a very wide applicability to many other medical diagnosis problems with similar characteristics, difficulties and complexities.

Further research is required for the design and evaluation of other structures based on the above Multiple Decision Tree Architecture (e.g. similar to the one shown in Fig. 12B), and also for the development of a systematic arbitration rule design methodology. This design of effective 'architectures' of such decision trees networks (i.e. with multiple Detection Decision Trees, Differentiation Decision Trees, Arbitration Decision Trees, etc.) for important categories of diagnostic problems is also an open research issue. Also further research is required in order to evaluate the applicability and suitability of such architectures for other significant problems in the area of heart sound diagnosis. Finally having the appropriate heart sound data set it would be very interesting to apply more sophisticated post-pruning approaches to the Decision tree structures.

References

Aggio S, Baracca E, Brunazzi C, Sgobino P, Aubert AE, Longhini C (1990) Clinical value of the pitch of the third heart sound in ischemic heart disease. *Cardiology* **77**(2): 86–92

Akay Y, Akay M, Welkowitz W, Kostis J (1994) Noninvasive detection of coronary artery disease. *IEEE Engineering in Medicine and Biology* **13**(5): 761–764

Criley SR, Criley DG, Criley JM (2000) Beyond heart sound: An interactive teaching and skills testing program for cardiac examination, Blaufuss Medical Multimedia, San Francisco, CA, USA. *Computers in Cardiology* **27**: 591–594

DeGroff CG, Bhatikar S, Hertzberg J, Shandas R, Valdes-Cruz L, Mahajan RL (2001) Artificial neural network-based method of screening heart murmurs in children. *Circulation* **103**: 2711–2716

Durand L, Guo Z, Sabbah H, Stein P (1993) Comparison of spectral techniques for computer-assisted classification of spectra of heart sounds in patients with porcine bioprosthetic valves. *Med Biol Eng Comput* **31**(3): 229–236

Han J, Kamber M (2001) *Data Mining: Concepts and Techniques*. Morgan Kaufman Publisher

Haghighi-Mood A, Torry JN (1995) A sub-band energy tracking algorithm for heart sound segmentation. *Computers in Cardiology* **Sept**: 501–504

Hebden JE , Torry JN (1996) Neural network and conventional classifiers to distinguish between first and second heart sounds. *IEE Colloquium (Digest)* 3/1–3/6

Hebden JE, Torry JN (1997) Identification of Aortic Stenosis and Mitral Regurgitation by Heart Sound Analysis. *Computers in Cardiology* **24**: 109–112

Kamber M, Winstone L, Gong W, Cheng S, Han J (1997) Generalisation and decision tree induction: Efficient classification in data mining. *Proceedings of Int'l Workshop on Research Issues on Data Engineering* (RIDE'97), pp 111–120

Koundourakis G (2001) *EnVisioner: A Data Mining Framework Based On Decision Trees*. Doctoral Thesis, University of Manchester Institute of Science and Technology.

Leung T, White P, Collis W, Brown E, Salmon A (1998) Analysing paediatric heart murmurs with discriminant analysis. *Proceedings of the 19th Annual conference of the IEEE/EMBS*, pp 1628–1631

Leung T, White P, Collis W, Brown E, Salmon A (2000) Classification of heart sounds using time-frequency method and artificial neural networks. *Proceedings of the 22nd Annual International Conference of the IEEE/EMBS* **2**: 988–991

Liang H, Lukkarinen S, Hartimo I (1997a) A heart sound segmentation algorithm using wavelet decomposition and reconstruction. *Proceedings of the 19th Annual International Conference of the IEEE/EMBS* **4**: 1630–1633

Liang H, Lukkarinen S, Hartimo I (1997b) Heart Sound Segmentation Algorithm Based on Heart Sound Envelogram. *Computers in Cardiology* **Sept**: 105–108

Liang H, Hartimo I (1998) A heart sound feature extraction algorithm based on wavelet decomposition and reconstruction. *Proceedings of the 20th Annual International Conference of the IEEE/EMBS* **3**: 1539–1542

Liu B, Xia Y, Yu P (2000) Clustering through decision tree construction. *Proceedings of the ACM International Conference on Information and Knowledge Management*

Livanos G, Ranganathan N, Jiang J (2000) Heart sound analysis using the S transform. *Computers in Cardiology* **Sept**: 587–590

Longhini C, Scorzoni D, Baracca E, Brunazzi MC, Chirillo F, Fratti D, Musacci GF (1996) The mechanism of the physiologic disappearance of the third heart sound with aging. *Japanese Heart Journal* **37**(2): 215–226

Lukkarinen S, Noponen A-L, Sikio K, Angerla A (1997) A New Phonocardiographic Recording System. *Computers in Cardiology* **24**: 117–120

Myint WW, Dillard B (2001) An Electronic Stethoscope with Diagnosis Capability. *Proceedings of the 33rd IEEE Southeastern Symposium on System Theory* **Mar**: 133–137

Sharif Z, Zainal MS, Sha'ameri AZ, Salleh SHS (2000) Analysis and classification of heart sounds and murmurs based on the instantaneous energy and frequency estimations. *Proceedings IEEE, TENCON* **2**: 130–134

Stasis AC (2003) *Decision Support System for Heart Sound Diagnosis, Using Digital Signal Processing Algorithms and Data Mining Techniques*. Phd Thesis, National Technical University of Athens

Stasis AC, Loukis EN, Pavlopoulos SA, Koutsouris D (2003) Using decision tree algorithms as a basis for a heart sound diagnosis decision support system. *Proceedings of the 4th Annual IEEE Information*

Technology Application in Biomedicine, pp 354–357

Tovar-Corona B, Torry JN (1998) Time-frequency representation of systolic murmurs using wavelets. *Computers in Cardiology* **Sept**: 601–604

Wang W, Guo Z, Yang J, Zhang Y, Durand LG , Loew M (2001) Analysis of the first heart sound using the matching pursuit method. *Med Biol Eng Comput* **39**(6): 644–648

Weiss SM, Indurkhya N (1994) Small sample decision tree pruning. *Proceedings of the 11th International Conference on Machine Learning*. Morgan Kaufmann, pp 335–342

White PR, Collis WB, Salmon AP (1996) Analysing heart murmurs using time-frequency methods. *Proceedings of the IEEE-SP International Symposium, Time-Frequency and Time-Scale Analysis* **June**: 385–388

Wu CH, (1997) On the analysis and classification of heart sounds based on segmental Bayesian networks and time analysis. *Journal of the Chinese Institute of Electrical Engineering, Transactions of the Chinese Institute of Engineers, Series E*, **4**(4): 343–350

Wu Y, Xu J, Zhao Y, Wang J, Wang B, Cheng J (1995) Time-frequency analysis of the second heart sound signals. *Proceedings of the IEEE/EMBS 17th Annual Conference* **1**: 131–132

Recorded Heart Sound Sources Educational Material:

Karatzas N, Papadogianni D, Spanidou E, Klouva F, Sirigou A, Papavasiliou S (1974) *Twelve Recorded Lessons with Heart Sound Simulator, Merk Sharp & Dohme Hellas*. Medical Publications Litsas, Athens

Criley JM, Criley DG, Zalace C (1995) *The Physiological Origins of Heart Sounds and Murmurs, Harbor UCLA Medical Center*. Blaufuss Medical Multimedia

Littman, *20 Examples of Cardiac & Pulmonary Auscultation*

Internet sites:

Cable C (1997) *The Auscultation Assistant*,
http: //www.wilkes.med.ucla.edu/intro.html.

Glass L, Pennycook B (1997) *Virtual Stethoscope*, McGill University, Molson Medical Informatics Project 1997,
http: //www.music.mcgill.ca/auscultation/auscultation.html.

VMRCVM Class of 2002, Virginia Maryland regional College of veterinary medicine,
http: //students.vetmed.vt.edu/2002/cardio/heartsounds.html.

Kocabasoglu YE, Henning RH, *Human Heart Sounds*,
http: //www.lf2.cuni.cz/Projekty/interna/heart_sounds/h12/index.html.

Frontiers in Bioscience, Normal and Abnormal EKGs and Heart Sounds,
http: //www.bioscience.org/atlases/heart/sound/sound.htm.

Student Internal Medicine Society, Class 2000 Cardiology Heart Sounds Wave Files, American College of Physicians, Baylor College of Medicine,
http: //www.bcm.tmc.edu/class2000/sims/HeartSounds.html.