ADAPTIVE RESONANCE THEORY: A FOUNDATION FOR "APPRENTICE" SYSTEMS IN CLINICAL DECISION SUPPORT?

by Robert F Harrison[*] , Simon S Cross[†] , R Lee Kennedy[‡] , Chee Peng Lim[*] and Joseph Downs[*]

Research Report 662

3 January 1997

Abstract

The idea of an "apprentice" system, in contrast to an expert system, is introduced, as one which continues, perpetually, to refine its knowledge-base. Neural networks appear to offer the necessary learning ability for this task, and the Adaptive Resonance Theory family is particularly suited to on-line (causal) learning. The ability of these networks accurately to represent decision problems and to disclose their acquired knowledge is discussed, and their practical application is assessed. Two problems of medical decision making are considered using the approach. The first is the early diagnosis of myocardial infarction from clinical and electrocardiographic data gathered at presentation. The second is the cytopathological diagnosis of breast lesions from fine needle aspirate samples. In both cases good performance is obtained along with sets of "if-then" rules which are in accordance with medical opinion. In the first case, examples of on-line learning are given and the system is seen to be behaving as expected, with performance improving with increasing sample size.

Introduction

apprentice *n*. Learner of a craft. [Old French *apprendre*]

In the field of clinical decision making, a decision aid which is able to continue to learn from "experience" is likely to have an advantage over one which isn't. For instance, a system which is developed from data gathered at one location should be able, safely, to tune-in to local conditions e.g. demography elsewhere. Similarly, as practice or technology changes, such changes should be accommodated by the device itself, rather than by having to involve statisticians, knowledge engineers etc. to re-derive algorithms. After all, when doctors change hospital they are not subjected to complete retraining. Neither should a computerised decision aid have to be. Of course, this assumes the need for such systems in the first place, which is a wider question not addressed here. We use the analogy of an apprentice to motivate development of systems which learn in perpetuity.

Expert systems are characterised by the processes of rule elicitation, rule-base development, and inference. Knowledge, in the form of rules, is built into the system *a priori* and, once embedded, remains unchanged throughout the lifetime of the system, unless a knowledge engineer intervenes. Conclusions are drawn by a deductive

[*] Department of Automatic Control and Systems Engineering, The University of Sheffield.
[†] Department of Pathology The University of Sheffield.
[‡] Department of Medicine, City Hospitals, Sunderland and School of Health Sciences, University of Sunderland

process. As a model of human expert behaviour this has some drawbacks because it presupposes that once individuals have achieved expert status they no longer continue to learn from their experience. In reality, experts are primed with knowledge, via schools, universities, on the job training etc., but ultimately become known as experts for what they know over and above what they have been taught, i.e. what experience has taught them or what they are able to deduce from their earlier knowledge. Indeed, expertise might well be thought of as that knowledge which does not exist in our primary repositories of knowledge (textbooks, lecture courses, etc.). Furthermore, it is well recognised that, even for a static expert system, the knowledge acquisition process is difficult and time-consuming .[1]

In contrast, we propose the idea of an "apprentice" system which attempts to model the human knowledge acquisition and inference process more closely, either by refining in-built, prior, knowledge or by developing a model of the problem domain from scratch and from example (i.e. by induction). In either case, the key feature is an ability to adapt, over time, in the light of experience. The ability of systems to "learn", incrementally, in this way is not something that expert systems in general possess. Neural networks, on the other hand, hold much promise for machine learning.

Looked at from a different perspective, expert systems have the advantage of being able to provide an explanation of their reasoning processes – an attractive property for the end-user – while neural networks have proven, in the main, unwilling to reveal the knowledge embedded within them, making potential beneficiaries of the technology wary of its adoption and raising a number of potential legal questions. [2] Some inroads have been made in addressing both of these problems: rule induction systems such as those based on "information gain" open the way towards automatic knowledge acquisition and update (see references 3-5), while rule extraction techniques attempt to "open the black box" of neural networks. [6-10]

As models of apprentice behaviour, mainstream rule induction and neural network techniques are hampered by the need, artificially, to suppress learning at some point prior to making the system operational. Thus the system, although apparently "learning" to solve the problem, in fact does not continue to learn into the future, that is, any learning that takes place is acausal (off-line). This is of course the conventional way of developing decision aids such as logistic regression models. Should, therefore, the problem characteristics change, perhaps owing to a change in practice (non-stationarity) or owing to differences between populations at different locations (inhomogeneity), or had there been an insufficient amount of representative information available at the time the system was established, the performance of these mainstream techniques may be severely compromised. Re-training on a new information set comprising both the original data, and any additional knowledge remains, by and large, the only solution, although techniques for incremental learning for both paradigms are beginning to emerge. The desirability of a system which can learn to improve its performance *in situ* and causally (on-line), without the intervention of a systems engineer, is evident.

The reason that learning must be suppressed derives from the so-called stability-plasticity dilemma. [11] This makes explicit the conflict between the need to retain previously learned knowledge (stability) and the ability to adapt to new information (plasticity), i.e. how can we prevent existing knowledge from being overwritten or corrupted by new information or noise? This problem is known as "catastrophic forgetting" [12] and besets the majority of machine learning paradigms.

Of those approaches which attempt to address this dilemma, the Adaptive Resonance Theory (ART) family of neural networks offers a number of significant advantages over the more common feedforward and competitive networks for the establishment of apprentice systems. These are:

- an ability to discriminate novelty from noise, and familiar (statistical) events from rare but important (outlier) ones;
- rapid learning based on predictive success rather than on predictive failure (mismatch);
- self-organisation, with few arbitrary parameters to tune, and automatic structure determination;
- linear rather than exponential scaling with problem size;
- straightforward revelation of embedded rule sets;
- inherently parallel implementation.

This is not to say that the establishment of ART-based systems is without its own problems, or indeed that ART is yet a mature technology. ART is under continual development and at present provides a way forward in this area. We shall explore some of ART's shortcomings at the appropriate points in the text.

Feedforward Neural Networks

Advances in neurocomputing have opened the way for the establishment of decision support systems which are able to learn complex associations by example. The main thrust of work in this area has been in the use of the feedforward networks (e.g. the multi-layer Perceptron (MLP) [13] or the Radial Basis Function networks (RBFN) [14]) to learn the association between evidence and outcome. Theoretical work in this area has led to the discovery of two important properties of feedforward networks:

- for one-from-many classification, their learning rules lead to an interpretation of their outputs as estimates of the posterior (class conditional) probability distribution, conditioned on a set of evidence, provided that "optimality" is attained; [15,16]
- architectures such as the MLP or the RBFN have been shown to be rich enough in structure so as to be able to approximate any (sufficiently smooth) function with arbitrary accuracy. [17,18]

It can be inferred from these facts that, given sufficient data, computational resources[1] and time[2], it is possible, using a feedforward network, to estimate the Bayes-optimal classifier to any desired degree of accuracy, directly and with no prior assumptions on the probabilistic structure of the data (e.g. independence). This is an attractive scenario and has been extensively exploited, although in the absence of a concrete set of design and validation criteria the establishment of such systems relies heavily on trial and error and cross validation. Indeed, it can be argued that the establishment of networks of the feedforward class is nothing other than non-linear regression, but, in the main, without the advantage of the extensive body of design, analysis and validation tools which have been developed within that branch of statistics, although this situation is changing. [19,20]

---

[1] The MLP, in particular, does not scale well with problem size.

[2] Non-linear optimization which is non-linear in the parameters may be time consuming to perform, numerically and much trial and error may be required in deriving an adequate network architecture.

However, contrasted with this must be the fact that the feedforward paradigm is intuitively appealing, straightforward to implement and has been taken up by a much wider community than has ever adopted non-linear statistics.

The inherent adaptability of feedforward neural networks may make it easier to tune-in to local conditions but would still require significant intervention and additional effort in data capture, retraining and revalidation. Indeed, the process of establishing such a system is precisely the same as that of establishing any other statistical classifier.

Feedforward networks are static devices in operation, and fail to cope with the stability-plasticity dilemma other than by suppressing learning after acceptable performance is attained. The system is then put into operation. Implicit in this is the assumption that a trained network both represents the problem adequately at the time of development and continues to do so into the future, or in remote locations. Should learning remain continuously active in feedforward networks, new data will be learned indiscriminately[3], with the attendant risk of serious performance degradation. [12]

Adaptive Resonance Theory

An entirely different approach, utilising a network comprising both feedforward and feedback components has been taken by Carpenter and Grossberg and colleagues, [11,21-25] which overcomes the stability-plasticity dilemma. This has resulted in the Adaptive Resonance Theory family of architectures which seek to model biological and psychological properties of the brain, rather than being derived from a data processing perspective. In their earliest manifestations these were unsupervised systems which autonomously learned to recognise categories of their own devising.

A schematic of a single ART module is shown in Figure 1. Here the intention is simply to describe the ART architectures in an informal way; the references [11,21-25] provide complete details. ART modules use feedback to compare the existing state of knowledge or long term memory (LTM or weights) of the system with the current set of evidence and either: (i) adjust the LTM which codes for a particular category, to account for the current situation if this is "similar" enough to other patterns in that category; or (ii) initiate a new category which codes for the unrecognised (current) pattern. Similarity is measured by comparing the stored representation of the class (prototype) with the current input pattern ascertain how close they are according to some measure of distance. This has a major advantage from a design view point in that there is no off-line "hand crafting" of network architecture to be done, i.e. one autonomous network can address any problem or, indeed, many problems simultaneously. Also, commonly occurring patterns have the effect of reinforcing their category's ability to recognise like examples, while categories representing spurious events are rarely, if ever, excited again and so do not corrupt previously learned information. Conversely, should a rare but valid event occur, it will reside in LTM until next recalled.

---

[3] More recent developments which enable feedforward networks to "grow" their own architectures and to learn causally are emerging although these do not in general overcome the problem of catastrophic forgetting and thus are not well suited to pattern recognition.

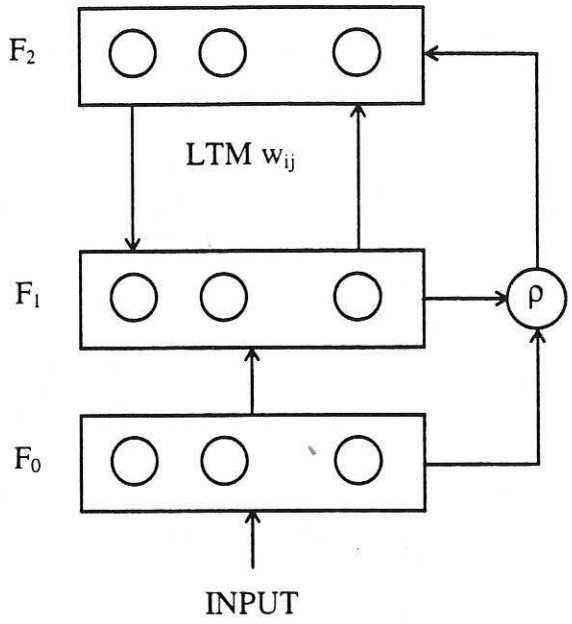

Figure 1 A single ART module comprising three layers, F_0, F_1 and F_2. F_1 and F_2 are fully interconnected in both directions via weighted links (w_{ij}) which form the LTM. ρ is the vigilance parameter which governs the coarseness of categorisation. F_0 buffers the input patterns so that they remain present during processing.

The ART architectures of interest here comprise two layers of nodes, fully interconnected in both directions, together with a layer which serves to distribute the input signal to the active components. These are the input/comparison field (F_1), and the output/recognition field (F_2) which latter implements a "winner-take-all" competition. F_0 acts merely as a buffer to register the current input during processing and comparison. Together F_1 and F_2 form an *attentional* subsystem which is complemented by an *orienting* subsystem which initiates search. ART takes its name from the interplay between learning and recall whereby signals reverberate between the two layers. When an input pattern is recognised, a stable oscillation (resonance) ensues and learning (adaptation) takes place. Categories are coded by the formation of templates in the competitive (F_2) layer (represented by the weight vector for a particular node) and these are refined as new information becomes available. During recall, when a given node is excited, a template is fed-back to the F_1 layer for comparison with the current input. The degree of match is assessed against the vigilance parameter (ρ) which is used to control the coarseness of categorisation. If the degree of match is not sufficiently good, parallel search is initiated until either an acceptable match is found (resonance) or the pattern is assigned to a new category (F_2) node.

ARTMAP

Single ART modules are restricted to unsupervised learning. This means that the autonomously selected categories are unlikely to correspond to meaningful categories in the problem domain. The so-called ARTMAP [26,27] family of architectures resolves this problem by providing a *mapping* network which is capable of supervised learning whilst retaining the desirable properties of the earlier ART networks. These networks comprise two ART modules (ART_a and ART_b) coupled via a *map* field. Each ART

module individually self-organises into categories representing data (evidence) and supervisory signal (target or outcome) and the association between categories is formed by the map field. Figure 2 presents the general ARTMAP configuration.

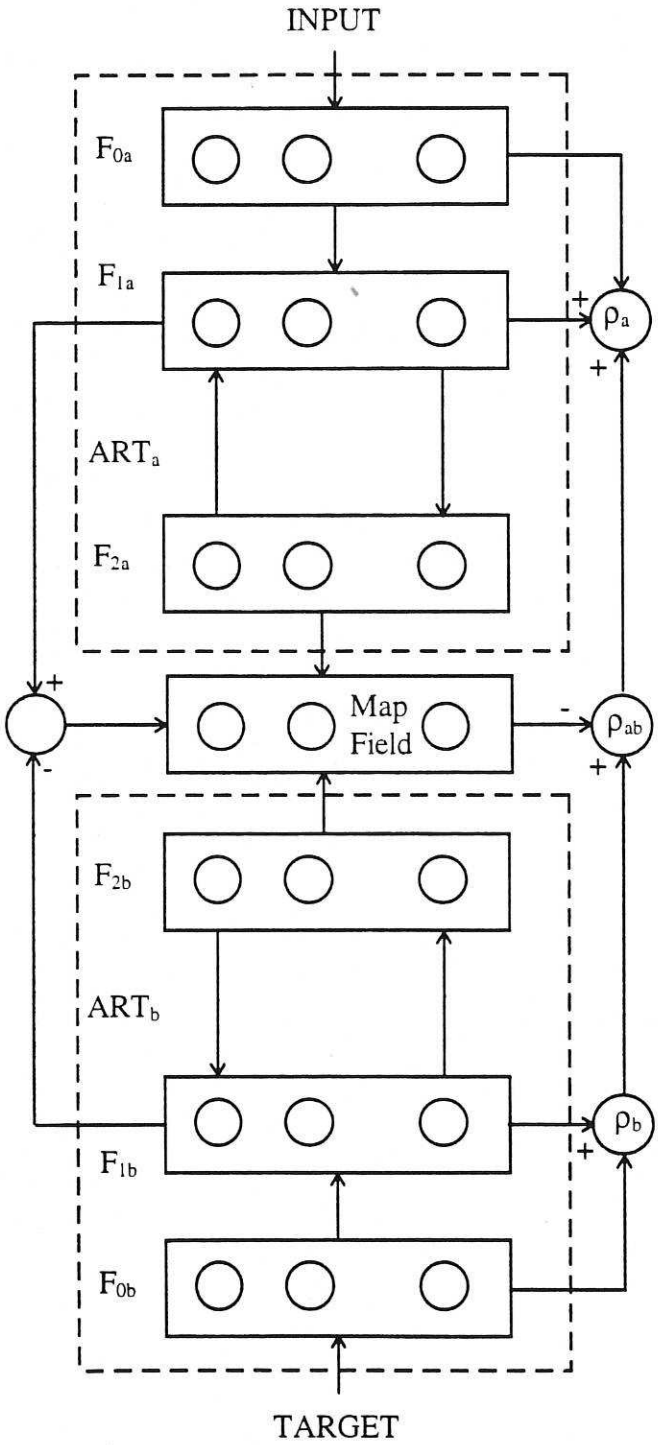

Figure 2 General ARTMAP configuration. This comprises two ART modules, labelled a and b, which self-organise the input and target data streams respectively. Categories formed for each of these are associated via the Map Field. Category size is determined for each module by its own vigilance parameter, and incorrect associations between ART_a and ART_b

categories are handled via the match tracking process, governed by the Map Field vigilance, ρ_{ab}.

In addition to the individual vigilance tests carried out for ART_a and ART_b, a further test is performed at the map field when both the ART modules are active (resonant). In this situation, a category prediction is sent from the winning node of the ART_a F_2 layer to the F_2 layer of ART_b and the so-called map field vigilance test is performed which determines whether or not the predicted class is equal to the actual class. If so, learning is permitted throughout ARTMAP (i.e. at ART_a, ART_b and in the map field). If not, an activity called *match tracking* will be triggered which initiates a search cycle in ART_a. The baseline ART_a vigilance is raised by this process by just enough to ensure that the ART_a vigilance test fails and the currently active node is thus deselected. A new winning node is selected from ART_a and a fresh prediction is sent to the map field. Match tracking therefore provides a means of selecting a node which satisfies both the ART_a and map field vigilance tests. If no such node exists the input is ignored. Full details of the ARTMAP learning procedure are given in. [26,27]

The basic ART and ARTMAP algorithms accept only binary valued inputs. However, by replacing the operations of bivalent logic (AND, OR) that take place in these networks, with their counterparts from fuzzy logic, a generalisation is obtained which accepts data on the interval zero to one. These networks are known as fuzzy ART and fuzzy ARTMAP (FAM) respectively. [22,26] Further developments which provide a Bayesian interpretation of ARTMAP operation in the sense that the outputs may be regarded as predictions of posterior or class conditional probabilities have recently been conducted. [28,29]

For computational efficiency, a simplified ARTMAP architecture results from noting that in one-from-many classification there is no need to self-organise the supervisory signal at ART_b because classes are predefined. [30]

ARTMAP networks are able to learn to improve their predictive performance on-line in non-stationary environments, using their entire memory capacities. Learning is driven by approximate (soft) match and takes place very rapidly as does recall or recognition – the basic theory, as opposed to the computational models allows for a full parallel implementation. Contrast this with the feedforward architectures. These learn off-line and assume a stationary environment. Learning must be suppressed to overcome the stability-plasticity dilemma and is: very slow, driven by mismatch; prone to spurious solutions; may scale poorly (e.g. exponentially) with problem size, and often requires lengthy cycles of "train and test" to arrive at a satisfactory solution. Recall, however, is very fast.

Two principal difficulties arise with the use of existing ART models: a local (as opposed to distributed) representation of information which arises through the adoption of a winner-take-all strategy in the competitive layer, and a sensitivity to the order in which stimulus data is encountered. The first is due to the assumptions made in deriving algorithms which are easily computed and owes nothing to the underlying theory. Indeed, a fully distributed ART model, dART, and its mapping equivalent has recently been proposed but its utility has yet to be evaluated. [31] The second is not, in fact, peculiar to ART but is rather a feature of all causal learning systems and is often present even in off-line training of feedforward networks, hence the need to "shuffle" the order of data presentation.

ARTMAP presents the prospect of an autonomous system capable of learning stably to categorise data whilst protecting the user from spurious predictions. This means that the system can safely continue learning *in situ*, whilst providing useful support. Thus, in clinical diagnosis, evidence would be presented. Should it excite a recognition category (from previous training) then a prediction is returned. Update of LTM can then be initiated if and when diagnosis is confirmed. If the current pattern is not recognised the user is so informed. Again adjustment of LTM is only initiated upon confirmation of the diagnosis. Provided diagnosis remains unconfirmed, no LTM adjustment takes place. This is a crucial issue in the development of a portable decision aid which should be able to adapt to local practice and to changing procedures, in much the same way as humans do.

Any decision making or diagnostic procedure where evidence is to be associated either with an objective outcome or with expert (subjective) opinion, is a potential application area for this approach and most importantly, it can put development (via, say, a fourth generation language) of decision aids into the hands of the domain expert, rather than the computing expert. This capability can be seen as crucial in overcoming resistance to the use of computational decision aids – the domain expert assumes "ownership".

Practical Strategies

Voting Strategy

As stated above the formation of category clusters in ARTMAP is affected by the order of presentation of input data items. [26] Thus the same data presented in a different order to different ARTMAP networks can lead to the formation of quite different clusters within the two networks. This subsequently leads to differing categorisations of novel data, and thus different performance scores. The effect is particularly marked with small training sets and/or high-dimensional input vectors.

A voting strategy can be used to compensate for the ordering problem (see reference 26). A number of ARTMAP networks are trained on different orderings of the training data. During testing, each individual network makes its prediction for a test item in the normal way. The number of predictions made for each category is then totalled and the one with the highest score (majority votes) is the final predicted category outcome. The voting strategy can provide improved performance in comparison with that of the individual networks. In addition it also provides an indication of the confidence of a particular prediction, since the larger the voting majority, the more certain is the prediction. Clearly strategies other than a simple majority can be used depending on the desired effect. Furthermore, recent work has indicated the effectiveness of other ways of combining outputs from multiple classifiers [32-34] such as via the Bayesian formalism [35] or the so-called Behaviour-Knowledge Space approach. [36]

Symbolic Rule Extraction

Most neural networks suffer from the opaqueness of their learned associations. [10] In medical domains, this "black box" nature may make clinicians reluctant to use a neural network based application, no matter how well it performs in a statistical sense. Thus, there is a need to supplement neural networks with symbolic rule extraction capabilities in order to provide explanatory facilities for the network's reasoning. ARTMAP

provides such a capability [37] as a result of its localised knowledge representation. Thus what is seen as a shortcoming from one angle becomes an advantage from another.

Rule extraction from feedforward networks has proved to be a difficult problem and, although some progress has been made, [6,7,10,38-40] it seems that the feedforward paradigm is not a natural one for semantic interpretation. The act of rule extraction is a straightforward procedure in ARTMAP compared with that required for feedforward networks since there are no hidden units with implicit meaning. In essence, each category cluster in ART_a represents a symbolic rule whose antecedents are the category prototype weights, and whose consequent is the associated ART_b category (indicated by the map field).

ARTMAP's symbolic rules also differ from those of conventional expert systems as regards the way they are matched to input features. Expert system rules are "hard" – an input must match to each and every feature in a rule's antecedent before the consequent will be asserted. In ARTMAP the rules are "soft". Recall that they are derived from prototypical category clusters which are in competition with each other to match to the input data. Exact matching between inputs and categories is not necessary, merely a reasonably close fit suffices. (The degree of inexactness that is tolerated being determined by the value of the ART_a vigilance parameter.) This provides greater coverage of the state space for the domain using fewer rules.

A drawback of the approach is that the rules are "correlational" rather than causal, since ARTMAP possesses no underlying theory of the domain but simply associates conjunctions of input features with category classes. Of course, this problem is not specific to ARTMAP but occurs with neural networks generally, being based upon an inductive rather than a deductive mechanism. Nonetheless, useful diagnostic performance can often be achieved from correlational features without recourse to any "deep" knowledge of the domain.

Category Pruning

An ARTMAP network often becomes over-specified to the training set, generating many low-utility ART_a category clusters which represent rare but unimportant cases, and subsequently provide poor-quality rules. The problem is particularly acute when a high ART_a baseline vigilance level is used during training. To overcome this difficulty, rule extraction involves a pre-processing stage known as category pruning. [37] This involves the deletion of these low utility nodes. Pruning is guided by the calculation of a confidence factor (CF) between nought and one for each category cluster, based equally upon a node's usage (proportion of training set exemplars it encodes) and accuracy (proportion of correct predictions it makes on a separate data sample, known as the prediction set). All nodes with a confidence factor below a user-set threshold are then excised.

The pruning process can provide significant reductions in the size of a network. In addition, it also has the very useful side-effect that a pruned network's performance is usually superior to the original, un-pruned net on both the prediction set and on entirely novel test data.

In the original formulation of the pruning process, a uniform CF threshold is used to select nodes for deletion, irrespective of their category class. [37] We have since generalised the pruning process to allow separate CF thresholds for nodes belonging to different category classes. [41,42] This allows us to vary the proportion of the state-space

covered by different categories and is useful for medical domains since it allows an ARTMAP network to be pruned so as to trade sensitivity for specificity and vice versa.

Generalisation of the category pruning process enabled us to devise a novel "cascaded" variant of the voting strategy to be employed as shown in Figure 3. [41,42] This comprises three layers, a set of voting networks pruned so as to maximise sensitivity, another set pruned so as to maximise specificity, and a third set of voters pruned so as to have approximately equal sensitivity and specificity (ESAS). The first two layers are intended to identify those cases which have a very high certainty of being classified correctly, with the sensitive networks being used to "trap" the negative cases and the specific networks capturing the positive cases. The intuition behind this is that a set of networks which displays very high sensitivity will rarely make false negative predictions and so any negative predictions made by the networks are very likely to be correct. Conversely, highly specific networks will make very few false positive predictions, and so their positive predictions have a high certainty of being correct.

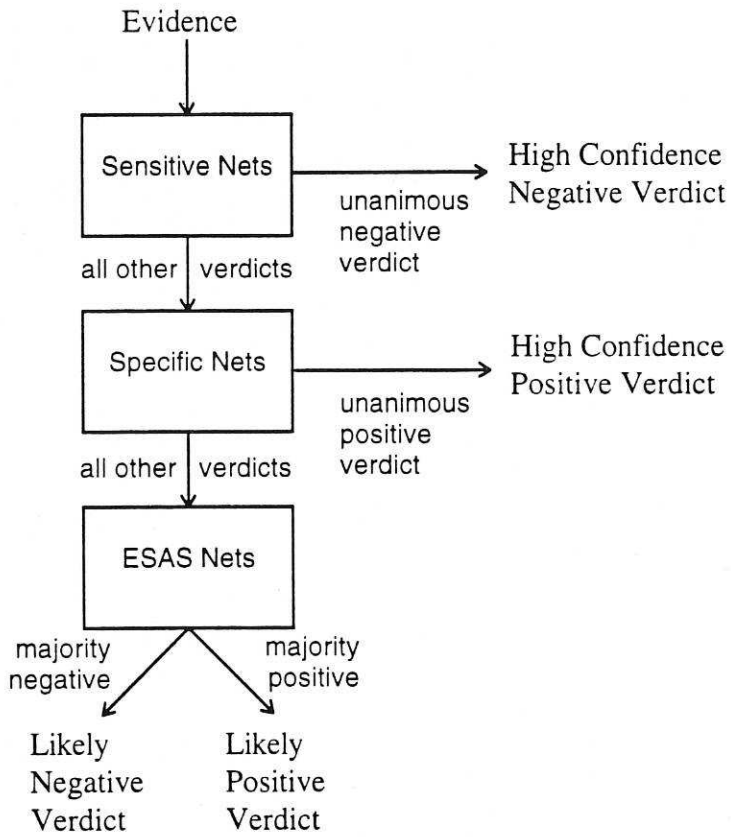

Figure 3 Cascaded ARTMAP voting strategy showing how high confidence decisions can be made by allowing cases to percolate through a pair of stringent voting systems tuned for high sensitivity and specificity, respectively. Those cases for which a unanimous decision cannot be made are treated by a majority voting system whose degree of confidence can be estimated from the relative numbers of votes for each diagnosis. Adapted from Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95. Pavia, Italy 1995; 355-366.

The cascaded voting strategy therefore operates as follows. An input data vector is first presented to the sensitive voting networks. If these yield a unanimous negative verdict, this is taken as the final category prediction. If not, the data item is next presented to the specific voting nets. If these yield a unanimous positive verdict, this is taken as the ultimate category prediction. Otherwise the final prediction of the category class of the input is obtained by majority verdict from the ESAS nets, with a lower certainty of the prediction being correct than with the previous two layers.

Case-Studies In Clinical Decision Support

Early Diagnosis Of Myocardial Infarction

The early identification of patients with acute ischaemic heart disease remains one of the great challenges of emergency medicine. The electrocardiograph (ECG) only shows diagnostic changes in about one half of acute myocardial infarction (AMI) patients at presentation. [43,44] None of the available biochemical tests becomes positive until at least three hours after symptoms begin, making such measurements of limited use for the early triage of patients with suspected AMI. [45] The early diagnosis of AMI, therefore, relies on an analysis of clinical features along with ECG data. A variety of statistical and computer-based algorithms has been developed to assist with the analysis of these factors (for review see reference 46), including the use of feedforward neural networks. [47-49] Although none of these has yet found widespread usage in clinical practice, this remains an important area of research not only owing to its clear potential to improve triage practices for the commonest of all medical problems, but also because of the light it may shed on techniques for the development of decision aids for use in other areas of medicine.

Patients and Clinical Data

The data used in this study were derived from consecutive patients attending the Accident and Emergency Department of the Royal Infirmary, Edinburgh, Scotland, with non-traumatic chest pain as the major symptom. The relevant clinical and ECG data (see below) were entered onto a purpose-designed proforma at, or soon after, the patient's presentation. The study included both patients who were admitted and those who were discharged. Nine hundred and seventy patients were recruited during the study period (September to December 1993). The final diagnosis for these patients was assigned independently by a Consultant Physician, a Research Nurse and a Cardiology Registrar. This diagnosis made use of follow-up ECGs, cardiac enzyme studies and other investigations as well as clinical history obtained from review of the patient's notes. Patients discharged from Accident and Emergency were contacted directly regarding further symptoms and, where necessary, their General Practitioners were also contacted and the notes of any further hospital follow-up reviewed. The final diagnosis in the 970 patients was Q wave AMI in 146 cases, non-Q wave AMI in 45, unstable angina in 69, stable angina in 271 and other diagnoses in 439 cases. The patients were 583 men and 387 women with a mean age of 58.2 years (range 14 – 92). Unstable angina was defined as either more than two episodes of pain lasting more than 10 minutes in a 24-hour period or more than three episodes in a 48 hour period, or as angina which was associated with the development of new ECG changes of ischaemia (either at diagnosis or in the subsequent three days).

The input data items for the ARTMAP model were all derived from data available at the time of the patient's presentation. In all, 35 items were used, coded as 37 binary inputs. The full list of the inputs is given in Appendix A, together with their feature names, used for symbolic rule extraction from the networks. For the purposes of this application, the final diagnoses were collapsed into two classes: "AMI" (Q wave AMI and non-Q wave AMI) and "not-AMI" (all other diagnoses). AMI cases were taken as positive, and not-AMI cases as negative, diagnoses. Informed consent was obtained from all patients participating in the study which was approved by the local Medical Ethics Committee.

Method

The 970 patient records were divided into three data sets; 150 randomly selected records formed the prediction set, a further 150 randomly chosen records formed the test set, and the remaining 670 comprised the training data. The prediction set consisted of 28 cases of AMI and 122 not-AMI; the test set of 30 AMI and 120 not-AMI.

The training data was randomly ordered in ten different ways, and each ordering applied to a different ARTMAP network using single-epoch training. The ART_a baseline vigilance was set to a medium level (0.6) for training, all other parameters were set to their standard values. [30] The performance of the ten trained ARTMAP networks was then measured on both the prediction and test sets. During this testing phase the ART_a baseline vigilance was relaxed slightly (to 0.5) in order to ensure that all test items were matched to an existing category cluster (i.e. forced choice prediction).

The performance of the trained networks on the prediction set alone was then used to calculate accuracy scores for the category nodes in each network, as a prerequisite of the category pruning process.

The "standard" form of category pruning [37] was performed on the original networks, such that all nodes with a CF below 0.5 were deleted from the networks in order to improve predictive accuracy. Performance of the resultant pruned networks was then measured on the prediction and test sets. Vigilance was further relaxed to 0.4 for testing these (and all other) pruned networks, again to ensure forced choice prediction.

The original networks were then pruned using different CF thresholds for the AMI and not-AMI nodes in order to produce pruned networks which maximised sensitivity. CF thresholds of 0.2 for AMI nodes and 0.95 for not-AMI nodes were employed, the criterion for setting the CF thresholds being to produce a mean sensitivity greater than 95% on the prediction set for the 10 pruned networks. Performance of the resultant nets was recorded for both the prediction and test sets. A similar procedure was then conducted to produce 10 networks which maximised specificity. CF thresholds of 0.7 AMI and 0.5 not-AMI were sufficient to yield a mean specificity greater than 95% on the prediction set.

The final pruning procedure was to produce 10 networks with approximately equal sensitivity and specificity (ESAS), the criterion for setting the CF thresholds being a performance on the prediction set where sensitivity and specificity were within 5% of each other. The performance of the pruned networks was again recorded on both the prediction and test sets.

Performance results using the voting strategy were then obtained for the un-pruned networks and all classes of pruned network. Three voters were used with all network

types, except the ESAS class, where five voters were used. Voters for the un-pruned, uniformly pruned, and ESAS network classes were selected on the basis of the networks with the highest accuracy on the prediction set. Selection criteria for the set of sensitive networks was maximum specificity, while maintaining a minimum sensitivity of 95% on the prediction set. The converse criteria were used for the set of specific networks.

Last, the cascaded variant of the voting strategy was employed utilising 3 sensitive nets, 2 specific nets and 5 ESAS nets (see Figure 3). The number of networks in each stage was chosen arbitrarily. The cascade operated as follows: data items were first applied to the sensitive voting nets. If these yielded a unanimous (3-0) verdict that the category prediction was not-AMI, this was taken as the final category prediction. If not, the input was presented to the specific voting nets. If these yielded a unanimous (2-0) verdict of AMI, this was taken as the final prediction. Otherwise the final prediction of the category class of the test item was obtained by majority verdict from the ESAS nets.

Results

The mean performance on the prediction and test sets for all classes of ARTMAP networks is shown in Table 1. As a baseline for comparisons, the Casualty Doctors showed an accuracy, sensitivity and specificity of 83.0%, 81.3% and 83.5% respectively over the entire data set.

Average accuracy for the un-pruned networks can be seen to be only slightly below this baseline. However this is largely an artefact of the unequal prior probabilities of the category distributions – specificity accounts for the majority of accuracy – and, although the networks' sensitivity is much poorer than the humans', this is compensated for by the superior specificity.

As expected, the uniformly pruned networks show an across-the-board increase in accuracy over the un-pruned nets, with a 2.7% increase on the test set, and a 7.3% increase on the prediction set. (The greater increase in performance on the prediction set is explained by the fact that pruning utilised the accuracy scores for this data, and the networks are consequently optimised for this data.) However, the increase in accuracy arises largely from an overall improvement in specificity rather than sensitivity, which actually drops on the test set.

Figures for the sensitive nets show that almost all AMI cases can be diagnosed by the network, while approximately 36% of the not-AMI cases are detected. Conversely, with the sensitive nets, almost all not-AMI cases are trapped while approximately 40% of the AMI cases are detected.

The performance of the ESAS class networks is most directly comparable with that of the Casualty Doctors, since they are not unduly biased towards specificity or sensitivity. It can be seen that the mean individual accuracy of such networks is approximately 7% worse than the human diagnoses.

	Prediction Set (%)			Test Set (%)		
Pruning Type	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
None	80.9	51.8	87.5	80.9	59.0	86.3
Uniform	88.2	60.7	94.5	83.6	52.0	91.5

	Prediction Set (%)			Test Set (%)		
Sens.	50.0	96.4	39.3	47.3	94.3	35.5
Spec.	86.9	41.8	97.2	84.7	39.7	96.0
ESAS	76.6	76.1	76.7	75.6	80.0	74.5

Table 1: Mean Performance of 10 Differently Pruned Networks. Adapted from Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95. Pavia, Italy 1995; 355-366.

When the voting strategy is employed the accuracy of all network types except the specific nets is improved, as shown in Table 2. Furthermore, unlike pruning, performance improvements owing to the voting strategy almost always result from increases in both sensitivity and specificity.

Accuracy for the ESAS nets is now much closer to that of the Casualty Doctors and sensitivity is slightly better. Accuracy for the un-pruned and uniformly pruned networks is now higher than the human diagnoses, particularly with the latter network class. However, this again results from the networks' very high specificity, while their sensitivity remains relatively poor.

	Prediction Set (%)			Test Set (%)		
Pruning	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
None	86.0	64.3	91.0	83.3	56.7	90.0
Uniform	92.0	78.6	95.1	88.0	56.7	95.8
Sens.	55.3	96.4	45.9	51.3	96.7	40.0
Spec.	88.7	46.4	98.4	84.7	33.3	97.5
ESAS	82.0	82.1	82.0	81.3	83.3	80.8

Table 2: Voting Strategy Performance of Differently Pruned Networks. Adapted from Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95. Pavia, Italy 1995; 355-366.

Use of the voting strategy with the sensitive networks on the test set results in increased coverage of the not-AMI cases while trapping more AMI cases than previously. However, the converse is not true for the specific nets, where a gain in not-AMI coverage is offset by poorer coverage of the AMI cases in comparison to the individual network means.

The best overall network performance was achieved by the cascaded voting strategy, shown in Table 3 below.

	Prediction Set (%)			Test Set (%)		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
High Certainty Voters	100.0	100.0	100.0	96.3	88.9	97.8

Lower Certainty Voters	71.0	73.7	70.3	72.9	81.0	70.7
Overall.	82.0	82.1	82.0	82.7	86.7	81.7

Table 3: Performance of the Cascaded Voting Strategy. Adapted from Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95. Pavia, Italy 1995; 355-366.

The cascade's overall performance can be seen to be almost identical to that of the Casualty Doctors. Moreover, the cascade provides a partitioning of input items into those with a high and a lower certainty of a correct diagnosis. Unanimous not-AMI decisions by the highly specific networks (i.e. the first stage of the cascade) are almost certain to be correct, similarly unanimous AMI decisions by the highly sensitive networks (the second stage of the cascade) are also almost certain to be correct. The ESAS class voters then provide lower certainty predictions for the remaining data items at the bottom of the cascade. High-certainty predictions accounted for 38% of items in the prediction set and 36% of items in the test set.

Perfect performance by the high-certainty voters on the test set was prevented by the occurrence of one false positive case and one false negative case. The false positive case displays most of the "barn-door" features of AMI, including ST-segment elevation, new pathological Q-waves and ST-segment or T-wave changes suggestive of ischaemia, while the false negative case displays almost no typical features (although the presence of old Q-waves should mean that a doctor would not entirely rule out AMI.

Symbolic Rule Extraction

The ability to extract symbolic rules from neural networks is an important enhancement to their use as decision-support tools in medical domains. Such symbolic rules provide two advantages which, taken collectively, should help to overcome reluctance to use a neural network decision-support tool.

First, a domain expert can examine the complete rule set in order to validate that the network has acquired an appropriate mapping of input features to category classes. Second, the symbolic rules provide explanatory facilities for the network's predictions during on-line operation. In the case of ARTMAP this corresponds to displaying the equivalent rule for the ART_a cluster node that was activated to provide a category decision. (In the case of the voting strategy, a number of such rules, one per voting network, would be displayed.) The diagnosing doctors are then able to decide whether or not to concur with the network's prediction, based upon how valid they believe that rule to be.

In this domain, each network retained, on average, 49 cluster nodes after uniform CF pruning. Space limitations therefore preclude the display of a typical complete rule set here. Instead, we provide a list of all rules for diagnosing AMI from nodes with a CF greater than 0.8 from the 10 original networks. In order to pass such a high threshold a node must encode a large proportion of the training exemplars and possess high predictive accuracy. Hence, these nodes are best in the sense of being the most useful

to their originating networks for the purpose of diagnosing AMI. In all, 18 such nodes occurred, their equivalent rules are shown in Table 4. See Appendix A for definitions of the terms in the rules.

IF retro THEN ami	IF retro sweat sttwave THEN ami	IF age=45-65 retro stelev THEN ami
IF age>65 retro sweat THEN ami	IF smokes retro sttwave THEN ami	IF retro newq sttwave THEN ami
IF age>65 retro sttwave THEN ami	IF age>65 retro alltight sweat THEN ami	IF age>65 retro larm sttwave THEN ami
IF retro larm sweat sttwave THEN ami	IF smokes retro alltight sttwave THEN ami	IF age=45-65 retro newq sttwave THEN ami
IF age>65 retro sweat likemi THEN ami	IF age=45-65 smokes retro sttwave THEN ami	IF age=45-65 smokes sweat nausea sttwave THEN ami
IF smokes retro larm nausea stelev THEN ami	IF age>65 retro alltight sweat nausea sttwave THEN ami	IF smokes retro alltight sweat nausea sttwave THEN ami

Table 4: Symbolic Rules for AMI Diagnosis Extracted from ARTMAP Networks. Adapted from Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95. Pavia, Italy 1995; 355-366.

Examination of the rules as a whole allows the following picture of a typical AMI case to be constructed: The patient is likely to be a smoker, aged over 45 (and most likely over 65), exhibiting central chest pain which possibly radiates to the left arm. The pain itself is likely to be described as "tight" or "heavy". Other physical symptoms may include sweating and nausea. ECG readings are very likely to show ST segment or T wave changes suggestive of ischaemia, and perhaps also new ST segment elevation and/or new pathological Q waves.

This picture closely corresponds to a "text-book" example of AMI, although it has been discovered by ARTMAP through self-organisation of the input data without any pre-specified knowledge of the domain. Thus the ARTMAP decision-support tool encodes rules which provide valid classifications for the domain, while bypassing the difficult and time-consuming knowledge-acquisition process found with rule-based expert systems. [1]

Causal (On-line) Learning

We now demonstrate the applicability of the ARTMAP variant, fuzzy ARTMAP (FAM) to the problem of the early diagnosis of AMI. [50] FAM achieves a synthesis of fuzzy logic and ART which enables it to learn and to recognise arbitrary sequences of analogue or binary input pairs, which may represent fuzzy or crisp sets of features. Here only twenty six features were abstracted from each patient record and these were coded into a binary-valued vector excepting real-valued data such as age, duration of pain etc. which was normalised in the range (0–1). [50,51] The need for FAM rather than its purely binary predecessor, ARTMAP, is evident, because interval data is now present and must be handled by the network.

In the assessment of on-line performance, a subset of 474 data were used both to train and to test the system; statistics being gathered prior to the verification of diagnosis at each stage. Thus the neural network starts out in a completely naïve state. The statistics of interest here are again the accuracy, sensitivity and specificity of diagnosis.

It should be noted that, whereas it is usual to select optimal decision thresholds by analysis of the Receiver Operating Characteristic (ROC) curve, [52] this technique is not appropriate here owing to the "all or nothing" predictions made by FAM. It will be seen that this inability to select optimal thresholds, hence counteract the effects of bias in the data, can result in an imbalance in the values of accuracy, sensitivity and specificity. Subsequent work has introduced a modification to FAM which has the capacity to achieve, on-line, very close to Bayes optimal classification rates for strongly biased data, and to deliver accurate estimates of the Bayesian (posterior or class conditional) probabilities. [28,29]

Figure 4 indicates the on-line performance of FAM for two separate cases. The first uses the technique of "sample replacement". Here, samples are drawn at random and are returned after use. Thus any individual sample may be chosen repeatedly. The second case is analogous to *in situ* or real-time learning when samples are taken in the order they occur and are not returned to the pool. Average values over ten runs are plotted with an indication of their standard deviations. There are three important points to note pertaining to on-line processing.

(i) Sometimes FAM fails to make a prediction (recognise a pattern). This is especially true in the early stages of learning when insufficient prototypes have been created. We have chosen to count such non-predictions as errors so that the performance indicators are biased downwards slightly.

(ii) Because statistics are gathered sequentially for each run, frequent poor (or non) predictions in the early stages are included in the long-run results. Again this has the effect of biasing the results downwards.

(iii) Although any given problem may itself be stationary, the learning procedure is inherently non-stationary owing to the build-up of knowledge. Thus to obtain truly statistically valid results averages should be taken over the ensemble of all possible realisations. For real problems this is often not feasible, as is the case here. To overcome this we have artificially created a small ensemble (of 10) by training 10 networks using different orderings of the data and averaging both across those, and also with time (see (ii) above).

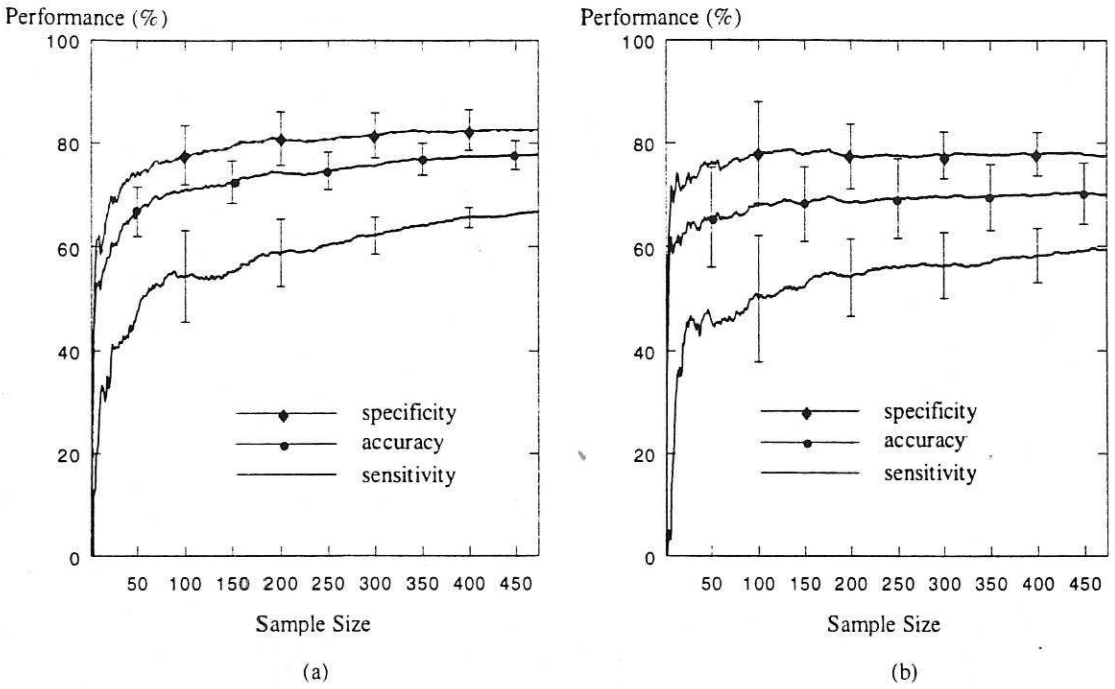

Figure 4 On-line fuzzy ARTMAP performance. Adapted from Harrison, RF, Lim, CP, and Kennedy, RL, Autonomously learning neural networks for clinical decision support. Proceedings of the International Conference on Neural Networks and Expert Systems in Medicine and Healthcare, Plymouth, UK 1994; 15-22.

In both cases the qualitative behaviour of FAM is as expected: broadly speaking, a monotone improvement in performance as the number of samples increases. Peaks and troughs in the early stages result from initial formation of poor templates and more frequent non-predictions. Sample replacement yields a better result owing to the relatively small sample size (relatively large probability of repetition).

This set of data comprises approximately equal proportions of infarction, angina and non-ischaemic heart disease sufferers and has a bias towards excluding a diagnosis of MI of 2.2:1. This bias manifests itself as favouring specificity over sensitivity. Clearly, the ability to predict a probability of class membership (as presented in reference 29) rather than the simpler binary decision would enable a user to control the types of misclassification to suit the domain, e.g. high sensitivity for initial screening, high specificity when deciding whether or not to thrombolyse.

Diagnosis Of Breast Cancer From Fine Needle Aspirate Samples

Breast cancer is a common disease affecting approximately 22,000 women yearly in England and Wales and is the commonest cause of death in the 35–55 year age group of the same population. [53] The primary method of diagnosis is through microscopic examination by a pathologist of cytology slides derived from fine needle aspiration of breast lesions (FNAB). [54] The acquisition of the necessary diagnostic expertise for this task is a relatively slow process. (A trainee pathologist in the UK requires at least five years study and experience before being allowed to sit the final professional pathology examinations for membership of the Royal College of Pathologists.)

Large studies of the cytopathological diagnosis of FNAB have shown a range of specificity of diagnosis of 90-100% with a range of sensitivities from 84-97%. [55] These

studies have been produced in centres specialising in the diagnosis of breast disease by pathologists with a special interest in breast cytopathology. In less specialised centres, such as district general hospitals, when a diagnostic FNAB service is being set up, the performance is in the lower range of those values with a specificity of 95% and a sensitivity of 87%. [56] There is thus scope for an artificial intelligence decision-making tool for this domain to assist in training junior pathologists and to improve the performance of experienced pathologists.

Data and Method

The data set consisted of 413 patient records, each comprising ten binary-valued features recorded from human-observation of breast tissue samples, together with the actual outcome for each case (i.e. whether a lesion proved to be malignant or benign). [57] The distribution of categories within the data was fairly even – 53% of cases were malignant, 47% benign. The features themselves are all claimed to have predictive value for the diagnosis task. [58,59] The following abbreviations: DYS, ICL, 3D, NAKED, FOAMY, NUCLEOLI, PLEOMORPH, SIZE, NECROTIC and APOCRINE are used here: full definitions of the features are provided in Appendix B.

As with almost all information gathered from a medical domain, the data set possesses a degree of "noise". Specifically, some feature-states do not always have the same outcome in every case. Analysis of the data set revealed the existence of 12 such states, which collectively account for 188 cases. Assuming that the most frequent outcome should always be chosen when an ambiguous feature-state occurs will result in 17 of these cases being misclassified. This represents approximately 4% of the data-set, and thus optimal performance in the domain is a diagnostic accuracy of 96%.

On this particular data-set, assigning malignant cases as "positive" and benign cases as "negative", an expert human pathologist (of Consultant status with 10 years experience in the field) performed with accuracy, 91%, sensitivity, 83%, and specificity, 100%, while a Senior House-Officer with 18 months experience achieved an accuracy of 71%, a sensitivity of 57% and a specificity of 98%.

Notice that these figures are biased towards specificity. The pathologist's prime concern is to avoid false positive predictions (i.e. diagnosing benign tumours as malignant), since these may result in unnecessary mastectomies. The resultant increase in false negatives (diagnosing malignant tumours as benign) is tolerated because, if the clinical suspicion of malignancy remains, the surgeon will then take further samples to be sent to the pathologist for additional testing.

One hundred records were randomly selected from the data to serve as test items in the evaluation of ARTMAP for the task. The remaining 313 records served as the teaching data. Ten ARTMAP networks were trained, each on a different random ordering of the teaching data. During training, the ART_a baseline vigilance parameter was set to 0.9 to ensure narrow category clustering; during testing this was relaxed to 0.6 to ensure that a category prediction (diagnosis) was made for all data items. (High vigilance during testing can lead to items failing to match sufficiently to any existing category clusters.)

Results

The subsequent performance of the 10 networks on the test set is shown in Table 5.

No. ART nodes	No. False +ve DX	No. False -ve DX	Accuracy (%)	Sensitivity (%)	Specificity (%)
60	5	2	93	96.2	89.6
61	4	4	92	92.3	91.7
59	3	1	96	98.1	93.8
58	3	2	95	96.2	93.8
58	5	2	93	96.2	89.6
60	5	1	94	98.1	89.6
61	5	2	93	96.2	89.6
60	3	2	95	96.2	93.8
68	2	4	94	92.3	95.8
65	5	1	94	98.1	89.6

Table 5 Performance of 10 ARTMAP networks on a 100 item test set. Adapted from Downs, J, Harrison, RF, and Cross, SS, A neural network decision support tool for the diagnosis of breast cancer. In: Hallam J ed., Hybrid Problems, Hybrid Solutions. Amsterdam: IOS Press, 1995: 51-60.

The mean performance of the 10 networks gives an accuracy of 93.9%, a sensitivity of 96.0% and a specificity of 91.7%. The five most accurate individual networks were then tested collectively, using the voting strategy described above. [26]

In this particular domain the voting strategy yields performance figures of accuracy, 95%, sensitivity, 96.1% and specificity, 93.8%. Although this may seem to be only a slight improvement on the individual ARTMAP results, it should be noted that diagnostic accuracy with the voting strategy is almost at the maximum possible for the domain.

Furthermore, when unanimous voting decisions only were considered, performance becomes near-perfect on a large subset of the test cases. Five-nil category decisions accounted for 91% of the test set and showed an accuracy of 99%, a sensitivity of 100% and a specificity of 98% on this subset of the data. Thus the voting strategy can provide a useful partitioning between data items with high and low certainty of outcome.

Symbolic Rule Extraction

Symbolic rule extraction [37] was then performed upon all 10 of the previously trained ARTMAP networks.

Severe pruning was performed upon the 10 trained ARTMAP networks, using a threshold confidence level of 0.7. The number of category cluster nodes remaining for each individual network after pruning ranged from three to nine. Thus the networks were reduced to a small number of ART_a category nodes of strong predictive power from which rules could be extracted. Before doing so however, the test data was re-applied to each of the pruned networks to check that pruning had not adversely affected performance. Since pruning necessarily reduces ARTMAP's coverage of the feature space, the baseline ART_a vigilance was this time relaxed further to 0.5. Despite this, some pruned networks were still unable to generate category predictions for all

test set items. The mean performance of the 10 networks after pruning gave an accuracy of 94.1% a sensitivity of 92.3% and a specificity of 97.9%.

It can be seen that pruning has virtually no effect upon overall diagnostic accuracy but has led to increased specificity and reduced sensitivity. The five most accurate pruned networks (excluding those which did not generate predictions on all test set items) were then tested using the voting strategy. This resulted in an accuracy of 95%, a sensitivity of 92.3% and a specificity of 97.9%, again confirming that the voting strategy allows the optimum accuracy for the domain to be closely approached.

Rule extraction from the 10 pruned nets yielded 14 distinct rules, 12 for malignant outcomes and 2 for benign. The full list of rules is shown in Table 6, ranked by how many of the 10 pruned networks a rule occurred in. Definitions of the terms in the rules can be found in Appendix B.

It can be seen that an absence of features, or the FOAMY feature present in isolation, leads to a benign diagnosis. PLEOMORPH and SIZE are found in all rules for malignant diagnoses, and NUCLEOLI is additionally present in all but two of these same rules (both of which have low frequency of occurrence). Thus these three features in combination seem to be the strongest indicators of malignancy. Other features are weaker indicators of malignancy, and indeed two input features, NAKED and APOCRINE, are conspicuous by their absence from any of the rules. We would conclude therefore that these two features are the least useful in forming a diagnosis, at least for this particular data set.

Rule 1 (10 occurrences) IF no symptoms THEN benign	*Rule 2 (8 occurrences)* IF 3D nucleoli pleomorph size THEN malignant	*Rule 3 (8 occurrences)* IF 3D foamy nucleoli pleomorph size THEN malignant
Rule 4 (7 occurrences) IF foamy THEN benign	*Rule 5 (4 occurrences)* IF icl 3D nucleoli pleomorph size THEN malignant	*Rule 6 (4 occurrences)* IF dys nucleoli pleomorph size THEN malignant
Rule 7 (3 occurrences) IF foamy nucleoli pleomorph size THEN malignant	*Rule 8 (3 occurrences)* IF nucleoli pleomorph size THEN malignant	*Rule 9 (2 occurrences)* IF 3D foamy nucleoli pleomorph size necrotic THEN malignant
Rule 10 (2 occurrences) IF 3D foamy pleomorph size necrotic THEN malignant	*Rule 11 (2 occurrences)* IF dys icl nucleoli pleomorph size THEN malignant	*Rule 12 (1 occurrence)* IF icl nucleoli pleomorph size THEN malignant
Rule 13 (1 occurrence) IF foamy nucleoli pleomorph size necrotic THEN malignant	*Rule 14 (1 occurrence)* IF icl 3D pleomorph size THEN malignant	

Table 6 Symbolic rules for FNAB diagnosis. Adapted from Downs, J, Harrison, RF, and Cross, SS, A neural network decision support tool for the diagnosis of breast cancer. In: Hallam J ed., Hybrid Problems, Hybrid Solutions. Amsterdam: IOS Press, 1995: 51-60.

An expert human pathologist confirmed the relative importance of the features listed above in making his own diagnoses, with the exception that he places no value on the presence or absence of the FOAMY feature. It should be noted that this feature has a somewhat ambiguous status within the ARTMAP rules. In isolation, it is indicative of a benign diagnosis. However, when it occurs in combination with other features, a malignant diagnosis results.

There is some disagreement between different domain experts as to the relative importance of the features in making diagnoses. Thus another pathologist states "I think the presence of bipolar naked nuclei and foamy macrophages can be taken as indicative of benignancy. This is not to say however, that when these features are combined with cells showing obvious features of malignancy, malignancy should not be diagnosed.". This accords with the self-discovered ARTMAP rules for the FOAMY feature.

Table 7 below summarises the performance figures for ARTMAP in comparison with human pathologists in this domain. It can be seen that in terms of diagnostic accuracy ARTMAP always performs at least as well as the human expert and much better than the novice. However, the weak spot in the un-pruned ARTMAP networks' performance is the lower specificity in comparison to the human pathologists. As pointed out earlier, it is vital that false positive cases (which reduce specificity) are avoided in this domain.

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Human Expert	91	83	100
Human Novice	71	57	98
Un-pruned ARTMAP (mean)	94	96	92
Un-pruned ARTMAP (voting)	95	96	94
Pruned ARTMAP (mean)	94	90	99
Pruned ARTMAP (voting)	95	92	98

Table 7 Relative performance of human pathologists and ARTMAP. Adapted from Downs, J, Harrison, RF, and Cross, SS, A neural network decision support tool for the diagnosis of breast cancer. In: Hallam J ed., Hybrid Problems, Hybrid Solutions. Amsterdam: IOS Press, 1995: 51-60.

The pruning procedure achieves this goal, by increasing specificity at the expense of sensitivity without changing overall diagnostic accuracy. The reason for this is that the category clusters formed at ART_a predominantly indicate positive (malignant) cases. (On average, 70% of ART_a category nodes in the un-pruned networks denote malignant outcomes.) Pruning therefore mostly deletes nodes with malignant outcomes, and so coverage of these cases in the state space is reduced disproportionately more than for benign cases. This effect of biasing the trade-off between sensitivity and specificity was achieved naturally in this domain as a side-effect of the rule-extraction process although such an effect can be achieved purposely in other domains by use of the generalised pruning procedure discussed earlier.

Conclusions

ART-based systems are clearly one candidate for providing the knowledge acquisition and inference engine in apprentice systems. Our studies have shown that in two different medical problem domains the ARTMAP neural network architecture provides solutions with performance which at least equals that of human experts and provides explicit rules which agree with those given by human experts. Continued on-line learning is possible and the networks can be implemented to run on standard personal computers. All these factors provide a suitable environment for the development of

apprentice systems which can be used for clinical decision-support. The use of such technology is in its very early stages and much research and development is needed to establish a truly autonomously learning decision aid which can operate safely in a medical environment.

References

[1] Hayes-Roth, F, Waterman, DA, and Lenat, DB, Building Expert Systems. London: Addison Wesley, 1983.

[2] Brahams, D and Wyatt, J, Decision aids and the law. *Lancet*, 1989; 632-634.

[3] Quinlan, JR, C4.5: Programs for Machine Learning. Morgan Kauffman, 1993.

[4] Quinlan, JR, Decision trees and decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 1990; 20: 339-346.

[5] Quinlan, JR, Induction of decision trees. *Machine Learning*, 1986; 1: 81-106.

[6] Andrews, R and Geva, S, RULEX & CEBP networks as the basis for a rule refinement system. In Hallam, J, Ed. *Hybrid Problems, Hybrid Solutions: 10th Biennial Conference on AI and Cognitive Science*. Amsterdam: IOS Press, 1995: 1-12.

[7] Ma, Z and Harrison, RF, GR2: a hybrid knowledge-based system using general rules. 488-493, Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, 1995.

[8] Saito, K and Nakano, R, Rule extraction from facts and neural networks. 379-382, Proceedings of the International Neural Network Conference, 1990.

[9] Shavlik, JW, Mooney, RJ, and Towell, GG, Symbolic and neural learning algorithms: an experimental comparison. *Machine Learning*, 1991; 6: 111-143.

[10] Towell, GG and Shavlik, JW, Extracting refined rules from knowledge-based neural networks. *Machine Learning*, 1993; 13: 71-101.

[11] Carpenter, G and Grossberg, S, A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*, 1987; 37: 54-115.

[12] Sharkey, NE and Sharkey, AJC, An analysis of catastrophic interference. *Connection Science*, 1995; 7: 301-329.

[13] Rumelhart, D, Hinton, G, and Williams, R, Learning representations by back-propagating errors. *Nature*, 1986; 323: 533-536.

[14] Moody, J and Darken, C, Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1989; 1: 281-294.

[15] Richard, M and Lippman, R, Neural network classifiers estimate Bayesian *a posteriori* probabilities. *Neural Computation*, 1991; 3: 461-483.

[16] Wan, EA, Neural network classification: a Bayesian interpretation. *IEEE Transactions on Neural Networks*, 1990; 1: 303-305.

[17] Cybenko, G, Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 1989; 2: 303-314.

[18] Park, J and Sandberg, I, Universal approximation using radial basis function networks. *Neural Computation*, 1991; 3: 246-257.

[19] Bishop, CM, Neural Networks for Pattern Recognition. Oxford: Clarendon Press, 1995.

[20] Ripley, BD, Pattern Recognition and Neural Networks. Cambridge, UK: Cambridge University Press, 1996.

[21] Carpenter, G, Grossberg, S, and Rosen, D, ART2-A: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, 1991; **4:** 493-504.

[22] Carpenter, G, Grossberg, S, and Rosen, D, Fuzzy ART: Fast, stable learning and categorisation of analogue patterns by an adaptive resonance system. *Neural Networks*, 1991; 4: 759-771.

[23] Carpenter, G and Grossberg, S, ART3: Hierarchical search using chemical transmission in self-organising pattern recognition architectures. *Neural Networks*, 1990; 3: 129-152.

[24] Carpenter, G and Grossberg, S, The ART of adaptive pattern recognition by a self-organising neural network. *Computer*, 1988; 21: 77-88.

[25] Carpenter, G and Grossberg, S, ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 1987; 26: 4919-4930.

[26] Carpenter, G, Grossberg, S, Markuzon, S, Reynolds, J, and Rosen, D, Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multi-dimensional maps. *IEEE Transactions on Artificial Neural Networks*, 1992; 3: 698-712.

[27] Carpenter, G, Grossberg, S, and Reynolds, J, ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, 1991; 4: 565-588.

[28] Lim, CP and Harrison, RF, Modified fuzzy ARTMAP approaches Bayes optimal classification rates: an empirical demonstration. *Neural Networks*, Accepted 1996. (In Press)

[29] Lim, CP and Harrison, RF, An incremental adaptive network for on-line, supervised learning and probability estimation. *Neural Networks*, Accepted 1996. (In Press)

[30] Kasuba, T, Simplified fuzzy ARTMAP. *AI Expert*, 1993; 8: 18-25.

[31] Carpenter, G, Distributed learning, recognition and prediction by ART and ARTMAP neural networks. Research Report CAS/CNS-96-004, Boston University, Boston, USA, 1996.

[32] Lim, CP and Harrison, RF, On-line pattern classification with multiple neural network systems: an experimental study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Submitted 1996.

[33] Lim, CP, Harrison, RF, and Kennedy, RL, Application of autonomous neural network systems to medical pattern classification tasks. *Artificial Intelligence in Medicine*, Submitted 1996.

[34] Lim, CP and Harrison, RF, A multiple neural network architecture for sequential evidence aggregation and incomplete data classification. *IEE Fifth International Conference on Neural Networks*, Submitted 1996.

[35] Xu, L, Krzyzac, A, and Suen, CY, Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 1992; 22: 418-435.

[36] Huang, YS and Suen, CY, A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995; 17: 90-94.

[37] Carpenter, G and Tan, A, Rule extraction, fuzzy ARTMAP and medical databases. 501-506, Proceedings of the World Congress on Neural Networks, 1993.

[38] Setiono, R, Extracting rules from pruned neural networks for breast cancer diagnosis. *Artificial Intelligence in Medicine*, 1996; 8: 37-51.

[39] Mahoney, J and Mooney, RJ, Combining neural and symbolic learning to revise a probabilistic rule-base. 107-114, Proceedings of the European Joint Conference on Artificial Intelligence 93, 1993.

[40] Ma, Z, Harrison, RF, and Kennedy, RL, A heuristic for general rule extraction from a multilayer Perceptron. In Hallam, J, Ed. *Hybrid Problems, Hybrid Solutions*. Amsterdam: IOS Press, 1995: 133-144.

[41] Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. 355-366, Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95, Pavia, Italy, 1995.

[42] Downs, J, Harrison, RF, Kennedy, RL, and Cross, SS, Application of the fuzzy ARTMAP neural network model to medical pattern classification tasks. *Artificial Intelligence in Medicine*, 1996; 8: 403-428.

[43] Adams, JE, Trent, R, and Rawles, J, Earliest electrocardiographic evidence of myocardial infarction: implications for thrombolytic treatment. *British Medical Journal*, 1993; 307: 409-413.

[44] Stark, ME and Vacek, JL, The initial electrocardiogram during admission for myocardial infarction. *Archives of Internal Medicine*, 1987; 147: 843-847.

[45] Adams, JE, Abendschein, DR, and Jaffe, AS, Biochemical markers of myocardial injury. Is MB creatine kinase the choice for the 1990s? *Circulation*, 1993; 88: 750-763.

[46] Kennedy, RL, Harrison, RF, and Marshall, SJ, Do we need computer-based decision support for the diagnosis of acute chest pain? *Journal of the Royal Society of Medicine*, 1993; 86: 31-34.

[47] Baxt, WG, Use of an artificial neural network for data analysis in clinical decision-making: The diagnosis of acute coronary occlusion. *Neural Computation*, 1990; 2: 480-489.

[48] Hart, A and Wyatt, J, Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. *Medical Informatics*, 1990; 15: 229-236.

[49] Harrison, RF, Marshall, SJ, and Kennedy, RL, A connectionist aid to the early diagnosis of myocardial infarction. 119-128, Proceedings of the Third European Conference on Artificial Intelligence in Medicine, Maastricht, 1991.

[50] Harrison, RF, Lim, CP, and Kennedy, RL, Autonomously learning neural networks for clinical decision support. 15-22, Proceedings of the The International

Conference on Neural Networks and Expert Systems in Medicine and Healthcare, Plymouth, UK, 1994.

[51] Lim, CP, An autonomous-learning system. MSc Dissertation: The University of Sheffield, 1993.

[52] Meistrell, ML, Evaluation of neural network performance by receiver operating characteristic (ROC) analysis: examples from the biotechnology domain. *Computer Methods and Programs in Biomedicine*, 1990; 32: 73-80.

[53] Underwood, JCE, Tumours: benign and malignant. In Underwood, JCE, Ed. *General and Systematic Pathology*. Edinburgh: Churchill Livingstone, 1992: 223-246.

[54] Elston, CW and Ellis, IO, Pathology and breast screening. *Histopathology*, 1990; 16: 109-118.

[55] Wolberg, WH and Mangasarian, OL, Computer-designed expert systems for breast cytology diagnosis. *Analytical and Quantitative Cytology and Histology*, 1993; 15: 67-74.

[56] Start, RD, Silcocks, PB, Cross, SS, and Smith, JHF, Problems with audit of a new fine-needle aspiration service in a district general hospital. *Journal of Pathology*, 1992; 167: 141A (Abstract)

[57] Downs, J, Harrison, RF, and Cross, SS, A neural network decision support tool for the diagnosis of breast cancer. In Hallam, J, Ed. *Hybrid Problems, Hybrid Solutions*. Amsterdam: IOS Press, 1995: 51-60.

[58] Trott, PA, Aspiration cytodiagnosis of the breast. *Diagnostic Oncology*, 1991; 1: 79-87.

[59] Koss, LG, Diagnostic cytology and its histopathologic basis. 1992.

Appendix A: Coding For AMI Data

Input Code	Meaning	Input Code	Meaning
age<45	Age less than 45 years	alltight	Pain described as "tight"
age=45-65	Age 45 – 65 years	allsharp	Pain described as "sharp"
age>65	Age greater than 65 years	sweat	Sweating
smokes	Smokes	s_o_breath	Short of breath
ex_smoker	Ex-smoker	nausea	Nausea
fam_ihd	Family history of IHD	vomit	Vomiting
diabetes	Diabetes mellitus	syncope	Syncope
hypertense	Hypertension	epis	Episodic pain
hyperlipid	Hyperlipidaemia	likemi	Worse than usual angina/similar to previous AMI
retro	Central chest pain	lvf	Fine crackles suggestive of pulmonary oedema
lchest	Pain in left side of chest	added_hs	Added heart sounds
rchest	Pain in right side of chest	hypoperf	Signs of hypoperfusion
back	Pain radiates to back	stelev	New ST segment elevation
larm	Pain radiates to left arm	newq	New pathological Q waves
jaw	Pain radiates to neck or jaw	sttwave	ST segment or T wave changes suggestive of ischaemia
rarm	Pain radiates to right arm	bbb	Bundle branch block
breathing	Pain is worse on inspiration	old_q	Old ECG features of myocardial infarction
posture	Pain related to posture	old_st	ECG signs of ischaemia known to be old
tender_cw	Chest wall tenderness		

Appendix B: FNAB Feature Definitions

DYS: True if majority of epithelial cells are dyhesive, false if majority of epithelial cells are in cohesive groups.

ICL: True if intracytoplasmic lumina are present, false if absent.

3D: True if some clusters of epithelial cells are not flat (more than two nuclei thick) and this is not due to artefactual folding, false if all clusters of epithelial cells are flat.

NAKED: True if bipolar "naked" nuclei in background, false if absent.

FOAMY: True if "foamy" macrophages present in background, false if absent.

NUCLEOLI: True if more than three easily visible nucleoli in some epithelial cells, false if three or fewer easily visible nucleoli in epithelial cells.

PLEOMORPH: True if some epithelial cell nuclei with diameters twice that of other epithelial cell nuclei, false if no epithelial cell nuclei twice the diameter of other epithelial cell nuclei.

SIZE: True if some epithelial cells with nuclear diameters at least twice that of lymphocyte nuclei, false if all epithelial cell nuclei with nuclear diameters less than twice that of lymphocyte nuclei.

NECROTIC: True if necrotic epithelial cells present, false if absent.

APOCRINE: True if apocrine change present in all epithelial cells, false if not present in all epithelial cells.

List of Captions

Figure 1 A single ART module comprising three layers, F_0, F_1 and F_2. F_1 and F_2 are fully interconnected in both directions via weighted links (w_{ij}) which form the LTM. ρ is the vigilance parameter which governs the coarseness of categorisation. F_0 buffers the input patterns so that they remain present during processing.

Figure 2 General ARTMAP configuration. This comprises two ART modules, labelled a and b, which self-organise the input and target data streams respectively. Categories formed for each of these are associated via the Map Field. Category size is determined for each module by its own vigilance parameter, and incorrect associations between ART_a and ART_b categories are handled via the match tracking process, governed by the Map Field vigilance, ρ_{ab}.

Figure 3 Cascaded ARTMAP voting strategy showing how high confidence decisions can be made by allowing cases to percolate through a pair of stringent voting systems tuned for high sensitivity and specificity, respectively. Those cases for which a unanimous decision cannot be made are treated by a majority voting system whose degree of confidence can be estimated from the relative numbers of votes for each diagnosis. Adapted from Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95. Pavia, Italy 1995; 355-366.

Table 1: Mean Performance of 10 Differently Pruned Networks. Adapted from Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95. Pavia, Italy 1995; 355-366.

Table 2: Voting Strategy Performance of Differently Pruned Networks. Adapted from Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95. Pavia, Italy 1995; 355-366.

Table 3: Performance of the Cascaded Voting Strategy. Adapted from Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95. Pavia, Italy 1995; 355-366.

Table 4: Symbolic Rules for AMI Diagnosis Extracted from ARTMAP Networks. Adapted from Downs, J, Harrison, RF, and Kennedy, RL, A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proceedings of the Fifth European Conference on Artificial Intelligence in Medicine, AIME-95. Pavia, Italy 1995; 355-366.

Figure 4 On-line fuzzy ARTMAP performance. Adapted from Harrison, RF, Lim, CP, and Kennedy, RL, Autonomously learning neural networks for clinical decision support. Proceedings of the International Conference on Neural Networks and Expert Systems in Medicine and Healthcare, Plymouth, UK 1994; 15-22.

Table 5 Performance of 10 ARTMAP networks on a 100 item test set. Adapted from Downs, J, Harrison, RF, and Cross, SS, A neural network decision support tool for the diagnosis of breast cancer. In: Hallam J ed., Hybrid Problems, Hybrid Solutions. Amsterdam: IOS Press, 1995: 51-60.

Table 6 Symbolic rules for FNAB diagnosis. Adapted from Downs, J, Harrison, RF, and Cross, SS, A neural network decision support tool for the diagnosis of breast cancer. In: Hallam J ed., Hybrid Problems, Hybrid Solutions. Amsterdam: IOS Press, 1995: 51-60.

Table 7 Relative performance of human pathologists and ARTMAP. Adapted from Downs, J, Harrison, RF, and Cross, SS, A neural network decision support tool for the diagnosis of breast cancer. In: Hallam J ed., Hybrid Problems, Hybrid Solutions. Amsterdam: IOS Press, 1995: 51-60.