Application of the Fuzzy ARTMAP Neural Network Mo
to Medical Pattern Classification Tasks

Joseph Downs, Robert F Harrison
Department of Automatic Control and Systems Engineering
The University of Sheffield

R Lee Kennedy
Department of Medicine
The University of Edinburgh

Simon S Cross
Department of Pathology
University of Sheffield Medical School
The University of Sheffield

Research Report #584
19 June 1995

Abstract

This paper presents research into the application of the fuzzy ARTMAP neural network model to medical pattern classification tasks. A number of domains, both diagnostic and prognostic, are considered. Each such domain highlights a particularly useful aspect of the model. The first, coronary care patient prognosis, demonstrates the ARTMAP voting strategy involving "pooled" decision-making using a number of networks, each of which has learned a slightly different mapping of input features to pattern classes. The second domain, breast cancer diagnosis, demonstrates the model's symbolic rule extraction capabilities which support the validation and explanation of a network's predictions. The final domain, diagnosis of acute myocardial infarction, demonstrates a novel category pruning technique allowing the performance of a trained network to be altered so as to favour predictions of one class over another (e.g. trading sensitivity for specificity or vice versa). It also introduces a "cascaded" variant of the voting strategy intended to allow identification of a subset of cases which the network has a very high certainty of classifying correctly.

Correspondence Address

R.F. Harrison
Department of Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street
Sheffield, S1 3JD
United Kingdom

Telephone:+44 (0)114 2825139
Facsimile: +44 (0)114 2780409
E-mail: r.f.harrison@sheffield.ac.uk

1 Introduction

Neural networks potentially have great value in medical decision-support applications. Unlike expert systems, they bypass the difficult and time-consuming knowledge acquisition process (Hayes-Roth, Waterman and Lenat, 1983) by learning complex associations directly from domain examples. This provides the opportunity for a neural network decision-support tool to adapt to perform the same task under varying conditions. This occurs, for example, because of differing demographic conditions or clinical procedures from region to region, or because procedures may vary over time owing to advances in medical knowledge or technology.

A large and ever-growing body of work now exists on applying neural networks to various medical classification tasks, e.g. the diagnosis of epilepsy (Apolloni et al., 1990), diagnosis of low back disorders (Bounds, Lloyd and Mathew, 1990), early diagnosis of myocardial infarction (Harrison, Marshall and Kennedy, 1991), classification of thyroid disorders (Egmont-Peterson et al., 1994), identification of Alzheimer's diseased tissue (Pizzi et al., 1995) etcetera.

The main thrust of this work has been in the use of feedforward networks to learn the association between evidence and outcome. Primarily, the Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) network classes have been employed. (See Rumelhart, Hinton and Williams, 1986, and Moody and Darken, 1989, respectively.) Both the MLP and the RBF have been shown to be rich enough in structure so as to be able to approximate any (sufficiently smooth) function with arbitrary accuracy (Cybenko, 1989; Park and Sandberg, 1991). Thus, given sufficient data, computational resources (the MLP, in particular, does not scale well with problem size) and time (non-linear optimization which is non-linear in the parameters may be time consuming to perform, numerically), it is possible 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. However, despite this attractive property, there are two serious drawbacks with these classes of feedforward networks in addition to the caveats given above.

First, these networks require artificial termination of training, since they are susceptible to new but irrelevant data over-writing useful existing associations and thus degrading general classification performance. However, this requirement seriously compromises the adaptivity of a neural network. New data is not always irrelevant, sometimes it reflects significant changes in the classification domain which requires new associations to be learned. This is termed the *stability-plasticity dilemma*: "How can a learning system be designed to remain plastic, or adaptive, in response to significant events, and yet remain stable in response to irrelevant events?" (Carpenter and Grossberg, 1988, p.77).

The MLP and RBF networks do not cope well with this dilemma. The termination of learning once a pre-determined level of performance has been achieved sacrifices plasticity for the sake of stability. In non-stationary classification domains (i.e. when the underlying statistics of the population are changing with time), these networks cannot incrementally acquire new associations as the environment changes. Instead, they must be completely retrained on new domain data, losing all previously learned associations even though some may still be useful (and will be reacquired alongside the new associations with retraining). Furthermore, when retraining with additional data there is no guarantee that the previous network's topology, learning parameters etc. will still provide a good solution. It is possible that significant changes to the network will be needed when it is re-derived. (For a detailed discussion on this issue with regard to feedforward networks see Sharkey and Sharkey, 1994.)

Many medical domains are non-stationary to a greater or lesser extent, for example, owing to changes in clinical procedures. Furthermore, the artificial termination of learning means that a neural network trained on data from one site is likely to perform the same task sub-optimally using data from another site because of variations in local conditions (e.g. Kennedy, Harrison and Marshall, 1994). Thus it would be desirable if such a network could be "fine-tuned" to its changed operating conditions by incremental learning of cases from the new site.

The second problem stems from a common general criticism of the neural network paradigm that the rules governing the predicted outcome are obscure. This can lead to a strong resistance to acceptance of a network's predictions by potential users. This is particularly true for medical domains. For example, a diagnosing clinician using a neural network decision-support tool has to be convinced that the underlying model captures the salient features of the domain and the system is further able to offer an explanation of its diagnoses in user-comprehensible (i.e.symbolic) terms. However, attempts to extract domain rules from feedforward networks have met with limited success, with, so far, no completely general method published (see Ma and Harrison, 1995).

In the research described here we provide an evaluation of a neural network model, fuzzy ARTMAP, which is not susceptible to these two criticisms, and has other desirable properties for medical classification tasks. The next section provides an overview of this powerful, but relatively little-known, model. We then describe the application of this model to three different medical pattern classification tasks, in each instance utilizing data gathered from hospitals in the UK. Section 3 describes a prognosis task, predicting the death or survival of patients admitted to a coronary care ward, section 4 concerns the diagnosis of breast cancer, and section 5 the diagnosis of acute myocardial infarction. With each task we highlight a different useful aspect of the fuzzy ARTMAP model; in section 3 the ARTMAP voting strategy; in section 4 the model's symbolic rule extraction capabilities; and in section 6 novel variants of the voting strategy and category pruning. Section 7 presents the findings and identifies areas for further research.

2 Fuzzy ARTMAP

Adaptive resonance theory, or ART (Carpenter and Grossberg, 1991) represents a family of neural network models originally developed from the competitive learning paradigm with the intention of overcoming the stability-plasticity dilemma (Grossberg, 1987). This was achieved by utilizing feedback between layers of input and category nodes in addition to the standard feedforward connections of competitive learning. Thus, in ART models, an input pattern is not automatically assigned to the category that is *initially* maximally activated by that input. Instead, if the feedback process rejects the initial categorisation, a search process is initiated which terminates when a category node with an acceptable match to the input is found. If no such node exists, a new category node is formed to classify the input.

It should also be noted that ART models usually employ a localist representation for category nodes owing to the so-called "winner-take-all" competitive learning dynamics. Although biologically implausible, this feature does have the advantage of facilitating symbolic rule extraction from a trained network (see section 4). Furthermore, localization results from a simplification used to obtain the computational models and is not inherent in adaptive resonance theory per se.

Since ART was an outgrowth of competitive learning, initial models developed from it employed unsupervised learning. Examples of such models include ART 1 (Carpenter and Grossberg, 1987) which is restricted to the classification of binary input patterns, and fuzzy ART (Carpenter, Grossberg and Rosen, 1991) which generalizes ART 1 so as to classify both analogue and binary patterns. More recently ART models employing supervised learning have been developed which are based upon these earlier models and so retain their self-organizing properties.

Fuzzy ARTMAP (Carpenter et al., 1992) is one such model, based upon fuzzy ART. It is thus a self-organizing, supervised learning, neural network model for the classification of both analogue and binary patterns. Fuzzy ARTMAP consists of three modules, two fuzzy ART systems called ART_a and ART_b, and a related structure called the map field. During training, input patterns are presented to ART_a together with their associated teaching stimuli at ART_b. Associations between patterns at ART_a and ART_b are then formed at the map field. During testing, supervisory inputs at ART_b are omitted, and instead the inputs at ART_a are used to recall a previously learned association with an ART_b pattern via the map field.

However, fuzzy ARTMAP does not directly associate inputs at ART_a and ART_b. Rather, such patterns are first self-organized into prototypical category clusters before being associated at the map field. Hence generalized associations are formed. If the ART_a category cluster selected through self-organization does not match with the teaching category at ART_b, the map field generates a re-set at ART_a, forcing the input to be re-classified to an appropriate ART_a category prototype. If no such prototype exists, a new cluster is automatically created for classification of the input. Thus it can be seen that supervision of learning is only employed when self-organization leads to a classification error.

Training in fuzzy ARTMAP almost always results in multiple category clusters forming at ART_a for each teaching category present at ART_b, with each such cluster encoding multiple input exemplars (i.e. each ART_a cluster represents a significant sub-region of the overall state space covered by a particular teaching category). Hence fuzzy ARTMAP instantiates a many-to-one mapping between ART_a input patterns and their actual classification. For full details on fuzzy ARTMAP see Carpenter et al. (1992).

Simplified fuzzy ARTMAP (henceforth abbreviated to SFAM) is a "streamlined" version of fuzzy ARTMAP intended to be more computationally efficient than a full implementation but with a minimal loss of computational power (Kasuba, 1993). Figure 1 gives a diagrammatic representation of the model; circled lines denote adaptive weight connections, arrowed lines show processing flow. The teaching stimulus has a dashed arrow to indicate its variable status—if it is present learning occurs, if it is absent prediction takes place instead.

The model does not self-organize teaching inputs at ART_b, but instead encodes these patterns directly. (Thus, unlike fuzzy ARTMAP, the ART_b module in SFAM is not a complete fuzzy ART system.) This is based on the observation that in most pattern classification tasks the teaching stimuli themselves do not need to be further categorised since they directly represent distinct, known classes, e.g. one from many classification.

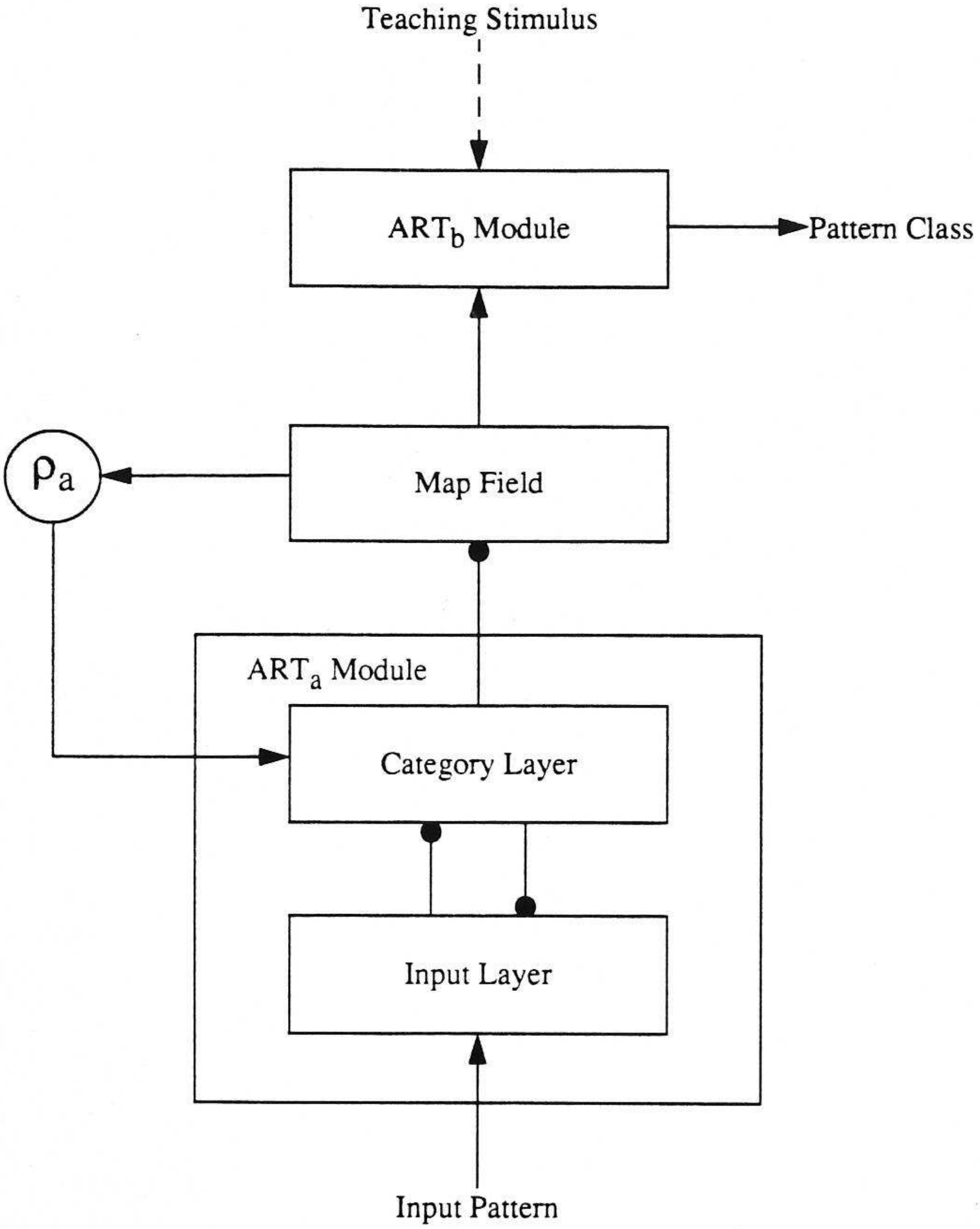

Figure 1: Simplified fuzzy ARTMAP

In addition, SFAM converts all but one of the three user-changeable parameters in fuzzy ARTMAP to constants whose values are the usual default settings of the original parameters. (For the benefit of those familiar with the ARTMAP models, the category choice parameter, α, is fixed to be near-equal to zero and the learning rate, β, is set to its maximum value of one—so-called fast learning.) The only remaining user-changeable parameter is the baseline vigilance for the ART_a module, $\bar{\rho}_a$. This determines how close a match is required between an ART_a input pattern and a category cluster prototype before accepting the input as a member of the cluster. This parameter (indirectly) controls the size of the category clusters that will form, since the higher it is set, the closer acceptable matches must be, and the smaller the coverage of the state space each cluster will have. Generally, higher vigilance provides better classification performance, although this must be balanced against the potential proliferation of category clusters, providing poor data compression and leading the network to become little more than a "look-up table" (Marriot and Harrison, In Press). Additionally, with small training sets and/or high-dimensional input vectors with many features, high vigilance can lead to incomplete coverage of the feature space by the network.

As well as its capabilities for continuous learning and symbolic rule extraction, SFAM has a number of other useful properties for medical pattern classification tasks.

First, as noted earlier, the model has but one user-changeable parameter, the baseline vigilance of the ART_a module. SFAM can thus be easily tuned to a particular task.

Second, successful learning can occur with only one pass through the data set (termed single-epoch training). This is demonstrated within this paper, since all three classification tasks we describe utilize such single-epoch training.

Third, the model does not perform optimization of an objective function and is not therefore prone to the problem of local minima as occurs with feedforward networks using backpropagation. Also, the problem of selecting the appropriate number of hidden units does not occur. This is because, as described previously, SFAM self-organizes its own structuring of the data, automatically creating new category clusters for itself as and when they become needed.

Fourth, the model is able to discriminate rare events from a "sea" of similar cases with different outcomes owing to the feedback mechanism based on top-down matching of learned categories to input patterns. This is again in contrast to feedforward networks using backpropagation where weights are refined by a process which effectively averages together similar cases and thus fails to acknowledge rare events. Therefore SFAM should be suitable for domains where the distribution of data items is highly skewed between different categories. Such an application domain is described in the next section.

3 Prognosis of Coronary Care Patients

3.1 Application Domain

The application task described in this section is the prediction of the death or survival of patients admitted to a coronary care unit. We highlight the ARTMAP voting strategy (section 3.2) using this domain. A more extensive description of our findings is provided in Downs et al. (1995).

Since the fifties there has been a progressive trend to reduce the length of hospital stay for coronary care patients, which has provided economic benefits without significantly increasing mortality rates (Parsons et al., 1994). However, continuation of this trend requires the accurate identification of low-risk patients soon after their admission to hospital. Neural networks have the potential to allow this.

The data used in this study consisted of 4200 complete records for patients admitted to the coronary care unit of Leicester Royal Infirmary (United Kingdom) over a five year period

(1987–1992). Each record consisted of 43 items of clinical or electrocardiographic data considered to be useful for patient prognosis, together with the outcome for the patient's stay in hospital—death or survival.

Of the three tasks described in this paper, this problem is the most difficult. First, the data is "noisy" in the sense there are no features which provide clear-cut delineation of category boundaries. (In other words, all features are very weak indicators of the actual pattern classification.) Second, the distribution of outcomes is highly skewed—only 7.1% of all patients admitted die while on the ward.

3.2 ARTMAP Voting Strategy

The formation of category clusters in the ARTMAP models is affected by the order of presentation of input data items (Carpenter et al., 1992). Thus the same data presented in a different order to distinct SFAM networks can lead to the formation of quite different clusters within the nets. This subsequently leads to different categorisations of test data, and thus different performance scores. This effect is particularly marked with small training sets and/or high-dimensional input vectors, where the data set may not be fully representative of the domain, and with single-epoch training.

This effect can be compensated for by the use of the ARTMAP *voting strategy* (Carpenter et al., 1992). This works as follows: a number of 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 (or the most "votes") is the final predicted category outcome. The voting strategy can provide improved performance in comparison with 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. In particular, this and subsequent, applications utilize unanimous verdicts to indicate predictions which have a high certainty of being correct.

The voting strategy potentially compromises the utility of SFAM for incremental learning in non-stationary environments, since randomization of the input disrupts their original temporal order. However, this should not be a problem if the training data is "batched" appropriately. Thus, in non-stationary domains, instead of randomizing across the entire set of training data, a number of subsets, each containing consecutively ordered data items would be taken. Each such batch would then be separately randomized and the voting networks trained incrementally on the batches, presented to the networks in the correct temporal order.

3.3 Method

The data were partitioned into a training set, comprising the first 3000 patient records, and a test set comprising the remaining 1200 records. Twenty different orderings of the training set were derived and served as input data to separate instances of SFAM. The vigilance parameter was set low (0.3) to avoid excessive cluster formation with the large training set. (This is a notable problem for ARTMAP models—see Marriott and Harrison, In Press.)

The voting strategy was also employed on the test data. A range of 3 to 13 inclusive odd numbered voters was used (odd numbers ensuring no tied decisions occurred), choosing those

SFAM instances from the pool of 20 that had achieved the highest individual accuracy scores.

3.4 Results

Initial performance on the test set proved disappointing. Accuracy for the individual SFAM networks ranged between 73.2% and 87.5% with a mean of 81.1%. This compares with a default accuracy of 92.9% for the simple assumption that all patients will survive. The reason for this is that SFAM over-represents the rare cases of patient deaths in excess of their actual frequency within the data set. (This is probably because such cases were not tightly clustered together but widely spread throughout the feature space.). Thus SFAM appears to suffer from the opposite problem to the feedforward networks—too much credence, rather than too little, is given to rare cases.

The general effect of the voting strategy was to increase accuracy to around 89–91%, still slightly below baseline performance. However, the voting strategy did provide useful results for the important special case of high-confidence predictions of patient survival. (A high confidence prediction being one upon which all voters agreed.) Such patients are the most suitable for early hospital discharge.

With 3 voters, a unanimous survival decision accounted for 911 of the data items and was proved wrong 44 times. This translates to 95.2% accuracy covering 75.9% of the 1200 test items. With extra voters, accuracy steadily improved at the cost of decreased coverage (see table 1), until at the 13 voter stage an accuracy of 99.3% covering 34.0% of the data was achieved. The figures for this latter case are nearly identical to those achieved by Parsons et al. (1994) using the statistical technique of logistic regression upon a different data set collected for the same purpose. (Parsons et al. achieved 99.2% accuracy in a third of all cases using a data set of 5746 training items and 1000 test items respectively.) However, the advantage of the unanimous voting strategy with SFAM is its ability to gain wider coverage of the data set with only a small decrease in accuracy by reducing the number of voters.

Table 1: Voting strategy performance for unanimous survival decisions

Number of Voters	Accuracy (%)	Coverage of Cases (%)
3	95.2	75.9
5	95.6	64.5
7	97.8	53.0
9	98.2	45.3
11	98.1	40.3
13	99.3	34.0

4 Diagnosis of Breast Cancer

4.1 Application Domain

The application task described in this section is the diagnosis of cancer from fine needle aspirates of the breast. The section provides a synopsis of research detailed in Downs, Harrison and Cross (1995, In Press). Within this domain we highlight the symbolic rule extraction capabilities of SFAM (section 4.2).

Breast cancer is a common disease affecting around 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 (Underwood, 1992). The primary method of diagnosis is through microscopic examination by a pathologist of cytology slides derived from fine needle aspiration of breast lesions. 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.) 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.

The most important performance metric in this domain is not overall diagnostic accuracy but specificity. This is because the pathologist's prime concern is to avoid false positive predictions (diagnosing benign lesions as malignant) since these may result in unnecessary surgery such as mastectomy or wide local excision of the lesion. False negatives are tolerated because, if the clinical suspicion of malignancy remains, the surgeon will then take further samples for additional testing by the pathologist. (Indeed, false negatives are inevitable within this domain since some aspirations fail to locate a malignant lesion and extract nearby healthy tissue.)

The data used in this study consisted of 413 patient records each comprising ten binary-valued features recorded from observation of breast tissue samples by an expert pathologist (of Consultant status with 10 years experience in the field). The samples were taken from patients referred to the Royal Hallamshire Hospital, Sheffield, UK with symptomatic breast lesions between 1989–1993. The distribution of categories within the data was fairly even—53% of cases were malignant, 47% benign. An additional data set was also employed comprising 82 malignant and 82 benign cases. This data was derived from tissue observations performed by a "neophyte" pathologist (Senior House Officer with 18 months experience of the field).

The ten data features used in the study are all claimed to have diagnostic value for the task (see Wells et al., 1994). Table 2 provides the definitions of each feature, together with the abbreviations by which they will be referred to throughout the remainder of this section.

4.2 Symbolic Rule Extraction

A common general criticism of neural networks is the opaqueness of their learned associations. In medical domains, this "black box" nature may make clinicians reluctant to utilize a neural network decision support tool, no matter how great the claims that are made for its performance. 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". In particular, if a clinician who routinely uses a decision-support tool becomes involved in litigation, the rules may serve as important legal evidence (Brahams and Wyatt, 1989).

Table 2: Abbreviation and definition of data features used in breast cancer diagnosis

Abbreviated Feature Name	Definition of Feature
DYS	True if majority of epithelial cells are dyshesive, 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 present 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 majority of epithelial cells, false if not present in majority of epithelial cells.

The ARTMAP models have been endowed with symbolic rule extraction capabilities (Carpenter and Tan, 1993; Tan, 1994). The act of rule extraction in SFAM is a straightforward procedure 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 antecedent is the category prototype weights and whose consequent is the associated ART$_b$ category (denoted via the map field).

These rule extraction facilities provide two advantages which, taken collectively, should help to overcome reluctance to utilize 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 SFAM 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 clinician is then able to decide whether or not to concur with the network's prediction, based upon how valid he or she believes the rule(s) to be.

The specific rules discovered for this domain will be presented in section 4.4. However, some discussion of their general nature is needed here since they differ somewhat from the production rules used in conventional expert systems. 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 models, 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 inexactitude 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.

Additionally, the rules are self-discovered though exposure to domain exemplars, rather than having been externally provided by a human expert. ARTMAP models are thus able to bypass the difficult and time-consuming knowledge-acquisition process found with rule-based expert systems (Hayes-Roth, Waterman and Lenat, 1983). However, collection of the data may itself be a non-trivial task in many medical domains.

A drawback of this approach is that the rules are "correlational" rather than causal, since SFAM 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 the ARTMAP models but occurs with neural networks generally.) However, this difficulty is probably not of great importance from an applications viewpoint since useful diagnostic performance can often be achieved from correlational features without recourse to any "deep" knowledge of the domain.

A final general point concerns the learning rule in SFAM which governs the formation of category clusters, and hence the rules that will be derived from these clusters. Under the "fast-learning" conditions with binary data used in this application, whenever an input is successfully matched to an existing category cluster node the new weights for that node are formed by taking the logical AND of the input pattern and the existing weights for that cluster. This has the effect of deleting all features from the category cluster weights that are not also present in the input pattern. Hence, the weights tend to denote progressively more general clusters as they encode more input patterns and more features are deleted. Additionally, all features that are still present in the weights for a cluster once training ceases are known to have been present in all input vectors encoded by that cluster.

4.2.1 Category Pruning

A SFAM network can often become "over-specified" on 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 as occurs in this domain (see section 4.3). To overcome this difficulty, rule extraction involves an initial stage of category pruning prior to that act of rule extraction per se. (With continuously-valued category weights, rule extraction is preceded by a second *quantization* stage—see Carpenter and Tan, 1993. However, the binary data under fast-learn conditions used in this domain yields purely binary category weights and subsequently provides rules of greater clarity. Quantization is therefore omitted from this description.)

Pruning is guided by the calculation of a *confidence factor* (CF) between nought and one for each category cluster, based upon a node's *usage* and *accuracy*. The usage score for an ART_a node is simply the number of training set exemplars it encodes, normalized through division by the maximum number of exemplars encoded by any node with the same category outcome. (Hence, there will be at least one node for each different category class which has a maximal usage score of one.) The accuracy score for a node is calculated as the proportion of predictions that are correct which the node makes on a prediction data set separate to the training data. This score is then normalized, similarly to the usage calculation, through division by the maximum proportion of correct predictions made by any node with the same outcome. (Thus there will be at least one node for every category class which has a maximal accuracy score of one.) The confidence factor for a node is then calculated as the mean of its usage and accuracy scores. All nodes with a confidence factor below a user-set threshold will be pruned. Full details of the process are given in Carpenter and Tan (1993) or Tan (1994).

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

4.3 Method

One hundred cases were randomly selected from the data set of the Consultant's observations to serve as a combined prediction and test set for the neural network model. The remaining 313 items served as training data. The neophyte's observations provided an additional test set.

Ten SFAM networks were trained on different orderings of the teaching data. Vigilance was set very high (0.9) during training in order to maximize classification performance. Performance on both test sets was recorded for each network. Performance using the voting strategy was also recorded on both test sets, using the 5 networks with the highest accuracy on each test set as the "voters". Vigilance was relaxed to 0.6 during testing to ensure all cases matched to an existing category cluster node (i.e. forced choice prediction).

All ten trained networks were then severely pruned using a confidence factor threshold of 0.7, based upon performance with the Consultant's data. The testing procedure was then repeated using the resultant pruned networks. Vigilance during testing for the pruned networks was relaxed further to 0.5 owing to their decreased coverage of the feature space. In addition, symbolic rules were extracted from each pruned network.

There is a flaw in the method for this domain, in that the pruned networks are tested upon the prediction set, which has been previously utilized to guide the actual pruning process (via calculation of the accuracy scores), rather than upon entirely novel test data. This problem occurs because the relatively small size of the data set did not allow separate training, prediction and test sets of reasonable size to be derived. The possibility therefore exists that diagnostic accuracy for the pruned networks will be optimized on the prediction set without necessarily providing improvements that generalize to novel data. However, this is not in actuality a problem, for reasons that will be provided in section 4.4.

4.4 Results

Table 3 shows the voting strategy and mean individual performance for both pruned and unpruned network types on the test set of the expert's observations, and comparative performance figures for that expert.

Table 3: Relative performance of senior pathologist and network types

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Consultant Pathologist	91	83	100
Unpruned SFAM—Individual Mean	94	96	92
Unpruned SFAM—Voting Strategy	95	96	94
Pruned SFAM—Individual Mean	94	90	99
Pruned SFAM—Voting Strategy	95	92	98

It can be seen that in terms of diagnostic accuracy SFAM always performs slightly better than the human expert. However, the weak spot in the unpruned SFAM networks' performance is their much lower specificity in comparison to the human pathologist. As pointed out in section 4.1, it is vital that false positive cases (which reduce specificity) are avoided in this domain. 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 $\mathrm{ART_a}$ predominantly indicate positive (malignant) cases. (On average, 70% of $\mathrm{ART_a}$ category nodes in the unpruned networks denoted 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 as a side-effect of the pruning process. In section 5.2.1 however we introduce a simple generalisation of the pruning algorithm which allows this effect to be achieved deliberately.)

The accuracy of both types of voting networks for this data appears to be very close to the optimum possible, since the existence of ambiguous feature-states means that approximately 4% of data will always be misclassified. This explains our previously noted lack of concern about the absence of a novel test set of Consultant's observations for the pruned networks. Recall that the difficulty is that pruning may optimalize accuracy on the prediction set without the improvements generalising to new data. However, in this case near-optimal accuracy has been achieved prior to pruning (i.e. from training alone), as shown by the unpruned voting networks' performance (the prediction/test set being entirely novel data for the unpruned networks). Subsequently therefore, pruning does not improve accuracy (indeed how could it?), it merely alters the balance between sensitivity and specificity.

Table 4 shows the voting strategy and mean individual performance for both pruned and unpruned network types on the test set of the neophyte's observations, and comparative performance figures for that pathologist.

Table 4: Relative performance of junior pathologist and network types

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Junior Pathologist	78.7	57.3	100.0
Unpruned SFAM—Individual Mean	73.7	66.7	80.7
Unpruned SFAM—Voting Strategy	75.0	74.4	75.6
Pruned SFAM—Individual Mean	76.0	57.6	94.5
Pruned SFAM—Voting Strategy	75.6	57.3	93.9

Previously with the Consultant's test data, it was observed that pruning had the effect of biasing network performance towards increased specificity and also that the voting strategy always gave improved performance (albeit slight) over the individual networks. With the data used here, it can be seen that the former effect still occurs, but the latter does not. More importantly, performance of all types of network is not significantly better than that of the junior pathologist. The unpruned networks show better sensitivity but possess unacceptable specificity. In comparison to the unpruned networks, the pruned networks achieve higher specificity but at the expense of reducing sensitivity to a very similar level to that of the junior pathologist.

Kappa statistics for the observations of each of the features, reflecting the level of agreement in feature assignment between the senior and junior pathologists, showed that for most of the features there was only a moderate level of agreement. (Indeed three features—NAKED, FOAMY and NECROTIC—had levels of agreement that were little better than chance. These features are the most difficult to identify since a high level of interpretation is required by the pathologist to identify cell type and biological viability.) It is highly likely that this lack of agreement in feature assignment was the cause of the reduction in network performance when using the junior pathologist's data.

Thus the performance results indicate that although the existing application could prove useful as a decision-support tool for use by senior pathologists, it is inadequate (without further modifications) for use with poor-quality input data provided by a junior pathologist.

Turning now to the symbolic rule extraction capabilities, 14 distinct rules were derived from the 10 pruned networks, 12 for malignant outcomes and 2 for benign. These are shown in table 5, ranked by how many of the 10 networks each rule occurred in. No single rule in the set can be taken as canonical, since it should be recalled that each rule is derived from a node which covers only a portion (albeit an important one) of the overall feature space for each diagnostic category. However, taking the rules as a whole, a picture of a typical benign or malignant case can be constructed.

Table 5: Symbolic rules extracted from pruned networks

```
Rule 1 (10 Occurrences)
IF
    NO-SYMPTOMS
THEN
    BENIGN

Rule 2 (8 Occurrences)
IF
    3D=TRUE
    NUCLEOLI=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
THEN
    MALIGNANT

Rule 3 (8 Occurrences)
IF
    3D=TRUE
    FOAMY=TRUE
    NUCLEOLI=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
THEN
    MALIGNANT

Rule 4 (7 Occurrences)
IF
    FOAMY=TRUE
THEN
    BENIGN

Rule 5 (4 Occurrences)
IF
    ICL=TRUE
    3D=TRUE
    NUCLEOLI=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
THEN
    MALIGNANT

Rule 6 (4 Occurrences)
IF
    DYS=TRUE
    NUCLEOLI=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
THEN
    MALIGNANT

Rule 7 (3 Occurrences)
IF
    FOAMY=TRUE
    NUCLEOLI=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
THEN
    MALIGNANT

Rule 8 (3 Occurrences)
IF
    NUCLEOLI=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
THEN
    MALIGNANT

Rule 9 (2 Occurrences)
IF
    3D=TRUE
    FOAMY=TRUE
    NUCLEOLI=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
    NECROTIC=TRUE
THEN
    MALIGNANT

Rule 10 (2 Occurrences)
IF
    3D=TRUE
    FOAMY=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
    NECROTIC=TRUE
THEN
    MALIGNANT

Rule 11 (2 Occurrences)
IF
    DYS=TRUE
    ICL=TRUE
    NUCLEOLI=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
THEN
    MALIGNANT

Rule 12 (1 Occurrence)
IF
    FOAMY=TRUE
    NUCLEOLI=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
    NECROTIC=TRUE
THEN
    MALIGNANT

Rule 13 (1 Occurrence)
IF
    ICL=TRUE
    NUCLEOLI=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
THEN
    MALIGNANT

Rule 14 (1 Occurrence)
IF
    ICL=TRUE
    3D=TRUE
    PLEOMORPH=TRUE
    SIZE=TRUE
THEN
    MALIGNANT
```

Benign cases are likely to display either no features, or the FOAMY feature in isolation. Malignant cases are almost certain to display a combination of NUCLEOLI, PLEOMORPH and SIZE. The 3D feature is also strongly implicated in malignancy. FOAMY, ICL, NECROTIC, and DYS may further be present, although with a lower likelihood. The senior pathologist in this study confirmed the validity of these rules and the relative importance of the features, with the exception that he places no value on the presence or absence of the FOAMY feature. This matter will be discussed later in this section.

Wells et al. (1994) provide a canonical list of diagnostic criteria for FNAB which includes all features used in this study, although no assessment of their relative importance or likelihood is given. In summary, they cite FOAMY, APOCRINE and NAKED as indicators of benignancy, and all other features used here as indicators of malignancy. The self-discovered rules of SFAM show good overall agreement with these criteria apart from two notable exceptions. First, APOCRINE and NAKED are conspicuous by their absence from any of the SFAM rules. Second, FOAMY has an ambiguous status, being present in rules for both benign and malignant outcomes.

The first discrepancy can be explained by reference to the way in which CFs are calculated for nodes in SFAM based equally upon both usage and accuracy. The high CF threshold for pruning in this application requires a node to be both highly accurate and to encode a large proportion of exemplars of a particular category in order to remain unpruned. It is thus possible for a node with very good predictive accuracy but low usage to be pruned. This indeed happens in the case of nodes containing the APOCRINE and NAKED features, which both occur rarely in the data. Examination of the unpruned networks revealed the frequent occurrence of nodes where these features, in isolation or conjunction with the FOAMY feature, indicate a benign diagnosis. Although such nodes usually have a perfect accuracy score, they also have a very low usage score and hence their overall CF value falls below the threshold for pruning.

It therefore seems likely that the CF threshold for pruning was set too high and the networks were "over-pruned", resulting in the loss of some nodes with useful predictive powers. (Further evidence for this is provided by the fact that some of the pruned networks were unable to make a definitive prediction on all test cases despite the employment of a reduced vigilance level, indicating that pruning had left some networks with incomplete coverage of the state space.)

In further work this anomaly could be corrected by using a lower CF threshold for pruning. However, some degree of care must be taken with selection of the value of this threshold since if it is set too low the opposite problem will occur—relatively unimportant nodes will be left unpruned, increasing the size of the rule set and so making their validation a more time-consuming task for a domain expert.

The status of the FOAMY feature is a more problematic issue. Wells et al. (1994) classify it as an indicator of benignancy. However, the senior pathologist in this study regards its occurrence as little more than "background noise" which is as likely to be found in malignant cases as benign. Its status in the SFAM rules is certainly ambiguous. In isolation, the FOAMY feature frequently indicates a benign outcome. However, it is also present, in conjunction with other features, in a number of rules with malignant outcomes. The frequent occurrence of this feature in the rules as a whole indicates that it is present in a large proportion of the data, regardless of outcome. (This follows from the nature of the SFAM learning rule and the employment of a usage factor in the CF calculation as described previously in sections 4.2 and 4.2.1 respectively.)

If the relative frequency of occurrence is considered, the FOAMY feature can be seen to be present in 1 of the 2 distinct rules for benignancy, and 5 of the 12 for malignancy. Alternatively, if occurrence without regard for distinctiveness is considered, it occurs in 7 out of 17 benign rules and 16 out of 39 malignant rules. By either calculation its distribution between outcomes is very similar. We therefore conclude that, at least for this particular data set, the FOAMY feature tends more towards being "background noise" than a useful indicator of benignancy.

This issue further illustrates an important tension in the SFAM application between *knowledge engineering* and *machine learning*. From the standpoint of knowledge engineers, we would like all the rules discovered by SFAM for a domain to be acceptable to experts in that domain, since this obviously enhances confidence in the use of the model as a decision-support tool. However, from the machine learning standpoint, we would like SFAM to teach us something new about the domain, such as providing supporting evidence to resolve disagreements between experts, establishing the relative importance of different diagnostic features, or even establishing novel diagnostic features. Of course however, such findings may be at odds with the "received wisdom" of domain experts.

5 Diagnosis of Acute Myocardial Infarction

5.1 Application Domain

The application task described in this section is the diagnosis of acute myocardial infarction (AMI) from information available at an early stage of hospital admission. This section is based upon research described in Downs, Harrison and Kennedy (In Press). Within this domain we introduce a generalization of the category pruning process (section 5.2.1) as well as a "cascaded" version of the voting strategy which is intended to allow the identification of a subset of test cases for which SFAM has a very high certainty of providing a correct classification (section 5.2.2).

The early identification of patients with acute ischaemic heart disease remains one of the greatest challenges in emergency medicine. Chest pain is the commonest reason for emergency medical referral in the western world and is a major symptom of the onset of AMI. Each year in the UK alone over 240 000 cases are confirmed. However, the ECG only shows diagnostic changes in about half of AMI patients at presentation (Adams, Trent and Rawles, 1993). 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 (Adams, Abendschein and Jaffe, 1993). The initial diagnosis of AMI, therefore, relies on an analysis of clinical features along with ECG data.

The data used in this study were derived from consecutive patients attending the Accident and Emergency Department of the Edinburgh Royal Infirmary, UK, with non-traumatic chest pain as the major symptom. 970 patients were recruited during the study period (September to December 1993). The final diagnoses for the patients was AMI in 191 cases (which includes both Q wave and non-Q wave AMI) and not-AMI in all other cases (which includes stable and unstable angina plus other diagnoses). The distribution of data items in this domain is thus moderately skewed with only 19.7% of cases being positive (AMI). The input data items for the SFAM model were all derived from clinical or ECG data available at the time of the patient's presentation. In all, 35 items were used, coded as 37 binary inputs.

5.2 Modifications to Simplified Fuzzy ARTMAP

In this section two modifications to the standard SFAM model are introduced. First, the category pruning process described previously in section 4.2.1 is generalized to allow different CF pruning thresholds for nodes with different category outcomes. The resultant differently pruned networks are then utilized in a "cascaded" version of the voting strategy which allows identification of those cases for which SFAM is almost certain to make the correct category prediction.

5.2.1 Generalized Category Pruning

In the original formulation of the pruning process, a uniform CF threshold is used to select candidate nodes for deletion, irrespective of their category class. In this application, the pruning process is generalized to allow separate CF thresholds for nodes belonging to different category classes. This allows the proportion of the state-space covered by different categories to be varied. For example by increasing the CF threshold for nodes with positive outcomes the relative proportion of such nodes is decreased and thus the sensitivity of the network is reduced. (The same effect can also be achieved of course by decreasing the CF threshold for nodes with negative outcomes.)

This modification is useful for medical domains since it allows a SFAM network to be pruned so as to trade sensitivity for specificity and vice versa. In particular, variable CF thresholds are used to produce networks whose performance shows near-perfect sensitivity, near-perfect specificity, and approximately equal sensitivity and specificity (implying the same value for accuracy).

5.2.2 Cascaded Voting Strategy

The generalization of the category pruning process described above allows a novel "cascaded" variant of the voting strategy to be employed as shown in figure 2. This consists of three layers, a set of voting networks pruned so as to maximize sensitivity, another set pruned so as to maximize 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.

The cascaded voting strategy therefore operates as follows: An input data item is first presented to the sensitive voting networks. If these yield a unanimous negative (not-AMI) 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 (AMI) verdict, this is taken as the ultimate category prediction[1]. 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

[1] Obviously, the order of presentation between the sensitive and specific voting layers is not crucial, although for efficiency reasons it is preferable to have the voters which capture the largest number of cases as the first layer.

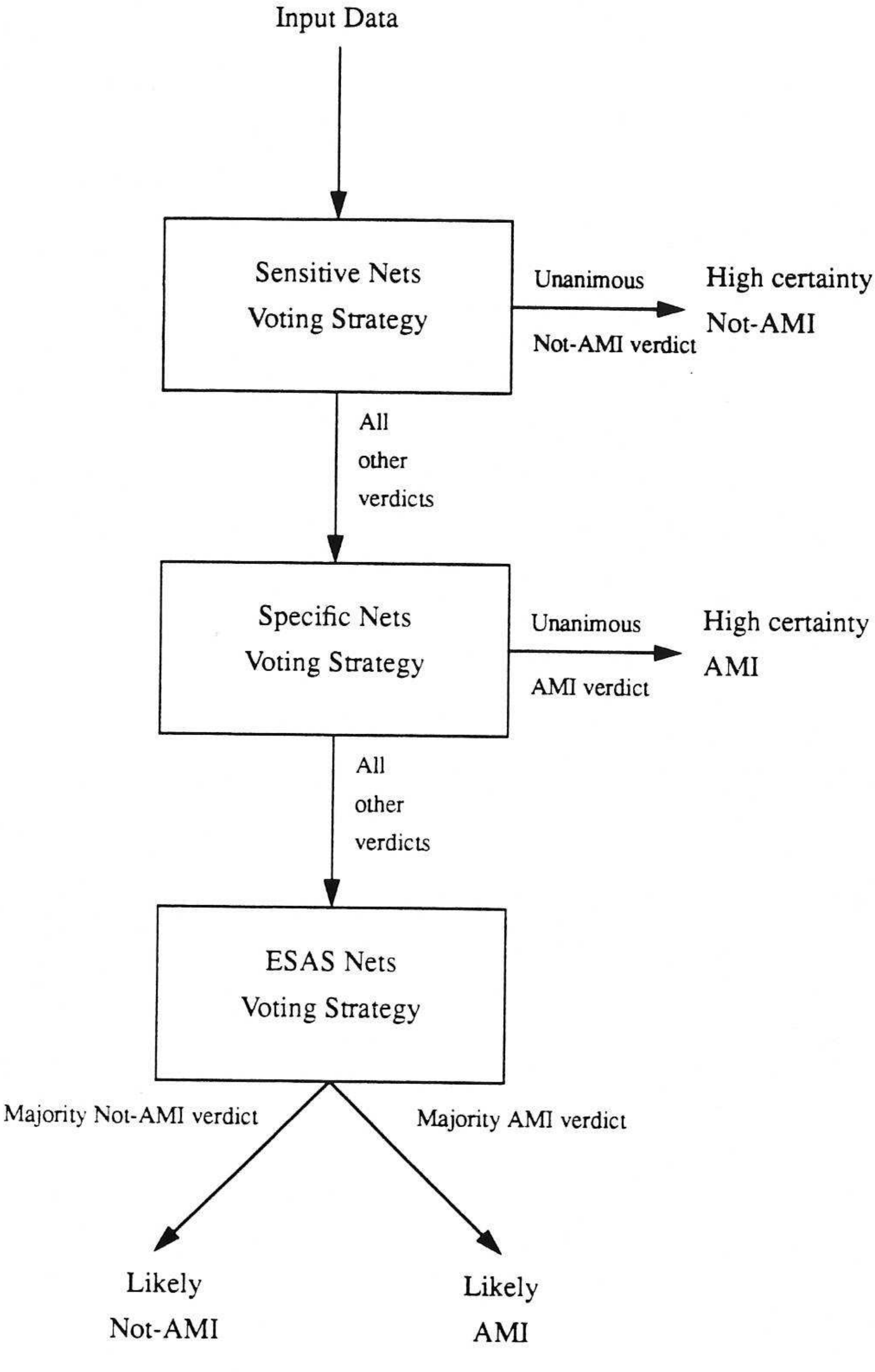

Figure 2: Voting strategy cascade for AMI diagnosis

correct than with the previous two layers.

5.3 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 (reflecting the prior distributions of outcomes.)

The training data was randomly ordered in ten different ways, and each ordering applied to a different SFAM network. The ART_a base-line vigilance was set to a medium level (0.6) for training. The performance of the 10 trained SFAM networks was then measured on the prediction set in order to calculate accuracy scores for the category nodes in each network, as a prerequisite to category pruning. During this testing phase the ART_a baseline vigilance was relaxed slightly (to 0.5) to ensure that all test items were matched to an existing category cluster (i.e. forced choice prediction).

The networks were then pruned in four different ways. First, the "standard" form of category pruning (Carpenter and Tan, 1993) was performed on the original networks, such that all nodes with a CF below 0.5 were deleted from the networks to improve predictive accuracy. The original networks were then pruned using different CF thresholds for the AMI and not-AMI nodes to produce pruned networks which maximized *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 all 10 pruned networks. A similar procedure was then conducted to produce 10 networks which maximized *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.

Performance results were then measured on both the prediction and test sets using the voting strategy with 5 networks. Voters for the unpruned, 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 specific networks. Vigilance was further relaxed to 0.4 for testing all pruned networks, again to ensure forced choice prediction.

Lastly, performance results on the prediction and test sets were recorded for the "cascaded" voting strategy. This employed 3 sensitive nets, 2 specific nets and 5 ESAS nets, the number of voters for the high-certainty prediction layers being selected on the basis of maximizing the number of cases "trapped" while maintaining perfect sensitivity or specificity on the prediction set.

5.4 Results

Table 6 shows the standard voting strategy performance for the different network types on both the prediction and test sets. The figures for the test set are of the most importance since, unlike the prediction set, this comprises entirely novel data not previously presented to the networks.

(Recall that initial performance on the prediction set is reflected during category pruning owing to the use of the derived accuracy scores for each node.) 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.

Table 6: Standard voting strategy performance of differently pruned networks

	Prediction Set			Test Set		
Network Type	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
Unpruned	86.0	64.3	91.0	83.3	56.7	90.0
Uniform Pruning	92.0	78.6	95.1	88.0	56.7	95.8
Pruning for Sensitivity	55.3	96.4	45.9	51.3	96.7	40.0
Pruning for Specificity	88.7	46.4	98.4	84.7	33.3	97.5
Pruning for ESAS	82.0	82.1	82.0	81.3	83.3	80.8

It can be seen that accuracy on the test set for the unpruned networks is very close to 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 slightly superior specificity.

As expected, the uniformly pruned networks show an across-the-board increase in accuracy over the unpruned nets, with a 4.7% increase on the test set, and a 6.0% increase on the prediction set. (The greater increase in performance on the prediction set is explained by the fact that pruning utilized the accuracy scores for this data, and the networks are consequently optimized for the prediction set.) However, the increase in accuracy is largely because of an overall improvement in specificity rather than sensitivity, which remains unchanged on the test set. Thus, although accuracy is now higher than for the human clinicians, this result remains an artefact.

Figures for the sensitive nets show that almost all AMI cases can be diagnosed by the network, while 40% of the not-AMI cases in the test set are detected by the network. Conversely, with the specific nets, almost all not-AMI cases are covered while approximately one-third of the AMI cases are also 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 accuracy and specificity of such networks is slightly worse than for the human diagnoses but sensitivity is slightly better.

The best overall network performance was achieved by the cascaded voting strategy, as shown in table 7.

Table 7: Performance of the cascaded voting strategy

	Prediction Set			Test Set		
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
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 Performance	82.0	82.1	82.0	82.7	86.7	81.7

The cascade's overall performance can be seen to be almost identical to that of the Casualty Doctors. Moreover, the cascade provided a successful partitioning of input items into those with a high and a lower certainty of a correct diagnosis. Unanimous not-AMI decisions by the highly sensitive networks (i.e. the first stage of the cascade) resulted in only one false negative prediction. Similarly unanimous AMI decisions by the highly specific networks (the second stage of the cascade) made only one false positive prediction. The ESAS class voters then provided lower certainty predictions for the data items reaching 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.

Examination of the input features for the two false predictions made by the high-certainty voters is revealing. The false positive case had the following features: age=45-65, smoker, family history of ischaemic heart disease, central chest pain radiating to the jaw, short of breath, nausea, new ST segment elevation, new pathological Q waves and ST segment or T wave changes suggestive of ischaemia. This exhibits almost all of the "classic" features of AMI, the latter three features being regarded as particularly strong AMI indicators. The false negative case had the following features: age<45, smoker, pain in left side of chest radiating to the left arm, pain described as sharp or stabbing, old ECG features of MI and ECG signs of ischaemia known to be old. This displays none of the classic features of AMI, although the existence of the latter two features should mean a human clinician probably would not entirely discount the possibility of AMI. We conclude therefore that these cases are idiosyncratic, particularly the false positive, and would cause most human experts to make the wrong diagnosis. (Unfortunately a direct comparison with the Casualty Doctor's performance cannot be made because the database did not include their diagnoses for individual cases.) Thus the general ability of the cascaded voting strategy to identify cases with high-certainty of a correct diagnosis is not greatly undermined by these cases.

6 Discussion

We believe the SFAM model to have definite promise as part of a decision-support tool for many medical pattern classification tasks. Useful performance figures with the voting strategy were obtained across all three domains demonstrated here, all of which are important medical tasks which potentially could benefit from computer-aided decision support. (See also Goodman et al., 1994, for an application of fuzzy ARTMAP to a further medical domain—prediction of length of hospital stay for patients with pneumonia.) Additionally, these results were obtained using single-epoch (and potentially incremental) learning without the need for extensive parameter tuning. Furthermore, the model's rule extraction facilities provide a highly valuable supplement to its predictive capabilities.

Nonetheless, we do not wish to claim that SFAM offers a panacea for medical decision-support. A number of limitations (and thus possible directions for future work) can be identified from our findings.

First, the claim that ARTMAP models are suitable for domains with skewed distributions of outcomes needs to be regarded with a moderate degree of caution. While useful performance was obtained with the skewed data of the AMI domain, performance with the heavily skewed data of the coronary patient prognosis domain was not entirely satisfactory. However, a modified version of fuzzy ARTMAP exists (Lim and Harrison, In Press) which gives superior performance for a single network with this data but does not perform well when used with the voting strategy (see Downs et al., 1995).

A more important limitation of SFAM for medical decision-support applications is that the model makes no provision for missing data items when generating predictions. However, a variant of fuzzy ARTMAP, known as fusion ARTMAP (Asfour et al., 1993), has been developed which it is claimed can cope with this problem. Fusion ARTMAP utilizes a modular approach clustering data from disparate sources locally and then passing the results to a global classifier. This enables the system to assign credit for successive predictions to those sources of information which have the highest predictive value. This would enable the system to make a reasonable guess even if some data were missing, or alternatively, to request additional information if insufficient is available. Thus, for example, in the diagnosis of myocardial infarction, the system would be able to make use of the most highly predictive data (the ECG codings, say) first and then request information on physical signs, associated symptoms, risk factors, clinical history etc. as required, until a confident prediction could be made. Further tests such as a chest X-ray might then be requested to reinforce the diagnosis. In this way, fusion ARTMAP begins to reflect human behaviour—building up a picture gradually using the least amount of data concomitant with confident diagnosis. In future research, therefore, we intend to investigate the ability of fusion ARTMAP to perform robustly in medical domains when data items are missing.

Another area for future work is to automate the CF threshold selection process for the differential category pruning described in section 5.2.1. In the present implementation, the CF thresholds were "hand-set" by the system's designer to achieve the desired changes in network performance. However, this was a rather laborious trial-and-error process (particularly for the ESAS networks, where each individual network required a different CF threshold) which contrasts poorly with the general ease of tuning of the basic SFAM model.

Additionally, we would like to achieve useful performance figures with the "noisy" data provided by the junior pathologist for the breast cancer domain (see section 4.4.). We conjecture that one way this might be achieved is to modify SFAM with a more sophisticated matching technique between input cases and category clusters. In SFAM, each true input feature contributes equally to the match with a category prototype. We envisage introducing a variable weighting for features, which attaches more importance to individual features that are considered to be (a) very strongly predictive for the domain and (b) most easily identified by an inexperienced pathologist.

The possibility also exists that revised and/or expanded version of the data sets for each domain may yield improved performance figures. We believe this is least likely to be true for the breast cancer domain and most likely to be true for the AMI domain. As noted earlier in section 4.4, network performance for the diagnosis of breast cancer is already very close to the optimum possible. (However, further data would allow the flaw in the method to be corrected that was noted in section 4.3.) In contrast, with the AMI data we believe that the prediction and test sets were probably too small, particularly given the unequal distribution of category classes with relatively few AMI cases. The small number of AMI cases in the prediction set is the cause of most concern, since optimum benefit from category pruning is achieved only if the prediction set is truly representative of the overall domain. Otherwise, pruning will optimize a net's performance on the prediction set, but not will not generalize well to novel test data. The generally lower performance of all network types on the test set in comparison to the prediction set for this domain (see section 5.4) leads us to believe that this is the case here.

However, performance results alone are not enough to ensure the acceptance of a decision-support tool based upon SFAM. Usability is (at least) an equally important factor. Thus the tool must provide the capability to interface in a straightforward manner between different medical databases and SFAM by providing standard database and SFAM manipulation procedures. For example, there should be facilities for partitioning a database into training, prediction and test sets; selecting particular input features to train a SFAM network on; setting the SFAM vigilance parameter; saving and loading a trained network's weights; extracting symbolic rules and so on. Such facilities should be as easy to use as possible, thus offering the possibility that a SFAM decision-support tool for a medical domain could be constructed with little or no intervention by an AI expert or knowledge engineer. A graphical user interface (GUI) therefore seems to be called for.

Finally, a cross-comparative study of SFAM with other techniques needs to be performed across a range of medical domains. Rival approaches for the comparison could be statistical (e.g. logistic regression), neural network (e.g. MLP or RBF), or symbolic machine learning (e.g. decision trees). Preliminary findings by us seem to suggest that SFAM is likely to show somewhat superior performance to logistic regression and the MLP with the breast cancer data, but slightly inferior performance with the AMI data, although these results should by no means be taken as definitive. Additionally, Goodman et al. (1994) demonstrate that fuzzy ARTMAP has superior performance to linear discriminant analysis in a pneumonia prognosis task.

Acknowledgement

Thanks to K.L. Woods of the Department of Pharmacology, University of Leicester, UK for providing the coronary care patient data used in section 3.

This research was supported by the Science and Engineering Research Council (SERC) of the UK, grant number GR/J/43233.

References

Adams, J., Abendschein, D.R. and Jaffe, A.S. (1993) Biochemical Markers of Myocardial Injury. Is MB Creatine Kinase the Choice for the 1990s?, *Circulation*, 88, 750–763.

Adams, J., Trent, R., and Rawles, J. (1993) Earliest Electrocardiographic Evidence of Myocardial Infarction: Implications for Thrombolytic Treatment, *British Medical Journal*, 307, 409–413.

Apolloni, B., Avanzini, G., Cesa-Bianci, N. and Ronchini, G. (1990) Diagnosis of Epilepsy via Backpropagation, *Proceedings of the International Joint Conference on Neural Networks*, Volume II, 571–574

Asfour, Y.F., Carpenter, G.A., Grossberg, S. and Lesher, G.W. (1993) Fusion ARTMAP: A Neural Network Architecture for Multi-Channel Data Fusion and Classification, *Proceedings of the World Congress on Neural Networks*, Volume II, 210–215.

Brahams, D, and Wyatt, J. (1989) Decisions Aids and the Law, *Lancet*, 632–634.

Bounds, D., Lloyd, P. and Mathew, B. (1990) A Comparison of Neural Network and Other Pattern Recognition Approaches to the Diagnosis of Low Back Disorders, *Neural Networks*, 3(5), 583–591.

Carpenter, G.A. and Grossberg, S. (1987) A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine, *Computer Vision, Graphics and Image Processing*, 37, 54–115.
Reprinted in Carpenter and Grossberg (1991) 316–382.

Carpenter, G.A. and Grossberg, S. (1988) The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network, *Computer*, 21(3), 77–88.

Carpenter, G.A. and Grossberg, S., eds (1991) *Pattern Recognition by Self-Organizing Neural Networks*.
Cambridge, MA: MIT Press.

Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H. and Rosen, D.B. (1992) Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps, *IEEE Transactions on Neural Networks*, 3(5), 698–712.

Carpenter, G.A., Grossberg, S. and Rosen, D.B. (1991) Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System, *Neural Networks*, 4(6), 759–771.

Carpenter, G.A. and Tan, A.H. (1993) Rule Extraction, Fuzzy ARTMAP, and Medical Databases, *Proceedings of the World Congress on Neural Networks*, Volume I, 501–506.

Cybenko, G. (1989) Approximations by Superpositions of a Sigmoidal Function, *Mathematics of Control, Signals and Systems*, 2, 303–314.

Downs, J., Harrison, R.F. and Cross, S.S. (1995) A Neural Network Decision Support Tool for the Diagnosis of Breast Cancer, in J.Hallam, ed., *Hybrid Problems, Hybrid Solutions*, 51–60.
Amsterdam: IOS Press.

Downs, J., Harrison, R.F. and Cross, S.S. (In Press) Evaluating a Neural Network Decision-Support Tool for the Diagnosis of Breast Cancer, to appear in *Proceedings of the 5th European Conference on Artificial Intelligence in Medicine* (AIME–95).
Berlin: Springer-Verlag.

Downs, J., Harrison, R.F. and Kennedy, R.L. (In Press) A Prototype Neural Network Decision Support Tool for the Diagnosis of Acute Myocardial Infarction, to appear in *Proceedings of the 5th European Conference on Artificial Intelligence in Medicine* (AIME–95).
Berlin: Springer-Verlag.

Downs, J., Harrison, R.F., Kennedy, R.L. and Woods, K. (1995) The Use of Fuzzy ARTMAP to Identify Low Risk Coronary Care Patients, in D. W. Pearson, N. C. Steele and R. F. Albrecht, eds., *Artificial Neural Networks and Genetic Algorithms*, 511–514.
Vienna: Springer-Verlag.

Egmont-Peterson, M., Talmon, J.L., Brender, J. and McNair, P. (1994) On the Quality of Neural Network Classifiers, *Artificial Intelligence in Medicine*, 6(5), 359–381.

Goodman, P.H., Kaburlasos, V.G., Egbert, D.D., Carpenter, G.A., Grossberg, S., Reynolds, J.H., Rosen, D.B., and Hartz, A.J. (1994) Fuzzy ARTMAP Neural Network Compared to Linear Discriminant Analysis Prediction of the Length of the Length of Hospital Stay in Patients with Pneumonia, in R.J. Marks, ed., *Fuzzy Logic Technology and Applications*, 424–429.
Piscataway, NJ: IEEE.

Grossberg, S. (1987) Competitive Learning: From Interactive Activation to Adaptive Resonance, *Cognitive Science*, 11(1), 23–63.

Hayes-Roth, F., Waterman, D.A. and Lenat, D.B. (1983) *Building Expert Systems*.
London: Addison-Wesley.

Harrison, R.F., Marshall, S.J. and Kennedy, R.L. (1991) A Connectionist Approach to the Early Diagnosis of Myocardial Infarction, *Proceedings of the 3rd European Conference on Artificial Intelligence in Medicine* (AIME–91), 119–128.
Berlin: Springer-Verlag.

Kasuba, T. (1993) Simplified Fuzzy ARTMAP, *AI Expert*, 8(11), 18–25.

Kennedy, R.L., Harrison, R.F. and Marshall, S.J. (1994) A Comparison of Logistic Regression and Artificial Neural Network Models for the Early Diagnosis of Acute Myocardial Infarction, Research Report 539, Department of Automatic Control and Systems Engineering, University of Sheffield.

Lim, C.P. and Harrison, R.F. (In Press) Modified Fuzzy ARTMAP Approaches Bayes Optimal Classification Rates: An Empirical Demonstration, to appear in *Neural Networks*.

Ma, Z. and Harrison, R.F. (In Press) A Heuristic for General Rule Extraction from a Multilayer Perceptron, in J.Hallam, ed., *Hybrid Problems, Hybrid Solutions*, 133–144.
Amsterdam: IOS Press.

Marriott, S. and Harrison, R.F. (In Press) A Modified Fuzzy ARTMAP Architecture for the Approximation of Noisy Mappings, to appear in *Neural Networks*.

Moody, J. and Darken, C. (1989) Fast Learning in Networks of Locally-Tuned Processing Units, *Neural Computation*, 1, 281–294.

Park, J. and Sandberg, I. (1991) Universal Approximation Using Radial Basis Function Networks, *Neural Computation*, 3, 246–257.

Parsons, R.W., Jamrozik, K.D., Hobbs, M.S.T. and Thompson, D.L. (1994) Early Identification of Patients at Low Risk of Death after Myocardial Infarction and Potentially Suitable for Early Hospital Discharge, *British Medical Journal*, 308, 1006–1010.

Pizzi, N., Choo, L.P., Mansfield, J., Jackson, M., Halliday, W.C., Mantsch, H.H. and Somorjai, R.L. (1995) Neural Network Classification of Infrared Spectra of Control and Alzheimer's Diseased Tissue, *Artificial Intelligence in Medicine*, 7(1), 67–79.

Rumelhart, D., Hinton, G. and Williams, R. (1986) Learning Representations by Back-Propagating Errors, *Nature*, 323, 533–536.

Sharkey, N.E. and Sharkey, A.J.C. (1994) Understanding Catastrophic Interference In Neural Nets, Research Report CS–94–4, Department of Computer Science, University of Sheffield.

Tan, A.H. (1994) Rule Learning and Extraction with Self-Organizing Neural Networks, in M. Mozer, P. Smolensky, D. Touretzky, J. Elman and A. Weigend, eds, *Proceedings of the 1993 Connectionist Models Summer School*, 192–199.
Hillsdale, NJ: Lawrence Erlbaum Associates.

Underwood, J.C.E. (1992) Tumours: Benign and Malignant, in J.C.E. Underwood, ed., *General and Systematic Pathology*, 223–246.
Edinburgh: Churchill Livingstone.

Wells, C.A., Ellis, I.O., Zakhour, H.D. and Wilson, A.R. (1994) Guidelines for Cytology Procedures and Reporting on Fine Needle Aspirates of the Breast, *Cytopathology*, 5, 316–334.