A Decision-Support Tool for the Diagnosis of Breast Cancer Based upon Fuzzy ARTMAP

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

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

R Lee Kennedy
Department of Medicine
The University of Edinburgh

Research Report #605
8 November 1995

Abstract

This paper presents research into the application of the fuzzy ARTMAP neural network model to the diagnosis of cancer from fine-needle aspirates of the breast. Trained fuzzy ARTMAP networks are differently pruned so as to maximize accuracy, sensitivity and specificity. The differently pruned networks are then employed in a "cascade" of networks intended to separate cases into "certain" and "suspicious" classes. This mimics the predictive behaviour of a human pathologist. The fuzzy ARTMAP model also provides symbolic rule extraction facilities and the validity of the derived rules for this domain is discussed. Additionally, results are provided showing the effects upon network performance of different input features and different observers. The implications of the findings are discussed.

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). For a general introduction to artificial neural network applications in medicine, see Cross, Harrison and Kennedy (1995), Baxt (1995) and Dybowski and Gant (1995).

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). In addition, it can be shown that, for one-from-many classification problems, attainment of the minimal value of a variety of cost functions with respect to the weights yields an estimate of the posterior (class conditional) probabilities required for the implementation of a Bayesian classifier (e.g. Richard and Lippmann, 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 adaptability 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. In general, causality dictates that it is not possible to know, a priori, whether or not a domain is stationary.

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 that 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 this paper we describe the application of a powerful, but relatively little-used, neural network model, fuzzy ARTMAP, to a medical decision-support task—assisting a pathologist in the diagnosis of breast cancer. Fuzzy ARTMAP is a neural network model, using both feedforward and feedback, which is not susceptible to the two criticisms cited above and has other desirable properties for this task (and medical classification tasks in general, see Downs et al., In Press). The paper will however demonstrate only the symbolic rule extraction capability of fuzzy ARTMAP, although some discussion will be made of its incremental learning ability in the conclusions.

The structure of the remainder of this paper is as follows. Section two defines some general medical terminology before progressing to a description of the breast cancer diagnosis task and previous work in this domain. Section three provides details of the fuzzy ARTMAP neural network model and its associated activities such as the voting strategy and symbolic rule extraction. Since fuzzy ARTMAP is relatively little-known this section provides quite detailed coverage of the model. Section four provides the basic performance results using fuzzy ARTMAP. Section five evaluates the symbolic rules derived from the trained networks. Section six provides further performance results investigating the effect of variations in input features. Section seven provides our conclusions and suggestions for further work.

2 Medical Background

2.1 Terminology

Medical domains inevitably have their own specialist terminology, the basics of which are defined in this subsection in order to facilitate comprehension of the remainder of this paper.

The presence of a particular disease or disorder is referred to as a *positive* case, while conversely the absence is referred to as a *negative* case. This in turn leads to four types of prediction when assessing the veracity of a diagnosis. A *true positive* is a case where the presence of a disease has been correctly diagnosed. Similarly, a *true negative* is a case where the correct diagnosis of the absence of a disease has been made. The two remaining types of prediction, false positives and false negatives, relate to incorrect diagnoses. A *false positive* is a case where a disease is diagnosed as being present, when in fact the patient does not have the disease. Conversely, a *false negative* is a case where a disease is diagnosed as being absent, when in actuality the patient does have the disease.

There are three commonly used metrics of diagnostic performance in medical domains. *Accuracy* is the proportion of all cases (positive or negative) correctly diagnosed, *sensitivity* is the proportion of positive outcomes correctly diagnosed, and similarly *specificity* is the proportion of negative outcomes correctly diagnosed.

Obviously, high accuracy requires both high sensitivity and high specificity. However, there is also a trade-off between sensitivity and specificity. If all cases are diagnosed as having a positive outcome, 100% sensitivity but 0% specificity will be achieved (and vice versa if all cases are classed as negative outcomes). Although this is an extreme example, it is often the case that diagnoses will be deliberately made that are to some extent biased towards one or the other outcome. (This will be demonstrated in the next subsection.) It should also be noted that false positive predictions lower specificity, while false negatives reduce sensitivity.

2.2 Breast Cancer

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 (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, FNAB, (Elston and Ellis, 1990). 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 cytopathologic diagnosis of FNAB have shown a range of specificity of diagnosis of 90–100% with a range of sensitivities from 84–97% (Wolberg and Mangasarian, 1993). These studies have been produced in centres specializing in the diagnosis of breast disease by pathologists with a special interest in breast cytopathology. In less specialized 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% (Start et al., 1992). 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[1]. 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.)

Table 1: 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 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 majority of epithelial cells, false if not present in majority of epithelial cells.

In the cytodiagnosis of FNAB there are some observable features which are cited as being important in the recognition of malignant cells. A "canonical" list is provided by Wells et al. (1994), although this publication does not attribute weights to these features or indicate the significance of combinations of these features. Ten of the features described by Wells et al. are utilized in this research. The medical definitions are shown in table 1, together with the abbreviations by which they will be referred to throughout this paper. As a general guideline, the features NAKED, FOAMY and APOCRINE are regarded as indicators of a benign outcome and all other features are indicative of malignancy. (However, some interactions

[1] However, in other contexts sensitivity can be more important than specificity. For example, the initial stage of the U.K.'s nationwide screening program (which uses mammography rather than FNAB) needs to avoid false negatives, so that all women with breast cancer will be referred to hospital for treatment.

between conjunctions of features are possible, see subsection 5.2.)

2.3 Previous Work

Some expert systems have been described which attempt to use human observations of features in FNAB and then apply computers to process these observations and attach weight to the presence and combination of features. Heathfield et al. (1990) describe a rule-based expert system with rules derived from cytopathological textbooks and discussions with pathologists but they do not give any results for the performance of the system on a test set of data. A Bayesian belief network has been developed by Hamilton et al. (1994). The conditional probability matrices relating each observed feature to the diagnosis were defined by a cytopathologist. The network was tested using 40 cases, it is difficult to assess the results because four categories of diagnosis were used (benign, malignant, atypical probably benign and suspicious) but 6% of the true benign cases and 9% of the true malignant cases were assigned to an equivocal category. Wolberg and Mangasarian (1993) have produced a large study with a 420 case training set and 215 case test set and they have used a user-modified computer-generated decision tree, the multisurface method of pattern separation and a connectionist system with a back-propagation learning algorithm. Nine cytological features were observed and given a scalar value of 1-10. On the test data set the decision tree method gave a specificity of 97% with a sensitivity of 93%, the connectionist network a specificity of 99% and a sensitivity of 97%, the multisurface separation method produced 100% specificity and sensitivity. However, some cases (such as cancer judged to have been missed by the aspirating needle) were excluded before analysis.

In previous work by the authors (Downs, Harrison and Cross, 1995a, 1995b) we applied the ARTMAP neural network model to this task using a 313 item training set and a 100 item test set. Various configurations of the model gave an accuracy of 94–95%, a sensitivity of 90–96%, and specificity of 92–99% (for full details see Downs, Harrison and Cross, 1995a). Atypical cases were not removed prior to measuring system performance. The model was shown to perform at least as well as an expert human pathologist and displayed diagnostic accuracy very close to the optimum possible for the domain. In this paper we present further results using a new ARTMAP configuration with a revised and expanded data set. The present work also overcomes some minor flaws in the methodology of our previous papers.

3 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 subsection 3.2). 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.

Additionally, input patterns are initially "preprocessed" by complement coding, such that each input feature also has its complement passed to the fuzzy ARTMAP input layer. (The fuzzy ARTMAP input layer is thus twice the length of the original input vector.) Complement coding is necessary to help reduce proliferation of category clusters in ART models. With purely binary data, complement coding ensures that the vector received by the fuzzy ARTMAP input layer always has a fixed number of true bits (equal to the length of the original input vector).

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 $\mathrm{ART_b}$, but instead encodes these patterns directly. (Thus, unlike fuzzy ARTMAP, the $\mathrm{ART_b}$ 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.

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 $\mathrm{ART_a}$ module, $\bar{\rho}_a$. This determines how close a match is required between an $\mathrm{ART_a}$ 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, 1995). 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 $\mathrm{ART_a}$ 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 the results we describe were achieved by means of 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. SFAM is therefore suitable for domains where the distribution of data items is skewed between different categories. This is demonstrated within the present application domain (see subsection 4.1).

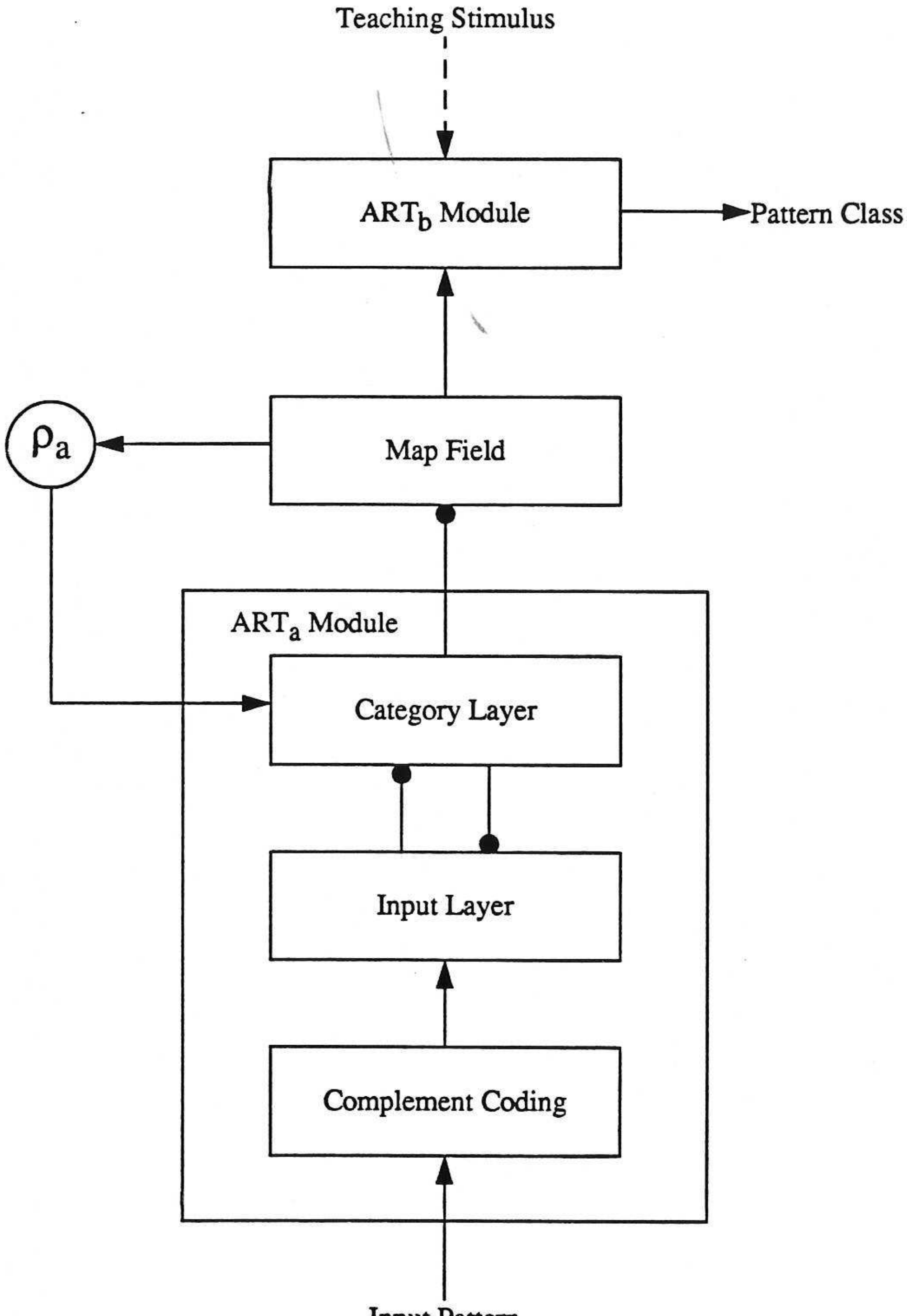

Figure 1: Simplified fuzzy ARTMAP

Additionally, there are a number of supplementary features of our approach which require some explication:

3.1 Voting Strategy

The formation of category clusters in 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 separate SFAM networks can lead to the formation of quite different clusters within the two networks. 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 input items 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 SFAM 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 SFAM performance in comparison with the individual networks.

3.2 Symbolic Rule Extraction

Most neural networks suffer from the opaqueness of their learned associations (Towell and Shavlik, 1993). In medical domains, this "black box" nature may make clinicians reluctant to utilise a neural network application, no matter how great the claims 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". The ARTMAP models have been endowed with such capabilities (Carpenter and Tan, 1993; Tan, 1994). The act of rule extraction 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).

3.3 Category Pruning

An SFAM network often becomes "over-specified" on the training set, generating many low-utility ART_a category clusters which may represent noise or 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 "preprocessing" stage of category pruning[2]. 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 upon a node's *usage* and *accuracy*. The usage score for an ART_a

[2] With continuously-valued category weights, rule extraction also involves a second preprocessing stage of *quantization* (see Carpenter and Tan, 1993). However, this application uses binary data under so-called fast-learn conditions (Carpenter et al.,1992) which yields purely binary category weights. Quantization is therefore omitted from this description.

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 independent of 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. In addition, it also has the very useful side-effect that a pruned network's performance is usually superior to the original, unpruned net on both the prediction set and on entirely novel test data. This is because the removal of the low-utility nodes caused by over-specification on the training set improves the general performance of the network.

In the original formulation of the pruning process, a uniform CF threshold is used to select nodes for deletion, irrespective of their category class. In this application, we have generalised the pruning process to allow separate CF thresholds for nodes belonging to different category classes. This allows us to vary the proportion of the state-space covered by different categories. 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 is useful for medical domains since it allows an ARTMAP network to be pruned so as to trade sensitivity for specificity and vice versa.

3.4 SFAM cascade

The generalization of the category pruning process described above allows a novel "cascaded" configuration of SFAM networks to be employed as shown in figure 2. This consists of three layers, a network pruned so as to maximize sensitivity, another pruned so as to maximize specificity, and a third layer consisting of a set of uniformly pruned networks (which maximize accuracy) operating by the voting strategy. (The network cascade described here is a slightly simplified version of that first presented in Downs, Harrison and Kennedy, 1995, which was used for the early diagnosis of heart attacks.)

The first two layers are intended to identify those cases which have a very high certainty of being classified correctly, with the sensitive network being used to "trap" the negative cases and the specific network capturing the positive cases. The intuition behind this is that a network which displays very high sensitivity will rarely make false negative predictions and so any negative predictions made by that network are very likely to be correct. Conversely, a highly specific network will make very few false positive predictions, and so its positive predictions have a high certainty of being correct.

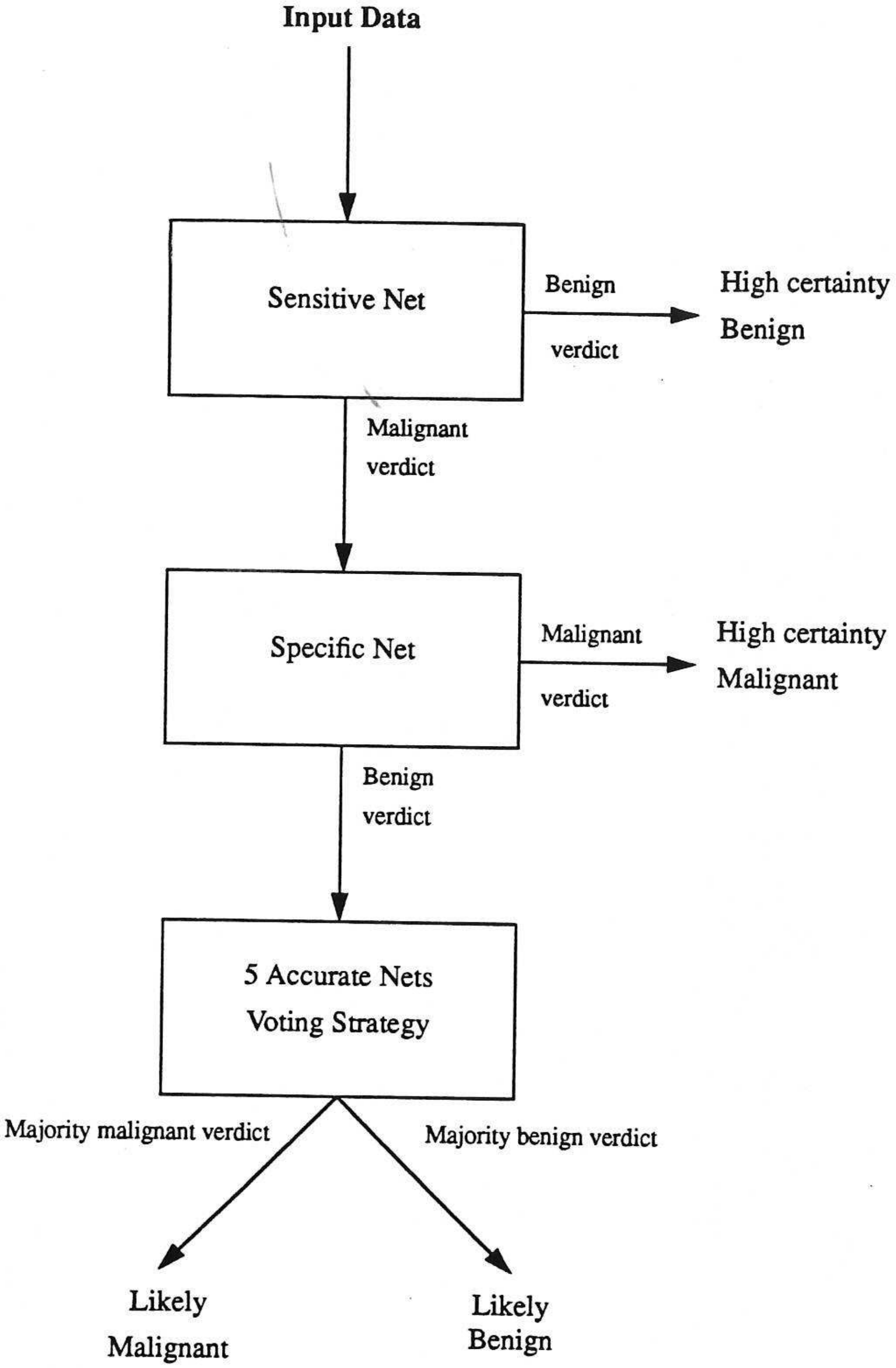

Figure 2: Cascaded SFAM configuration for breast cancer diagnosis

The cascade therefore operates as follows: An input data item is first presented to the sensitive network. If this yields a benign (negative) verdict, this is taken as the final category prediction. If not, the data item is next presented to the specific network. If this yields a malignant (positive) verdict, this is taken as the ultimate category prediction[3]. Otherwise the final prediction of the category class of the input is obtained by majority verdict from the uniformly pruned nets, with a lower certainty of the prediction being correct than with the previous two layers. The cascade mimics the behaviour of a human pathologist in this domain in that some diagnoses are reported as "certainly" correct, while others are labelled as "suspicious" (and will thus require a second opinion and/or further test samples).

4 Basic Findings

4.1 Data

The total data set for the application comprised 600 patient records each composed of 10 binary-valued features (see table 1). Each record was derived from microscopic observation of FNAB specimens 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 and screening-detected breast lesions in 1993. The distribution of categories within the data was slightly skewed but represents the prior probabilities of adequate specimens received in the laboratory—215 cases were malignant, the remaining 385 benign (i.e. 35.8% malignant, 64.2% benign). This data differs therefore from that used in our previous work (Downs, Harrison and Cross, 1995a, 1995b) in that the previous data set was artificially biased to represent approximately equal numbers of each outcome, rather than the actual distribution across outcomes within the domain. (Additionally, the extra data in the current work allows performance to be measured on a test set that is independent of the prediction data set used to guide category pruning.)

In many medical situations the final diagnosis or outcome is difficult to confirm without unnecessary invasive procedures or a long period of time has to elapse to allow further manifestations of the disease to appear. In FNAB the final outcome is relatively easy to confirm within a few months of the initial procedure. In this study the majority of cases where a malignant diagnosis was made on FNAB had further excision of tissue (e.g. lumpectomy or mastectomy) and the final diagnosis was made by histological examination which has a specificity and a sensitivity very close to 100%. In a few elderly patients the tumour was not removed and the patient was treated by radiotherapy and chemotherapy, in these cases the clinical diagnosis of malignancy was secure since all were large tumours with clinical features of malignancy such as invasion through the skin. A benign outcome was confirmed by benign clinical features recorded at the time of aspiration, mammographic examination and absence of subsequent malignant specimens from the same area of the breast.

It should also be noted that, as with almost all information gathered from a real-world medical domain, the data set possesses a degree of "noise". Specifically, some combinations of features do not always have the same outcome in every case (notably owing to flawed aspirations of malignant lesions). Analysis of the data set revealed the existence of 16 such states, which collectively account for 365 cases. Assuming that the most frequent outcome should always be chosen when an ambiguous feature-state occurs will result in 22 of these cases being misclassified. This represents 3.7% of the total data-set, and thus the maximum possible diagnostic accuracy with this data is 96.3%.

[3] Obviously, the order of presentation between the sensitive and specific networks is not crucial, although for efficiency reasons it is preferable to have the network which captures the largest number of cases as the first layer.

4.2 Method

The total data were partitioned into three subsets; 125 randomly selected items formed a *prediction set* of 44 malignant and 81 benign cases (i.e. 35.2% malignant, 64.8% benign), a further 100 randomly selected items formed a *test set* of 35 malignant and 65 benign cases, the remaining 375 items formed the *training set* of 136 malignant and 239 benign cases (i.e. 36.3% malignant and 63.7% benign).

Ten SFAM networks were then trained on different random orderings of the teaching data. Vigilance was set very high (0.9) during training in order to maximize classification performance. Vigilance was relaxed to 0.6 for all predictions to ensure that all cases were matched to an existing category cluster node (i.e. forced choice prediction). Performance on the prediction set was recorded for each network in order to calculate the accuracy ratings for each category node as a prerequisite to category pruning.

The ten trained networks were then pruned in three separate 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 malignant and benign nodes to produce pruned networks which maximized *sensitivity*. CF thresholds of 0.05 for malignant nodes and 0.95 for benign nodes were employed, the criterion for setting the CF thresholds being a performance on the prediction set such that no network made more than one false negative prediction. A similar procedure was then conducted to produce 10 networks which maximized *specificity*. CF thresholds of 0.65 for malignant and 0.5 for benign nodes were sufficient to yield no more than one false positive for all networks on the prediction set.

The cascade configuration shown in figure 2 was then derived from the pruned networks. Selection criteria were as follows: From the ten networks pruned for sensitivity, the one which had the highest specificity while maintaining 100% sensitivity on the prediction set was selected to form the first layer of the cascade. Similarly, from the 10 networks pruned for specificity, the one which had the highest sensitivity while maintaining 100% specificity on the prediction set was chosen to form the second layer of the cascade. Finally, from the uniformly pruned networks, the five with the highest individual accuracies on the prediction set were selected to form the third (voting strategy) layer of the cascade. Performance of the cascade was then measured on the test set.

4.3 Results

The performance of the network cascade on the test set is shown in table 1, alongside the performance of the human consultant on the same set of cases. It can be seen that the overall performance of the cascade is very similar to that of the human pathologist—in terms of numbers of cases, the cascade made two false positive predictions avoided by the consultant, but also made only three false negative predictions, compared with six made by the consultant. Additionally, the two false positive decisions were confined to the "suspicious" layer of the cascade. The overall accuracy of the cascade can also be seen to be very close to the maximum possible for the domain (see subsection 4.1).

The top two layers of the cascade (representing "certain" decisions) showed almost perfect performance, and covered a large proportion of the total test data (71%). Perfect performance was marred only by the occurrence of one false negative decision. Further examination of the details of this case revealed that the sample was very likely to have resulted from a flawed aspiration, which, as noted earlier, is an unavoidable feature of the domain.

It should also be noted that standard category pruning produced an mean reduction in network size of 70.3%; from a mean of 91 category nodes in the unpruned networks to an average of 27 nodes in the pruned networks.

Table 2: Performance of SFAM voting strategy cascade on 100 item test set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
"Certain" Decisions	98.6	96.7	100.0
"Suspicious" Decisions	86.2	60.0	91.6
Overall Performance	95.0	91.4	96.9
Consultant's Performance	94.0	82.9	100.0

5 Symbolic Rules

5.1 General Discussion

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 5.2. However, some discussion of their general nature is needed here since they differ somewhat from the production rules used in conventional decision support aids such as 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; 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 (Carpenter et al., 1992). 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.

5.2 SFAM Rules for the Domain

Rule extraction from the uniformly pruned networks yielded 44 distinct rules, 12 with benign outcomes (shown in table 3) and 32 with malignant outcomes (shown in table 4). Rules are ranked within each table by number of occurrences within the five networks and by the mean value of their certainty factor. The rules which the Consultant pathologist involved in this study did not consider to be in accord with standard diagnostic criteria are shown hatched in grey within the tables.

Table 3: Benign rules extracted from uniformly pruned SFAM networks

Rule 1. 5 Occurrences
Mean CF = 1.00
IF
NO_SYMPTOMS
THEN
BENIGN
	Rule 2. 5 Occurrences
	Mean CF = 0.57
	IF
	NAKED = TRUE
	THEN
	BENIGN
		Rule 3. 5 Occurrences
		Mean CF = 0.52
		IF
		FOAMY = TRUE
		NUCLEOLI = TRUE
		THEN
		BENIGN

Rule 4. 4 Occurrences
Mean CF = 0.53
IF
APOCRINE = TRUE
THEN
BENIGN
	Rule 5. 4 Occurrences
	Mean CF = 0.50
	IF
	PLEOMORPH = TRUE
	THEN
	BENIGN
		Rule 6. 4 Occurrences
		Mean CF = 0.50
		IF
		3D = TRUE
		FOAMY = TRUE
		THEN
		BENIGN

Table 3: Benign rules extracted from uniformly pruned SFAM networks

Rule 7. 3 Occurrences
Mean CF = 0.52
IF
FOAMY = TRUE
APOCRINE = TRUE
THEN
BENIGN
	Rule 8. 3 Occurrences
	Mean CF = 0.52
	IF
	PLEOMORPH = TRUE
	SIZE = TRUE
	APOCRINE = TRUE
	THEN
	BENIGN
		Rule 9. 2 Occurrences
		Mean CF = 0.77
		IF
		FOAMY = TRUE
		THEN
		BENIGN

Rule 10. 2 Occurrences
Mean CF = 0.53
IF
NAKED = TRUE
FOAMY = TRUE
THEN
BENIGN
	Rule 11. 2 Occurrences
	Mean CF = 0.52
	IF
	FOAMY = TRUE
	PLEOMORPH = TRUE
	SIZE =TRUE
	APOCRINE = TRUE
	THEN
	BENIGN
		Rule 12. 2 Occurrences
		Mean CF = 0.50
		IF
		3D = TRUE
		NAKED = TRUE
		NECROTIC = TRUE
		THEN
		BENIGN

Table 4: Malignant rules extracted from uniformly pruned SFAM networks

Rule 1. 5 Occurrences
Mean CF = 0.94
IF
3D = TRUE
NUCLEOLI = TRUE
PLEOMORPH = TRUE
SIZE = TRUE
THEN
MALIGNANT
	Rule 2. 5 Occurrences
	Mean CF = 0.60
	IF
	DYS = TRUE
	ICL = TRUE
	NUCLEOLI = TRUE
	PLEOMORPH = TRUE
	SIZE = TRUE
	NECROTIC = TRUE
	THEN
	MALIGNANT
		Rule 3. 5 Occurrences
		Mean CF = 0.59
		IF
		3D = TRUE
		NUCLEOLI = TRUE
		PLEOMORPH = TRUE
		SIZE = TRUE
		NECROTIC = TRUE
		THEN
		MALIGNANT

Rule 4. 4 Occurrences
Mean CF = 0.99
IF
ICL = TRUE
3D = TRUE
NUCLEOLI = TRUE
PLEOMORPH = TRUE
SIZE = TRUE
THEN
MALIGNANT
	Rule 5. 4 Occurrences
	Mean CF = 0.75
	IF
	DYS = TRUE
	NUCLEOLI = TRUE
	PLEOMORPH = TRUE
	SIZE = TRUE
	THEN
	MALIGNANT
		Rule 6. 4 Occurrences
		Mean CF = 0.73
		IF
		NUCLEOLI = TRUE
		PLEOMORPH = TRUE
		SIZE = TRUE
		THEN
		MALIGNANT

Table 4: Malignant rules extracted from uniformly pruned SFAM networks

Rule 7. 4 Occurrences
Mean CF = 0.64
IF
PLEOMORPH = TRUE
SIZE= TRUE
THEN
MALIGNANT
	Rule 8. 4 Occurrences
	Mean CF = 0.59
	IF
	ICL = TRUE
	3D = TRUE
	FOAMY = TRUE
	NUCLEOLI = TRUE
	PLEOMORPH = TRUE
	SIZE = TRUE
	NECROTIC = TRUE
	THEN
	MALIGNANT
		Rule 9. 4 Occurrences
		Mean CF = 0.58
		IF
		ICL = TRUE
		FOAMY = TRUE
		PLEOMORPH = TRUE
		SIZE = TRUE
		THEN
		MALIGNANT

Rule 10. 4 Occurrences
Mean CF = 0.58
IF
ICL = TRUE
NUCLEOLI = TRUE
PLEOMORPH = TRUE
SIZE = TRUE
THEN
MALIGNANT
	Rule 11. 4 Occurrences)
	Mean CF = 0.56
	IF
	DYS = TRUE
	3D = TRUE
	FOAMY = TRUE
	NUCLEOLI = TRUE
	PLEOMORPH = TRUE
	SIZE = TRUE
	NECROTIC = TRUE
	THEN
	MALIGNANT
		Rule 12. 4 Occurrences
		Mean CF = 0.55
		IF
		3D = TRUE
		NAKED = TRUE
		NUCLEOLI = TRUE
		PLEOMORPH = TRUE
		SIZE = TRUE
		THEN
		MALIGNANT

Rule 13. 3 Occurrences
Mean CF = 0.79
IF
3D = TRUE
FOAMY = TRUE
NUCLEOLI = TRUE
PLEOMORPH = TRUE
SIZE = TRUE
THEN
MALIGNANT
	Rule 14. 3 Occurrences
	Mean CF = 0.69
	IF
	FOAMY = TRUE
	NUCLEOLI = TRUE
	PLEOMORPH = TRUE
	SIZE = TRUE
	THEN
	MALIGNANT
		Rule 15. 3 Occurrences
		Mean CF = 0.68
		IF
		FOAMY = TRUE
		NUCLEOLI = TRUE
		PLEOMORPH = TRUE
		SIZE = TRUE
		NECROTIC = TRUE
		THEN
		MALIGNANT

Rule 16. 3 Occurrences
Mean CF = 0.62
IF
3D = TRUE
FOAMY = TRUE
PLEOMORPH = TRUE
SIZE = TRUE
NECROTIC = TRUE
THEN
MALIGNANT
	Rule 17. 3 Occurrences
	Mean CF = 0.60
	IF
	ICL = TRUE
	THEN
	MALIGNANT
		Rule 18. 3 Occurrences
		Mean CF = 0.57
		IF
		ICL = TRUE
		3D = TRUE
		NUCLEOLI = TRUE
		PLEOMORPH = TRUE
		SIZE = TRUE
		NECROTIC = TRUE
		THEN
		MALIGNANT

Table 4: Malignant rules extracted from uniformly pruned SFAM networks

Rule 19. 3 Occurrences
Mean CF = 0.53
IF
ICL = TRUE
NAKED = TRUE
FOAMY = TRUE
NUCLEOLI = TRUE
PLEOMORPH = TRUE
SIZE = TRUE
THEN
MALIGNANT
	Rule 20. 2 Occurrences
	Mean CF = 0.67
	IF
	ICL = TRUE
	3D = TRUE
	NUCLEOLI = TRUE
	THEN
	MALIGNANT
		Rule 21. 2 Occurrences
		Mean CF = 0.62
		IF
		DYS = TRUE
		ICL = TRUE
		NUCLEOLI = TRUE
		PLEOMORPH = TRUE
		SIZE= TRUE
		THEN
		MALIGNANT

Rule 22. 2 Occurrences
Mean CF = 0.62
IF
NUCLEOLI = TRUE
PLEOMORPH = TRUE
SIZE = TRUE
NECROTIC = TRUE
THEN
MALIGNANT
	Rule 23. 2 Occurrences
	Mean CF = 0.62
	IF
	ICL = TRUE
	FOAMY = TRUE
	NUCLEOLI = TRUE
	PLEOMORPH = TRUE
	SIZE = TRUE
	THEN
	MALIGNANT
		Rule 24. 2 Occurrences
		Mean CF = 0.60
		IF
		DYS = TRUE
		NUCLEOLI = TRUE
		PLEOMORPH = TRUE
		SIZE = TRUE
		NECROTIC = TRUE
		THEN
		MALIGNANT

Rule 25. 2 Occurrences
Mean CF = 0.57
IF
ICL = TRUE
PLEOMORPH = TRUE
SIZE = TRUE
THEN
MALIGNANT
	Rule 26. 2 Occurrences
	Mean CF = 0.59
	IF
	ICL = TRUE
	FOAMY = TRUE
	THEN
	MALIGNANT
		Rule 27. 2 Occurrences
		Mean CF = 0.55
		IF
		DYS = TRUE
		FOAMY = TRUE
		NUCLEOLI = TRUE
		PLEOMORPH = TRUE
		SIZE = TRUE
		NECROTIC = TRUE
		THEN
		MALIGNANT

Rule 28. 2 Occurrences
Mean CF = 0.53
IF
DYS = TRUE
ICL = TRUE
3D = TRUE
NUCLEOLI = TRUE
PLEOMORPH = TRUE
SIZE = TRUE
THEN
MALIGNANT
	Rule 29. 1 Occurrence
	Mean CF = 0.63
	IF
	SIZE = TRUE
	THEN
	MALIGNANT
		Rule 30. 1 Occurrence
		Mean CF = 0.59
		IF
		DYS = TRUE
		3D = TRUE
		NUCLEOLI = TRUE
		PLEOMORPH = TRUE
		SIZE = TRUE
		THEN
		MALIGNANT

Table 4: Malignant rules extracted from uniformly pruned SFAM networks

Rule 31. 1 Occurrence
Mean CF = 0.56
IF
ICL = TRUE
3D = TRUE
NAKED = TRUE
FOAMY = TRUE
NUCLEOLI = TRUE
SIZE = TRUE
THEN
MALIGNANT
	Rule 32. 1 Occurrence
	Mean CF = 0.53
	IF
	DYS = TRUE
	ICL = TRUE
	3D = TRUE
	NUCLEOLI = TRUE
	PLEOMORPH = TRUE
	SIZE = TRUE
	NECROTIC = TRUE
	THEN
	MALIGNANT	

Of the 44 rules extracted from the trained SFAM networks 39 were in complete agreement with published canonical lists (Wells et al., 1994). The majority of the rules for a malignant diagnosis consisted of combinations of the features ICL, 3D, NUCLEOLI, PLEOMORPH and SIZE with other features appearing less frequently. Most cytopathologists agree that the nuclear features of increased nuclear size (SIZE), multiple prominent nucleoli (NUCLEOLI) and variation in nuclear size (PLEOMORPH) are the most important diagnostic features of malignancy in FNAB (Bottles et al., 1988; Trott, 1991; Wells et al., 1994). Rule 17 for malignant diagnoses is interesting in that it only contains one feature, the presence of intracytoplasmic lumina; one specific type of breast carcinoma, lobular carcinoma, often produces cells which do not have prominent nuclear abnormalities but which do have these abnormal cytoplasmic inclusions (Quincey et al., 1991). Rule 29 for malignant diagnoses gives only the feature of increased nuclear size but most cytopathologists require other nuclear abnormalities to be present before making an unequivocally malignant diagnosis, with increased nuclear size alone a cytopathologist would probably assign the case to a suspicious category. In the benign rules the features PLEOMORPH and SIZE appeared together twice which taken alone would be suggestive of malignancy but in these rules they were in combination with the feature APOCRINE. Apocrine change in breast epithelial cells produces large cells with abundant cytoplasm but the nucleus is also enlarged which would produce positive observations of the features PLEOMORPH and SIZE without prominent multiple nucleoli (NUCLEOLI). Apocrine change could also have led to the production of benign rule 5 where PLEOMORPH is the only feature since a few cells with apocrine change in the specimen would trigger this positive observation without also triggering APOCRINE because the majority of epithelial cells did not show apocrine change. This circumstance arises because of the binary nature of the definitions used in this study, a coding scheme with more gradation for each feature could circumvent this problem. Benign rules 3, 6 and 12 are not in complete accord with published cytopathological knowledge and these cases would probably have been assigned to a suspicious category by a cytopathologist.

6 Further Results

This section presents results concerning investigations into the effects upon SFAM performance of employing both different input features and different observers to derive the features. It follows up issues raised in Downs, Harrison and Cross (1995b).

6.1 Effect of Different Input Features

In this subsection we consider three variations of the "standard" set of input features used to obtain the results described in section 4:

1) The removal of the "FOAMY" feature (representing the presence of foamy macrophages in the background of the slide.) In previous work (Downs, Harrison and Cross, 1995b) it was observed that there was disagreement as to the utility of this feature in making a diagnosis. This variation therefore attempts to resolve this issue by determining if removal of the "FOAMY" feature has an adverse effect upon network performance.

2) The addition of an "AGE" feature, representing the patient's age. (The age feature was encoded for input to SFAM by dividing the patient age in years by 100 in order to yield a value in the range 0–1.) The motive for including this feature is that after menopause the breast atrophies and thus the proportion of epithelial cells drops. Aspirates are expected therefore to show fewer and smaller epithelial cells. Hence, patient age is a potentially useful diagnostic indicator.

3) The addition of a binary "NOISE" feature, derived from whether or not the laboratory number of the sample was odd or even. This variation is intended to investigate the robustness of SFAM in the face of irrelevant data.

In order to compare the effect of different input features, the SFAM network configuration and experimental method were changed from the procedure described in subsection 4.2. This earlier method was intended to optimize SFAM performance for use of the model as a decision-support tool. This goal was achieved, with a diagnostic accuracy very close to the maximum possible for the domain. However, comparison of the effects of different input features requires a non-optimal baseline performance in order to detect any improvements caused by input feature variations. The following changes in the experimental method were therefore adopted, all for the purpose of lowering the baseline SFAM performance with the standard set of input features:

1) Category pruning was omitted. This also allowed an expanded test set to be created, by merging the prediction and test sets described in subsection 4.2 into a single 225 item test set.

2) Vigilance was set at zero during training of the networks. This change was further necessitated by the fact that a uniform non-zero vigilance level can have a slightly different effect upon two SFAM networks which have different sized input vectors (see Stork, 1989).

3) The full SFAM cascade was not employed in measuring performance. Instead a simpler three network voting strategy was used.

The full description of the experimental method is as follows:

Ten SFAM networks were trained on different random orderings of the 375 item teaching data set described in subsection 4.2, using single-epoch training with a vigilance level of zero. These networks used the 10 "baseline" standard input features described previously in table 1. Performance of each individual network was then recorded on the new 225 item test set. The three networks with the highest individual accuracy were then selected to form a three network voting strategy. The voting strategy performance of these networks was then recorded on the test set.

The procedure described in the above paragraph was then repeated using the three variations in the network input features outlined previously. SFAM results with the three input feature variations were then compared with the baseline results across all performance metrics (i.e. accuracy, sensitivity and specificity) using McNemar's test (see Bland, 1987).

The performance with the three input feature variations in comparison to the standard input feature set is shown in table 5. It can be seen that the deletion of the "FOAMY" feature led to a small drop in overall accuracy because of reduced specificity. However, none of the differences across the three metrics was statistically significant (McNemar's test, p>0.1). The addition of the "AGE" feature led to a increase across all performance metrics in comparison to the baseline. These increases were all statistically significant (McNemar's test; accuracy p<0.001, sensitivity p<0.05, specificity p<0.01.). The addition of the "NOISE" feature did not cause a deterioration in performance in comparison to the baseline. Indeed, it led to a slight increase in accuracy owing to higher specificity. However, neither difference was statistically significant (McNemar's test, p>0.1).

Table 5: Effect of different input features on SFAM performance

Input Features	Accuracy (%)	Sensitivity (%)	Specificity (%)
BASELINE 10	85.3	84.8	85.6
-FOAMY	82.2	87.3	79.5
+AGE	94.2	94.9	93.8
+NOISE	87.6	84.8	89.0

6.2 Effect of Different Observers

In previous work (Downs, Harrison and Cross, 1995b) performance of the SFAM network appeared to degrade badly in the face of "noisy" input data provided by an inexperienced junior pathologist. It is thus possible that SFAM's utility as a decision-support tool is limited by the quality of the input data, being suitable for use by senior pathologists but vulnerable in the face of incorrect feature assignments made by junior pathologists. In this subsection therefore, this issue is investigated further by studying the performance of SFAM using input feature observations provided by several different pathologists.

A test set of 50 cases (25 malignant and 25 benign) was employed which was entirely independent of the 600 item data set described previously in subsection 4.1. The presence or absence of the ten standard input features (see table 1) for each case was then rated by six different pathologists, all working separately and with no knowledge of clinical data apart from patient age. One of these pathologists had provided the original 600 item data set. Of the six pathologists, two were of consultant status (each with over 10 years experience of interpreting FNAB), three were senior registrars (each with 6 years experience) and one a senior house officer (2 years experience). All pathologists reported FNAB in their daily work but those below consultant status would have had their reports scrutinized by a consultant pathologist who might modify the report after examining the cytological preparations.

Voting strategy performance of the five uniformly pruned networks trained previously to form the third layer of the SFAM cascade (see section 4) was recorded for each of the six pathologist's observations. The results with the observations from the pathologist who had provided the original training data served as the baseline for comparison with the other 5 pathologists' observations (observer 1 in table 6). The results for all three performance metrics were compared using McNemar's test.

The performance of SFAM with the six sets of observations is shown in table 6. It should be noted that the small size of the data set used here means that the differences between observers may appear somewhat exaggerated since, in percentage terms, a single false case alters specificity or sensitivity by 4% and accuracy by 2%.

General performance of the SFAM networks held up well across all observers, with no more than four misclassifications for all but one observer's data (seven misclassifications with observer 4). Somewhat surprisingly, the networks performed slightly better than the baseline with two observers (numbers 3 and 5). Furthermore, the observation set showing the worst performance had only four extra misclassifications in comparison to the baseline. There were no statistically significant differences between the baseline observation set and any of the other observation sets across any performance metric (McNemar's test, p>0.1).

Table 6: Effect of different observers on uniformly pruned SFAM performance

Observer	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	94	100	88
2	92	92	92
3	96	100	92
4	86	84	88
5	96	100	92
6	94	92	96

7 Discussion

The performance of the SFAM cascade described in subsection 4.3 indicates that the architecture has useful potential value as a decision-support tool for this domain. Performance is very close to the domain optimum, despite only utilizing single-epoch training on a skewed data set. Furthermore, the cascade served to isolate a large proportion of cases where the network's predictions were very likely to be correct, and, very importantly, avoided any false positives within this subset of the test data.

Additionally, the model was shown to have self-discovered a generally valid set of rules for the domain, with only five rules from a set of 44 considered not to be "canonical". It is interesting to note that all but one of these five rules had a mean CF factor very close to the threshold for pruning. We therefore believe it is possible that the networks were "under-pruned" in this application and that a slightly higher CF threshold for pruning (e.g. 0.55) would have almost entirely eliminated all dubious rules, as well as providing a more compact overall rule set. However, balanced against this is the need to avoid "over-pruning" where useful category nodes are deleted resulting in degraded system performance and incomplete coverage of the state space for the domain. This is therefore an area for future work. A further interesting related piece of future work might be to use expert knowledge as a "post-processor" for pruning, by deleting only those rules above the standard pruning threshold which are considered to be invalid by the domain expert, and recording any subsequent effects upon system performance.

The rule discovery aspect of SFAM can be seen as a *knowledge engineering* facet of the model—validating that a network behaves in a way that is acceptable to a domain expert in order to enhance confidence in the use of the model as a decision-support tool. Conversely, our experiments with variations in the input features illustrate a *machine learning* facet—establishing the utility of different diagnostic features which may be at odds with the "received wisdom" of domain experts. Specifically our results indicate that patient age should be regarded as a useful diagnostic feature, whereas the presence of foamy macrophages is not perhaps as useful as is conventionally claimed. Furthermore, the SFAM was shown to be robust in the face of an irrelevant input feature which did not degrade system performance.(See also Goodman et al., 1994, for another experiment concerning fuzzy ARTMAP's noise resistance in a different medical domain—prediction of the length of hospital stay of pneumonia patients.)

Our findings with different observers providing the network inputs also showed the SFAM network to be quite robust across potential variations caused by subjectivity in feature assignments made by different pathologists. This is somewhat at odds with our previous findings (Downs, Harrison and Cross, 1995b). The previous study, which showed a marked decrease in performance with an inexperienced observer, only used two observers and the inexperienced observer had 18 months experience compared with a minimum of 2 years in this study. This is therefore an area where more research is needed in order to establish the limits of and reasons for SFAM's apparently variable robustness, probably requiring a larger volume of data providing feature ratings from a variety of pathologists of different levels of experience. The aptitude of individuals for pattern recognition and visual discrimination tasks varies and larger numbers of inexperienced observers are required before any firm conclusions about the system's performance with very junior pathologists can be made. Unfortunately, the collection of data for such a study is a non-trivial task. Evidently, an objective method for feature extraction would enhance the utility of the current system and this remains an area for future work.

A problem with studies where human observers are used to make observations which are then used in a decision-support system is the separation of observational and interpretative processes. In the cytodiagnosis of FNAB in a routine (non-teaching) setting, pathologists rarely express their decisions in component observations and the diagnosis is made by a process which is largely subconscious and is the result of training the natural neural networks of the human brain by observation of past specimens and knowledge of their diagnosis/outcome. Since the pathologists in this study knew the "canonical" lists of features for the diagnosis of FNAB it is possible that their observations may have been biased if they made a diagnosis of the specimen whilst recording the observed features. The presence of 5 extracted rules which do not correspond with the "canonical" lists is evidence against this as is the evidence that the SFAM voting strategy cascade produced a better overall performance than the human who made the observations which the SFAMs used. The separation between observations and diagnosis could be improved by selecting microscopic fields of view in specimens, digitizing them and then displaying them to observers in random order so that fields from the same specimen were not in an unbroken sequence, but this would require numerous images for each specimen (since some of the definitions are based on the presence or absence of a feature in all cells of the specimen) and the resolution of the best digitized images is still not as great as those viewed directly through the microscope.

In this study verbal definitions of observed features were used and the observers were just given these with no additional training and no sample images. The reliability of observations, especially among trainee pathologists, could be improved by a more visually-based input system such as a range of sample images for each feature displayed on a computer screen with a cursor controlled input scale (Hamilton et al., 1995).

We are also considering the development of a modified version of 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. This should increase the robustness of an SFAM-based decision-support tool when used by an inexperienced pathologist, by effectively building a priori expert knowledge into SFAM.

Another possible area for future work is to automate the CF threshold selection process for the differential category pruning described in subsection 3.3. 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 is a rather laborious trial-and-error process which contrasts poorly with the general ease of tuning of the basic SFAM model.

Finally, an important area for future work not so far addressed is to investigate the claimed potential for incremental learning with the SFAM model. The relatively small size of the data set did not allow such a study to be made in this application. However, in theory, were further data for the domain to become available, the existing networks could be trained on this new data without the necessity also to retrain with the original data. It should be noted though that there are a few caveats to this statement. Specifically, the voting strategy and category pruning are essentially "off-line" learning processes which cannot be employed in conjunction with continuous (case-by-case) learning. The voting strategy requires randomization of the ordering of input data which obviously disrupts its original temporal order. Similarly, the category pruning process requires a "batch" of input data to form the prediction set.

However, these features can still be feasibly employed in the less stringent circumstance of incremental learning on a batch of new data, rather than on-line learning on a case-by-case basis. For example, suppose a further 125 of the most recent hospital cases were to become available for the domain. This could form a new prediction set and the existing 125 item prediction set would be freed to serve as further training data. This new training data could then be randomized and applied to the ten existing unpruned networks. Category pruning could then occur on the basis of performance on the new prediction set. This would yield new pruned networks to be employed in the SFAM cascade which should be adapted to the most recent data while still retaining useful older associations, without having to have undertaken extensive retraining from scratch.

Acknowledgement

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

References

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

Baxt, W.G. (1995) Application of Artificial Neural Networks to Clinical Medicine, *Lancet*, 346, 1135–1138.

Bland, M. (1987) *An Introduction to Medical Statistics*.
Oxford: Oxford University Press.

Bottles, K., Chan, J.S., Holly, E.A., Chiu, S. and Miller, T.R. (1988) Cytologic Criteria for Fibroadenoma, *American Journal of Clinical Pathology*, 89, 707–713.

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.

Cross, S.S., Harrison, R.F. and Kennedy, R.L. (1995) Introduction to Neural Networks, *Lancet*, 346, 1075–1079.

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. (1995a) 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. (1995b) Evaluating a Neural Network Decision Support System for the Diagnosis of Breast Cancer, in P. Barahona, M. Stefanelli and J. Wyatt, eds., *Proceedings of the 5th Conference on Artificial Intelligence in Medicine Europe* (AIME'95), 239–250.
Berlin: Springer-Verlag.

Downs, J., Harrison, R.F. and Kennedy. R.L. (1995) A Prototype Neural Network Decision Support Tool for the Early Diagnosis of Acute Myocardial Infarction, in P. Barahona, M. Stefanelli and J. Wyatt, eds., *Proceedings of the 5th Conference on Artificial Intelligence in Medicine Europe* (AIME'95), 355–366.
Berlin: Springer-Verlag.

Downs, J., Harrison, R.F., Kennedy, R.L. and Cross, S.S. (In Press) Application of the Fuzzy ARTMAP Neural Network Model to Medical Pattern Classification Tasks, to appear in *Artificial Intelligence in Medicine.*

Dybowski, R. and Gant, V. (1995) Artificial Neural Networks in Pathology and Medical Laboratories, *Lancet*, 346, 1203–1207.

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.

Elston, C.W. and Ellis, I.O. (1990) Pathology and Breast Screening, *Histopathology*, 16, 109–118.

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 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.

Hamilton, P.W., Anderson, N., Bartels, P.H. and Thompson, D. (1994) Expert System Support Using Bayesian Belief Networks in the Diagnosis of Fine Needle Aspiration Biopsy Specimens of the Breast, *Journal of Clinical Pathology*, 47, 329–336.

Hamilton, P.W., Bartels, P.H., Montironi, R., Anderson, N. and Thompson, D. (1995) Improved Diagnostic Decision-Making in Pathology: Do Inference Networks Hold the Key? *Journal of Pathology*, 175, 1–6.

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

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

Heathfield, H.A., Kirkham, N., Ellis, I.O. and Winstanley, G. (1990) Computer Assisted Diagnosis of Fine Needle Aspirate of the Breast, *Journal of Clinical Pathology*, 43, 168–170.

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.

Ma, Z. and Harrison, R.F. (1995) 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. (1995) A Modified Fuzzy ARTMAP Architecture for the Approximation of Noisy Mappings, *Neural Networks*, 8(4), 619–641.

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.

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.

Quincey, C., Raitt, N., Bell, J., and Ellis, I.O. (1991) Intracytoplasmic Lumina—A Useful Diagnostic Feature of Adenocarcinomas, *Histopathology*, 19, 83–87.

Richard, M.D. and Lippmann, R.P. (1991) Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities, *Neural Computation*, 3, 461–483.

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.

Start, R.D., Silcocks, P.B., Cross, S.S. and Smith, J.H.F. (1992) Problems with Audit of a New Fine-Needle Aspiration Service in a District General Hospital, *Journal of Pathology*, 167, 141A.

Stork, D.G. (1989) Self-Organization, Pattern Recognition, and Adaptive Resonance Networks, *Journal of Neural Network Computing*, 1(1), 26–42.

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.

Towell, G. and Shavlik, J.W. (1993) Extracting Refined Rules from Knowledge-Based Neural Networks, *Machine Learning*, 13(1), 71–101.

Trott, P.A. (1991) Aspiration Cytodiagnosis of the Breast, *Diagnostic Oncology*, 1, 79–87.

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.

Wolberg, W.H. and Mangasarian, O.L. (1993) Computer-Designed Expert Systems for Breast Cytology Diagnosis, *Analytical and Quantitative Cytology and Histology*, 15, 67–74.