Monograph:
Downs, J., Harrison, R.F. and Cross, S. (1994) Evaluating a Neural Network Decision-Support Tool for the Diagnosis of Breast Cancer. Research Report. ACSE Research Report 553 . Department of Automatic Control and Systems Engineering

---

Evaluating a Neural Network Decision-Support Tool for the Diagnosis of Breast Cancer

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

Research Report #553
29 November

Abstract

This paper describes the evaluation of an application of the ARTMAP neural network model to the diagnosis of cancer from fine-needle aspirates of the breast. The network has previously demonstrated very high performance when used with high-quality data provided by an expert pathologist. New performance results are provided for its use with "noisy" data provided by an inexperienced pathologist. Additionally, ARTMAP supports the extraction of symbolic rules from a trained network and the validity of these autonomously-acquired rules is discussed. It is concluded that the symbolic rules provide an appropriate mapping of input features to category classes in the domain. However, the network in its present form is only suitable for use as a decision-support tool by a senior pathologist, since its performance deteriorated greatly with poor-quality data provided by a junior pathologist. The implications of the findings are discussed.

Correspondence Address

R.F. Harrison
Department of Automatic Control and Systems Engineering
The University of Sheffield
PO Box 600, Mappin Street
Sheffield, S1 4DU
United Kingdom.

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

1 Introduction

Carcinoma of the breast is a common disease which is diagnosed in about 22 000 women in England and Wales each year and is the commonest cause of death amongst women in the 35–55 years age group in the same population (Underwood, 1992). Early detection and treatment gives a better prognosis and a Breast Screening Program has been introduced in the National Health Service using mammography as the primary detection modality. The primary method of diagnosis of breast carcinoma, with distinction from benign lesions causing mammographic abnormalities or clinically-detected masses, is cytopathological examination of fine needle aspirates of the breast, FNAB, (Elston and Ellis, 1990). In this technique cells are aspirated from the breast lesion using a fine gauge needle attached to a plastic syringe, the aspirated cells are either applied directly to a glass slide or put into transport medium and a cytological preparation is produced. The cytology slide is examined by light microscopy by a medically-qualified doctor who has been trained in cytopathology and a diagnosis is made. FNAB is more acceptable to the patient than a needle core biopsy or an open tissue biopsy since it involves much less damage to tissue with less pain. It is also cheaper and more rapid than tissue biopsy methods. Interpretation of the specimen is more difficult for the pathologist than tissue biopsies because the architecture of the lesion is lost and important information, such as invasion at the boundary of a tumour, is therefore not available. 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). The most important performance parameter is the specificity, since a malignant diagnosis on FNAB, combined with clinical assessment, will be the sole diagnostic step before definitive treatment such as mastectomy or wide local excision of the lesion and a false positive result may lead to unnecessary surgery. The acquisition of diagnostic expertise is a relatively slow process in pathology with at least five years study and experience of pathology required before medically-qualified doctors in Britain are allowed to sit the final professional pathology examinations (Membership of the Royal College of Pathologists). During this period, trainee pathologists are supervised by fully-qualified colleagues but it would be expected that their performance without supervision in FNAB cytodiagnosis would fall well below the figures in published studies. There is thus scope for an artificial intelligence decision-making tool in the cytodiagnosis of breast FNAB to assist in training junior pathologists and to improve the performance of experienced pathologists.

1.1 Background

The process of human diagnosis in pathology is not fully characterized but has been divided into pattern recognition and heuristic algorithms (Underwood, 1987). In pattern recognition an image is observed and compared with memories of past observations, the pathologist then makes decisions as to whether the present image is sufficiently similar to past observations of a particular diagnostic category to be assigned to that category. This process of decision-making may take a very short period of time with no awareness of the process if the lesion falls into a well-recognised diagnostic grouping, such as basal cell carcinoma of the skin. Heuristic algorithms may be applied to lesions which are not classified by pattern recognition and usually consist of branching trees with mutually exclusive branching points. Disadvantages of these are that a wrong branch selected near the top of the tree will lead to a lesion being misclassified

into a diagnostic group very distant from the actual diagnosis and that some lesions have features which lie on the borderline between criteria for the branching points. In the cytodiagnosis of FNAB there are some observable features which are cited as being important in the recognition of malignant cells, such features include variation in size of the epithelial cell nuclei (pleomorphism), increase in the size of epithelial cell nuclei, prominent multiple nucleoli within epithelial cell nuclei and intracytoplasmic lumina (Bottles et al., 1988; Trott, 1991; Quincey et al., 1991). However these publications do not attribute weights to these features or indicate the significance of combinations of these features.

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) using the observed features of bare nuclei, cellularity, cohesion, pleomorphism, cell arrangement, nuclear size, nucleoli, intracytoplasmic lumina, apocrine cells and mucinous background. 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 but some cases (such as cystosarcomas and cancer judged to have been missed by the aspirating needle) were excluded before analysis.

In previous work by the authors (Downs, Harrison and Cross, 1994) we applied a powerful, but little-known, neural network model (termed ARTMAP) to this task. 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, 1994). The model was shown to perform at least as well as an expert human pathologist. However, these results were achieved using the high-quality feature assignments provided by the human expert. Less experienced pathologists are more likely to make incorrect feature assignments and thus provide "noisier" input data to the model. This paper provides performance results for ARTMAP under such conditions. Additionally, ARTMAP possesses symbolic rule extraction capabilities which support the validation and justification of its diagnostic predictions. A detailed discussion of the ARTMAP rules used in this domain is therefore provided.

The structure of the remainder of this paper is as follows therefore. Section 2 describes ARTMAP and justifies the selection of this particular model for the task. Section 3 describes the data used in the study, and the trials performed with ARTMAP. Section 4 details the results. Section 5 describes and evaluates the symbolic rules extracted from ARTMAP that are used to make diagnosis decisions. Section 6 discusses the findings and suggests directions for further research.

2 ARTMAP

2.1 Motivation for use of ARTMAP

Advances in neurocomputing have opened the way for the establishment of decision support systems which are able to learn complex associations by example. The main thrust of work in this area has been in the use of the so-called feedforward networks to learn the association between evidence and outcome. Examples of such networks include the MultiLayer Perceptron, MLP, (Rumelhart, Hinton and Williams, 1986) or the Radial Basis Function networks, RBFN (Moody and Darken, 1989).

The MLP or the RBFN have been shown to be rich enough in structure so as to be able to approximate any (sufficiently smooth) function with arbitrary accuracy (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. This is an attractive scenario and has been extensively exploited in medical diagnosis.

A common criticism of the neural network approach is that the rules governing the predicted outcome are obscure, leading to a strong resistance to acceptance amongst potential users who wish to be convinced that the underlying model captures the salient features of the domain and is able to offer an explanation of its diagnosis in terms understood by the user. Attempts to extract domain rules from feedforward networks have met with limited success, with, so far, no completely general method published (Towell and Shavlik, 1993; Ma and Harrison, 1994).

The feedback architecture, ARTMAP (Carpenter, Grossberg and Reynolds, 1991), possesses a number of attractive features not found in feedforward networks such as: a dynamic architecture which "designs" itself; the ability to distinguish rare from frequent events, and more recently it has been demonstrated that, in a modified form, it can classify data optimally, in a Bayesian sense (Lim and Harrison, In Press).

For full details of the advantages provided by ARTMAP for medical domains generally see Harrison, Lim and Kennedy (1994) and Downs, Harrison and Kennedy (1994). However, in this work, ARTMAP was selected primarily for two reasons. First, it has been demonstrated to provide superior performance to both statistical and rival neural network approaches. With the same data used in Downs, Harrison and Cross (1994), logistic regression achieved an accuracy of 92%, sensitivity of 90%, and specificity of 94%; a MLP had accuracy, sensitivity and specificity of 92% (Cross et al., In Press). In comparison (see section 1.1), ARTMAP showed both superior sensitivity and specificity. Second, ARTMAP provides explicit symbolic rules which can be easily understood by a human user. This capability will be discussed in detail within this paper (see section 2.2.2 and section 5).

2.2 Overview of ARTMAP

ARTMAP (Carpenter, Grossberg and Reynolds, 1991) is a self-organizing, supervised learning, neural network model for the classification of binary patterns. It is one of a series of models based upon Adaptive Resonance Theory, or ART, (Carpenter and Grossberg, 1991) an outgrowth of competitive learning which overcomes the stability problems of that paradigm

(Grossberg, 1987). This is 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 the input. It should also be noted that most ART models, including ARTMAP, employ a localist representation for category nodes owing to the so-called "winner-take-all" competitive learning dynamics.

ARTMAP itself consists of three modules, two ART 1 systems (Carpenter and Grossberg, 1987) termed ART_a and ART_b, and a related structure termed 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. ARTMAP does not directly associate inputs at ART_a and ART_b. Instead, such patterns are first self-organized into prototypical category clusters before being associated at the map field. Hence generalized associations are formed.

Training in ARTMAP almost always results in multiple category clusters forming at ART_a for each teaching category present at ART_b. Each such ART_a cluster thus represents a significant sub-region of the overall state space covered by a particular teaching category. It can be seen therefore that ARTMAP instantiates a many-to-one mapping between ART_a input patterns and their actual classification.

For the purposes of this paper, three further features of ARTMAP are of particular note, the *voting strategy*, *symbolic rule extraction* and *category pruning*. These are described in detail next.

2.2.1 Voting Strategy

The formation of category clusters in ARTMAP 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 ARTMAP networks can lead to the formation of quite different clusters within the two 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 "wide" 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 ARTMAP networks are trained on different orderings of the training data. During testing, each individual network makes its prediction for a test item in the normal way. The number of predictions made for each category is then totalled and the one with the highest score (or the most "votes") is the final predicted category outcome. The voting strategy can provide improved ARTMAP 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.

2.2.2 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". ARTMAP has recently been endowed with such capabilities (Carpenter and Tan, 1993; Tan, 1994). The act of rule extraction is a straightforward procedure in ARTMAP compared with that required for feedforward networks since there are no hidden units with implicit meaning. In essence, each category cluster in ART_a represents a symbolic rule whose antecedent is the category prototype weights and whose consequent is the associated ART_b category (denoted via the map field).

2.2.3 Category Pruning

An ARTMAP 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. To overcome this difficulty, rule extraction involves a "preprocessing" stage of category pruning[1]. This involves the deletion of these low utility nodes. Pruning is guided by the calculation of a *confidence factor* (CF) between nought and one for each category cluster, based equally upon a node's usage (proportion of training set exemplars it encodes) and accuracy (proportion of correct predictions it makes on a separate prediction set). All nodes with a confidence factor below a user-set threshold are then 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.

3 Patients and Methods

3.1 Study Population

The total data set composed cytological specimens from 413 FNAB prepared by a cytocentrifuge method and stained by the Papanicolaou method. (Dundas et al., 1988) The final outcome of benign disease or malignancy was confirmed by open biopsy where that result was available. In benign aspirates with no subsequent open biopsy a benign outcome was assessed by clinical details on the request form, mammographic findings (where available) and by absence of further malignant specimens. A malignant outcome was confirmed by histology of open biopsy or clinical details where the primary treatment modality was chemotherapy or hormonal therapy. Idiosyncratic cases were not removed prior to use with the neural network.

3.2 Human Observations

Ten observable features were defined to give a binary value. The features were cellular dyhesion, intracytoplasmic lumina, "three-dimensionality" of epithelial cell clusters, bipolar "naked" nuclei, "foamy" macrophages, multiple prominent nucleoli, nuclear pleomorphism, nuclear size, necrotic epithelial cells and apocrine change. The precise definitions of these

[1] With continuously-valued category weights, rule extraction also involves a second preprocessing stage of *quantization* (see Carpenter and Tan, 1993). However, we prefer to use binary data under so-called fast-learn conditions (Carpenter et al.,1992) which yields purely binary category weights and subsequently provides rules of greater clarity. Quantization is therefore omitted in this application.

features are given in Appendix 1, together with their abbreviated names used for symbolic rule extraction. The observations on the specimens were made independently by a senior pathologist with 10 years experience of interpreting FNAB and a junior pathologist with 18 months experience. The observations were made blind to clinical details or outcome and the pathologists recorded their diagnosis for each case. The interobserver agreement between the two pathologists was assessed using kappa statistics (Silcocks, 1983).

3.3 Method

Ten ARTMAP networks had been trained previously on 313 data items using the senior pathologists feature assignments. A severely pruned version of each network had also been derived using the remaining 100 items as a prediction set and a CF threshold for pruning of 0.7. (Downs, Harrison and Cross, 1994, gives full details.)

An independent test set was derived from the junior pathologist's feature assignments for 82 randomly selected malignant cases and 82 benign cases. Performance results on this test set were recorded for each individual pruned and unpruned network, as well as for the voting strategy using five unpruned nets and five pruned nets. ART_a baseline vigilance for testing was set to 0.6 for the unpruned nets and 0.5 for the pruned nets ensuring forced choice prediction. This closely replicated the original test procedure which had used data from the senior pathologist (Downs, Harrison and Cross, 1994).

4 Results

Table 1 below shows the junior pathologist's performance on the test set in comparison with that of the various ARTMAP networks. (Full details of the individual ARTMAP networks' performance are given in tables 2 and 3 in Appendix 2).

Table 1: Relative Performance of Junior Pathologist and Network Types

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

In Downs, Harrison and Cross (At Review) it was observed that pruning had the effect of biasing network performance towards increased specificity (an essential requirement for the domain, see section 1), 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. 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 between the senior and junior pathologists, show that for most of the features there was only a moderate level of agreement (about 0.40 for the raw kappa scores, see table 2) and this lack of agreement will be the cause of the reduction in network performance when using the junior pathologist's data. Three of the features, "naked" nuclei, "foamy" macrophages and necrotic epithelial cells, had low levels of agreement that were little better than that expected by chance. Each of these features require a high level of interpretation by the pathologist to identify the cell type ("naked", "foamy" or epithelial) and to assess its apparent biological viability or non-viability at the time of sampling (necrosis). The feature with the highest level of agreement, nuclear size, had the clearest definition requiring least interpretation - the observer simply had to assess whether any epithelial cell nuclei had diameters greater than twice that of adjacent lymphocyte nuclei. The other two features which are prominent in the extracted ARTMAP rules (see section 6) are multiple nucleoli and nuclear pleomorphism and both had reasonable levels of agreement. The second column in table 2 gives the ratio of the raw kappa statistic to the maximum possible kappa value in each particular confusion matrix, providing correction for uneven marginals. However, it should be noted that the uneven marginals are themselves caused by lack of agreement between observers so the first column may be the best reflection of interobserver agreement.

Table 2: Kappa statistics for the confusion matrices of observations of each feature by a senior and junior pathologist

Feature	Kappa Statistic	Ratio Kappa/Kappa Max
Dyshesion	0.43	0.50
Intracytoplasmic lumina	0.34	0.60
"3D" epithelial cell clusters	0.38	0.40
"Naked" nuclei	0.15	0.38
"Foamy" macrophages	0.18	0.56
Multiple nucleoli	0.49	0.69
Nuclear pleomorphism	0.40	0.71
Nuclear size	0.55	0.71
Necrotic epithelial cells	0.16	0.21
Apocrine change	0.44	0.99

5 Symbolic Rules

As mentioned previously in section 2.2, the ability to extract symbolic rules from neural networks is an important enhancement to their use as decision-support tools in medical domains. Such symbolic rules provide two advantages which, taken collectively, should help to overcome reluctance to 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 ARTMAP this corresponds to displaying the equivalent rule for the ART_a cluster node that was activated to provide a category decision. (In the case of the voting strategy, a number of such rules, one per voting network, would be displayed.) The diagnosing clinician is then able to decide whether or not to concur with the network's prediction, based upon how valid they believe that rule to be.

Before discussing the specific rules discovered by ARTMAP for this domain, some discussion of the general nature of the rules is needed. These are of a somewhat different nature from those found in conventional rule-based 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 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, ARTMAP rules are self-discovered though exposure to domain exemplars, rather than having been externally provided by a human expert. ARTMAP is thus able to bypass the difficult and time-consuming knowledge-acquisition process found with rule-based expert systems (Hayes-Roth, Waterman and Lenat, 1983)[2]. A drawback of this approach is that the rules are "correlational" rather than causal, since ARTMAP possesses no underlying theory of the domain but simply associates conjunctions of input features with category classes. (Of course, this problem is not specific to ARTMAP but occurs with neural networks generally.) 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 ARTMAP which governs the formation of category clusters, and hence the rules that will be derived from these clusters. Under the "fast-learning" conditions 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.

Rule extraction from the 10 pruned nets used in this domain yielded 14 distinct rules, 12 for malignant outcomes and 2 for benign. The full list of rules is shown in table 3, ranked by how many of the 10 pruned networks each rule occurred in. No single rule in the set should be taken as canonical, since each is derived from a node which covers only a portion (albeit an important one) of the overall state space covered by each diagnostic category. However, taking the rules

[2] However, collection of the data may itself be a non-trivial task in many medical domains.

as a whole, a picture of a typical benign or malignant case can be constructed.

Table 3: Symbolic Rules Extracted from Pruned ARTMAP 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
ICL=TRUE
NUCLEOLI=TRUE
PLEOMORPH=TRUE
SIZE=TRUE
THEN
MALIGNANT

Rule 13 (1 Occurrence)
IF
FOAMY=TRUE
NUCLEOLI=TRUE
PLEOMORPH=TRUE
SIZE=TRUE
NECROTIC=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 ARTMAP 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 ARTMAP 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 ARTMAP based equally upon both usage and accuracy (see section 2.3). 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. 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 original, 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.

In future work this anomaly might be corrected by using a different weighting for the CF calculation, so as to bias the overall CF score more towards accuracy than usage. However, this has the risk that the resultant networks will possess incomplete coverage of all possible cases in the domain owing to the absence of high usage nodes encoding general cases.

The status of the FOAMY feature is more problematic. 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 ARTMAP 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. (Otherwise, the feature would usually be deleted from the node weights by the learning rule during training.) If the relative frequency of occurrence is considered, the 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 of the 17 benign rules and 16 of the 39 malignant rules. By either calculation the proportions are 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 conclusion may be tested in future work by training new networks which omit the FOAMY feature from the inputs and observing whether performance is subsequently degraded.

6 Discussion

The findings in section 4 indicate that although the existing ARTMAP application should prove

useful as decision-support tool for senior pathologists (Downs, Harrison and Cross, 1994), its performance is inadequate with poor-quality input data provided by a junior pathologist. The results further suggest that initial feature identification rather than subsequent diagnostic decision-making is the key criterion which distinguishes expert and neophyte performance in this domain. If this hypothesis is verified by further research, it obviously has implications for the training of junior pathologists in this field.

Further studies are required with a larger number of pathologists to evaluate the levels of agreement in identification of the observed features used by the network. Sets of cases should be selected for each feature which, in the opinion of an experienced pathologist, have the feature present in 50% of the cases. This means that kappa statistics give a more realistic view of agreement on each feature since the interlinkage that occurs when using one set for all features is lost and features which are rare in unbiased data sets are less likely to produce uneven marginals in the confusion matrices. The junior pathologist in this study was given a brief training session in the identification of the features using a microscope which allows simultaneous viewing of a slide by two observers but they did not have reference material available during their coding of the features which was spread over several weeks. Photographic examples of each feature could be provided to be used in visual comparison during the coding of features and this might produce less noisy data.

It would of course also be desirable if the network could be modified so as to improve performance on the "noisy" data provided by the junior pathologist. However, as yet we have no substantial ideas as to how this might be achieved, if indeed it can.

The symbolic rule extraction process described in section 5 provides more positive immediate results. ARTMAP has been shown to have acquired autonomously a valid mapping from input features to category classifications for the domain. This mapping is made explicitly available by means of the symbolic rules, and thus the "black box" criticism common to neural networks is alleviated.

From a purely AI viewpoint, we would further hope that ARTMAP's symbolic rules might serve another purpose beyond validation and justification of predictions—the discovery of novel information about the domain and/or the resolution of disagreements between domain experts about diagnostic criteria. For example, the ARTMAP rules provide an indication of the relative importance of different indicators of malignancy, based upon both frequency of occurrence and predictive accuracy, a matter on which no canonical information seems to be available. Furthermore, the rules suggest that the FOAMY feature should not perhaps be regarded as an important indicator of benignancy.

Acknowledgement

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

References

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.

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, pp.54–115.
Reprinted in Carpenter and Grossberg (1991), 316–382.

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 Reynolds, J.H. (1991) ARTMAP: Supervised Real-Time Learning and Classification of Nonstationary Data by a Self-Organizing Neural Network, *Neural Networks*, 4(5), 565–588.

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., Stephenson, T.J., Diez, Y., Harrison, R.F., Underwood, J.C.E. and Downs, J. (In Press) Diagnosis of Breast Fine Needle Aspirates Using Human Observations and a Multi-Layer Perceptron Neural Network, to appear in *J. Pathol.*.

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. (1994) A Neural Network Decision Support Tool for the Diagnosis of Breast Cancer, Research Report 548, Department of Automatic Control and Systems Engineering, University of Sheffield.

Downs, J., Harrison, R.F. and Kennedy, R.L. (1994) A Prototype Neural Network Decision Support Tool for the Diagnosis of Acute Myocardial Infarction, Research Report 552, Department of Automatic Control and Systems Engineering, University of Sheffield.

Dundas, S.A.C., Sanderson, P.R., Matta, H. and Shorthouse, A.J. (1988) Fine Needle Aspiration of Palpable Breast Lesions: Results Obtained with Cytocentrifuge Preparation of Aspirates, *Acta Cytologica*. 32, 202–206.

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

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, *J. Clin. Pathol.* 47, 329–336.

Harrison, R.F., Lim, C.P. and Kennedy, R.L. (1994) Autonomously Learning Neural Networks

for Clinical Decision Support, in E.C. Ifeachor and K.G. Rosen, eds *Proceedings of the International Conference on Neural Networks and Expert Systems in Medicine and Healthcare* (NNESMED–94), Portsmouth, UK, 15–22.

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, *J. Clin. Pathol.*, 43, 168–170.

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

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. (1994) A Heuristic for General Rule Extraction from a Multilayer Perceptron, Research Report 549, Department of Automatic Control and Systems Engineering, University of Sheffield.

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

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

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

Silcocks, P.B. (1983) Measuring Repeatability and Validity of Histological Diagnosis—A Brief Review with Some Practical Examples, *J. Clin. Pathol.*, 36, 1269–1275.

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, *J. Pathol.*, 167, 141A.

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, *Diagn. Oncol.*, 1, 79–87.

Underwood, J.C.E. (1987) *Introduction to Biopsy Interpretation and Surgical Pathology.* London: Springer-Verlag.

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, *Anal. Quant. Cytol. Histol.*, 15, 67–74.

Appendix 1: Definition of Input Features

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 in background, false if absent.

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

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

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

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

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

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

Appendix 2: Performance of ARTMAP networks using Junior Pathologist's Feature Classifications

Table 4: Unpruned Network Performance

Network Number	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	76.2	76.8	75.6
2	68.9	58.5	79.3
3	76.8	63.4	90.2
4	74.4	73.2	75.6
5	75.0	56.1	93.9
6	75.0	74.4	75.6
7	68.3	62.2	74.4
8	75.6	80.5	70.7
9	75.0	57.3	92.7
10	72.0	64.6	79.3
Mean	73.7	66.7	80.7

Table 5: Pruned Network Performance

Network Number	Accuracy (%)	Sensitivity (%)	Specificity (%)	No Diagnosis Possible
1	75.6	57.3	93.9	0
2	76.8	58.5	95.1	1
3	75.6	57.3	93.9	1
4	75.6	54.9	96.3	1
5	75.6	57.3	93.9	1
6	76.8	56.1	97.6	1
7	76.8	57.3	96.3	0
8	76.2	61.0	91.4	1
9	75.6	58.5	92.7	1
10	75.6	57.3	93.9	1
Mean	76.0	57.6	94.5	