# Breast Cancer Diagnosis and Prognosis

**Patrick Pantel**
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada R3T 2N2
ppantel@cs.umanitoba.ca

**Abstract**

Breast cancer accounts for the second most cancer diagnoses among women and the second most cancer deaths in the world. Until recently, evasive surgical biopsy was the only accurate and reliable diagnosis and prognosis tool. Of late, machine learning techniques have been applied to these tasks. We present a series of papers applying machine learning techniques to both problems of diagnosis and prognosis. One neural network diagnosis tool achieves 97.5% predicted accuracy and 100% observed accuracy. Prognosis poses a much more difficult problem due to the censored nature of the data. For most patients, we know only the disease-free survival time as opposed to the time of recurrence. One prognosis tool we present shows that good prognosis can be achieved without lymph node dissection.

## 1 Introduction

Only skin cancer accounts for more cancer diagnoses among women than breast cancer [Parker]. As with most cancers, early diagnosis of the disease radically increases survivorship. Consequently, breast cancer awareness campaigns have flourished in the recent decades. Furthermore, cancer researchers have had the laborious task of increasing the accuracy and precision of their diagnoses. Only recently have machine learning techniques been applied to this task. Such techniques can provide highly accurate and precise diagnoses that are much less evasive than surgical biopsy. Also, these techniques usually provide much quicker diagnoses.

Breast cancer also accounts for the second most cancer deaths, second to only lung cancer [Parker]. Once a patient is diagnosed with breast cancer, the prognosis gives the anticipated long-term behavior of the ailment. The prognosis is the principal factor in determining the treatment that will immediately follow the diagnosis of the disease. Prognosis is a much more difficult problem to solve because the data is censored [Street,

1998]. For only a small subset of the patients diagnosed with breast cancer do we have a time of recurrence for the disease. We will call this the time to recur (TTR). As for the others, all we have is the time of their last medical check-up. We will call this the disease-free survival time (DFS) [Street, 1995]. Data of the latter sort is often unusable by machine learning techniques resulting in a considerable loss of information.

The remainder of this paper is organized as follows. The next section presents three papers that describe machine learning techniques applied to breast cancer diagnosis. Then, section three concentrates on the harder problem of applying machine learning techniques to prognosis. Two different approaches will be presented in this section. Finally, section four describes future work and summarizes the contributions made in breast cancer diagnosis and prognosis.

## 2 Diagnosis

The goal of diagnosis is to distinguish between malignant and benign breast lumps. The three methods currently used for breast cancer diagnosis are mammography, fine needle aspirate and surgical biopsy. Mammography has a reported sensitivity (probability of correctly identifying a malignant lump) which varies between 68% and 79% [Mangasarian]. Taking a fine needle aspirate (i.e. extracting fluid from a breast lump using a small-gauge needle) and visually inspecting the fluid under a microscope has a reported sensitivity varying from 65% to 98% [Mangasarian]. The more evasive and costly surgical biopsy has close to 100% sensitivity and remains the only test that can confirm malignancy. The goal of machine learning techniques is to have the sensitivity of surgical biopsy without its evasiveness and cost.

In this section, we present three solutions to this problem. The first two utilize backpropagation neural networks (one single-stage and the other multi-stage) to recognize patterns in mammograms [Tsai, 1993b] [Zheng]. In the third solution presented, linear programming has been used to automatically inspect fine needle aspirates [Mangasarian]. This approach achieves the upper bound of the visual inspection of fine needle aspirate and is currently used in practice. It achieves the better results than the aforementioned two papers.

## *2.1 Mammography Screening Using Backpropagation Neural Networks*

Mammograms are currently the most used method for breast cancer screening. Although they are not highly accurate, they are very cost effective. Most of the early papers in breast cancer diagnosis using machine learning dealt with recognizing patterns in mammograms. In this section, we will compare two different papers which did just that. The first method uses a classic backpropagation neural network with feature selection [Tsai, 1993b]. The second uses a multi-stage backpropagation neural network [Zheng]. Although the former claims an accuracy of 100%, the experimental results are questionable. The latter paper claims a sensitivity of 100% without even mentioning specificity. As it will be discussed in more detail later, one could easily generate a system that has 100% sensitivity.

### 2.1.1 Single-Stage Approach

In this study, two features are calculated from a digitized mammogram and then passed through a backpropagation neural network for diagnosis [Tsai, 1993b]. The neural network used has two input units, one output unit, and one hidden layer with a maximum of 80 units. More precisely, the authors experimented with 20, 30, 40, …, 80 hidden units. Each unit uses a sigmoid (more closely a logistic) activation function and delta learning.

#### 2.1.1.1 Training

The training process was very poor. Their data consisted of 20 mammograms with a known malignant tumor and 20 with a known benign tumor. Ten of each class was selected as the training data. The remaining 20 cases were used as the testing set. Clearly this is a very poor training scheme. With so little training data, leave-one-out cross-validation should evidently have been used.

We expect that the system would not perform very well (since it is only training on 20 cases). Furthermore, we anticipate that the predicted error will be well off the true error of the system because of the poor validation scheme.

#### 2.1.1.2 Experimental Results

After 3000 epochs, the authors claim that their best network correctly classified 100% of the testing set. From our above expectations, this is very surprising. It is almost as if the experiment was tainted. One probable cause is that the testing set was not chosen independently of the training set. That is, the 20 cases of each class must be very similar. They might all be classic cases of malignant vs. benign tumors. If this is the case, the system would be skewed to only recognize classic patterns. It would probably be the case that, given a new randomly selected set of data, the system would not perform so well.

The authors experimented with a different number of hidden units. With 20 and 80 hidden units, they noticed that the system did not converge during training. On the other hand, networks with 30, 40, 50, 60, and 70 hidden units each achieved a 100% recognition rate. They also tried switching their training set with their testing set. After doing so, they noticed exactly the same results! This supports our above presumption that the 40 cases must be very related.

One final experiment was conducted. The authors claim that changing the order of the training data, as presented to the network during an epoch, did not affect the overall performance of the system. If all this was true and that the training and testing sets were independent, then this would be an amazing accomplishment.

It is interesting to discuss the fact that the input layer has only two units. This is explained by the fact that the authors extract only two features from each mammogram. A digitized mammogram is a 1024 x 1024 10-bit gray scale image. These were obtained from a digital radiographic system in Japan. A region of interest of 256 x 256 was selected and the gray scale was reduced to 3 bits. In a previous paper by the same authors, they claim that they could achieve only a 75% accuracy rate by passing the 256 x 256 x 3 bits of information to the input unit [Tsai, 1993a].

Consequently, the authors calculated two features from each mammogram and used these as the input to the input unit. This greatly reduced the number of connections in the network and at the same time gave a recognition rate of 100% (so they claim). They do not discuss how they selected their features but they do state the two features:

the standard deviation and the entropy of the image (i.e. the information contained in the image). It would have been advantageous to explain how they came about deciding on these two features and why they did not use more features.

To summarize, the authors used a single-stage 2-[30|40|50|60|70]-1 backpropagation neural network and achieved a predicted recognition rate of 100% in breast cancer diagnosis. In previous work, by using no calculated features (only the image bits), the authors attained a 75% recognition rate. Our discussion leads us to believe that the true recognition rate must not be 100%. In the next section, we will show that a combination of the methods seen in this section will result in a better system.

### 2.1.2 Multi-Stage Approach

In the previous section, the authors recognized that using a set of calculated features from the image as input gives better performance than using all image information bits. In this section, we present a multi-stage backpropagation network that uses a combination of the calculated features and the image information bits to learn to recognize malignant tumors in mammograms [Zheng].

#### 2.1.2.1 Network Architecture

The network presented by the authors is a two-stage network as illustrated in figure 2.1. It is used to recognize microcalcification clusters in mammograms. The presence of such clusters is often indicative of a malignant tumor. The network consists of two independently trained stages. Both stages are in fact backpropagation neural networks with one hidden layer.

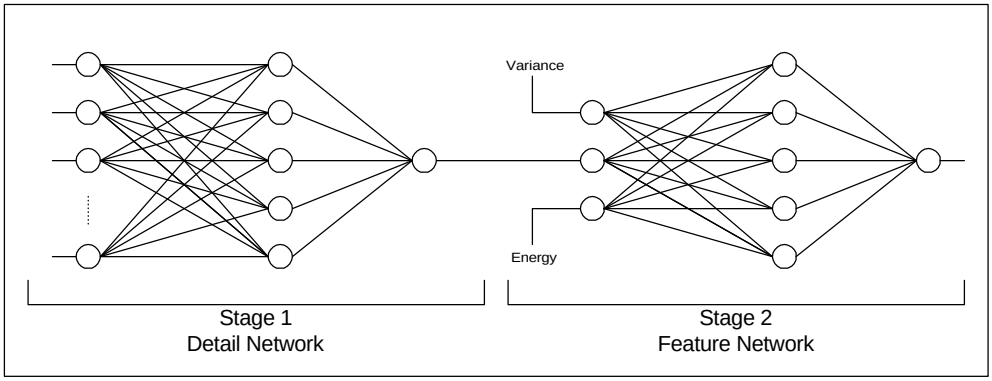


**Figure 2.1** - Multistage backpropagation neural network for classifying breast tumors in mammograms [Zheng].

As in the previous paper, a region of interest is extracted from the mammogram. The first stage, called the detail network, takes as its inputs all information bits of the region of interest. In their initial experiment, the authors used a 64-bit region of interest. Thus, they had 64 input units. They state that they achieved best results with five hidden units in the hidden layer. Finally, one output unit is used representing the two classes of malignant and benign.

The second stage takes three inputs (i.e. it has three input units): the output from the first stage and two calculated features. These features are both calculated from the region of interest extracted from the mammogram. The first feature is the variance and is obtained by the following formula [Zheng]:

$$\text{var}\,iance = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{i,j} - \quad ^2}$$ , where $x_{i,j}$ are the inputs and  are their average.

This feature contains the same information as the standard deviation used in section 2.1.1. The average  is defined as [Zheng]:

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{i,j}$$ , where $x_{i,j}$ is as before.

The second feature is the energy and is defined by [Zheng]:

$$energy = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{i,j}}$$ , where $x_{i,j}$ is as before.

The energy contains the same information as the entropy used in section 2.1.1. These two features combined with the output from stage one form the inputs for stage two. As in stage one, stage two contains five hidden units. Its single output unit represents the final classification of the network: malignant or benign.

#### 2.1.2.2 Training

Each stage is sequentially trained. Initially, the first stage is trained with the training data. Then, its output, along with the two calculated features, will be used as input to train the second stage. The training data consists of an unmentioned amount of mammogram images of 512 x 512 pixels. Twenty sub-images where selected from each mammogram and where tagged, by expert oncologists, as indicating malignant vs. benign tumors. An accumulation of more than two microcalcifications per cubic centimeter indicated a probable malignant tumor.

#### 2.1.2.3 Experimental Results

Experimental results show that this multi-stage network resulted in a sensitivity of 100%. Unfortunately, the authors forgot to mention their achieved specificity. It is meaningless to mention one without the other. One could easily build a network that has 100% sensitivity. Simply, make the network always classify inputs as malignant. All true malignant tumors will surely be correctly classified! However, all benign tumors will be misclassified.

Furthermore, the authors neglect to describe their validation scheme. They do not even mention how many mammograms were used during training. Unfortunately, it is difficult to understand the true effectiveness of this network. The authors did, however, give a comparison to a single-stage network. They claim that their best single-stage network achieved a sensitivity of 81%. Again, they do not mention specificity. However, assuming that they kept the specificity between the single-stage and multi-stage networks close, clearly the multi-stage network is an improvement over the single-stage network.

To summarize, section 2.1.1 claims that using calculated features from a mammogram image, instead of using the actual information bits of the image, as input to

a single-stage neural network performs better. In this section, the authors presented a multi-stage network that takes advantage of both the calculated features and the information bits. However, both approaches remain inconclusive in their experimental results (even though they claim 100% accuracy and 100% sensitivity respectively).

## *2.2 Fine Needle Aspirate Screening Using Linear Programming*

Although mammography is the most popular primary screening tool, it only has a reported accuracy of 68% to 79% [Mangasarian]. On the other hand, visual inspection of fine needle aspirates has a reported accuracy of up to 98%. Unfortunately, its reported accuracy varies widely to as low as 65% [Mangasarian]. In the paper that we introduce in this section, the authors applied linear programming techniques to learn to diagnose breast lumps in images obtained from fine needle aspirates [Mangasarian]. Their goal was to achieve the upper bound of the reported accuracy of the method of visual inspection of fine needle aspirates.

### 2.2.1 Features

The system that they used is called Xcyt, which was written by one of the co-authors in his Ph.D. dissertation [Street, 1994]. A fine needle aspirate is taken directly from a lump in a patient's breast. The extracted fluid is then stained to emphasize the nuclei of the cells in the fluid. Then, a digital image of the fluid is taken.

In the previous two papers that used mammograms for diagnosis, the authors simply computed two features from the digital images. In this paper, the authors computed 30 features from each image. These features are: "area, radius, perimeter, symmetry, number and size of concavities, fractal dimension (of the boundary), compactness, smoothness (local variation of radial segments), and texture (variance of gray levels inside the boundary)" [Mangasarian]. For each of these ten features, the authors calculated the mean value, extreme value, and standard error, totaling 30 features. These 30 features will serve as input to the diagnosis tool.

### 2.2.2 Training

The data set collected by the authors consisted of 569 patients for which the diagnosis was known. These cases formed the training set used to train the system. The linear programming method used by the authors is called multisurface method-Tree (MSM-T) [Mangasarian]. MSM-T will generate decision boundaries around the two classes of malignant and benign tumors. Figure 2.2 gives an example of decision boundaries that could be created by MSM-T. The darker objects exemplify the class of benign tumors while the lighter objects symbolize the malignant tumors.

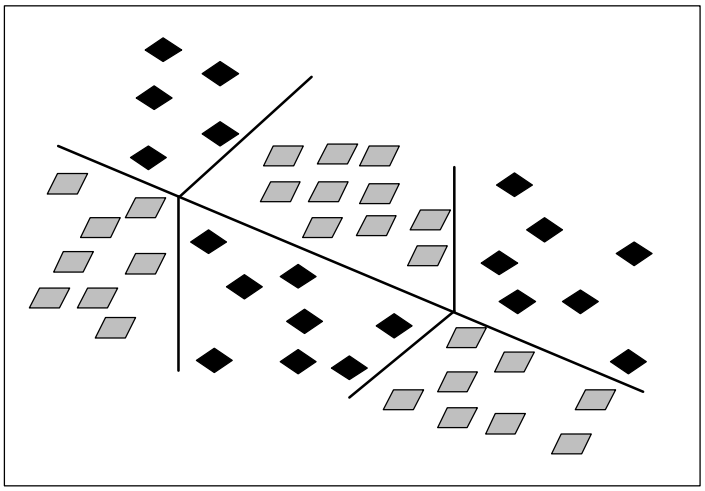

**Figure 2.2** - Decision boundaries generated by MSM-T. Dark objects represent benign tumors while light object represent malignant ones.

The authors addressed the problem of overtraining in their system by including a feature selection algorithm. They claim that better generalization is achieved by reducing the number of features actually used in the system. As opposed to the other papers discussed earlier in this survey, the authors actually describe their feature selection algorithm.

During training, many MSM-T classifiers were built. One was built for each subset of one, two, three and four features (from the thirty calculated features). The remaining combinations that performed well on the testing set were then re-trained using 10-fold cross-validation in order to estimate its real-world performance.

### 2.2.3 Experimental Results

After the experiment described above, the authors affirm that the best classifier used only the three following features: extreme area, extreme smoothness and mean texture. Experimental results indicate a predicted accuracy of 97.5%, achieving the upper bound of the visual inspection of fine needle aspirates.

Xcyt has been used in practice since 1993. Although a 97.5% predicted accuracy is reported, in practice the system has yet to misdiagnose a case. More precisely, on 131 (94 benign and 37 malignant) consecutive new patients, Xcyt has achieved 100% accuracy.

Using simple Bayesian probability estimations, Xcyt can also give a probability of malignancy for new patients. That is, if the fine needle aspirate image clearly shows malignancy, then the system will give a higher probability of malignancy than an image that is not as clear. The advantage of this is two-fold. First, we can plot a probability density graph of the malignant and benign classifications. Then, we can show a new patient where her particular case lies on the graph. This can help explain a diagnosis to a patient. Secondly, probability gives a certain confidence interval for a diagnosis. If a patient's diagnosis lies in a "gray" area on the graph (i.e. where it is hard to determine whether the tumor is benign or malignant), the oncologists may suggest a surgical biopsy to confirm the diagnosis. Thus, the system can also recognize and report cases that are more difficult for it to classify.

To summarize, the authors present a system that diagnoses breast cancer using a fine needle aspirate image. The whole process of diagnosis (including the extraction of tumor fluid) takes around 15 minutes. With an accuracy of 97.5%, this quick non-evasive diagnosis procedure is surely more desirable than surgical biopsy. Although some patients have opted for a surgical biopsy to confirm a malignant diagnosis by the system, rarely benign diagnoses need to be confirmed. Also, the system provides not only a diagnosis, but also a probability of the diagnosis. This can help oncologists decide whether surgical biopsy is needed.

# 3 Prognosis

In the previous section, we presented three different methods in which machine learning has been applied to breast cancer diagnosis. Once a patient is diagnosed with breast cancer, the malignant lump must be excised. During this procedure, or during a different post-operative procedure, physicians must determine the prognosis of the disease. This is simply a prediction of the expected course of the disease. Prognosis is important because the type and intensity of the medications are based on it. Currently, the most reliable method of determining the prognosis is by axillary clearance (the dissection of axillary lymph nodes) [Choong]. Unfortunately, for patients with unaffected lymph nodes, the result is unnecessary "numbness, pain, weakness, swelling, and stiffness" [Choong].

Prognosis is an example of a particular class of problems called "analysis of *survival* or *lifetime* data" [Street, 1998]. It poses a more difficult problem than that of diagnosis since the data is *censored*. That is, there are only a few cases where we have an observed recurrence of the disease. In this case, we can classify the patient as *recur* and we know the *time to recur* (TTR). On the other hand, we do not observe recurrence in most patients. For these, there is no real point at which we can consider the patient a non-recurrent case. So, the data is considered censored since we do not know the time of recurrence. For such patients, all we know is the time of their last check-up. We call this the disease-free survival time (DFS).

In fact, specific to our problem of prognosis, the data is right censored [Street, 1998]. This is because it is the right endpoint of time that we sometimes do not know. Another problem that adds to this is the fact that patients move, change doctors or die from cancer unrelated causes. Again here, we do not have a time of recurrence for the disease.

In previous work on prognosis, researchers have simply chosen a subset of the available training data such that only cases with recur times of less than $x$ years were included. Then, machine learning techniques were used to create classifiers that

determine whether or not a patient will recur within *x* years. Unfortunately, this approach does not utilize the training data containing no recurrence time.

In the following two sections, we present two solutions to breast cancer prognosis. The first paper we present uses maximum entropy estimation methods to create a new network architecture called *Entropy Maximization Network* [deSilva] [Choong]. This network is then used to predict the presence of axillary lymph node metastases in breast cancer patients. The second paper uses a backpropagation neural network to ascertain prognosis [Street, 1998]. It will attempt to classify inputs as "good" vs. "bad" prognosis and also generate survival curves for individual patients.

## *3.1 Entropy Maximization Network*

One of the biggest problems in breast cancer prognosis is the insufficient data available. In this section, maximum entropy estimation methods are utilized to create a novel network that requires only small data sets to be trained [deSilva] [Choong]. This is a probabilistic network. Entropy, the amount of information (or uncertainty), is maximized in the network by a "constrained gradient ascent algorithm" [deSilva].

### 3.1.1 Learning Rule Using Maximum Entropy Estimation

We now present a brief review of maximum entropy estimation [deSilva]. Given a finite set *S* of *n* elements, let $p = \{ p_0, p_1, p_2, \ldots, p_n\}$represent the probability distribution for *S*, where $p_i$ is the probability of occurrence of the $i^{th}$ element in *S*. We define the entropy *H* of *p* as:

$$H(p) = -\sum_{i=0}^{n} p_i \log(p_i)$$

Now, let $f_k$ be a set of *c* functions on *S* and $m_k$ be the set of values of the means of these functions. Maximum entropy estimation is simply the problem of finding a probability distribution *p* on *S* such that the distribution has the specified mean values $m_k$ and *H(p)* is maximized.

The authors claim that determining $p$ is a constrained maximization problem. In fact, using the gradient of the entropy, $\nabla H$, we can solve it, where

$$\nabla H = \left(-(\log(p_0)+1), -(\log(p_1)+1), \ldots, -(\log(pn)+1)\right)^T$$

Now, the sum of an initial probability distribution $p$ and a linear combination of $n - c$ vectors forms the solution space. We can then find a basis vector $b = (b_1, b_2, b_3, \ldots, b_{n-c})$ which spans the "affine subspace" defined by the $c + 1$ constraint equations [deSilva]. Note that the additional constraint equation comes from the normalization condition that must be satisfied by all probability distributions. That is,

$$\sum_{i=0}^{n} p_i = 1$$

Now, we have the knowledge to build our gradient ascent learning algorithm. Given a probability distribution $p$, the new probability distribution $p'$ will be defined as $p' = p + \varepsilon h$, where $\varepsilon$ is a small learning rate and $h$ is the projection of the gradient onto the above subspace given by

$$h = \sum_{i=1}^{n-c} (\nabla H \bullet b_k) b_k \text{, [deSilva].}$$

This development is very similar to the development of the backpropagation gradient descent algorithm [Mitchell, 1997].

### 3.1.2 Entropy Maximization Network Architecture

The Entropy Maximization Network is a three layer circular network [deSilva]. Figure 3.1 illustrates the network architecture. The layers are named as follows: probability layer, gradient layer, and constraint layer. The former two layers contain $n + 1$ units and the latter layer contains $n - c$ units. Each probability unit produces an output that becomes an input to one (and only one) unit of the gradient layer. The output of a particular gradient unit will be passed on to every unit in the constraint layer. Finally, to finish the circle, the output of a unit in the constraint layer will be passed on to each unit in the probability layer.

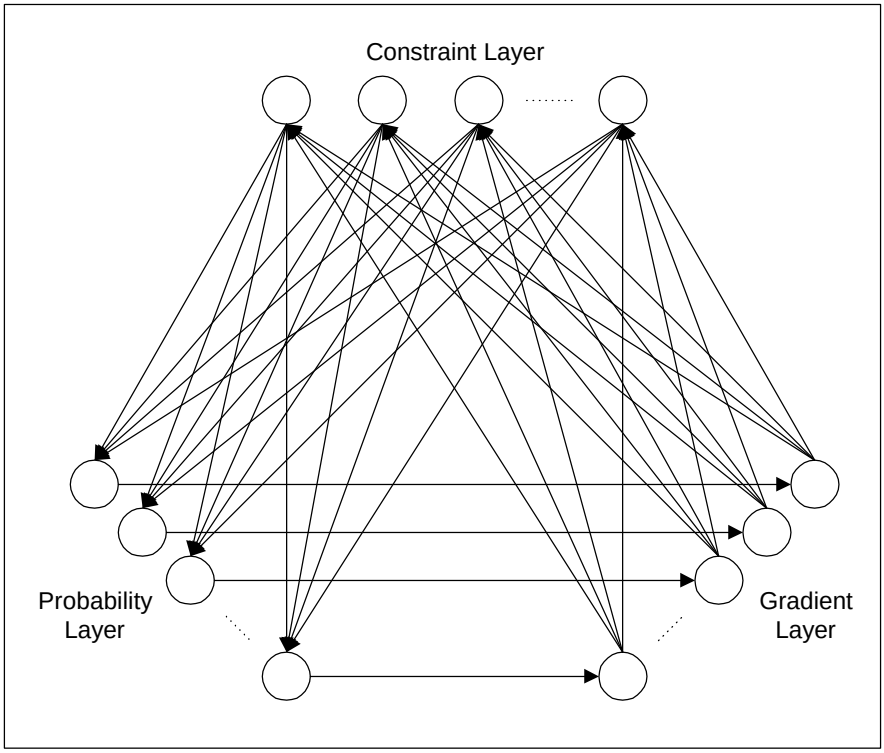


**Figure 3.1** - Entropy Maximization Network [deSilva].

The network architecture presents a nice representation of the theory presented in section 3.1.1. A particular unit in the constraint unit exemplifies one basis vector *b*. The weights coming into this unit form the coordinates of *b*. As for the weights on the links between a particular constraint unit and a probability unit, these represent the coordinates of a basis vector multiplied by the learning rate ε.

The initial probability distribution will be a solution to the constraint equations within the unit cube. The system is then solved for a set of basis vectors whose components will become the weights in the network (as defined above).

### 3.1.3 Experimental Results

In a related paper, the authors have applied their Entropy Maximization Network to breast cancer prognosis [Choong]. Their system attempts to “predict the occurrence of axillary lymph node metastases in breast cancer patients” [Choong].

The data used for this study was obtained from the Department of Pathology at Sir Charles Gairdner Hospital in Australia. It consists of 176 breast cancer patients. The

number of lymph nodes metastases for all 176 patients is known. Table 3.1 describes the clinical and histopathological features available from the data bank.

| **Risk Factor** | **Values** |
|---|---|
| Age | Continuous |
| Mitotic Count | 0, 1, 2 |
| Tubule | 0, 1, 2 |
| Nuclear Size | 0, 1, 2 |
| Nuclear Pleomorphism | 0, 1, 2 |
| Tumor Grade | 0, 1, 2 |
| Tumor Size | Continuous |
| Vascular Invasion | 0, 1 |

**Table 3.1** – Clinical and histopathological features available from the data bank [Choong].

The data set was randomly separated into a training set and a testing set. The testing set contains 42 patients with axillary lymph node metastases and 42 without. The testing set contains 42 patients with affected lymph nodes and 50 patients without. The network was evaluated by training the system and testing it on the testing set. Unfortunately, for the size of data available, this training method is poor. The authors should have used *k*-fold cross-validation. Probably 10-fold cross-validation would have been appropriate in this case since they do have more than 100 cases.

Initially, the system was tested by using single risk factors from table 3.1 as input. The risk factor *nuclear size* obtained the highest sensitivity. It correctly classified 85.7% of patients having axillary lymph node metastases. Unfortunately, this risk factor also obtained the lowest specificity. It correctly classified only 30% of the patients without axillary lymph node metastases. From this, we can conclude that *nuclear size* alone does not give a good classification. It is very easy to have a system correctly classify all patients of a particular class. All you have to do is always assign that classification to any input. A good system will have both a high sensitivity and a high specificity. However, it is preferable to have a higher sensitivity then specificity. That is, we prefer occasionally to misclassify patients that do not have axillary lymph node metastases than to misclassify the other case. The latter case is a much more serious error.

Using a single risk factor, *vascular invasion* achieved the highest specificity and accuracy. Its specificity reached 84% and its accuracy 71.7%. However, its sensitivity

was at a low 57.1%. Again, this system is unusable in practice since it is unrealizable to misclassify so many patients with affected lymph nodes.

Since using one risk factor did not generate a very good network, the authors tested combinations of risk factors from table 3.1. This yielded much better systems. The best combinations are illustrated in table 3.2. Although both have the same accuracy, we would like to claim that the first combination, *nuclear pleomorphism*, *tumor size*, and *vascular invasion*, is better than the second. This can be simply derived from the above argument. That is, we prefer a system that has higher sensitivity than specificity.

| Risk Factor | Sensitivity | Specificity | Accuracy |
|---|---|---|---|
| Nuclear Pleomorphism<br>Tumor Size<br>Vascular Invasion | 83.3% | 80.0% | 81.5% |
| Tumor Grade<br>Tumor Size<br>Vascular Invasion | 81.0% | 82.0% | 81.5% |

**Table 3.2** – Experimental results of the most accurate models involving a combination of risk factors [Choong].

To summarize, the authors used a novel network architecture, *Entropy Maximization Network*, to classify breast cancer patients as having affected lymph nodes or not. They determined that the most useful features were nuclear pleomorphism, tumor size, and vascular invasion. They achieved a sensitivity of 83.3%, a specificity of 80%, and an overall accuracy of 81.5%. A better testing mechanism, such as *k*-fold cross-validation, would give a better approximation of the true performance of the system. Although they achieved good performance, unfortunately this system is still not good enough to be used in practice.

## *3.2 Survival Curves Using Backpropagation Neural Networks*

In most attempts at applying machine learning to *survival* or *lifetime* data, researchers have utilized only uncensored data. Regretfully, most data available is censored and does indeed contain some valuable information. In this section, the author presents a method that utilizes all available training data, including these censored cases.

In previous work, the author attempted to use recurrence surface approximation techniques (using linear programming) to solve the problem of prognosis [Mangasarian]

[Street, 1995]. Their system has been used in clinical practice with much success. However, the linearity of the system causes some problems. We present in this section the most recent work in this domain. The author has created a backpropagation neural network capable of not only assessing prognosis but also apt to generate disease-free survival curves for individual breast cancer patients [Street, 1998].

### 3.2.1 Data

The network was tested on two different sets of data. The first data set is the Wisconsin Prognostic Breast Cancer Data (WPBCD) from which the author extracts 32 features using Xcyt [Street, 1994] (which was described in section 2.2). These features include: “area, radius, perimeter, symmetry, number and size of concavities, fractal dimension (of the boundary), compactness, smoothness (local variation of radial segments), and texture (variance of gray levels inside the boundary)” [Mangasarian]. For each of these ten features, the authors calculated the mean value, extreme value, and standard error totaling 30 features. The final two features are long proven prognosis predictors: tumor size and number of involved lymph nodes. WPBCD consists of 227 cases. Only 61 of these cases have recurred. This data set is characterized as a complete test set with a small amount of test cases. That is, all 32 features are available for every test case.

The second data set is the Surveillance, Epidemiology, and End Results (SEER) data set. Each entry consists of five features: “histological grade, tumor size, tumor extent, number of positive lymph nodes, and number of nodes examined” [Street, 1998]. SEER consists of over 38,000 cases. It is characterized as an incomplete test set with a large amount of test cases. 1200 of these cases do not even contain any values for these features.

### 3.2.2 Network Architecture

The author used a standard backpropagation neural network with one hidden layer. The activation function used was the hyperbolic tangent function. Thus, the output of a particular unit will be in the range [-1, 1]. Figure 3.2 illustrates the network used in this paper.

The input layer consists of 32 units for the tests with the WPBCD data set. That is, there is one unit for each feature in the data set. As for the SEER data set, five input units were used. For both backpropagation networks, three hidden units were used. Also, in both networks, ten output units were utilized.

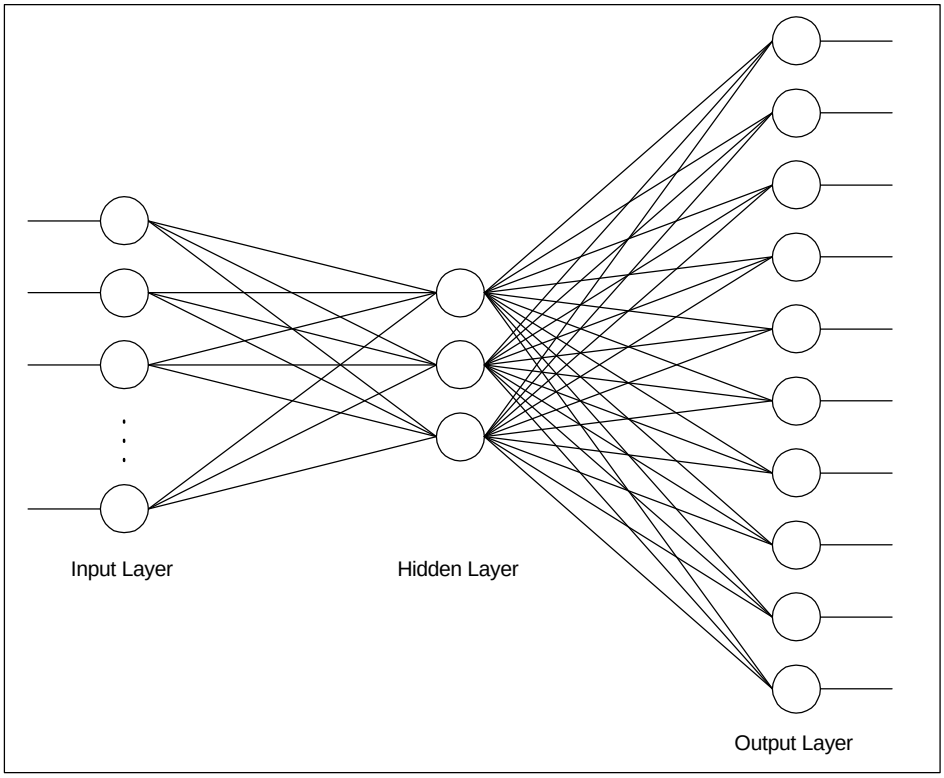


**Figure 3.2** - Backpropagation network architecture [Street, 1998].

Let $o = (o_1, o_2, o_3, \ldots, o_{10})$ be the outputs of the neural network. Let $o_i$ represent the probability that a class of cases will recur between $i - 1$ and $i$ years. Since the hyperbolic tangent activation function is used, the probabilities will be scaled from [-1, 1]. More precisely, *activation* = 2 * probability – 1. An output of +1 represents a 100% chance of non-recurrence and an output of –1 represents a 100% chance of recurrence. Now we can describe how the output of training cases may be encoded. Non-censored (i.e. recurrent) cases will be encoded differently than censored ones.

First, we describe how recurrent cases will be encoded. Say that a given case recurs between $i - 1$ and $i$ years. Then, the output vector will have values of +1 for the first $i - 1$ components and –1 thereafter. For example, say that a given case recurs after 43

months. The output vector associated with this case would be $o$ = (1, 1, 1, -1, -1, -1, -1, -1, -1, -1).

It is for censored cases that values between –1 and +1 will be seen. Remember that all we have for censored cases is the disease-free survival time (DFS). Say that the DFS for a particular case is $i$ years. Then, the probability of non-recurrence for the first $i$ years will be 100%. Thus, the first $i$ components in $o$ will be +1. After the last available check-up time, a variation of the Kaplan-Meier maximum likelihood approximation will be used to estimate the true probability of recurrence. Let $risk_t$ be the conditional probability that a breast cancer patient will recur after $t$ years given that they have not recurred in the previous $t$ - 1 years. Let $S$ be the Kaplan-Meier estimation of the disease-free survival curve. That is, $S$ consists of the cumulative probability of DFS at any year $t$. $S_t$ is defined as [Street, 1998]:

$$S_t = \begin{cases} 1, & 0 \le t \le DFS(i) \\ S_{t-1}(1 - risk_t), & t > DFS(i) \end{cases}$$

That is, the probability of non-recurrence is 100% up until the observed disease-free survival time. After that, the probability is estimated using the Kaplan-Meier estimation. For example, say our training set consists of 100 patients. If 20 patients recurred within the first year, then $risk_1$ = 0.2. Now, say that the training set has five censored cases in the first year and ten recurrences within the second year. Then $risk_2$ = 0.1/0.8 = 0.125. Now, we can calculate $S_0$, $S_1$, and $S_2$: $S_0$ = 1; $S_1$ = 1(1 – 0.2) = 0.8; $S_2$ = 0.2(1 – 0.125) = 0.175.

For a particular output $o_i$, the value of $o_i$ represents the given case's probability of membership in the class of recurrent cases between $i$ – 1 and $i$ years. On the other hand, collectively, the output vector $o$ represents an expected disease-free survival curve for the given input.

This particular probabilistic network architecture is beneficial for at least three reasons. First, the output units can be further abstracted to classify inputs as prognosis "good" and prognosis "bad". For example, this study considers that if an input is not

expected to recur in the first five years, then it is classified as prognosis “good”. Otherwise, it is classified as prognosis “bad”. Secondly, we can plot a customized disease-free survival curve for every patient. This can greatly help in explaining the prognosis to a particular patient. Finally, we can determine the expected year of recurrence as being the first output $o_i$ such that the probability of recurrence during year $i$ is greater than 0.5.

### 3.2.3 Training

The WPBCD and SEER data sets were trained on two different networks separately. Each system was trained for 1000 epochs. While the WPBC data set was tested using 10-fold cross-validation, the SEER data set was tested using leave-one-out validation.

In order to test the true efficacy of this network, only those cases from the WPBCD and SEER data set for which the prognosis is difficult to ascertain were used. That is, the author removed all test cases with distant metastasis (i.e. prognosis is poor) and carinoma *in situ* (i.e. prognosis is good).

### 3.2.4 Experimental Results

Since problems dealing with the analysis of survival of lifetime data do not fit well in traditional machine learning function approximation or classification problems, it is difficult to find a nice way to evaluate this network. We may still take advantage of the primary goal of this study: flawless prediction of prognosis [Street, 1998].

The author presents two criteria used to evaluate the network: the accuracy of predicted recurrence rates and the ability to classify cases as prognosis “good” and prognosis “bad”. We will present these results for both data sets. Recall that the WPBCD is characterized by a complete but small set of data from which 32 features may be extracted. Conversely, the SEER data set is a highly incomplete large data set from which only five features are extracted.

#### 3.2.4.1 Accuracy of Survival Curves

The author presents two graphs depicting the predicted and actual survival-curves for each test set. It is clear that the curves for the WPBCD data set are very similar. It is thus concluded that the accuracy of the survival curve estimated using the WPBCD data set is very close to the true survival curve.

On the other hand, the curves for the SEER data set are not at all similar. This is what is expected based on previous research that has determined that the SEER data set is just too incomplete and has too little features [Mangasarian] [Street, 1995]. Thus it is concluded that the accuracy of the predicted survival curve using the SEER data set is not reliable.

#### 3.2.4.2 Good vs. Bad Prognosis

To be used in practice, the system must accurately classify cases as “good” and “bad” prognosis. Recall that a prognosis is “good” if the expected time of recurrence is greater than five years. Using either data set, the system well classified cases as “good”

One of the features used in both data sets was the number of affected lymph nodes. The goal of using machine learning techniques to determine prognosis is to avoid having to dissect lymph nodes (i.e. determine the number of affected lymph nodes). So, the author re-trained the system without using this feature. The result was very encouraging. The system was able to perform almost as well without using this crucial feature. Thus, prognosis can be achieved without lymph node dissection.

To summarize, this paper presents a backpropagation neural network used to determine prognosis of breast cancer patients by learning disease-free survival curves. Experimental results show that the system is accurate in classifying new cases as “good” vs. “bad” prognosis. More substantial, it can do so without using the number of affected lymph nodes as a feature (thus eliminating the need for post-operative lymph node dissection).

# 4 Conclusion

We presented a series of papers applying machine learning techniques to breast cancer diagnosis and prognosis. In diagnosis, the third paper we presented is currently used in clinical practice with an observed accuracy of 100% and a predicted accuracy of 97.5%. Instead of using surgical biopsy to make the diagnosis, this system simply inspects images of fluids extracted from fine needle aspirates. The experimental results of the other two papers discussed in diagnosis were inconclusive.

The problem of diagnosis is mostly solved and current research deals much more with the problem of prognosis. Prognosis is a difficult because the data is censored. That is, for most patients, only the time of the last check-up is known. We only have a recurrence time for a subset of the patients. Also, patients move, change doctors or die of cancer unrelated causes. Machine learning has been making great strides in the prediction of prognosis. Already, reliable separation of “good” and “bad” prognosis is achievable without the need for lymph node dissection. In future studies, we need to determine the sensitivity and specificity of our approaches and the way that particular features influence them.

In future work, the same methods applied to breast cancer prognosis could be applied to other difficult problems dealing with analysis of survival of lifetime data. Other such applications could include survival characteristics of electronic components or even discovering characteristics of long-lasting marriages.

# References

[ACS] American Cancer Society, *http://www.cancer.org/frames.html*.

[Choong] P. L. Choong and C. J. S. deSilva, “Breast Cancer Prognosis using the EMN Architecture”, *1994 IEEE International Conference on Neural Networks*, Orlando, Florida, 1994.

[deSilva] C. J. S. deSilva and P. L. Choong, “A Network Architecture for Maximum Entropy Estimation”, *1994 IEEE International Conference on Neural Networks*, Orlando, Florida, 1994.

[Mangasarian] O.L. Mangasarian et al, “Breast cancer diagnosis and prognosis via linear programming”, *Operations Research*, 43(4), pages 570-577, July-August 1995.

[Mitchell, 1997] Tom Mitchell, “Machine Learning”, McGraw Hill, 1997.

[Parker] Parker S. L. et al, “Cancer Statistics”, 1997 *CA-A Cancer Journal for Clinicians*, 47:5-27, 1997.

[Street, 1998] W. N. Street, “A neural network model for prognostic prediction”, *Proceedings of the Fifteenth International Conference on Machine Learning*, Madison, Wisconsin, Morgan Kaufmann, 1998.

[Street, 1995] W. N. Street et al, “An inductive learning approach to prognostic prediction”, *Proceedings of the Twelfth International Conference on Machine Learning*, San Francisco, California, Morgan Kaufmann, 1995.

[Street, 1994] W. N. Street, *Cancer Diagnosis and Prognosis via Linear-Programming-Based Machine Learning*, Ph.D. dissertation, University of Wisconsin-Madison, 1994.

[Telfer] B. A. Telfer et al, “Neural Network Prediction of Mortality”, *1993 International Joint Conference on Neural Networks*, Nagoya, Japan, 1993.

[Tsai, 1993a] Du-Yih Tsai et al, “Breast Tumor Classification by Neural Networks Fed with Sequential-Dependence Factors to the Input Layer”, *IEICE Trans. Information & Systems*, 1993.

[Tsai, 1993b] Du-Yih Tsai et al, “Classification of Breast Tumors in Mammograms using a Neural Network: Utilization of Selected Features”, *1993 International Joint Conference on Neural Networks*, Nagoya, Japan, 1993.

[Wolberg] W. H. Wolberg et al, “Machine Learning Techniques to Diagnose Breast Cancer From Image-Processed Nuclear Features of Fine Needle Aspirates”, *Cancer Letters*, 77:163-171, 1994.

[Xing] G. Xing and R. Feltham, “Pyramidal Neural Networking for Mammogram Tumor Pattern *1994 IEEE International Conference on Neural Networks*, Orlando, Florida, 1994.

[Zheng] Baoyu Zheng et al, “Multistage Neural Network for Pattern Recognition in Mammogram *1994 IEEE International Conference on Neural Networks*, Orlando, Florida, 1994.