Part IV

Conclusions

Chapter 8
SUMMARY & CONCLUSIONS

This thesis examines the applicability of neural networks to clinical decision-making problems in psychiatry.

8.1 Summary of Findings

Examination of the literature on clinical decision-making (**Chapter 1**) found that the empirical evidence suggests the most accurate basis for clinical decision-making is the use of statistical decision-making. The least accurate was the use of human clinical judgement, which research has found to be subject to number of heuristics and biases that collectively have adverse effects on accuracy. Despite this body of research and its strong conclusions in favour of the use of statistical decision-making, many clinicians have a strong personal faith in their own judgement and pronounced lack of faith in the efficacy of statistical decision-making techniques. Neural networks, a relatively new form of computation, inspired by the functioning of neural systems such as the brain, are beginning to be applied to clinical decision-making problems, as well as to a broad range of problems which involve pattern recognition and pattern classification. This raises questions about their applicability of neural networks to clinical decision-making problems in psychiatry.

Chapter 2 examines Multi-Layer Perceptron (MLP) neural networks in detail. Our summary of the relevant theoretical and empirical work concluded that MLP type neural networks are capable of approximating non-linear decision boundaries in classification problems using piecewise linear approximation of the non-linear boundary. The accuracy of this approximation for any given data set can be arbitrarily increased towards it's maximum possible (the Bayesian Decision Boundary) by systematically increasing the number of hidden units in an MLP and applying appropriate optimisation techniques (such as Back-Error Propagation).

However, in the context of solving clinical decision-making problems the objective is not to maximise classification accuracy on a given data set but to maximise classification accuracy on a population, from which that given data set is drawn. This is the problem of generalisation from a sample to a population. Pursuit of this objective requires the consideration of issues related to the bias-variance tradeoff which places some constraints upon the capacity of MLPs to generalise. By understanding the relationship of bias and variance to generalisation we can undertake MLP model development practices which enhance generalisation. Importantly consideration of the bias-variance tradeoff framework leads to the conclusion that it is possible for an MLP type neural network model to classify better than a Logistic Regression (LR) model (or more generally for one model to classify better than another model) provided the MLP model offers a better tradeoff between bias and variance than does the LR model.

Our empirical review of the application of MLPs to clinical decision making problems concludes that there is evidence in favour of such use, but that the literature also contains a pervasive publication bias in favour of neural networks, which produces an overly optimistic picture of their practical usefulness. As with model development consideration of the relationships between bias, variance and generalisation error, provides a methodological framework for assessing classifier performance and comparing classifiers.

Chapter 3 is concerned with issues and methodology for "Measuring the Accuracy of a Diagnostic Classifier". It introduces and discuses the concept of a "Gold Standard". It describes and discusses a number of commonly used indices of classification accuracy. It describes and compares the various methodologies for generalisation assessment. And it discusses the need to consider the impacts of "Spectrum Effects", upon classification.

Chapter 4 Integrating considerations discussed in Chapters 2 & 3, this chapter details the methodology that will be used to implement, compare and evaluate Logistic Regression (LR) models and MLP type Neural Network models, as they are applied to clinical datasets. The methodology has the following major components:

- Use of the same software to implement all models (LR and MLP).
- Use of early stopping and weight decay to prevent overfitting by all models.
- Use of the Akaike Information Criteria (AIC) as a criteria for model selection amongst MLP models.

- Use of Area under the ROC Curve (A_z) as the basis for model comparison (deciding if one model classifies better than another).
- Use of a bootstrap correction procedure to obtain estimates of the classification accuracy (measured using A_z), on future cases, of all models (LR and MLP3).
- Use of a significance test proposed by Hanley & McNeal [1983] for Model comparison.
- Use of an independent test set for model evaluation (estimating the likely classification accuracy on future cases), where possible.

In **Chapter 5** we applied MLPs to the diagnosis of the Melancholia subtype of Depression. Traditional endogeneity symptoms of Melancholia, and a set of items from the CORE scale (which measures psychomotor disturbance), were used as the predictors. This problem, and this data set, was previously investigated by Parker et al [1995] using standard linear statistical decision-making techniques and tools. Comparisons of MLP (2 hidden units) with a logistic regression, found that the MLP-2 classified better than a logistic regression on only one of the nine comparisons

In **Chapter 6**, we extended the previous work of Levy (Levy & Hobbes, 1981, Levy 1997) which demonstrated that the Continuous Performance Test (CPT) is able to discriminate children with a diagnosis of ADHD, from those without, using linear discrimination techniques. We compared an MLP (3 hidden units) with a logistic regression for the prediction of response to treatment with stimulant medication, of children with Attention Deficit Hyperactivity Disorder (ADHD). Age, sex and the child's

pre-treatment response to the Continuous Performance Test (CPT) were used as the basis for prediction. Compared to a logistic regression, the MLP type neural network with 3 hidden units classified significantly better. The degree of shrinkage between the training data set A_Z and the Bootstrap A_Z was very large (0.167 A_Z units). This suggests substantial overfitting and that a much larger data set (with many more non-responders) is required to obtain a better solution to this particular classification problem.

Clinical application cannot yet be recommended, despite significantly better accuracy by the MLP for two reasons. Firstly the magnitude of the shrinkage indicates the training data set sample size was much too small, and that the absolute level of accuracy of the MLP can be improved. Our discussion of the bias–variance tradeoff (Chapter 2), predict that with increasing training data set sample size, the magnitude of the shrinkage will decrease and the level of cross-validated accuracy will increase.

Secondly, the criterion used as the "Gold Standard", clinical judgement by a single clinician, lacks external validity. Further study, using a prospective data collection and a criterion with better external validity, would be a natural next step for this clinical decision-making problem.

In **Chapter 7**, we applied MLPs to the diagnosis of DSM-IV Autistic Disorder in Children and Adolescents with an Intellectual Disability. Age, sex, IQ range and parent/carer ratings of behaviour on 40 items selected from the 96 item Developmental Behaviour Checklist were used as the basis for classification. We found that an MLP with

3 hidden units classified better than a logistic regression. This MLP-3 classifier was then validated using a second totally independent test data set (Sydney Test data set) and found to have good classification performance.

Other studies conducted and reported in Chapter 7 demonstrated that the posterior probability of a diagnosis of Autistic Disorder assigned to an individual by the MLP-3 diagnostic classifier is an accurate probability. As well, it was found that MLP-3 assigned diagnoses of Autistic Disorder were stable, over a 5-year period.

Finally for the MLP-3 Autistic Disorder diagnostic classifier developed in Chapter 7, the overall level of classification accuracy (test dataset A_Z = .88) is "good" (see Table 3.2 Chapter 3). These accuracy levels, the high quality of the "gold standard" used, and the real world clinical settings in which the data sets were obtained, recommend that this MLP can be applied clinical settings.

Chapter	Main Finding(s)
1	Statistical Decision-Making is superior to Clinical Judgement for Clinical Decision-Making, but clinicians seem to prefer Clinical Judgement.
2	MLP type Neural Networks can solve classification problems involving a non-linear decision boundary. They have been successfully applied to clinical decision making in medicine and have been demonstrated, in some cases, to classify better than a Logistic Regression. We extend the systematic review of Sargent [2001] adding studies published after his review and conclude that though there is evidence of a publication bias in favour of neural networks, there is also evidence that, in some applications to clinical datasets, MLPs can offer better classification than a Logistic Regression. The bias-variance tradeoff is a central consideration in the application of MLPs and other classification techniques to clinical decision making problems and provides a framework for guiding decisions and in designing a methodology for evaluating classifiers.
3	A classifier can only by evaluated against another "Gold Standard" classifier. There are several measures of classifier accuracy, which can be used. For Clinical Decision-Making problems, classification accuracy needs to be assessed with respect to a clinical population rather than the sample use to develop the classifier, so as to measure generalisation to future cases.
4	Based upon considerations discussed in Chapters 2 and 3 a Methodology for implementing LD and MLP models, comparing models and evaluating models is outlined.
5	In 1 out of 9 comparisons the MLP classified cases as Melancholic or Non-Melancholic Depression better than a Logistic Regression. In the remaining 8 comparisons MLPs and Logistic Regression classified equivalently. In the one comparison where the MLP classified better, the amount of better classification was not of an amount which would be clinically useful or significant.
6	An MLP classified cases as Responders or Non-Responders, to Stimulant Medication for ADHD, better than a Logistic Regression. Shrinkage was large for both models indicating Overfitting and a need for a larger dataset. The level of classification accuracy indicated for the MLP was not a level that would be clinically useful.
7	An MLP classified cases as Autistic Disorder or Not Autistic Disorder better than a Logistic Regression. The level of classification accuracy by the MLP is good and the MLP was demonstrated to have useful attributes as a clinical decision making tool in this domain of clinical practice, It is suggested that the most practical application of this system would be for clinicians to use it as an independent second opinion for the diagnosis of Autistic Disorder.

Table 8.1 Overview of main findings by Chapter.

8.2 Can Neural Networks be applied to Clinical Decision-Making problems in Psychiatry?

This is the core question which was investigated in this thesis. Collectively the findings presented in Table 8.1 support the hypothesis that MLP type neural networks can be fruitfully applied to clinical decision-making problems in psychiatry to produce practical solutions. More generally the findings also confirm the hypothesis that the field of psychiatry does contain some clinical decision-making problems, which are better conceptualised as non-linear rather than linear.

The theory of the bias-variance tradeoff predicts that whether or not a more complex non-linear model classifies better or worse than a less complex linear model depends upon the bias-variance dynamics of each individual problem and the amount of data available for training. Thus for some problems more complex models, such as an MLP based model, may be able to classify cases better than a less complex models, such as an LR based model, whilst in other problems the converse is true.

Following on from this another prediction, of the bias-variance tradeoff framework, is that a review of the kind carried out by Sargent [2001] and extended in Chapter 2, should probably locate some studies with large training and test dataset sample sizes which find in favour of MLPs in comparison to a logistic regression, as well there should exist similar sized studies where they perform equivalently or with logistic regression performing better. And this is what we found in Chapter 2.

Having shown that MLPs can potentially solve some classification problems in psychiatry better than linear modelling, the next step was to apply MLPs to a range of psychiatric clinical decision-making problem to see if any of these, demonstrated better classification by an MLP. Three problems were explored, Diagnosis of Melancholia, Prediction of Response to Treatment with Stimulant Medication in children with ADHD, and the Diagnosis of Autistic Disorder. In three (3) out of eleven (11) classification problems, an MLP was found to classify better than a Logistic Regression. These findings confirm the hypothesis that MLP type neural networks can be fruitfully applied to some clinical decision-making problems in psychiatry.

It is important to note that of the three classification problems in which the MLP classified better, only one (Diagnosis of Autistic Disorder) produced a result that could be exploited by further development to produce a useful clinical decision making application. This one in eleven success rate indicates that MLPs are a possible solution worthy of consideration, rather than a universally applicable solution, for the exploration of difficult clinical decision making problems. This kind of result is in line with expectations predicted both by our discussion the bias-variance tradeoff, and by our empirical review of the literature.

8.3 The Importance of training data set sample size

A necessary condition for successful application of MLP type neural networks to clinical decision-making problems in psychiatry is the availability of large enough datasets. An MLP with two hidden units will require approximately twice as many cases in a training data set as an equivalent linear modelling classification technique applied to the same problem to produce a similar level of parameter estimation accuracy, which in turn results in similar magnitude of error due to Variance. An MLP with 3 hidden units will require three times as many cases in a training data set as the equivalent linear modelling classification technique, and so on. Training dataset sample size requirements increase arithmetically with respect to increases in the model complexity of MLP classifiers

Having an adequate training data set sample size, allows the investigator to confidently explore the hypothesis that non-linear MLP classifier can classify better than a linear classifier, because it biases the bias-variance tradeoff in favour of bias. If a more complex MLP classifier is a better classifier (relative to a less complex classifier), then this occurs a result of a reduction in error due to a reduction in bias (relative to a less complex classifier) which outweighs the increase in error due to an increase in variance (relative to a less complex classifier). The magnitude of the decrease in error due to Bias is related to the fit between the model and the phenomena being modelled. However the magnitude of the error due to variance is inversely related to training dataset sample size. With a larger training data set sample size, a smaller change in fit (and bias) is needed to outweigh the smaller error due to variance. Thus a smaller change in fit can potentially be detected.

8.4 Proto-types, Subgroups and Clinical Entities

The classification of disorders in psychiatry is the ongoing subject of much investigation, theorisation and debate. There are many competing, compelling and overlapping concepts, such as categorical diagnosis, dimensional assessment and multi-axial assessment. Commonly used systems such as DSM and ICD are revised, sometimes with quite radical changes to the definitions of a particular disorder or sets of disorders, at intervals of about every 10 years or less. Allen [1998] has called Neural Networks "a new microscope" for studying the relationship between symptoms and disorders. As such, they may be able to make future contributions to the ongoing investigation of disorders and the development of diagnostic systems.

An LR linear model is able to optimally solve classification problems that consist of two multivariate groups, where each group consists essentially of a proto-type (the group mean) and some degree of variation around that proto-type. If this is a clinical decision-making problem, the clinical entities (e.g. diagnostic groups, responders and non-responders to treatment, poor prognosis vs. good prognosis, etc) would each be best described in terms of their proto-typical symptoms and the degree of possible variation. Figure 8.1 below presents this concept.

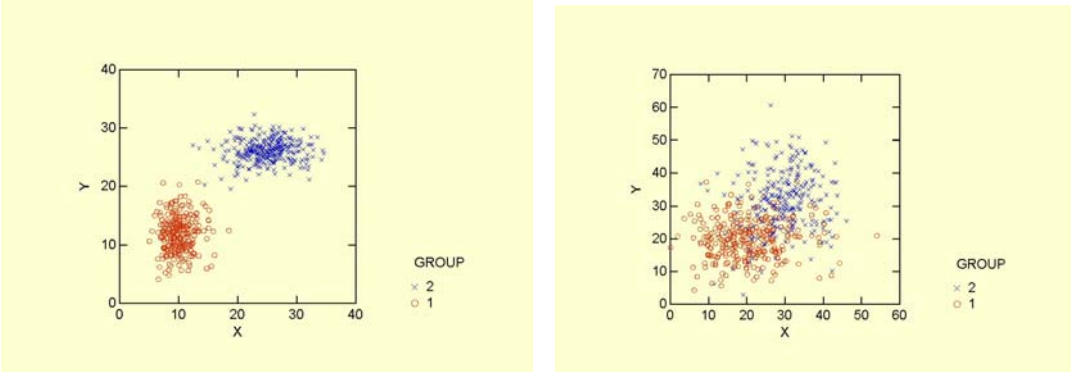

Figure 8.1 Classification problems in which the two entities being discriminated consist of cases that are variations of a class proto-type. In each of the above scatterplots, the cases belonging to two clinically defined groups, the X and Y axes indicate scores on two symptom measures

Given enough data, the MLP non-linear model on the other hand, is able to optimally solve classification problems, where at least one (or possibly both), group does not have a simple proto-type structure, but is instead a complex amalgam of subgroups. The subgroups are each prototype structured in themselves, but the group is a composite entity. When the two groups are closely packed in the data space (in clinical terms, when the two clinical entities have overlapping symptomatology), but their distributions do not significantly overlap, then a MLP model is able to classify cases into groups. Figure 8.2.presents this concept.

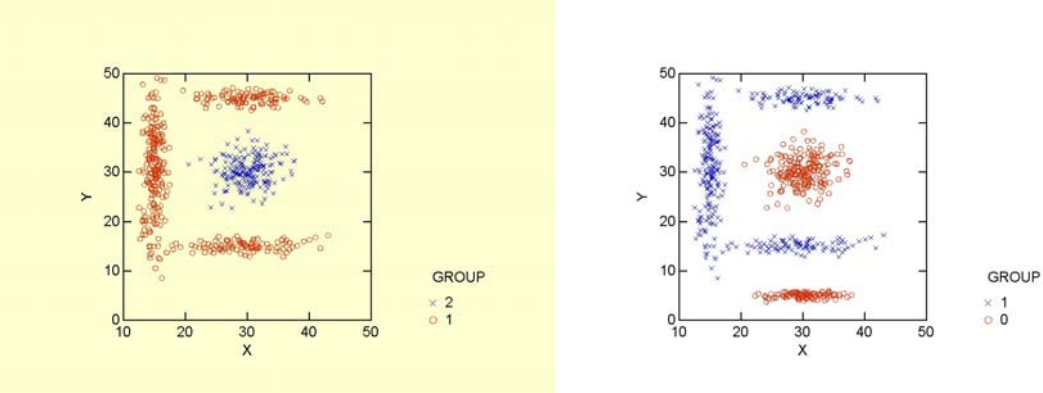

Figure 8.2 Classification problems in which one (or both) of the two entities being discriminated consist of subgroups and is not a simple prototype distribution.

This difference in specialisation between LR models and MLP models according to the types of clinical entities represented by the groups being analysed, provides a possible mechanism for examining clinical entities and concluding if the entity conforms to a simple prototype model or a complex subgroup model.

8.5 The use of MLP-type Neural Networks for Research related to Psychiatric Theories and Psychiatric Taxonomy

The kinds of tools available to psychiatric researchers play an important role in the development of psychiatric theory and psychiatric taxonomy. Allen [1998], commenting on the publication of the Diagnosis of Melancholia study [Florio et al. 1998], points out that MLP type neural networks offer up a new tool for studying psychiatric taxonomy and predicts that they will provide new insights into the relationships between symptoms and disorders.

The relationship between sets of symptoms and a disorder(s) need not be linear. However the exclusive use of linear techniques such as Linear Discriminant Function Analysis or Logistic Regression for classification in clinical decision-making studies implicitly assumes that the inherent relationship is a linear one.

However, it is important to consider that linear classifiers are robust [Dawes et al., 1989, 2000]. That is, they will often provide good classification when the underlying assumption that the classification boundary is linear is not true. This indicates that, with respect to the size and quality of many datasets obtained in medicine, the bias-variance tradeoff is such that linear models reduce bias without a great tradeoff in variance. Paradoxically this “robustness”, often listed as one of the advantages of linear modelling techniques, is a disadvantage when linear modelling techniques are used in the development of psychiatric theory and psychiatric taxonomy. Researchers may mistakenly assume that an underlying relationship in a data set is linear, when in fact it is

not, because a linear model was able to provide predict or classify well. In such a situation, the robustness of the linear models will lead researchers to draw incorrect conclusions about the nature of relationships and base their theoretical constructs and taxonomies upon fundamentally incorrect assumptions. This inturn will corrupt the accuracy of these theories and taxonomies, lead to conflicting findings, and generally retard progress in our understanding.

The advent of MLP type neural networks provides a new tool for examining relationships. Being able to confirm that a relationship in a data set is non-linear will improve our understanding and help to provide more accurate insights. As an example, in Chapter 7, when examining for Spectrum Effects amongst cases with PDD-NOS, we found that contrary to the prevailing view that PDD-NOS is a milder form of Autism, the distribution of PDD-NOS cases was bi-modal (rather than unimodal as predicted). This suggests that whilst some cases of PDD-NOS (one of the modes) are a milder form of Autism, other cases (the other mode) are non-autistic, despite having equivalent symptomatology on DSM-IV criteria.

In Chapter 1, it was pointed out that improvements to clinical decision-making practice can occur in two ways: by improvements in our understanding of individual clinical entities or by improvements in our general technology for clinical decision-making. This thesis has largely focused upon the latter, by attempting to find evidence that their exist some clinical decision-making problems in psychiatry which can be better solved as classification problems by a non-linear MLP-type neural networks than by the

employment of a more traditional linear classifier such as logistic regression. However, neural networks can also be used to help to improve our theoretical understanding of an individual clinical entity or group of related clinical entities (eg Depressive Disorders, Autism Spectrum Disorders).

Disorders in psychiatry are largely defined as phenomenological syndromes. That is as collections of symptoms which have been observed to co-occur. They are not strongly understood in terms of pathogenesis, underlying pathology, or pathological process. As such, in respect of many psychiatric disorders, we have only relatively weak indications as to the correct actions needed for prevention, diagnosis, treatment, prognosis and rehabilitation in individual cases. As well with our current understanding of psychiatric disorders, we are currently unable to resolve fundamental issues such as: what is a psychiatric disorder and what is not?

By providing a tool which potentially can help to identify and define some of the prototypical pathological groupings present amongst populations of patients with currently defined psychiatric disorders, MLP type neural network models may be able to aid in progressing our understanding of specific psychiatric disorders.

8.6 Suggested Further Work

This thesis compared MLPs to the traditionally employed linear classification technique, of Logistic Regression, as classifiers used in developing clinical decision-making practices in psychiatry. The rationale being that the discipline of psychiatry is likely to contain problems, which have been hitherto difficult to solve using traditional linear techniques, and which might be better re-conceptualised as non-linear boundary classification problems. It was found that in a specific niche and under a very specific set of conditions, MLPs could be fruitfully applied to clinical decision-making practices in psychiatry.

Moving forward, there are number areas for research and development, which are suggested by the work carried out in this thesis.

8.6.1 Transforming inputs

In our comparison between an MLP and an Logistic Regression, we did not consider the application of logistic regression (or MLPs for that matter) to transformations of the input variables. These transformations can include raising input variables by powers or creating new variables as products of the initial sets of variables. The application of logistic regression using these transformed variables in the input set creates a non-linear decision surface in the original untransformed input space. As such the resulting models might have fewer parameters than an equivalent MLP model and therefore potentially less error due to variance, but also may have a reduction in error due to bias relative to the

untransformed input set. In which case, it might classify better than both a logistic regression and an MLP applied to the original input dataset.

Though in this thesis the focus was on comparing logistic regression and MLPs, in a larger sense, this is really a non-issue. The more important conceptual and practical issue, which we have come to understand mainly through a theoretical exploration of the bias-variance tradeoff, is how do the differences between alternate models under consideration, differentially effect error as a result of differences in bias and variance.

Interestingly, the hidden layer of an MLP can be conceptualised as transforming the inputs, with the transformations being fed to a logistic regression like output unit. Viewed this way MLPs could be potentially used as exploratory tools for suggesting transformations of inputs which might improve an LR model.

8.6.2 Generic Classifiers

In same vein as we discussed in the foregoing section, the same principles apply to all classifiers. MLPs and LR are not the only techniques capable which can be applied to classification problems. There are a range of techniques with similar capabilities, such as Machine Learning, Nearest Neighbour Algorithm, Classification and Regression Trees (CART), Latent Class Analysis, Support Vector Learning Machines, Projection Pursuit Regression and Multivariate Adaptive Regression Splines (MARS), to mention a few. But there is no clear indication that any one method should, in general, be preferred [Michie et al, 1994, Ripley 1994, Bishop 1995, Sarle, 1994, Sarle 2002]. As Ripley

[1996, 1997] points out '*every method has its day',* meaning that empirically some methods are shown to be better suited to some problems (better fit to the inherent Bayesian classification decision boundary) whereas other methods are found to classify better on other problems. We do not yet fully understand why this so, or how to tell apriori which method(s) is best for which classification problem(s). As well, there are emerging new techniques for optimising MLPs and/or overcoming the problem of local minima [Bishop 1995, Sarle 2002], which need to also be considered.

Going forward, one approach might be to develop a multi-method generic classifier, which automatically investigates a range of the above non-linear methods, optimisation algorithms and linear methods, in parallel on the same data set, and then makes a decision as to the best classification method for that data set. The general classifier evaluation methods used in this thesis can be automated and extended to a wider range of classifiers. The size of the data set could be used as an initial filtering heuristic criterion, which selects classification methods to be applied. Other characteristics of the data set might also be used to rule in or rule out the application of various methods. Automated evaluation of the classification methods could be based upon the kinds of principles outlined in Chapters 3 and 4, and would need to involve a cross-validation methodology, to rule out the spurious favouring of methods which are better able to capitalise upon sample error. If two or more of the classification methods are equivalently the best, then a predetermined hierarchy of preferences amongst methods can be used to select a final classifier.

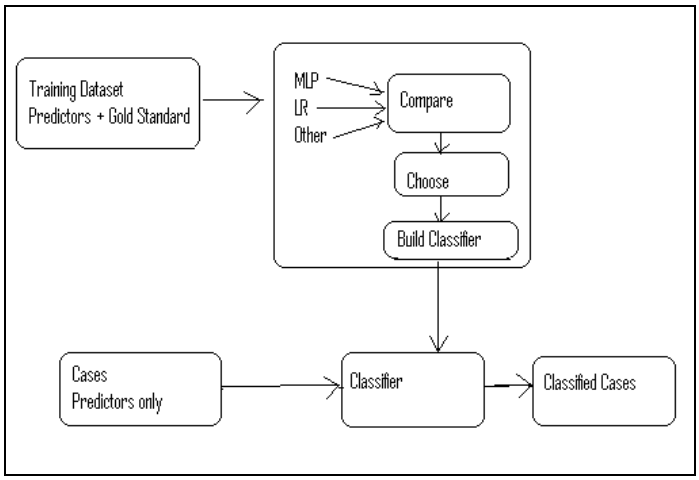

Figure 8.3 Block Diagram of a Generic Classifier

These generic classifiers, could be used as opaque "black boxes" by those who are interested only in tackling a local clinical decision-making problem, and not interested in the specifics of classification, but are concerned to ensure that the best possible classification accuracy is obtained. Clinicians, in general, are pragmatic. PAPNET, which is a proprietary neural network based system for the screening and re-screening of PAP Smears for the diagnosis of cervical cancer, is in wide use in the US, the UK and Australia. It has approval from the US Food and Drug Administration and from the UK National Health Service. As a commercial product, specific details of the how the neural networks are used to classify PAP smears as normal or abnormal are not available. However it increasingly being used by clinicians, mainly because it has been shown to be effective in number of large scale trials [Cenci et al 2000, Halford et al 1999].

8.6.3 Power Analysis for MLPs

The issue of sample size requirements for MLP- type neural networks is not well understood. It is clear that MLPs require larger training data set samples sizes than is currently the norm in psychiatric researcher. But it is not clear, at the design stage of an investigation, exactly how large a data set needs to be collected. During training shrinkage can be used a guideline for training data set sample size adequacy.

Ripley [1996] has suggested the use a subject to parameter ratios of at least 5 as a conservative guideline (i.e. one that will give larger than actually required sample sizes). While Schwarzer et al [2000] suggest that "*traditional rules of thumb in the statistical literature, requiring 5 or 10 observations per parameter to be estimated*" (p554).

Another factor which complicates matters is that the effective parameterisation of MLPs (or any model) trained with weight decay, early stopping or other regularization techniques, can be lower than its actual parameterisation [Bishop 1996], Presumably this will lead to over-estimates of required sample size in such cases. More research in this area will be of assistance to practitioners who want to apply MLPs (or other multi parameter models) to clinical decision-making.

8.6.4 Bias – Variance Tradeoff for Classification Problems

Hastie et al [2001] point out that the original analysis by Geman et al [1992] of bias and variance in neural networks does not directly apply to classification problems. Geman et al [1992] studied the bias-variance tradeoff for MLP type neural networks using Mean Square Error (MSE) as their measure of error. MSE is a measure of the performance for regression models. That is models which are attempting to estimate a continuous dependent variable. However with classification problems we are attempting to estimate class membership, in which case error is better measured using a 0-1 loss function.

Hastie et al [2001] show that using 0-1 loss, an increase variance can in some instances result in a reduction in error (as opposed to always an increase in error, as it does for MSE). The upshot of this is that the relationship of error to complexity and other variations in a model for classification models can differ, in important ways, from that of a regression model. Importantly, model changes, such as variations in complexity, need not necessarily always result in a tradeoff between bias and variance, because with some changes there will be a reduction in error due to a reduction in bias and a reduction in error due to an increase in variance, resulting in an additive reduction (rather than a tradeoff) with respect to 0-1 loss. Importantly, the minimum of generalisation error for classification problems can be located differently from that of regression problems with respect to complexity or point in training.

Clearly this behaviour in respect of classification problems has important ramifications for the application of a wide variety of classification techniques, including MLPs, to clinical decision making problems. As such, more investigation of this phenomena is warranted.

8.6.5 Fast track bridging between psychiatry and classifier research

Medicine and psychiatry are major consumers of statistical technology, including methods for classification which can be used to develop clinical decision making applications or to investigate taxonomic theories. There has been an accelerating rate of development in the field of classifiers and its related sciences, over the past 10 to 20 years. All indications are these current rapid rates of development will continue into the near future. All this begs consideration of mechanisms whereby medicine and psychiatry can more quickly come to understand and incorporate improved classifier technology.

For example, Schiavo and Hand [2000], who review recent development in the assessment of classifier performance, conclude that there has been much improvement brought about by improvements to old estimators and the introduction of some new estimators. For example in relation to the methodology employed in this thesis, Schiavo & Hand's [2000] review, clearly suggests that the basic Bootstrap estimation procedure we used to gauge classifier performance should be replaced with the more accurate 632 Bootstrap estimator.

The practical constraint which prevented upgrading our methodology to the new 632 Bootstrap estimator, is that the basic Bootstrap estimator we used is built into the NevProp software we used. Upgrading to the 632 estimator would have required a software re-write or a change to different software.

The faster track bridge in this case could be provided by the software which quickly incorporates new developments, such as a better performance estimator (or an increasing range of estimators). The ongoing development of the NevProp software stalled in the late 1990s. Generic statistical software packages such as SPSS and SAS, now include a variety of classifiers, including MLP type neural networks, but their substantial lag time for incorporation of new refinements does not really provide a solution.

A possible solution could be a specialist Internet site, which acted as a classifier server, offering a range of classifiers relevant to medicine (produce results in formats easily understood by medical researchers), with a full and expanding range of options (e.g. a range of error rate estimators). Medical researchers – users of this site could submit data for training over the Internet and obtain a set of results in return. Because the software exists as only one copy at one site, it can be frequently updated, without the need to redistribute copies to users.

The intersection of growth of interest in classifiers amongst medical researchers, a rapid pace of development of new classifiers and of new classification evaluation methodologies, and the recent advent of a ubiquitous fast Internet may have just now

created the right 'market conditions' for such a site to exist and function as a much needed bridge.

8.6.6 Transparency of MLP solutions

One of the great advantages of Logistic Regression is that the solution can be readily interpreted, even by non-statisticians. A property of models normally referred to as "transparency". This has advantages in that it allows investigators to gauge the relative importance of variables and to refine their models by eliminating those that make little or no contribution. It has also allowed theoreticians to make predictions about the relative importance of factors and be able to test them in studies, which are designed to give quantitative weightings to variables in models. A good example of this is the work of Parker et al [1995a, 1995b, 1995c,], which explored their theory that Psychomotor Disturbance is necessary and sufficient component of the Melancholia subtype of Depression.

A Logistic Regression can provide Odds Ratios (probability of occurrence divided by probability of non-occurrence) for predictor variables, which can be used to calculate relative risk. These indicators can be used to determine which variables affect the probability of occurrence of a particular outcome. This is information which can be used for decision making with respect to designing interventions.

By contrast MLP type neural networks are relatively "opaque". That is their solutions cannot be readily interpreted. The main reason for this is that their solutions are naturally

more complex. The single column of weights, which signify a linear model solution, is relatively easy to scan and gain information from. The large and irregularly shaped table of MLP weights which signify an MLP solution is much more difficult to perceptually and conceptually deal with, and they do not directly provide information which can be used for decision making with respect to interventions.

Earlier in this chapter in the classification problems presented in Figures 8.1 and 8.2, we could visualise the solutions of both the linear models and the MLPs, and this provided insights into all of the solutions, including the MLPs. But this was only possible because we were operating in a 2 dimensional input space, which is within the human range of visualisation. 17, 18, 25, 35 and 43 dimensional input spaces (the dimensionality of the input spaces of our clinical studies) are outside of our visual perceptual range. Clearly, there is a need for the development of techniques that allow us to visualise and understand MLP solutions, so that we can use them in ways similar to the way we use linear modelling solutions.

A good candidate for the visualisation of MLP solutions is the “Hinton Diagram”, which is used by Dayhoff [1990]. It presents the set of weights of MLP as a set of small squares, which vary in size, according to the absolute size of the weight it represents. Positive weights are represented by solid squares and negative weights by open squares. This system presents the entire weight table of an MLP but in a more easily digestible format.

Another visualisation technique used by Dayhoff [1990] is the "Firing Diagram". This diagram of all the connected units in an MLP shows which units "fire" and which units do not fire in response to a specific input pattern. This can allow investigators to see how the MLP responds to a particular kind of case (e.g. diseased vs. non-diseased, one subgroup vs. another subgroup(s), correctly classified vs. incorrectly classified, etc). This sort of visualisation could lead to important insights and might allow for operations such as variable culling and hypothesis testing.

Alternately the firing diagram could be converted into a short animation, using computer multimedia tools. Such a system would show how an MLP fires in response to an individual case, or to a series of cases. This form of transparent output may be of value to clinicians, who can mentally compare it to previous cases or proto-type "textbook" cases, and therefore better understand the clinical decision-making tool they are using.

It is interesting to note that brain PET Scans and EEG scans (the kind which overlay levels of EEG activity as colour on a physical presentation of the brain), which provide clinicians and researchers with a visual map of how an individual's brain responds to a particular stimulus, illustrate essentially the same concept, as the firing diagram outlined above. The main difference is that in artificial neural networks the diagram or animation records the firing of individual units, whilst in the PET and EEG scans the recording is of the firing of assemblies of neurons.

8.6.6 Deployment over the Internet

Computers are increasingly being used for psychiatric assessment [Alexandar & Andrews 1999, Garb 2000]. In first world countries, the Internet is becoming ubiquitous. Many clinicians have, or will in the near future have, a desktop Internet connection. This emerging situation makes possible a new kind of software application, which processes information over the Internet. Deployment over the Internet is another possibility for classifiers, which are essentially information processors. An example of this is Prostate Calculator (http://www.prostatecalculator.org), which is a neural network based website which allows clinicians to enter clinical data about a patient with cancer of the prostate and obtain the following predictions:

- Cancer spreading outside the prostate
- Cancer spreading into the lymph nodes
- PSA recurrence after surgery
- Survival (with drug treatment)

For all predictions, MLP type neural networks are used to calculate probabilities of the above future events directly from clinical data entered onto the Internet site by the clinician.

The Autistic Disorder Diagnosis Neural Network, developed in Chapter 7, is at a stage of development, where deployment over the Internet, for clinicians to use as an independent second opinion in diagnosis, is a viable next step.

As well as providing easy access to clinicians, deployment over the Internet can also expand opportunities for data gathering, and for incremental continuous training of a classifier. This could address the problem of obtaining large-size training data sets, highlighted earlier in this chapter. As well as providing diagnoses (or other clinical decisions or predictions), an Internet deployed application can be used to collect data (inputs and "Gold Standard" criterion), store the data and periodically retrain itself. This is particularly feasible when the Neural Network is used as a second opinion, and "Gold Standard" diagnoses are also available at the time of use. In this way, an Internet deployed neural network diagnostic classifier could progressively learn and increase its accuracy over time, by training on progressively larger data sets, naturally accumulated as part of its operation. Superficially at least, this mode of operation is not very different to that of a clinician, who learns by experience and feedback over a career lifetime. Except that an Internet deployed neural network, would be able to "see" thousands (perhaps 10s or 100s of thousands) of patients and in a wide (global) range of clinical settings. A block-flow diagram for an Internet deployed Neural Network diagnostic classifier based upon the Autistic Disorder Diagnosis MLP developed in Chapter 7 is presented in Figure 8.4 below.

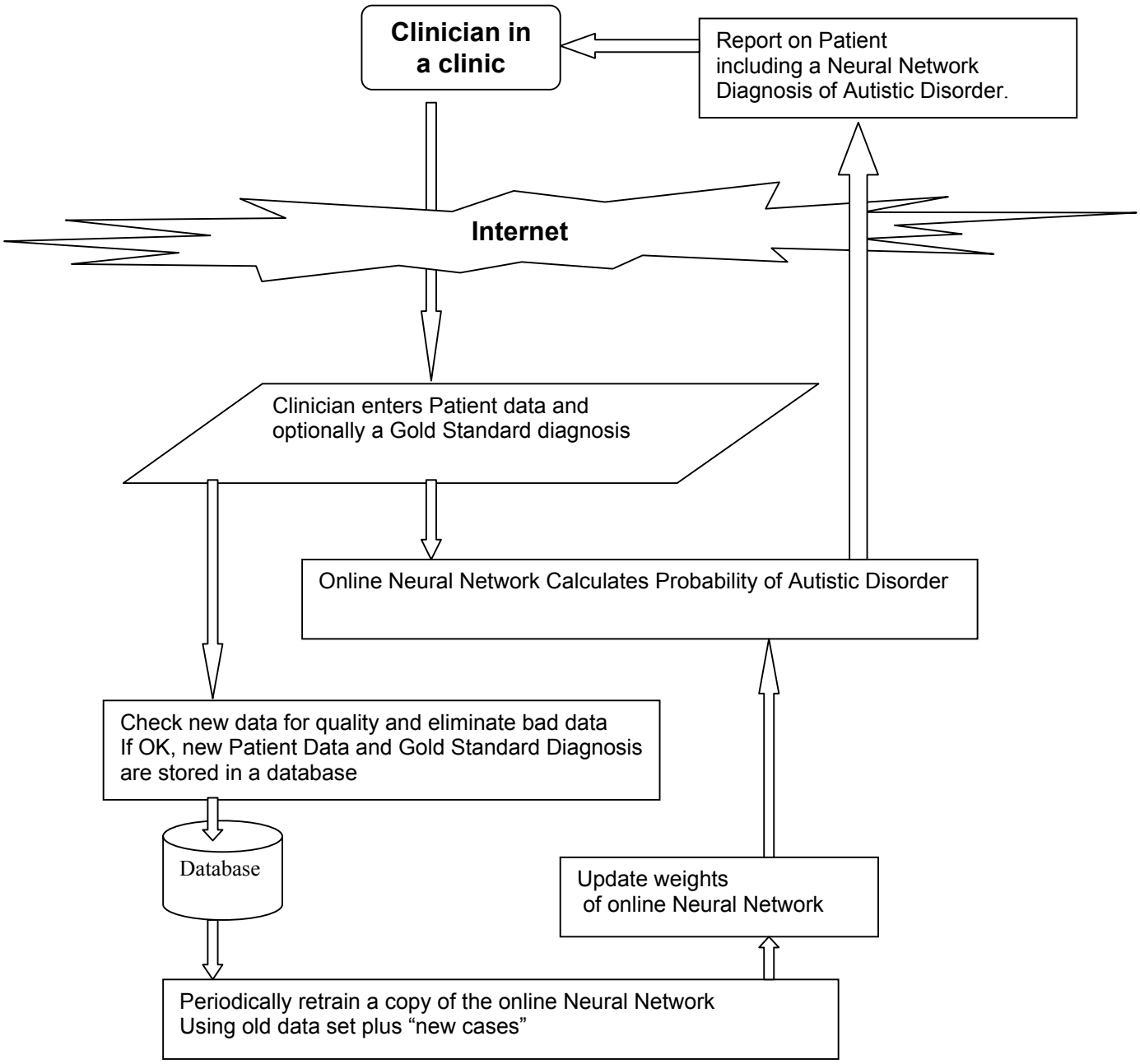

Figure 8.4 Block-flow diagram for an Internet deployed Neural Network diagnostic classifier based upon the DBC-NN developed in Chapter 7

8.6.7 Factors Affecting Uptake by Clinicians

Finally, there is the tangential but very important issue of acceptance by clinicians. Despite clear evidence that Statistical Decision-Making is generally superior to Clinical Judgment, there seems to be an irrational hesitation by clinicians which has prevented them from incorporating Statistical Decision-Making into their practice, even when they are available [Swets, Dawes & Monahan 2000; Garb 2000]. In the final analysis, MLP-type Neural Networks are a subset of Statistical Decision-Making. They are statistical classifiers that specialize in classifying in situations where the Bayesian classification decision boundary is non-linear. Will clinicians ignore diagnostic systems based upon Neural Networks, in much the same way as they have ignored practices based upon Statistical Decision-Making? We really don't know. Garb [2000] points out, that more research is needed upon the factors which lead clinicians to make particular choices or adopt particular practices. In the long run, the results of such research may assist researchers interested in improving clinical practices, in a similar way to research which compares and evaluates different practices.