

A NEURAL NETWORK APPROACH TO THE
PREDICTION OF VIOLENCE

By

JOLENE SCULLY GORDON

Bachelor of Arts
University of Missouri - Kansas City
Kansas City, Missouri
1978

Master of Science in Education
University of Kansas
Lawrence, Kansas
1983

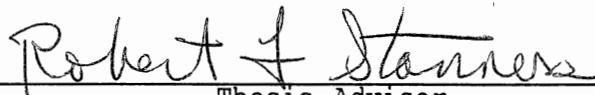
Master of Science
Oklahoma State University
Stillwater, Oklahoma
1989

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 1992

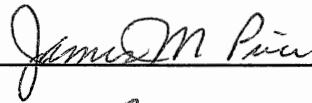
Thesis
1992D
6663n

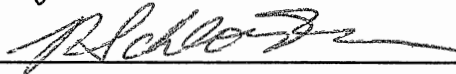
A NEURAL NETWORK APPROACH TO THE
PREDICTION OF VIOLENCE

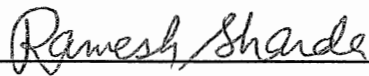
Thesis Approved:

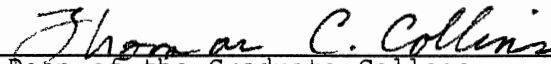


Thesis Advisor









Dean of the Graduate College

ACKNOWLEDGMENTS

I wish to express sincere appreciation to my major advisor, Dr. Robert F. Stanners, for his mentorship, thoughtful insights, and support throughout my four years of graduate study. He has been an outstanding model for me in learning to conceptualize the abstraction of cognitive theory and the logical pursuit of answers to scientific questions. I also wish to thank Dr. James M. Price, for serving on my dissertation committee, for his guidance and tutelage in research design, and for adding great depth and breadth to my ability to conceptualize statistical methods and mathematical concepts. The influence of both Dr. Stanners and Dr. Price will be carried throughout my research career. Further appreciation and thanks are extended to Dr. Robert Schlottmann and Dr. Ramesh Sharda for serving on my dissertation committee.

This research would not have been possible without the cooperation and assistance of the Oklahoma State Department of Corrections. I especially wish to thank Dr. David Adkins, clinical psychologist at Joseph Harp Correctional Center, for opening all the right doors to get me started, and for his assistance and thoughtful contributions; and Warden Jack Cowley of Joseph Harp Correctional Center, whose complete cooperation and enthusiasm for this research greatly facilitated its completion. Further appreciation is extended to Bill Chown, Administrator of Research and Evaluation; Jim Rabon,

Coordinator of Sentence Administration and Offender Records; Bud Clark, Systems Analyst; Marianne Benton, Offender Records Manager; and Lea Beatty, Closed Records; all of the Department of Corrections.

Most of all, I wish to thank my family: Richard, for his inspiration that helped me realize my aspiration to pursue this degree, for his undying support and belief in me, and for filling in the void in maintaining our home and nurturing our daughter while I devoted my life to achieving this goal; and my daughter, Ashley, whose mature understanding and support of my ambition rose well beyond her years, and for always knowing just what to say when the waters became rough. Her hugs gave me the strength to face any obstacle.

I also want to thank Steve Carver, for his encouragement and programming assistance; Sue Lykins, Joni Mihura, and Margarita Hernandez-Boyer, for their friendship and generous hospitality while I was commuting between four cities to collect data.

Finally, I wish to note that my research was aided by a Dissertation Research Award from the American Psychological Association and by a grant from the Social Concerns and Action Committee of the Oklahoma Psychological Association.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION	1
Traditional Approaches to Prediction	7
Clinical Prediction Methods	7
Statistical Prediction Methods	10
A Neural Network Approach to Prediction	16
Mechanics	18
Potential Advantages of the Neural Network Prediction Approach	26
Relative Accuracy of Prediction Methods	33
Measures of Predictive Accuracy	33
Comparative Studies	41
Issues in the Prediction of Violence	45
Definitions of Violence	47
Predictor Variables	48
Demographic Variables	48
History of Violence	49
Psychometric Variables	51
Substance Abuse	54
Situational Variables	55
Statement of Purpose	57
II. METHOD	58
Subjects	58
Demographic Characteristics of Sample	59
Procedure	60
Selection of Predictor Variables	60
Resampling Procedure	60
Data Collection	61
Neural Network Training	62
Apparatus	62
Architecture	63
Input Layer	63
Hidden Layer	65
Output Layer	66
Learning Parameters	67
Extent of Training	67
Weight Matrix Analysis	68
Criterion Variable	69
Design	70

Chapter	Page
III. RESULTS	72
Base Rates	72
Selection of Subsets of Predictor Variables	72
Neural Network Training Characteristics	73
Total Group Hit Rates	74
Concept of Chance	74
Entire set of Predictor Variables	76
Subsets of Predictor Variables	76
Selected by Stepwise Discriminant Analysis	76
Selected by Neural Network Analysis	77
Conditional Probabilities	79
False Positive and False Negative Ratios	80
Increasing Decision Thresholds	81
IV. DISCUSSION	84
"Best-One-Wins"	84
Increasing Decision Thresholds	88
Predictor Variables	90
Limitations of a Neural Network Approach	97
Future Research	98
REFERENCES	99
APPENDIXES	116
APPENDIX A - OFFENSES CONSIDERED VIOLENT	116
APPENDIX B - PREDICTOR VARIABLES	119
APPENDIX C - PREDICTOR VARIABLES SELECTED BY STEPWISE DISCRIMINANT ANALYSIS	122
APPENDIX D - PREDICTOR VARIABLES RANKED BY NEURAL NETWORK WEIGHT MATRIX ANALYSIS	123
APPENDIX E - OKLAHOMA DEPARTMENT OF CORRECTIONS APPROVAL FORMS	126

LIST OF TABLES

Table	Page
1. Parole Violation by Salient Factor Score	128
2. Performance Comparison of Neural Net and Discriminant Analysis with Two Different Base Rates, on 10 Test Files of Random Input	129
3. Frequencies of Four Classification Outcomes for Neural Network and Discriminant Analysis Models	130
4. Total Group Hit Rates	131
5. Conditional Probabilities by Model Type: Neural Network and Discriminant Analysis	132

LIST OF FIGURES

Figure	Page
1. Layers of a Backpropagation Neural Network	133
2. Activation Function	134
3. Proportion Correct Classification at Increasing Decision Thresholds	135
4. Comparative Performance of Two Prediction Methods vs. Base Rate	136
5. Total Group Hit Rates by Model Type	137
6. False Positive and False Negative Ratios	138
7. Proportion Correct Positive Predictions at Increasing Decision Thresholds	139
8. Proportion Correct Negative Predictions at Increasing Decision Thresholds	140

Abstract

A backpropagation neural network and discriminant analysis were compared for their efficacy in the prediction of violent behavior. Forty-eight predictor variables including demographic data, criminal history, psychometric data, substance abuse history, and situational factors were collected from official records of male criminal offenders (N = 392) and used to predict the violent or nonviolent nature of the offense for which each subject was incarcerated. Both neural network (NN) and discriminant analysis (DA) models showed statistically significant prediction accuracy of about 77% total hits on cross-validation. As decision thresholds for classification were made increasingly stringent, however, the NN models held their accuracy better than the DA models. The highest levels of accuracy were achieved for both NN and DA models with a collection of 17 variables that included demographic data (age, income, race, unskilled labor), criminal history (probation and parole status, previous violent arrests), psychometric data (MMPI scales 1, 3, 8, 0; IQ), situational factors (being married, living with a mate, irregular work history, supporting a family), and substance abuse (benzodiazepines).

A NEURAL NETWORK APPROACH TO THE
PREDICTION OF VIOLENCE

Artificial neural networks, otherwise known as parallel distributed processing models (Rumelhart & McClelland, 1986) or connectionist models (Feldman & Ballard, 1982), are a form of adaptive computer information processing system that associates input patterns with output patterns. This association, or "mapping," is said to be "learned" by the network as the input-output associations are formed by induction, that is, by repeatedly processing examples of input-output pairs and gradually adjusting a set of numerical weights, until the network can generate the correct output for each input used in the "training" process.

Neural networks differ in fundamental ways from traditional forms of artificial (computer) information processing systems. Unlike traditional artificial intelligence (AI), such as expert systems, neural networks contain no separate knowledge base of rules; in fact, the understanding of the rules for mapping inputs to outputs is not required of the programmer--they are generated in the training phase by the network. Many problems in AI have been intractable because of the lack of knowledge necessary for constructing explicit rules, even though large sets of examples based on experience exist. Furthermore, rule-based systems often fail when applied to real-world data that is corrupted with noise (Hecht-Nielson, 1988). Neural networks

have offered an alternative means of solving such problems, without the need for explicit rules. In contrast to traditional computing in general, neural networks are comprised of many simple distributed processing units, rather than a single complex central processing unit. Furthermore, the result of processing is not stored in a specific memory location, but consists of the overall state of the network (matrices of weights) after it has converged to a criterion condition of equilibrium (Caudill, 1987).

Neural networks were inspired by the neural architecture of the human brain, originally conceived by McCulloch and Pitts (1943) in a paper entitled, "A logical calculus of the ideas immanent in nervous activity." The adaptive nature and, hence the learning capabilities, were added 15 years later by Rosenblatt (1958). The computational units are highly interconnected, arranged in hierarchical layers, and operate in a metaphorical sense as neurons connected together into a functioning whole (Klimasauskas, 1991a). That is, each "neurode" sums the excitatory (+) or inhibitory (-) input received from each neurode in the preceding layer via a weighted "synaptic" connection, transforms that input, and produces an output, which is then received as input by each neurode in the successive layer, and processed in a similar fashion.

Many different types of neural network architectures exist (at least 50, 13 of which are in common usage; Hecht-Nielson,

1988), differing in topology (number of layers, number of neurodes per layer, degree of interconnectivity among and between neurodes in different layers); "learning" algorithms (specifying how the weights are to be adjusted); and transfer, or "activation" functions (for the within-neurode transformation). This study proposes the most popular form of network for pattern classification, a "backpropagation" neural network (Werbos, 1974; Parker, 1982), as an appropriate model for individual behavioral prediction.

Implementation of neural networks may be realized in several different forms. Hardware implementations operate in parallel at very fast speeds via simple processors and parallel circuitry. "Neurocomputers" combine partially parallel hardware and software which simulates the parallel processing of its elements. Strictly software forms of neural networks, such as the one employed in this study, simulate the parallel processing of elements, but run on conventional serial computers (Kinoshita & Palevsky, 1987).

The last decade has seen a surge of interest in neural networks on the part of researchers in a highly diverse range of disciplines, including artificial intelligence, computer science, electrical engineering, physics, neurobiology, philosophy, linguistics, and psychology. This excitement, evident by the hundreds of talks and papers on the subject each year (Caudill, 1989), may be attributed to the widespread and

often dramatic success recently achieved by applying neural networks to an impressive variety of pattern recognition, classification, nonlinear feature detection, and prediction problems (White, 1989a, 1989b), many of which had previously been intractable, or solved only by very difficult conventional approaches. Hornik, Stinchcombe, and White (1989) have provided a theoretical foundation which establishes that these successes are not just "flukes," rather they reflect the capabilities of backpropagation networks as general universal approximators of unknown nonlinear regression functions (p. 364).

Backpropagation neural networks are potentially applicable to any situation that requires the acquisition of a complex nonlinear mapping (Simpson, 1990). Successful applications have included speech processing (e.g., Elman & Zipser, 1987), image recognition (e.g., Cottrell, Munro, & Zipser, 1987), temporal processing (e.g., Elman, 1988), knowledge processing (e.g., Hinton, 1986; Pollack, 1988), text and sentence processing (e.g., Sejnowski & Rosenberg, 1987), optical character recognition (e.g. Becker & Hinton, 1991; Caudill, 1988), medical diagnosis (e.g., Weiss & Kulikowski, 1991), as well as diagnostics and robotic control. These examples are by no means exhaustive, but were selected to illustrate the tremendous diversity of recent work encompassed by the field (see Simpson, 1990, for an extensive bibliography).

Another area that has seen many successful neural network

applications is prediction. Prediction applications include Latin American conflict (Werbos & Titus, 1978), corporate bond rating (Moody & Utans, 1991), bankruptcy (Odom & Sharda, 1990), cancer recurrence (Weiss & Kulikowski, 1991), time series prediction (Sharda & Patil, 1990), time series of sunspots (Weigend, Rumelhart, & Huberman, 1991), solar flares (Fozzard, Bradshaw & Ceci, 1989), and Mackey-Glass chaotic time series (Crowder, 1991; Lapedes & Farber, 1987; Sanger, 1991).

Prediction is one of the most fundamental objectives of basic and applied science. Survival of early civilization depended on such problems: the prediction of weather cycles for planting and harvest, and of animal migration among the earliest examples. Success at prediction is taken as validation of theoretical explanations of phenomena.

One of the goals of psychological science is to predict human behavior. Literature on the applied prediction of human behavior reveals essentially three types of behavior that psychological science has tried to predict (Meehl, 1954): success in some type of training or schooling, recovery from psychological disorders, and criminal recidivism.

Although neural networks have been used in psychology to model perceptual, cognitive, and neurobiological processes, there has been no previous psychological study done, to this author's knowledge, which has applied neural networks to individual behavioral prediction. Examples of psychological

modeling efforts include models of word recognition and context effects (McClelland & Rumelhart, 1981; McClelland, 1991), memory (McClelland & Rumelhart, 1985), human categorization (Kruschke, 1991), speech (McClelland & Elman, 1986), cerebral cortical processing (Crick & Asanuma, 1986; Sejnowski, 1986), place recognition and goal location (Zipser, 1986), and neural plasticity (Munro, 1986). Thus, although neural networks have proven quite useful in psychology, and in other types of prediction, they have not yet been evaluated as a tool for predicting individual behavior, a primary objective in psychology.

A pressing, long-term problem in behavioral prediction, which has thus far proven intractable (Monahan, 1981; Wenk, Robison, & Smith, 1972) with traditional techniques is the prediction of violence. This problem was selected as a test case for a neural network approach to behavioral prediction for several reasons. First, it is an old problem (e.g., Burgess, 1928), which has a history of previous attempts (to be reviewed in a later section), that can serve as a baseline for comparison with a neural network approach. Second, not much progress has been made in the more than sixty years of documented attempts, thus, the potential for improvement over previous attempts is feasible. Third, violent behavior occurs with a very low base rate (proportion of the population that actually commits violent acts), a characteristic that plagues prediction attempts (Meehl,

1954), and one that, it will be argued, may be more tractable with neural networks than with traditional prediction methods. Fourth, any potential improvement realized would carry a very high societal value, as this problem is still a very important concern of the public, as well as of the criminal justice system; any contribution that would at least lead in a positive direction could eventually help solve some very serious practical problems. Fifth, a large data base exists in the official records of incarcerated offenders.

The focus of this research is on two fundamental issues. First, the aim is to empirically evaluate the potential contribution of neural network technology to an area important to psychology--behavioral prediction. A secondary aim is to attempt to predict, in a practical sense, an instance of low base rate behavior--violent behavior.

Traditional Approaches to Prediction

Traditional approaches to predicting criminal behavior have relied upon two general modes of combining data--clinical and statistical/actuarial methods. This section will give an overview of the processes involved in clinical and statistical prediction.

Clinical Prediction Methods

Clinical prediction involves hypothesis formulation concerning the structure and dynamics of the particular individual for whom the prediction will be made (Meehl, 1954).

This method entails an intuitive or subjective combination of factors deemed relevant by the clinician (Elstein, 1976). Such relevance is often determined per individual case from a study of occurrences in the individual's life (Meehl, 1954; Monahan, 1981). Factors are selected from interview impressions, case history, and psychometric information, often in the absence of any exact knowledge of the statistical relationships between predictive information and the criterion behavior (Meehl, 1954).

Accuracy of clinical prediction rarely exceeds accuracy obtainable by chance (Meehl, 1954). This method is particularly prone to overpredict, that is, to generate many "false positives," cases predicted to exhibit the criterion behavior which in fact do not display such behavior. This "leniency error" (Sarbin, 1942) has been demonstrated in the prediction of grade point averages, and virtually every study predicting success on parole (Meehl, 1954; Monahan, 1981; Steadman, 1980).

Overprediction is not unique to the clinical method of prediction, but stems from a problem of base rate in the criterion behavior, which plagues any attempt to predict a behavior that occurs only rarely. Base rate refers to the proportion of cases exhibiting a particular criterion behavior in a given population. This rate is critical in prediction, with the likelihood of maximal prediction accuracy occurring in criterion behaviors with a base rate of 50% (Meehl & Rosen,

1955).

Blind guessing, in a criterion distribution with a base rate of 50%, results in 50% correct decisions. In this case, a prediction method with only weak or moderate validity is likely to improve upon this base rate accuracy. Blind prediction in a skewed distribution, however, with a base rate, for example, of 20%, can achieve 80% correct decisions simply by predicting all cases to belong to the more frequent class (Meehl, 1954; Meehl & Rosen, 1955). Therefore, considerably higher levels of predictive validity are required for discrimination above base rate accuracy, as the base rate deviates from 50%.

The extreme manifestation of the base rate problem in clinical prediction results from the fact that it is often ignored in this method of prediction (Meehl, 1954; Meehl & Rosen, 1955). This tendency has been documented by Tversky and Kahneman (1974), who labelled it the "representativeness heuristic," the tendency to predict the outcome that appears most representative of the available evidence even when that outcome is statistically less likely than others. This heuristic is especially prominent when case-specific information, the sole basis for much clinical prediction, is present. In spite of this lack of accuracy inherent in clinical prediction, the criminal justice system has relied heavily on the clinical judgment of psychologists and psychiatrists for predictions of dangerousness (Monahan, 1981).

Statistical Prediction Methods

In contrast, statistical methods of prediction determine expectancies about future behavior on the basis of class membership, resulting in a probability figure that is an empirically determined relative frequency (Meehl, 1954). The data are mathematically combined by mechanistic decision rules for the purpose of classification (Meehl, 1954; Monahan, 1981). Actuarial tables containing the distribution of frequencies in cells represent complex conjunctions of data (Meehl, 1954). In contrast to the clinical method of selecting relevant factors on a per case basis, statistical methods dictate precisely the factors to be considered for every instance of a specified type of case (Monahan, 1981).

Statistical prediction can often be more efficient than clinical prediction, taking less time, less effort, and requiring lower level personnel to carry out (Meehl, 1954). In addition to greater efficiency, virtually all studies comparing the relative efficacy of the two methods find statistical prediction more accurate than the clinical approach (Meehl, 1954; Steadman, 1980; Monahan, 1981). Despite findings of vastly improved accuracy, reliability, and consistency, statistical methods have been neglected in the prediction of violent behavior (Shah, 1978; Monahan, 1981).

The most commonly used statistical methods for prediction are additive linear models. Two such methods of historical

significance, the Burgess method (1928) and the Glueck method (1950) have held up relatively well in the prediction of criminal recidivism. A third linear model, the standard tool for prediction, is multiple regression analysis, or its variant for use with dichotomous criterion variables, discriminant function analysis.

The Burgess method is a simple point scoring method, in which each predictor variable is dichotomized at the median. If an individual's status on a given predictor variable falls into the category associated with success on the criterion, his score is incremented by 1 point; if in the category associated with failure on the criterion, the individual scores 0 on that variable (Wilbanks, 1985). A total score is obtained by summing points for each predictor variable; thus the maximum possible score is equal to the number of predictor variables included. Scores for all subjects in a construction sample are cross-tabulated with the criterion variable to yield the proportion of successes and failures associated with each possible score, and appropriate categories of risk are thereby assigned (Wilbanks, 1985).

The Salient Factor Score, a modern variant of the Burgess method is used by the United States Board of Parole as an aid in predicting success on parole (Wilbanks, 1985). Possible scores range from zero (high likelihood of violation, hence poor risk) to ten (low likelihood of violation, hence good risk). Hoffman

and Beck (1985) used the Salient Factor Score to predict serious parole violation within a five-year follow-up period.

Recidivism was correctly predicted in 40% of individuals classified as "poor risk," whereas only 14% of those classified as "good risk" seriously violated parole (see Table 1).

Insert Table 1 about here

A second statistical method of historical importance in criminological prediction was developed by Glueck and Glueck (1950) in a well-known study of juvenile delinquency. The Gluecks compared 500 institutionalized juvenile males with 500 unconvicted juvenile males, studied at an average age of 14 - 15 years (Farrington & Tarling, 1985). A prediction table based on five factors concerning discipline, supervision, affection, and cohesiveness among family members showed remarkable discrimination. Of those scoring in the high risk range, 98.1% were delinquent and in the low risk range, 97.1% were nondelinquent. There were many serious flaws with the Gluecks' study, however, such as the use of extreme groups, an unrealistically high proportion of delinquents (50%), interviewer bias, and the absence of a validation sample (Farrington & Tarling, 1985). Although the Gluecks' results must be discounted due to these flaws, their method has held up in comparison to other methods and is therefore worthy of

mention.

The Glueck method is similar to the Burgess method, but more precise in the weighting of predictor variables (Wilbanks, 1985). The weight assigned to each dichotomized predictor variable is equal to the proportion of subjects in a construction sample who fail on the criterion variable and possess that attribute. Thus total scores for all subjects, derived by summing these percentage weights across all predictor variables, are divided into intervals associated with increasing levels of risk (Wilbanks, 1985). Wilbanks (1985) applied both the Glueck method and the Burgess method to a criterion of parole success based on twenty predictor variables. He found the methods to produce very similar results: 108 and 100 errors made, respectively, in the construction sample; 100 errors made, by both methods, in the validation sample. Copas and Tarling (1984) demonstrated that both the Burgess and Glueck models are, in fact, the same simple loglinear model in which all predictor variables are treated as independent.

Familiar multiple regression techniques rely on an ordinary least squares method (Tarling & Perry, 1985) to derive weights for each predictor variable based on its relative contribution to the explained variance, while holding constant the effects of other predictor variables in the equation (Wilbanks, 1985). Unlike the simpler point methods of Burgess and Glueck, multiple regression takes intercorrelations between predictor variables

into account. A subject's score is the linear combination of weighted scores on each predictor variable and some constant. Two variations of the multiple regression approach include discriminant analysis, for use with dichotomous criterion variables, and logistic regression. It has been shown (e.g., Copas, 1985) that multiple regression, with dichotomous criterion variables, is mathematically equivalent to discriminant analysis. Weiss and Kulikowski (1991) cite empirical comparisons of discriminant analysis and logistic regression and conclude that they usually give similar results. With a large number of categorical predictor variables, however, it was suggested that logistic regression may produce a slightly more optimal (in terms of greater classification accuracy) model (Weiss & Kulikowski, 1991).

Regression models, including discriminant analysis, have been the standard tools for prediction studies. It has been asserted (e.g., Lippmann, 1987; Weiss & Kulikowski, 1991) that these models, in contrast with neural network models, require fairly restrictive assumptions about the distributions of both criterion and predictor variables--normal distribution underlying the error component of the criterion variable; joint multivariate normal distribution of the predictor variables; and homoscedascity of variance, or constant error variance across different levels of the predictor variables (e.g., Neter, Wasserman, & Kutner, 1989). These assumptions apply, however,

only when the model will be used for purposes of making inferences to populations, by attaching probability values to inferential statistics. It is not on this basis that the neural network and regression models in this study will be compared. These models will be evaluated in strictly a descriptive sense, that is, in terms of their respective accuracy in deriving a prediction model equation which can be applied to new cases for the purpose of predicting membership in one of two classes. Even if one did intend to use regression methods in an inferential sense, the F test has been shown to be robust with respect to violations of these assumptions, except in extreme cases, especially when large sample sizes are used (Cohen, 1968; Hair, Anderson, & Tatham, 1987).

Of more significant concern for behavioral prediction is the number of, and intercorrelations among, the predictor variables used in the model. It is a common finding that more error is generated and little predictive power is gained by the inclusion of more than the first several variables in the linear model equation (Farrington, 1985; Gottfredson & Gottfredson, 1985; Tarling & Perry, 1985). That is, little predictive power is gained when variables, intercorrelated with those already in the equation, are added. Each additional variable adds a further increment of error that is unique to the construction sample and cannot be expected to exist in a new sample, and thus adds to the shrinkage (reduction in explained variation) of the

equation when applied to this new sample. This is a significant problem in the prediction of criminal behavior which involves a large number of potential predictor variables.

Finally, data relevant to the prediction of criminal behavior are potentially ridden with multilevel interactions, that is, nonadditive combinations of variables, although these have yet to be empirically demonstrated (Beverly, 1964). It is theoretically plausible that this lack of evidence for significant interactions in criminological data is inherent in the statistical method which requires that each potential interaction be specified and included in the equation as a separate term. When the number of predictor variables is large, theoretical knowledge of interactions, lacking in criminology, is necessary to guide a systematic investigation of such interactions (Palmer & Carlson, 1976). Without knowledge of which variables interact, and the nature of the combinatorial process, one faces a combinatorial explosion of the number of possible interactions. For example, if there were a total of ten predictor variables, all possible combinations involving 1 to 10 variables would result in a total of 1,023 possible combinations of predictor variables. Obviously, it would be feasible to empirically investigate only a few of these possibilities.

A Neural Network Approach to Prediction

Neural networks offer a fundamentally different statistical

approach to prediction problems. White (1989) is one of the few statisticians involved in analyzing the learning procedures of feedforward neural networks. He concluded that the method of backpropagation can be viewed as an application of the Robbins-Monro (1951) stochastic approximation procedure to solving a novel class of multidimensional nonlinear regression problems (p. 449). "Approximations" are used in place of the true response function of a nonlinear least squares framework (White, 1981). White (1989b) further suggests that neural networks are applicable to regression problems requiring some type of "flexible function form" (p. 1011).

Gallinari, Thiria, Badran, and Fogelman-Soulie (1991) have recently analyzed the relations between discriminant analysis and neural networks, analytically for linear neural networks, and empirically for the nonlinear case. The empirical investigation compared the two models on problems increasing in degree of nonlinearity. Their results showed an advantage for nonlinear networks over the discriminant analysis models that increased in magnitude as the nonlinearity of the problem increased. Furthermore the advantage of the neural network models extended to generalization on new cases. They established that each layer of weights in a network performs a nonlinear discriminant analysis from the states obtained in the previous layer. Thus each layer increases the separation and the clustering of the different classes and the last layer

classifies the final projection (p. 357).

Thus, although neural networks employ several preexisting concepts from the statistical literature, it is the **combination** of these that is novel (White, 1989b). The net input to a given hidden unit in a neural network is a familiar linear discriminant function which, when subjected to a nonlinear transformation within the hidden unit, acts as a nonlinear feature detector. The outputs of all feature detectors in the hidden layer are then inputs to another linear discriminant function and another nonlinear transformation at each unit in the output layer. "The approximation benefits from the use of nonlinear feature detectors, while retaining many of the advantages of linearity in a particularly elegant manner" (White, 1989b, p. 1004).

Mechanics

Backpropagation neural networks "learn" to classify a pattern through induction, by repeatedly processing examples of each class. The network is arranged in successive layers of simple computational devices called neurodes, or simply "units." The network consists minimally of three such layers of neurodes: an input layer, a hidden layer, and an output layer (see Figure 1). The intermediate layer consists of neurodes which receive neither direct input from the outside world, nor a direct training signal, and hence are "hidden." The number of neurodes that can be contained in any layer of the network, and the

number of hidden layers used, are constrained only by the power of the particular software package used, computational limits of a given hardware system, and practical considerations of training time.

Insert Figure 1 about here

Examples are coded as input-output pattern pairs, in the form of two n-dimensional vectors. The input vectors represent patterns of "activation" values distributed across all neurodes in the input layer. The output vectors represent the correct output for each corresponding input pattern. The pattern of activation on the input layer is propagated in a forward direction (hence, a "feedforward" network) to the hidden layer. The resulting pattern of activation on the hidden layer is then propagated on to the next layer, the output layer if it is a three-layer network. Each neurode, or unit, receives inputs from all neurodes in the previous layer, each of which is weighted by a value representing the "connection strength" between each pair of between-layer neurodes. The receiving neurode computes a linear combination of these inputs, resulting in a scalar value, or net input, which is then subjected to a nonlinear transformation, or "activation function." The backpropagation algorithm requires the activation function to be continuous and differentiable at all points (Rumelhart, Hinton,

& Williams, 1986). Typically a sigmoidal, or logistic, function, which meets this requirement, is used (Equation 2 below).

The net input to "receiving" unit j , for input/output pattern pair p is:

$$net_{pj} = \sum_i w_{ji} o_{pi} + \theta_j \quad (1)$$

where $i = 1$ to the number of sending units;

$j = 1$ to the number of receiving units;

w_{ji} = the connection weight between sending unit i and receiving unit j ;

o_{pi} = the output of sending unit i , produced by the presentation of input pattern p ;

θ_j = a bias, which functions as a threshold, in the form of a weight to receiving unit j , from an "extra" sending unit that always has an output = 1.

The output of receiving unit j , or its "activation" value, for input pattern p is:

$$o_{pj} = f_j(net_{pj}) = \frac{1}{1 + e^{-(net_{pj})}} \quad (2)$$

where f_j = a nonlinear function, sigmoidal in form.

This activation value is then output to all neurodes in the next layer (see Figure 2). The nonlinear activation function serves to constrain the output of each neurode to a value between 0 and 1, filtering out noise (very low values), and preventing output values from reaching very large magnitudes (Carpenter, 1989).

Insert Figure 2 about here

Knowledge is represented in the values of weights assigned to the connections between neurodes on different layers. These connection weights are initially set to small random values in the range $[-0.1, +0.1]$. Upon presentation of a single input pattern, the forward propagation through the network proceeds as described, resulting in a final activation value for each output node. This output value (o) is compared to a target value (t) for that node, that is, the correct output for the input pattern. The difference between the output and the target ($t - o$) is thus the error measure for the network's processing of the input pattern.

"Learning", via weight modification, takes place as this error is propagated *backward* through the network in a recursive fashion. The magnitude of the error in classifying input pattern p is used to determine the amount of change (Δ) needed in each weight (w) in order to reduce the error on the next presentation of pattern p . Each connection weight is modified

according to the "generalized delta rule" (Rumelhart, Hinton, & Williams, 1986a):

$$\Delta_p w_{ji}(n+1) = \eta(\delta_{pj} o_{pi}) + \alpha \Delta w_{ji}(n) \quad (3)$$

where $\Delta_p w_{ji}$ = the change in weight from unit i to j after processing input pattern p ;
 n = the presentation number for input pattern p ;
 η = learning rate, a constant of proportionality;
 α = momentum term, a constant that determines the magnitude of the effect of past weight change on current weight change;

$$\delta_{pj} = (t_{pj} - o_{pj}) o_{pj}(1 - o_{pj}) \quad (4)$$

where δ_{pj} = the error signal;
 t_{pj} = target, or correct output for unit j , for input pattern p ,
 for the weights connecting the **output** layer and **hidden** layer units, and

$$\delta_{pj} = o_{pj}(1 - o_{pj}) \sum_k \delta_{pk} w_{kj} \quad (5)$$

where k = number of units in the layer above unit j ,

for weights connecting the **hidden** layer and **input** layer units. All weights in the connection matrix are thus updated according to this "learning rule" (Equation 3), a procedure which is recursive by layers, in such a way as to improve performance of the network on the next occasion it receives similar input. Over many, perhaps thousands, "epochs" (one complete presentation of all input/output pattern pairs in training file) the total error, summed over the entire set of example patterns, is reduced to a minimal level in this implementation of a local gradient descent procedure, and the network is said to be "trained." Although the generalized delta rule does not guarantee that this minimum is the global minimum, and not a local one, empirical tests have demonstrated that convergence to a local minimum is quite rare (Rumelhart & McClelland, 1986a; Weiss & Kulikowski, 1991).

The trained network produces a matrix of connection weights, a complex mathematical model underlying the patterns of association inherent in the training data. Once trained, the learning mechanism is disabled, and the network can receive any pattern as input, from the training set or otherwise, and classify it according to the model developed from all the connection weights. Although several have referred to this matrix as a "black box" (e.g., Bailey & Thompson, 1990; Garson, 1991), meaning its weights are opaque to interpretation, researchers are actively seeking methods for interpreting the

connection matrix in terms of the phenomenon being modelled. Garson (1991), for example, suggests a method for using the connection weights to partition the relative share of the output associated with each input variable, by which the relative importance of input variables in a model can be analyzed. Klimasauskas (1991b) suggests using a nonparametric statistical technique, "sensitivity analysis," to investigate the relative importance of each input to a given output. These and other recently published methods (Arnaldo, Miller, & Gonzalez, 1990; Howell, 1990; Nelson & Illingworth, 1991; White, 1989) suggest that weight matrix analysis has the potential to contribute to the theoretical knowledge underlying the fitted model.

For behavioral prediction, a neural network can be trained with a construction sample of pattern pairs, with each input pattern representing the values of all predictor variables for one individual, and each output pattern representing the correct classification for that individual. Consider an example prediction net comprised of 50 inputs, 10 hidden units, and 2 output units. The input units might represent measurements on, for example, 20 predictor variables. The less than one-to-one representation of predictor variables on input nodes results from a "distributed" coding scheme, in which the value of a single predictor variable may be coded across several binary units, each representing a different category of a given variable. Marital status, for example, might be represented by

three binary units, each coding the presence or absence of one of three categories: single, married, or divorced. The use of three units allows the option of representing "unknown" as the absence of all categories. "Local" coding, using a single node to represent a single variable, may be incorporated as well. Current age might, for example, be represented by only one input unit, continuously valued. Input values are normalized, based on the dynamic range of values for a particular input, to values on a scale of 0 - 1. In this manner, the input units may represent variables of any level of measurement, categorical or continuous.

The two output units might represent the two levels of classification, "A" or "not A." Once the network is trained, new cases from the validation sample, not processed by the network in the training procedure, can be given as input, and the value of each output node, ranging from 0 to 1, may be interpreted as representing the conditional probability of membership in each class (White, 1989), or a continuous gradation of "certainty" of the classification decision (Jones & Hoskins, 1987; Williams, 1986). Using a "Best-One-Wins" decision rule, the output node with the higher value represents the network's classification of the given input pattern.

The hidden units are the unique feature of a neural network prediction scheme. The input patterns are mapped to (i.e., associated with) the output patterns via this layer of units,

which represent the inputs at a higher level of abstraction (than the level of the input units), and may be conceptualized as representing salient features of the data (Rumelhart, Hinton, & Williams, 1986b). In other words, new "hidden" variables are created from combinations of the input variables.

Potential Advantages of the Neural Network Prediction Approach

Neural networks offer some potential advantages over traditional statistical prediction methods. The first advantage lies in the interconnectivity of the network architecture. Each input neurode is connected to each hidden neurode, which is connected to each output neurode. This between-layer interconnectivity allows the network the opportunity to assign weights to any combination of variables necessary to reduce the output error, in the process of mapping input values to hidden units, and hidden unit values to output units. There is no counterpart to these hidden units in multiple regression or discriminant analysis.

Second, whereas traditional methods have generally restricted their models to linear relationships, this restriction is somewhat arbitrary (Thorndike, 1918) and it seems implausible to assume that the factors influencing human behavior combine in only a linear fashion. Neural network activation values are subjected to nonlinear transformation locally at each neurode in the network. Inherent in this transfer function is the nonlinear combination of many predictor

variables. This nonlinear processing that occurs within the neurodes gives neural networks the capability of forming nonlinear separations of classes in the multidimensional decision space created by the network (Lippmann, 1987). It has been well-established that backpropagation networks with only a single hidden layer can approximate any arbitrarily complex nonlinear mapping, to any desired degree of accuracy, provided a sufficient number of hidden units are used (Hecht-Nielsen, 1988; Hornik, Stinchcombe, & White, 1989; Lippmann, 1987; Simpson, 1990; White, 1989). Thus, a neural network has the potential for outperforming a linear discriminant function in classifying a criterion behavior which is an unknown nonlinear function of a given set of predictor variables. Lapedes and Farber (1987) have shown that the backpropagation learning algorithm provides a natural extension of linear methods into a nonlinear domain.

Third, rather than developing a prediction equation based on central tendencies and variability derived from the simultaneous processing of the training data (Lippmann, 1987), neural networks gradually fit a complex model by trial and error, as they process one example at a time, and adjust the connection weights in very small increments (Gallinari et al., 1991).

Fourth, neural networks have been demonstrated to be quite robust with regard to handling input corrupted by random noise, both in training and in generalization (Hartzberg, Stanley, &

Lawrence, 1990; Lippmann, 1987; Weiss & Kulikowski, 1991).

Features that appear noisy, as a result of measurement error, when considered individually, may prove to be highly predictive when combined with other features and mapped to a new set of higher order features (Weiss & Kulikowski, 1991). The same may be true of features that, individually, are only weakly correlated with the criterion. Neural networks are able to accurately generalize, that is, to classify new patterns, not seen in the training procedure, by interpolating between training examples (Gallinari et al., 1989; Hartzberg, et al., 1990; Lapedes & Farber, 1987), or in the case of noisy data, approximating the surface function between data points (i.e., where there are no examples; Poggio & Girosi, 1990).

Generalization accuracy is a function of the number of hidden units used and number of examples in the training set, and thus is a criterion by which the appropriate number of hidden units is determined (achieving an optimum number of hidden units is the object of the complexity fit procedure, to be described in a subsequent section). Increased generalization ability suggests that neural networks can reduce the size of shrinkage (that occurs when applying a model developed on a construction sample to a validation sample) inherent in statistical prediction methods. Reduced shrinkage results in greater predictive accuracy for the validation sample and hence greater external validity of the model.

Fifth, there is some evidence from preliminary simulation work (Gordon, 1991a) to suggest that neural networks may excel over linear discriminant models with increasingly stringent thresholds, or decision rules, for class membership. A decision threshold refers to a cutting score, a minimum score which must be reached or exceeded for classification into one of two classes. In this case, two decision thresholds were used in each simulation; a lower score at or below which a case was classified Nonviolent, and an upper score at or above which a case was classified as Violent.

Simulation data (N = 200) were generated randomly and then transformed, to have intercorrelations among ten inputs comparable to those found among the ten clinical scales of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) for males (Hathaway & McKinley, 1989, p. 99). Correlations between each of the ten inputs and the criterion varied from $-.25$ to $+.25$, thus reflecting varying degrees of weak relationship with the criterion violence.

A neural network with ten input neurodes, ten hidden neurodes, and two output neurodes, was trained with one-half of the simulated data (N = 100; 50 Violent, 50 Nonviolent), and tested with the other half (N = 100; 50 Violent, 50 Nonviolent). A discriminant analysis model was similarly developed on the same training data set, and applied to the same testing data set.

The output from the neural net and the discriminant analysis models on the testing data set, were compared in terms of proportion correct classifications, at decision thresholds (cutting scores) of .50/.50, .40/.60, .30/.70, .20/.80, .10/.90, for inclusion in the respective predicted classes, Nonviolent and Violent. Neural net and discriminant analysis accuracy, in terms of proportion correct decisions, were quite comparable at thresholds of .50 to .70, but beyond a .80 classification threshold, the neural network maintained its accuracy on a validation sample, while discriminant function analysis fell at a steep decline (see Figure 3). Both the network output and discriminant model output, at thresholds more stringent than .50/.50, result in a band of undecidable cases, with probability near .50, much as human decision makers have been found to do (Meehl, 1954).

Insert Figure 3 about here

Additional pilot work (Gordon, 1991b) compared the accuracy of a backpropagation net and discriminant analysis on problems with decreasing base rates. A training file of 100 input patterns, each composed of three randomly selected values, in the range [0,1], was created. Each input pattern was then randomly assigned a target value of 1 (Violent) or 0 (Nonviolent), in such a way to assure either a 50% base rate of

violent targets, or a 20% base rate of violent targets. Thus, two nets could be trained and two linear discriminant models could be developed, with different training pattern files, each having a different base rate of the target behavior.

Note that the random generation of input patterns and the random assignment of target patterns to input patterns, resulted in a near zero correlation ($M = -.06$ across the three inputs, $R^2 = .02$) between the input patterns and targets. That is, there was virtually no linear information present for the net to learn. Ten additional sets of randomly generated inputs were similarly created for running the trained net and applying the discriminant analysis model. The ten test sets consisted of input patterns only--no targets were provided. The objective of this simulation was to investigate the proportion of outputs in each of the two classes generated by each model, relative to the base rate of the training set, rather than assessing the accuracy of classifying patterns for which the targets were known.

In both base rate conditions, the neural nets classified a similar proportion of cases in the validation samples (test patterns) as was present in the construction sample (training patterns), as belonging to the Violent class (see Table 2). Discriminant analysis, on the other hand, performed well with a base rate of .50, but with the small base rate of .20, classified all test patterns as belonging to the more frequent,

Nonviolent, class. Thus, in the condition of greater interest (due to the low base rate), the neural network outperformed the discriminant analysis model, even in the absence of significant linear information.

Insert Table 2 about here

One further step was taken, to determine if the neural network had developed a bias based solely on the distribution of target signals, independent of the input patterns, or rather had learned the mapping from the small amount of information, linear or nonlinear, present in the training set, which discriminant analysis was unable to learn. Another set of 100 training patterns was created. This time, however, all input values were held constant, at a value of .50. This was done to ensure that absolutely no information was present from which the net could learn. The base rate of Violent targets was held at the same low value of 20%. The result was that now the neural net, as discriminant analysis had done before, classified all patterns as Nonviolent, the more frequent class.

The results of this pilot work, although preliminary, would imply a sixth unique advantage, in that the neural network was capable of discriminating between classes, even with a base rate as low as 20%. Furthermore, this discrimination was based on a set of three predictor variables which contained virtually no

linear information. It was determined however, by removing *all* linear information, the network performed at merely base rate level, as the discriminant analysis had performed in the earlier condition. These preliminary results provide support for the notion that neural networks may outperform discriminant analysis on problems of predicting low base rate behaviors.

Finally, the neural network literature (Hartzberg et al., 1990) suggests that there is no significant disadvantage, other than length of training time, in including a large number of predictor variables. The network will "disregard" variables that are not associated with the output, by not adjusting the weights connected to the inputs representing those variables, hence leaving them at or near their initial near zero values. Furthermore, it suggests that intercorrelation among predictor variables does not detract from the goodness of fit. This would seemingly make neural networks a suitable tool for use on problems where the number of potential predictor variables is great, and the intercorrelation among those variables is high.

Relative Accuracy of Prediction Methods

Measures of Predictive Accuracy

Any classification model results in four possible outcomes for a given case: correct positive, correct negative, false positive, and false negative. A Correct Positive (CP) is a case which is predicted to exhibit the criterion behavior, and in fact does so. A Correct Negative (CN), on the other hand, is a

case which is predicted not to exhibit the behavior, and in fact does not. In most classification problems, the objective is to maximize these two cases. Errors are committed when either of the following cases occurs: a case predicted to exhibit the criterion behavior does not do so (False Positive, or FP), or a case predicted not to display the behavior, does display the behavior (False Negative, or FN). Thus, predictions of violent behavior, when implemented, result in errors of either restricting an individual's freedom without cause (FP), or releasing an individual who will bring harm to an innocent member of the community (FN). Depending on one's perspective (community at risk vs. civil liberty) the relative value placed on these errors may be quite different.

Many researchers (e.g., Steadman, 1980; Monahan, 1981) from the civil liberty perspective, have focused on the ratio of False Positives to Correct Positives. Predictive accuracy from this viewpoint is dismal (Steadman, 1980). This ratio is directly a function of two parameters in the prediction scheme: base rate and selection ratio (proportion predicted positive of the total sample $[(CP + FP) / N]$; Meehl & Rosen, 1955; Brown, 1976). Selection ratio is determined by the particular decision threshold employed. Farrington and Tarling (1985) point out that only when base rate and selection ratio are equal can every case be correctly predicted. As base rate and selection ratio diverge, the maximum number of correct predictions decreases.

False positives will occur to the extent that the selection ratio exceeds the base rate. For example, if the base rate for violence were .60, and 60% of the total sample were selected to be predicted violent (selection ratio = .60) on the basis of a test with perfect validity, 100% correct predictions could be made. If, on the other hand, with the same base rate and perfect validity, the selection ratio were .80, the maximum proportion of correct violent predictions would be only .75 (Brown, 1976, p. 117).

Many different measures of predictive accuracy appear in the literature (see Farrington & Tarling, 1985, for a review of 14 such measures). One category of accuracy measures is the degree of association between the predicted outcomes and the actual outcomes. This is a measure of the internal validity of the model. A second category of accuracy measures, is the error rate of the prediction model. The error rate may be considered the measure with the most practical significance, that is, a measure of whether a prediction model should be considered for implementation. It is conceivable that a significant degree of association may be measured for a given model, but that the false positive error rate is too large to consider implementation of the model.

Error rate analysis may be approached from two different perspectives. The total error rate $[(FP + FN)/N]$ is the simplest and most comprehensive measure, giving equal weight to

the two different types of errors, false positives and false negatives. Error can also be analyzed separately by type (FP or FN), for the purpose of weighting these types differentially. False positive error rates, as previously mentioned, have received much attention in the criminology literature because of their implications for restricting individual liberty. It is this measure that has been the primary focus in studies that have concluded that violence cannot be predicted (e.g., Wenk, Robison, & Smith, 1972).

Many and varied measures of association appear in the criminological prediction literature, each with some advocates (Loeber & Dishion, 1983; Copas, 1985; Farrington & Tarling, 1985; Tarling & Perry, 1985; Wilbanks, 1985). Each correlational measure makes certain assumptions about the nature of the variables, and this must be considered in selecting a measure of association. Other measures have been shown to be very closely related mathematically, such as Mean Cost Rating, Kendall's r , Receiver Operating Characteristic Curve, and Goodman & Kruskal γ (Tarling, 1982).

The variety of reported accuracy measures, along with gross differences in definitions of the criterion behaviors, as well as varying lengths of follow-up periods, make direct comparisons of accuracy across studies a difficult task. Follow-up periods range from six months (Klassen & O'Connor, 1988) to five years (Hoffman & Beck, 1985). Another salient weakness in the

criminological prediction literature is the lack of cross-validation of many of the prediction models. Klassen and O'Connor, for example, report 93% correct classification (22% CP, 71% CN), with an accompanying .1 false positive for every one correct positive. This is an outstanding result, but there was no attempt to cross-validate the discriminant analysis model on a new sample. This weakness, along with a large number of predictors (64), renders these results in the prediction of short-term violence in non-schizophrenic mental patients somewhat meaningless.

It is essential to measure the accuracy of a prediction model, not only on the construction sample, but on a validation sample as well. Weiss and Kulikowski (1991) refer to the two resulting categories of error as "apparent" error and "cross-validation" error, respectively. Apparent error is the error as measured on a particular construction sample. Cross-validation estimates true error, that is, the expected value of error in the population from which the samples are drawn. A prediction model may be "overfitted" (with a large number of predictors) to the construction sample and produce a very small measure of apparent error. This fit reflects not the potential usefulness of the model on a new sample, or external validity, but capitalization on the measurement error and random fluctuations in the particular sample used. Any model with a high enough level of complexity (number of free parameters) can

closely fit these idiosyncratic characteristics in a single sample.

A more meaningful test of a model is one that is applied to a new sample for cross-validation. Most models do not hold up well on new samples, that is, there is a tremendous reduction in accuracy over the accuracy attained on the construction sample. It is this reduction in performance from construction sample to validation sample that is referred to as "shrinkage" in regression analysis, and it is a function of the number of predictor variables used in the model, combined with the construction sample size (Copas, 1985). For the purposes of clarity and evaluation of methods, this discussion will be restricted to those relatively few studies which include both a construction sample, on which the predictive equation is developed, and a validation sample, on which the efficacy of the equation is assessed.

As described previously, base rate affects the accuracy of any prediction scheme. Base rate is affected by definitions of criterion behaviors, sampling procedures, and length of the follow-up periods.

A series of three studies by Wenk, Robison, & Smith (1972) is often cited as evidence that violence cannot be predicted. Wenk and colleagues defined violence quite restrictively, as reconviction and reimprisonment for a violent parole violation. The base rates across these 3 studies ranged from 0.3% to 5%.

With such small targets, observed over follow-up periods of only 12 to 15 months, their models could produce no more than 0.42% correct positives (CP / Total predictions) in their total population of offenders, and 14% correct positives [CP / (CP + FP)] of those predicted to be "Most Violent" (less than 3% of the total population). Extreme false positive rates accompany these low base rates, with 6, 8, and 19 false positives per one correct positive, for base rates of 5%, 2.5%, and 0.3%, respectively.

Other studies demonstrate a linkage between definition of violence and base rate. A considerably less restrictive definition of violence was adopted by the State of Michigan (1978) in their study of parole violation. The criterion was defined as arrest for violent crime. Within a follow-up period of 14 months, 10.5% of the sample was arrested for a violent crime. Parolees had been previously classified into risk categories on the basis of a decision tree with six binary decision nodes, derived from an analysis of 350 predictor variables. Recidivism rates during the 14 month follow-up period were calculated for each risk category: Very high (40%), High (20.7%), Middle (11.8%), Low (6.3%), and Very low (2%).

A third study conducted by the U.S. Parole Commission, and replicated by Hoffman and Beck (1985), defined violation as any new commitment of 60 days or longer, or return to prison for parole violation (including technicalities). The follow-up

period was two years. This definition yielded a base rate of 26% violators.

As a further illustration of the effects of length of follow-up period, and definition of recidivism, Megargee and Bohn (1979) followed a prisoner cohort of 1,011 prisoners (entering prison within a two-year interval) as they were released into the community. The follow-up period for determination of recidivism was extremely variant, ranging from 18 to 67 months (mean = 42.8 months, standard deviation = 10.7 months). Recidivism rates were assessed according to three definitions of recidivism: rearrest (for any cause), reconviction (for a new offense), and reincarceration (for parole violations as well as new convictions). The base rates of recidivism for these three categories were 52.6%, 26.7%, and 26.2%, respectively. As is apparent from this and previous examples, the range of base rates across studies is quite large, from 0.3% (Wenk et al., 1972) to 52.6% (Megargee & Bohn, 1979).

The decision threshold, or cutting score, for membership in the "Predicted Violent" class, also has direct implications for the magnitude of the two types of prediction error, and hence social implications in the desired balance of community risk and individual liberty. The proper decision threshold for implementation of any prediction method must be determined for each specific application--its concomitant base rate in the specific population, and any externally imposed constraints on

selection ratio (Brown, 1967).

Comparative Studies

Two bodies of research exist which compare the relative accuracy of prediction methodologies, as applied to the same data. Farrington and Tarling (1985) have edited a collection of these studies, evaluating statistical methods of prediction in criminology. The methods include the simple point scoring methods of Burgess and Glueck, multiple regression models and a variety of less-known methods of clustering, or binary segmentation techniques.

The earliest such study was conducted by Simon (1971) to assess the relative efficacy of seven prediction methods, including simple point scoring methods, multiple regression, and five other statistical techniques. The study employed a sample of 539 prisoners released on probation, divided into equal construction and validation samples, to predict reconviction in a three-year follow-up period. Simon's best procedure (stepwise multiple regression) resulted in a multiple correlation of .17, prediction error of 42% $[(FP + FN) / \text{total predictions}]$, and a Goodman and Kruskal γ of .24, as measured on the validation sample. No other method produced significantly different results. She concluded that no method was superior to any other method in identifying the 43% actual recidivism.

Tarling and Perry (1985) extended this study, using Simon's

data and a more recent sample with a base rate of 58%, to include three additional methods, one of which was logistic regression. They also concluded equal efficiency in the performance of the ten methods, reporting no results significantly different from those of Simon.

Wilbanks (1985) used five methods, also including Burgess, Glueck, and multiple regression, to predict failure on parole. Parole failure occurred with a high base rate of 67% in a follow-up period of five years. He attained 77% correct classifications (prediction error of 23%), with an average false positive ratio of .21 to 1, and a multiple correlation of .57. Once again, no method was found consistently superior to any other in cross-validation.

Other researchers have reached the same conclusion; all of the methods involved in the comparisons perform about equally well when applied to the same data and cross-validated (Gottfredson & Gottfredson, 1985; Farrington, 1985; Farrington & Tarling, 1985). These consistent results support using the most mathematically developed of these methods, multiple regression, or rather its variant, discriminant analysis, as a single baseline measure for comparison with a neural network approach.

The second collection of comparative studies used Fisher's (1936) method of linear discriminant analysis, a variant of the linear regression approach, as a baseline for comparison with newer methods that take advantage of the increased processing

and searching capabilities of modern computers. Weiss and Kulikowski (1991) used four sets of data for an empirical comparison of statistical pattern recognition, neural networks, and machine learning systems. The four data sets include Fisher's original iris problem, a standard test for discriminant analysis, which discriminates between three classes of iris' using four predictor variables representing physical characteristics of the flowers. The second data set involved a prediction of appendicitis from seven laboratory tests. A third data set was based on nine tests for breast cancer recurrence, each of weak predictive value. Fourth, data collected on 22 medical tests were used to diagnose hypothyroidism, which occurs with a very low base rate of 8% in individuals suspected of the disease. Extensive cross-validation was conducted on each of five statistical methods, two neural network methods, and two classes of machine learning methods. It is beyond the scope of this project to include the details of many of these methods. Therefore, this discussion will be restricted to the results of extensive cross-validation obtained in the comparison of the linear discriminant model and a backpropagation neural network.

Figure 4 displays the error rates of each of these two methods and base rate prediction, across all four data sets. Base rate error is the degree of error resulting from predicting all cases to fall into the modal class (Hair et al., 1987; Meehl & Rosen, 1954). The iris data are clearly the most

discriminable data, resulting in striking improvement over base rate with either method. This is hardly surprising for the linear model, since these data are those used originally by Fisher in his development of the linear discriminant model.

Notice that the accuracy of the linear model and neural net model was quite similar with the iris, cancer, and appendicitis data. The most interesting result was found with the thyroid data. This data set had an extremely low base rate, of only 8%. The linear model was only slightly superior to base rate accuracy, whereas the neural net model was substantially more accurate, with an error rate of .0146. Another difference in the thyroid problem is that it used a much larger set of predictor variables (21) than did the iris problem (4), the cancer problem (7), or the appendicitis problem (9). This supports the contention that neural networks are potentially useful for fitting a model to data with very low base rates and a large set of predictor variables, both characteristic of the prediction of violence.

Odom and Sharda (1990) compared the efficiency of a backpropagation neural network and discriminant analysis on the prediction of bankruptcy. Using five financial ratios as inputs, five hidden units, and one output unit, the neural net was found to outperform the linear discriminant model on three samples, varying in base rate of bankrupt firms from .50 to .10. The superiority of their neural net model held on all measures

of performance, demonstrating greater robustness, higher consistency across decreasing base rates, and lower false positive rates. They concluded that neural nets hold promise for problems of prediction.

The relative performance of a neural network model and a Box-Jenkins automatic forecasting expert system was conducted by Sharda and Patil (1990). For a set of 75 time series predictions, the two methods performed about equally as well, which was an important result considering the high level of complexity and expertise involved in the Box-Jenkins forecasting system, and the relative simplicity of the neural net procedure. Given the potential theoretical advantages, and empirical support for the promise of neural nets as a prediction methodology, there is sufficient support for proposing a neural net approach to the prediction of human behavior as well.

Issues in the Prediction of Violence

The prediction of violence has received the attention of researchers across several disciplines for more than sixty years (e.g., Borden, 1928; Burgess, 1928; Walker, Hammond, & Steer, 1971; Jones, Beidleman, & Fowler, 1981; Black & Spinks, 1985; Klassen & O'Connor, 1988), yet it remains "the greatest unsolved problem the criminal justice system faces" (Rector, 1973; cited in Monahan, 1981, p. 21). The problem of judging the likelihood that an individual will engage in future violence is ubiquitous in the criminal justice system.

Shah (1978) delineates fifteen different occasions in the legal process at which such likelihood judgments must be considered, including, for example, decisions concerning bail, sentencing, parole, and involuntary commitment. The system relies heavily on the judgment of mental health professionals for such estimates of dangerousness, in spite of the fact that these professionals generally acknowledge their lack of ability to reliably make such judgments.

Equally impressive is the sheer volume of cases requiring such decisions. Consider that in 1988, nearly 14 million arrests were made in the United States (Flannagan & Maguire, 1990), each requiring decisions of detention or release, prosecution or not. In 1975, 1.5 million adults were placed on trial, of which 1 million were convicted (Gottfredson, Hindelang, and Parisi, 1978), each requiring a decision of penalty. Of those convicted, 190,014 were incarcerated, requiring many placement and security decisions (Megargee & Bohn, 1979). Decisions regarding parole for those already in prison, 600,000 on an average day in 1988, add to this number, as do decisions regarding the appropriate conditions for the 316,326 prisoners released into the community the same year (Flannagan & Maguire, 1990).

Thus, any tool for use in this massive number of decisions must be practical in terms of its applicability, efficiency, and economy (Megargee & Bohn, 1979). A valid and reliable

instrument for the prediction of violent behavior that meets these practical criteria would be of immeasurable societal value.

Definitions of Violence

There is much discussion and no agreement upon a definition of violence in the criminology literature (reviewed by Farrington, 1982; Megargee, 1979, 1982; Monahan, 1981). Definitional issues include terminology ("violence" vs. "dangerousness"), the scope of behaviors considered violent, types of violence ("angry," "instrumental," "criminal"), legality, intentionality, and targets (persons, property, animals).

For the purposes of this project, the following operational definition of violence will be adopted. Violence is operationally defined, for ease of implementation, according to the type of crime of which an individual has been convicted, and for which he was incarcerated (the "instant offense"). A subset of offenses previously selected from the National Crime Information Center Uniform Offense Codes by Megargee (1982), was selected for this project. This list of violent offenses includes offenses committed against one or more other persons, that carry a high probability of relatively serious physical injury, or actual physical harm to the person(s). Six categories of such offenses include all forms of homicide (except negligent manslaughter), kidnapping, sexual assault

(except nonforcible statutory rape), robbery, aggravated assault, and those forms of arson which endanger life (see Appendix A for a complete list of specific offenses).

Predictor Variables

Although there exist many potential variables (e.g., testosterone levels, genetic variables, skin conductivity, EEG abnormalities) that have been found significantly related to violent behavior, this discussion will be restricted to those variables most likely available in official records, and therefore plausible for use in this application.

Demographic variables

Gender is somewhat trivial in consideration; violent crime is almost exclusively a male phenomenon (Monahan, 1981). Whereas males comprise 48% of the general population, 95% of the prison population is male (Langan, 1991), and nine out of ten violent crimes are committed by males (Webster, 1978). This factor will be held constant in the present study, which will use exclusively male subjects.

Age is one of the most powerful predictors of violence (Monahan, 1981; Petersilia, Greenwood, & Lavin, 1977; Black & Spinks, 1985; e.g.). The relationship of age to violence is an inverted U-shaped function, heavily skewed to the young (Hoffman & Beck, 1985). Males in their twenties comprise 24% of the population, and 50% of the prison population (Langan, 1991). With regard to homicide, in particular, 59.3% of all arrests in

1973 were of males aged 15 to 29 (Shah, 1978).

Race, although a sensitive factor with regard to the implementation of any prediction device, must not be ignored as a factor in research. Silberman (1978) found that the racial difference (nonwhite vs. white) was at least four times greater for violent offenses than for nonviolent offenses, across all ages. Blacks, in particular, comprise 11% of the population, 48% of the prison population, and 46% of all arrests for violent crimes (Langan, 1991).

Other demographic/socioeconomic factors gleaned from the literature include preprison income level (Wolfgang, Figlio, & Sellin, 1972; 8 of 9 studies reviewed by Pritchard, 1977), occupation (Wolfgang & Ferracuti, 1967), and geographic location (Newman, 1979).

History of violence

By far the most ubiquitous factor found significant in the prediction of violent crime, is an individual's history of violence (American Psychiatric Association, 1974; Shah, 1978; Steadman et al., 1978; Wolfgang, 1978). Historical factors that have been considered include past convictions, in terms of both frequency (Monahan, 1981) and type (Black & Spinks, 1985), and number of previous arrests (for any cause; Shah, 1978; Monahan, 1981; Klassen & O'Connor, 1988). Wolfgang (1978) found that for individuals with four previous arrests (for any reason), the probability of a subsequent arrest was 80%; for ten previous

arrests, the probability was 90%. Shah (1978) determined that with five previous arrests, the probability of a future arrest approached unity.

Wolfgang (1972; 1978) followed a birth cohort, comprised of all males born in a single year in the city of Philadelphia. He followed these males until they were 30 years of age, and found that 6% of the cohort were chronic criminal offenders. Moreover, this 6% of the sample accounted for 71% of all homicides, 77% of all rapes, 70% of all robberies, and 69% of all aggravated assaults committed by the age cohort as a whole. This suggests that violence is somewhat concentrated in a small subset of offenders. Are there distinctive markers that distinguish this group from other criminal offenders?

Prior convictions also predict subsequent conviction (Walker, Hammond, & Steer, 1967; Hirschi & Hindilang, 1977; Farrington, 1982; Hoffman & Beck, 1985): 40% probability with two priors, 44% with three, and 55% with four or more prior convictions (Walker et al., 1967). The number of previous commitments of more than 30 days in either a juvenile or adult institution (Wenk et al, 1972; Hoffman & Beck, 1985), length of the most recent commitment free period, and criminal status (Hoffman & Beck, 1985) such as probation, parole, confinement, or escape at time of current offense, all merit consideration as relevant predictors.

Further factors concerning history of violence are age at

first police contact; before age 15 is highly predictive of future violence (Wolfgang, 1972). The mean age at first arrest was found to be 14.4 years for violent offenders (Hamparian, Schuster, Dinitz, & Conrad, 1978; State of Michigan, 1978). Finally, with regard to history variables, child abuse and parents engaging in physical fights are considered relevant by Klassen and O'Connor (1988).

Psychometric variables

Two general categories of psychometric factors have been found significantly associated with violent behavior: intelligence and personality traits. Low intelligence (Wolfgang et al., 1972; Hirschi & Hindelang, 1977; Farrington, 1982) and mental retardation (Klassen & O'Connor, 1988) are correlated with violence.

The Minnesota Multiphasic Personality Inventory (MMPI) has been used as the basis of a typology which differentiates ten different types of criminal offenders (Megargee & Bohn, 1979). The resultant ten types were subsequently found to differ significantly on five measures of recidivism (Megargee & Bohn, 1979). When the types were used to segregate predatory inmates from those most likely to be victimized, significant reductions in the overall amount of violence in the prison resulted, and assaults that did occur were isolated to predictable areas of the prison (Bohn, 1978). Although the ten types did not differ significantly in the violence of the offenses for which they

were incarcerated, they did differ on other measures of criminal behavior patterns (Megargee & Bohn, 1979). These results would suggest that at least some markers or traits that distinguish inmates with violent tendencies can be detected on the basis of MMPI profiles.

The MMPI is an empirically derived inventory, researched over five decades. It consists of ten clinical scales measuring various personality dimensions, and three validity scales measuring test taking attitudes that could influence the validity of scores on the clinical scales (Dahlstrom, Welsh, & Dahlstrom, 1972). Characteristics of individuals with elevated scores are well known. MMPI research concerning the prediction of aggressive behavior distinguishes between two categories of the clinical scales. Scales 4, 6, 8, and 9, are thought to suggest lack of impulse control. Scales 1, 2, 3, 5, 7, and 0, suggest control and inhibition of impulses (Graham, 1977). The classic "49" code (scales 4 and 9 most elevated) has long been associated with impulsive, hedonistic, and delinquent behavior--generally asocial or antisocial tendencies (Graham, 1977; Megargee & Bohn, 1979). Yet more recent research has revealed that the "49" code does not necessarily suggest physical harm to others. This evidence points to a "43" code as most associated with violent, assaultive behavior (Davis & Sines, 1971; Persons & Marks, 1971). A person with this profile is expected to be excessively inhibited until hostility reaches

a level such that these inhibitions are overwhelmed, resulting in bursts of aggressive, assaultive behavior (Graham, 1977; Megargee, 1973). This characteristic is referred to as "Overcontrolled-Hostility" and an auxiliary scale, the O-H scale (Megargee, Cook, & Mendelsohn, 1967), can be scored from the MMPI items to measure this characteristic.

A study conducted by Jones, Beidleman, and Fowler (1981) was successful in accounting for 34.9% of the variance between prisoners who were violent while in prison, and those prisoners who were not violent, on the basis of 22 MMPI scales (basic & auxiliary scales) and demographic data. The following basic scales contributed most to group membership (discriminant load values greater than .40): F, Pa (6), Pt (7), and Sc (8); followed by Ma (9), and auxiliary scales PaV, HOS, and FAM (load values greater than .35). PaV is a parole violation scale developed by Panton (1962). Manifest Hostility (HOS) and Family Problems (FAM) are two of Wiggin's (1969) content scales. Jones and colleagues report correct classification of 72.9% of the violent, and 80.6% of the nonviolent. It should be noted, however, that this fit was obtained on the construction sample, and not cross-validated.

The three validity scales have also been found related to antisocial or criminal behavior. Megargee and Bohn (1979) suggest that the F scale is second only to scale 4 + .4K (correction term) in such prediction. They further suggest that

the "?" scale (Cannot say), in which items are marked both True and False, or omitted, is relevant for criminal offenders. Finally, Black and Spinks (1985) also determined the F scale to be a significant correlate of criminal recidivism.

Substance abuse

Although the abuse of substances is not intrinsically criminogenic, it is thought to interact with socioeconomic factors (Mednick, Pollock, Volavka, & Gabrielli, 1982), and has been found a significant correlate of criminal violence in numerous studies. Heroin or opiate use was significant in nine of nine studies reviewed by Pritchard (1977), and at least six other studies reviewed by this author. Alcohol abuse has received a similar level of attention in the literature in its relationship to criminal violence (Farrington, 1982; Mednick et al., 1982; Petersilia et al., 1977; Pritchard, 1977; Wolfgang, 1958). Monahan (1981) and Mednick et al. (1982) suggest that both opiate and alcohol abuse may suppress factors that would otherwise inhibit violence.

Other substances considered relevant in promoting violence include amphetamines (Ellinwood, 1971; Moyer, 1976), phencyclidine, or PCP (Smith, 1980), barbiturates (Mednick et al., 1982), and benzodiazepines (Moyer, 1976). Therefore it would seem reasonable to include any information that is available regarding inmates' history of substance abuse.

Situational variables

The most recent trend in the prediction of violence is to emphasize the need to include situational variables (Klassen & O'Connor, 1988; Monahan, 1981), especially interactions between situational and personality factors (e.g., Bem & Allen, 1974; Mischel, 1973). Although there has been extensive discussion (Monahan, 1981, e.g.), there has been very little empirical effort. Such information is simply very difficult to obtain through traditional means (official records), and very expensive, if available, through alternative means (extensive interviews, etc.).

Theoretically, an individual, predicted to be dangerous, if released into a stable, supportive environment is likely to become a false positive (Cohen, Groth, & Siegel, 1978; Waller, 1974). The same individual, on the other hand, if released into a stressful environment will often recidivate (Klassen & O'Connor, 1988; Monahan, 1981).

It is apparent that some situations serve as an environmental stimulus leading to a violent response in some individuals, while the same situation does not instigate violence in other individuals. This would argue against radical situationalism. The purpose of this project is to identify traits that differentiate these two classes of individuals. Thus, as a means to this end, it would seem appropriate to glean from the available records any information that is present,

representing situational factors that pertained to individual inmates at the time of their offense.

Situational factors may be roughly categorized into three categories: family factors, peer group factors, and employment factors. Some information regarding these three categories may be obtained from an instrument named the Checklist for Analysis of Life History of Adult Offenders (CALH), developed by Quay (1984) as part of a battery of instruments designed to classify offenders into five groups, for the purpose of institutional custodial and program placement decisions. Information for this instrument is obtained by a case manager, who utilizes information contained in a presentence report to complete the checklist.

The following CALH items would seem to reflect situational information in the aforementioned categories.

Family Factors:

- 15. Claims offense was motivated by family problems.
- 20. Single marriage

Peer Group Factors:

- 1. Has few, if any, friends.
- 16. Close ties with criminal elements

Employment Factors:

- 11. Irregular work history (if not student)
- 14. Supported wife and children

- ___ 23. Suffered financial reverses, prior to
commission of offense for which
incarcerated.

Each of these factors, or at least a similar one, has been found significant in one or more studies: current relationships with parents and siblings (Klassen & O'Connor, 1988); marital status (State of Michigan, 1978); peer influence (Bandura, 1969), and anti-social peer groups (West & Farrington, 1977), or gang involvement (Redl & Wineman, 1957; Wheeler & Caggiula, 1966; Monahan, 1981); and preprison employment instability (Klassen & O'Connor, 1988; Pritchard, 1977; West & Farrington, 1975).

Statement of Purpose

The general purpose of this dissertation is to evaluate the potential contribution of artificial neural networks to problems in the prediction of human behavior. Specifically, it is designed to answer two questions:

- a. Will a backpropagation neural network model offer higher relative efficacy in the prediction of a low base rate criterion, violent behavior, than a linear discriminant model?
- b. Will the resultant weight matrix from the trained neural network model offer information of value concerning the relative contribution of individual predictor variables?

METHOD

Subjects

A preexisting database maintained by the Oklahoma Department of Corrections (DOC) served as a subject pool for this study. The database comprised a random sample ($N = 1233$) of all prisoners who were received for assessment and placement in state prison facilities during the 1983 calendar year. Each of the prisoners in the database was administered a battery of achievement, intelligence, and personality assessments during a 10-day routine assessment procedure conducted upon reception into the prison system. Only the results of the MMPI were preassembled, the rest of the assessment results were contained in the personal files for each individual subject, located at 32 different state correctional facilities. Access to the personal files was obtained, by permission of DOC (see Appendix E), to three of the 32 facilities: Joseph Harp Correctional Center (Lexington, Oklahoma), a medium-security facility; Mabel Bassett Correctional Center (Oklahoma City, Oklahoma), which housed the "Closed" classification files for subjects discharged from the prison system (via termination of sentence, parole, or death); and the Assessment and Reception facility (Lexington, Oklahoma), which housed the "Closed" medical files for discharged subjects. A subset of the DOC database was selected according to the following criteria: (a) male; (b) currently either incarcerated at Joseph Harp Correctional Center ($n = 28$),

OR discharged from the prison system ($n = 788$).

Subjects ($N = 400$) were randomly selected from the DOC database subset for use in this study by the DOC identification number, without knowledge of the prisoners' status on the criterion measure. Subjects were eliminated from this sample, and replaced from the remaining subject pool, if (a) their MMPI was invalid ($n = 39$) (Lachar, 1974; validity criteria: "?"[raw score] ≤ 30 , F minus K [raw scores] ≤ 16 , F[t-score] < 100), or (b) files were misplaced or incomplete ($n = 4$). Eight additional subjects were eliminated from the sample due to missing data on five or more of the selected predictor variables ($n = 4$), or ambiguous status on the criterion variable ($n = 4$), resulting in a total sample size (N) of 392 subjects.

Demographic Characteristics of Sample

Demographic characteristics of the sample include a mean age of 28 years ($SD = 9.38$), and mean education (highest grade completed) of 11 years ($SD = 1.86$). Race of subjects was 77% Caucasian, 17% Black, 5% Native American, and 1% Other. Income of the sample subjects was distributed as Less than \$10,000, 74%; \$10,000 - 19,999, 21%; \$20,000 - 29,999, 4%; \$30,000 and over, 1%. Occupation was distributed as 19% Unemployed, 47% Unskilled labor, 30% Skilled labor, 1% Professional, 2% Other, 2% Unknown. Subjects came from residential communities with populations of less than 4,000, 18%; 4,000 - 15,999, 18%; 16,000 - 49,999, 12%; 50,000 - 300,000, 9%; and over 300,000, 42%.

Procedure

Selection of Predictor Variables

An initial sample of 20 inmate files was examined to determine the degree to which each potential predictor variable, discussed above in the review of the violence prediction literature, was consistently available. Forty-eight predictor variables (see Appendix B) were selected, based on their previous significance in prediction models and their availability from DOC official records. Dummy-coding of categorical variables (creating a separate dichotomous variable for each level of the variable) resulted in a total of 60 predictor variables, which were used to develop both neural network and discriminant analysis models. Subsets of the total set of 60 predictor variables were selected by two methods: stepwise discriminant analysis and neural network weight matrix analysis. Each subset of predictor variables was also used to develop both types of models.

Resampling Procedure

A 3-fold cross-validation resampling technique was employed (Weiss & Kulikowski, 1991). The resampling technique makes optimal use of the sample to (a) estimate the true (population) hit rate and (b) use as much of the sample as possible to construct and validate the prediction models. The total sample ($N = 392$) was randomly divided into three test samples: test1 ($n = 131$), test2 ($n = 131$), and test3 ($n = 130$). Three training

samples were then constructed by forming all possible combinations of two of the test samples: train1 ($n = 261$) was formed by combining test2 and test3, train2 ($n = 261$) was formed by combining test1 and test3, and train3 ($n = 261$) was formed by combining test1 and test2 (and randomly selecting one observation from train3 to move to test3 to equalize the subsample sizes). Thus three pairs of Train-Test files were created, and each model (neural network and discriminant analysis) was replicated three times, using a different Train-Test pair to construct and subsequently cross-validate the model for each replication. All results reported are averaged across the three replications.

Data Collection

The procedure for data collection consisted of examining two files per subject: a medical file containing substance abuse and psychometric data; and a classification file containing demographic information, FBI/OSBI "rap sheets" (criminal history), a consolidated record card including offense and incarceration history, presentence and parole eligibility investigation reports. A total of 62 observations was recorded for each subject using a data collection form constructed to accommodate all observations for a single subject. The observations included demographic information (8 items), psychometric measures (22 items), criminal/violence history (11 items), situational factors (8 items), substance abuse (12

items), and the violent/nonviolent nature of the instant offense (1 item).

Neural Network Training

Apparatus

All neural network computer simulations were conducted on an 80386 personal computer, with a math coprocessor. A commercially available neural network software package was used to run the simulations. The software uses the standard backpropagation feedforward architecture and training algorithm. Neurodes on adjacent layers (e.g., Input Layer and Hidden Layer, Hidden Layer and Output Layer) are fully interconnected, that is, there exists a numerical weight representing the strength of the connection between each possible between-layer pair of neurodes.

Connection weights are adjusted according the generalized delta learning algorithm, which is affected by two learning parameters, the learning rate (η) and the momentum term (α). The software allows the user to set and adjust these parameters to reduce training time and increase the likelihood of convergence to a global minimum error value.

A third parameter, the training tolerance bandwidth, is set by the user to determine the degree of error that is tolerated for each input-output training pattern, in order for the software to count that pattern as "correct" during the training procedure. A training tolerance of .10, for example, will count

an output for a single training pattern as correct, if the absolute value of the difference between the output and target is less than or equal to .10. Outputs counted as correct cause no adjustment of the weights. Following each epoch, the numbers of "correct" and "incorrect" patterns are displayed.

Convergence of the network is attained when all patterns are designated as correct. The update per epoch of this display allows the user the ability to monitor the network's progress as it trains.

The software interface also provides on-screen histograms, displaying the frequency distributions of connection weights across their range of -8.0 to +8.0, separately for each layer of connections. These histograms also serve as visual aids by which to monitor the progress of training in terms of capacity for further weight adjustment, and hence improvement in accuracy of classification.

Architecture

Input Layer. The input layer of each neural network comprised one neurode for each **continuously valued** predictor variable, taking on the value of the predictor measure, such as intelligence test score or number of violent arrests. Missing values were replaced with the mean value of that measure for the entire sample (see Appendix B for a list of predictor variables, and the number of cases lacking values on each predictor variable). **Rank-ordered categorical** predictor variables, such

as income, were similarly represented with one neurode per variable, taking on values indicating the number of the interval category in which the observation fell, or the mean value of that measure for the entire sample if missing. **Discrete categorical** variables, such as marital status, were represented with a group of neurodes, one neurode per level of the variable, with each neurode taking a value of 1 to indicate category membership, 0 to indicate non-membership, and 0 on all neurodes in the group if the value were missing. **Dichotomous** predictor variables were represented with one neurode, taking values of 1 ("present"), 0 ("absent"), or overall sample mean if missing.

The decision to represent missing data with the overall sample mean of each variable was made based on the results of a two-tailed test for significant differences of two proportions, the proportion of data missing for Violent cases and the proportion of data missing for Nonviolent cases. Of the ten variables for which there were missing data, three significant differences in proportions per category were found: child abuse ($p < .05$), irregular work history ($p < .05$), and Beta IQ ($p < .01$). The mean of each variable used to replace cases with missing values was calculated on the entire sample ($N = 392$), rather than on the separate groups, to avoid biasing the results in favor of the category for which there was a smaller proportion of data missing.

Different networks were trained with input layer sizes of

60 neurodes representing the entire set of predictor variables, and subsets of 53, 29, 17, and 10 neurodes representing selected subsets of the entire set of predictor variables. Methods for selection of subsets of variables are described in a subsequent section of this report.

Hidden Layer. Networks with input layers of 60, 53, 29, 17, and 10 neurodes were each trained with a single hidden layer. The number of neurodes in the hidden layer for each network of different-sized input layers, was determined by an empirical "complexity fit" procedure. The objective of the complexity fit was to find the smallest number of hidden units necessary to yield the best generalization, as measured by proportion correct classification on the test samples.

Preliminary network construction, using the entire set of 60 input neurodes, indicated rather small differences in hit rates with very different sized hidden layers. Therefore, a wide range of values for the number of hidden neurodes was tested for each value of input neurodes. If a network produced a substantially better fit (e.g., $M + 2 SD$), then smaller steps were used in the procedure in an attempt to close in on an optimum number. The results of this procedure were that for 60-input networks, 15 different hidden layer values were tested (ranging from 5 to 200 neurodes), producing a mean hit rate of 0.71 ($SD = 0.04$). Hidden layer sizes of 25 and 10 neurodes produced the best performance on test1, with 79.4% and 73.3%

correct classifications, respectively. Therefore, for replications 2 and 3, only hidden layer sizes of 25 and 10 neurodes were trained and tested. A similar search procedure was used to achieve the best complexity fit for 17-input networks, resulting in an optimum hidden layer size of 12 neurodes.

The networks with 53, 29, and 10 inputs were trained with hidden layer sizes equal to one-half the number of inputs, or 26, 15, and 5 hidden neurodes, respectively. The use of the "one-half the number of inputs" rule to determine the hidden layer sizes of these latter networks was based on the relatively small variation in accuracy found in the previous two complexity fits (when averaged across the three replications), and to reduce the overall number of networks to be trained in order to evaluate differences in networks with varying number of inputs.

The effect of adding a second hidden layer to a network with the same 60 inputs was tested by training nine additional networks, with varying combinations of numbers of neurodes in each of two hidden layers. Networks with the following combinations of hidden layer sizes (Hidden Layer 1/Hidden Layer 2) were tested on Replication-1: 30/15, 25/10, 15/10, 10/10, 10/5, 12/8, 12/6, 13/9, and 6/3. The three combinations which produced the best results on test1 data (15/10, 13/9, and 12/8) were then replicated with test2 and test3 data.

Output Layer. The output layer of every network comprised

two neurodes, one for each class of the criterion measure, as recommended by Weiss and Kulikowski (1991) and others (e.g., Lippmann, 1987; Rumelhart & McClelland, 1986). During training, the outputs of these two neurodes were compared to target values of 1 and 0, respectively, for Violent cases, and 0 and 1 for Nonviolent cases.

Learning Parameters

No consistent rule-of-thumb exists for determining the optimal values for the two learning parameters; the learning rate, and the momentum term. The choice for learning rate typically affects only training time, not whether convergence is actually achieved. The momentum term is included to reduce the likelihood of convergence to a local minimum. Optimization is particular to the application. Therefore, some preliminary networks were trained using the entire data set ($N = 392$), for the purpose of investigating values of these parameters. Values of .5 for the learning rate, and .9 for the momentum term, were found to lead to rapid convergence and stable training characteristics. Several other combinations were tried, with no improvement in training, therefore the values of .5 and .9 were used throughout training for all networks.

Extent of Training

Each network was trained until it converged at its minimum training tolerance level. The minimum training tolerance for most of the networks was .02, that is, the network converged at

.02 but failed to do so at .01 training tolerance. With extended training, some of the networks converged at a training tolerance of .01. Each network was saved nine or ten times at decreasing training tolerances, beginning with .49, and saving at successively smaller training tolerances of .40, .30, .20, .10, .05, .04, .03, .02, and occasionally .01. The purpose of these successive "saves" was to allow for testing after each network's training was complete, and to capture the optimum degree of training in terms of the network's performance on the test data set, without "overfitting" the network model to the training data. Overfitting of a neural net model may be thought of as the network's "memorization" of the training data, to the detriment of performance on the test data (Klimasauskas, 1991c). Each saved state of the network was later tested on the test data; the state which yielded the highest cross-validation accuracy was selected for further analysis.

Weight Matrix Analysis

A separate network was trained, using the entire data set ($N = 392$) and all 60 predictor variables, for the purpose of analyzing the trained weight matrix for information regarding the relative influence of each input variable on the output of the network. The analysis was conducted according to Garson's (1991) technique for partitioning the weights connecting each hidden neurode to the output neurodes into the relative proportion, or "share," contributed by each input neurode

according to the following equation (p. 50):

$$share_{I_v} = \frac{\sum_j^{n_H} \left(\frac{I_{vj}}{n_v} O_j \right)}{\sum_i^{n_v} \left(\sum_j^{n_H} \left(\frac{I_{vj}}{n_v} O_j \right) \right)} \quad (6)$$

where n_v = the number of input variables,

n_H = the number of hidden units,

I_{vj} = the weight connecting input unit i and hidden unit j ,

O_j = the weight connecting hidden unit j and the output

unit.

Criterion Variable

The criterion measure of violent behavior was determined by the offense for which each subject was convicted and incarcerated at the time of assessment in 1983. This "instant offense" was classified as Violent according the previously described categories of violent offenses (see Appendix A for a list of offenses classified as Violent). Any other offense was classified as Nonviolent. The mean time elapsed between the date of arrest for the crime and date of assessment was 175 days ($SD = 250$) for the entire sample; 164 days ($SD = 239$) for the nonviolent cases, and 204 days ($SD = 275$) for the violent cases. The difference between mean time elapsed for the two groups was

nonsignificant, $t(390) = 1.392, p > .05$.

Design

For reasons of practical necessity, this study employed a retrospective design. That is, rather than actually predicting future occurrences of violent behavior, the models were developed to "postdict" violence for cases in which the outcome was already known.

A retrospective design is not without inherent weaknesses, the most important of which is retrospective bias. This source of bias, associated with the knowledge of the outcome of the criterion behavior, however, was minimized in this study in three ways. First, by restricting the source of predictor variables to official records, one can presume that the personnel entering data in the records did so in a clerical fashion, independent of the nature of the conviction. Second, predictor variables were composed of pre-offense data, or data for which an individual's status remains unchanged as a result of the nature of the offense. Third, it is not in the nature of machine learning to process the validation sample data any differently depending on the outcome, of which the machine is, of course, ignorant.

A second potential weakness of a retrospective design is one of sampling bias. This source of bias was controlled by selecting subjects randomly from the prison population, without concern for their status on the criterion. Rather, random

subject selection was affected only by the presence or absence of the set of predictor variables of interest. There was no *a priori* reason to suppose that this availability was in any way contingent upon the violence status of the instant offense for a given subject.

Although these reductions in retrospective bias do not equate the validity of the design to that of a prospective design, this design had the potential for extracting useful variables that might be incorporated in the future, in a longitudinal prospective study. To this extent, the retrospective design is defensible.

RESULTS

Base Rates

The overall base rate of violent cases for the entire sample ($N = 392$) was .27. Although divided randomly, the three test data sets ($n = 131$) resulted in very similar proportions of violent cases: .27, .28, and .25 for test1, test2, and test3 data sets, respectively. Please note that, unless specifically stated otherwise, all results reported are averaged across these three replications.

Selection of Subsets of Predictor Variables

A stepwise discriminant analysis of the entire data set, yielded a subset of 17 variables, using an F statistic ($p \leq 0.15$) as the criterion for selection. The discriminant model developed by the stepwise procedure using the 17 predictor variables accounted for 31.7% of the total variance in the criterion variable. This subset of 17 predictor variables was then used to develop both neural network and discriminant analysis models.

Three further subsets of predictor variables were selected based on the Garson (1991) method of analyzing the weight matrix from a neural network. The result of the analysis was a rank-ordering of the 60 predictor variables, based on the relative influence of each variable on the output of the network. Subsets of 53, 26, and 10 variables were chosen based on cutoff values of the proportions, greater than or equal to .01, .015,

and .02, respectively. Each of these subsets was also used to develop both neural network and discriminant analysis models.

For ease of presentation, the notation adopted for referencing individual models will take the form: $Mn_I - n_{H1} - n_{H2}$, where M represents the Model type (D=Discriminant Analysis, N=Neural Net); n_I = number of Inputs for the model; n_{H1} = number of units in Hidden Layer 1, and n_{H2} = number of units in Hidden Layer 2, if the model is a neural network. For example, "D60," "N60-25," and "N60-15-10", refer to Discriminant Analysis Model with 60 inputs, Neural Network Model with 60 inputs and 25 hidden units, and Neural Network Model with 60 inputs, 15 units in Hidden Layer 1, and 10 units in Hidden Layer 2, respectively.

Neural Network Training Characteristics

A total of 112 neural networks was trained during the course of this study, each of which was saved and tested at approximately ten different training tolerance levels. As a result of training each network to produce the highest Total Group Hit Rate based on cross-validation (test1) results, the extent of training across the "best" networks of different sizes varied considerably. The mean number of epochs necessary for achieving this criterion was 2,549 (minimum [N60-15-10] = 138, maximum [N17-12] = 6,042). The mean number of epochs for each network to achieve the maximum hit rate **and** to reach the *minimum* training tolerance was 3,270 epochs (minimum [N53-26] = 1,244; maximum [N17-12] = 6,568). The training tolerance levels that

produced the best test results also varied over the three cross validations. For N60-25, optimum training tolerance levels were .04, .05, and .02; for N53-26, .10, .02, and .02; for N29-15, .10, .10, .05; for N17-12, .05, .10, and .10; and for N60-15-10, .40, .40, and .20; for test1, test2, and test3 of each model, respectively.

Total Group Hit Rates

Concept of Chance

For classification systems with classes of unequal proportions, the meaning of "chance" accuracy of prediction is ambiguous. Two different criteria have been recommended (Huberty, 1984), neither of which seems completely appropriate for judging the improvement of a particular prediction model over the accuracy one could expect by chance alone. Therefore, a discussion of each criterion, and the position on this issue adopted for purposes of this study seems warranted.

The first criterion, the *maximum* "chance" criterion (Huberty, 1984; Meehl & Rosen, 1955; Weiss & Kulikowski, 1991) is equal to the accuracy one could achieve by simply applying the base rate alone to the problem, and predicting all cases to belong to the more frequent class, .73 in the present study. The problem with adopting this criterion is that it is not based on a prediction system at all, and is useless in any practical prediction situation, where the cost of false negative predictions would be far too high to consider its application.

Consider the results if one were to predict all parole candidates to be Nonviolent, and therefore suitable for release; or all patients suspected of cancer to be negative, and therefore unnecessary to conduct a biopsy. The maximum criterion is thus a hypothetical entity, which *overestimates* a more realistic chance criterion.

The second criterion, recommended by Huberty (1984), is a *proportional* "chance" criterion. By this criterion the total-group chance hit rate is equal to the sum (over groups) of the products of the sample proportion for each group, and the number of sample cases in each group, divided by the total sample size, or $[(.73 \times 287) + (.27 \times 105)] / 392 = .61$ in this study. This level of accuracy would result from a prediction system which had a classification bias equal to the base rate, but *no* valid information on which to base its predictions. This criterion also constitutes a hypothetical entity; it would produce worse total-group accuracy than the maximum criterion, but at least it would detect some of the members of each class.

Although, neither the maximum nor the proportional criterion seems appropriate as a proper baseline for comparison of the prediction models in this study, the more realistic alternative is not available. On this basis, both criteria will be used to "bracket" the decision space in which one could expect a prediction system to operate, and will thus be referred to as the "expected hit rates" per each criterion. The standard

error of measurement calculated on the base rate information, was used to calculate normal (z) statistics to quantify, separately, the differences between each model's total group hit rate and the two criteria. The standard error of measurement [$SE = \sqrt{[E(1-E)]/N}$, where $E = .27$ as defined by the base rate] is 0.022.

Entire Set of Predictor Variables

The neural network (N60-25) which produced the best results, given the entire set of predictor variables represented by 60 inputs, used a single hidden layer of 25 neurodes, and produced a Total Group Hit Rate of 1.0 on the training data, and .756 on the test data. (The Hit Rate on the training data was 1.0 for all neural net models, therefore, will not be reported in future results.) The best 2-hidden layer network (N60-15-10) did not perform any better (Hit Rate = .751) than the 1-hidden layer network, therefore its results and other 2-hidden layer networks were excluded from further analyses. A discriminant analysis model (D60), developed on the basis of the identical 60 inputs, produced a Total Group Hit Rate of .843 on the training data, and .730 on the test data. This model, D60, accounted for 38% of the variance in the criterion.

Subsets of Predictor Variables

Selected by Stepwise Discriminant Analysis

A stepwise discriminant analysis ($p \leq .15$ to enter, $p \leq .15$ to stay) produced a subset of 17 variables of the original set

of 60 predictor variables (see Appendix C for a list of this subset of variables). When the 17 variables thus selected were used to construct a second discriminant analysis model (D17), a Total Group Hit Rate of .920 was achieved on the training data, .791 on the test data. A second neural network (N17-12), using the same set of 17 inputs produced a Total Group Hit Rate on the test data was .761.

Of this subset of 17 predictor variables, ten were positively associated with violent behavior (income; violent arrests; MMPI clinical scales 3, 8, and 0; two levels of marital status--married and "live with"; supported family; irregular work history; and use of benzodiazepines). The remaining seven predictor variables were negatively associated with violent behavior (age; one level of race--Native American; unskilled labor; two levels of criminal status--probation and parole; Beta-IQ; and MMPI clinical scale 1).

Selected by Neural Network Weight Matrix Analysis

The neural network weight matrix analysis produced a list of the 60 inputs, rank-ordered with respect to the relative contribution ("share") of each input to the output of each of the two output neurodes (see Appendix D).

Neural network and discriminant analysis models were developed using each of three subsets of the 60 inputs: 53 inputs (share \geq .01), 29 inputs (share \geq .015), and 10 inputs (share \geq .02). Total Group Hit Rates for the neural network

models were .720 (N53-26), and .743 (N29-15). The network trained with 10 inputs (N10-5) failed to converge with a training tolerance of .49 after extended training, hence its results were excluded from further analyses. Discriminant analysis models developed with the same subsets of 53 variables, 29 variables, and 10 variables, produced Total Group Hit Rates of .743 (D53), .740 (D29), and .740 (D10). Table 4 summarizes the Total Group Hit Rates, on both the training data and test data, for each of the four neural network models and five discriminant analysis models. All models produced cross-validation hit rates that were significantly greater than the proportional criterion (.608), $p < .0001$, one-tailed. One model, D17, produced a hit rate that significantly exceeded the maximum criterion (.732), $p < .005$, one-tailed. Tests for differences between correlated proportions (McNemar, 1947) were nonsignificant for the difference between the NN and DA hit rates for a particular number of inputs, $\chi^2(1, N=392) = 1.184$ for the 60-input models, and .007 for the 29-input models, $p > .05$, one-tailed. All other NN - DA hit rates were opposite of the hypothesized direction. Among the DA results, D17 produced a significantly higher hit rate than D60, $\chi^2(1, N=392) = 12.5$, $p < .0005$; D53, $\chi^2(1, N=392) = 6.56$, $p < .005$; D29, $p < .005$; or D10, $p < .005$; all one-tailed tests. Tests for differences between N17 and N60, and between N17 and N29, were nonsignificant. Thus among the NN models, none of the subsets

of variables produced better classification results than the entire set of predictor variables. Table 3 displays the frequencies of the four possible classification outcomes for each of the models, from which the Total Group Hits Rates were calculated. Figure 5 displays these results in relationship to the Total Group Hit Rates expected as per the maximum and proportional criteria.

Insert Table 3 about here

Insert Table 4 about here

Insert Figure 5 about here

Conditional Probabilities

The probability of a correct classification can be conditionalized in four ways: (a) a case was actually violent [$p(\text{corr}|V) = CP/(CP+FN)$], (b) a case was actually nonviolent [$p(\text{corr}|NV) = CN/(CN+FP)$], (c) a case was predicted violent [$p(\text{corr|"V"}) = CP/(CP+FP)$], or (d) a case was predicted nonviolent [$p(\text{corr|"NV"}) = CN/(CN+FN)$]. Conditional probabilities (a) and (b) are equivalent to what Huberty (1987) calls Separate Group Hit Rates for the violent and nonviolent

groups, respectively.

Table 5 displays these conditional probabilities as derived from the classification results of each of the nine (4 neural net, 5 discriminant analysis) models. Also included in Table 5 are the expected hit rates (proportional criteria) for separate groups (Huberty, 1987). The probabilities correct for the Violent and Nonviolent groups were significantly greater (z statistic) than the respective expected separate group hit rates (.27 and .73), $p < .05$ or $.01$ (see Table 5), for all models, with the exception of D10, which failed to meet the expected value for the Violent group and was greater than the expected value for the Nonviolent group. There were no distinguishable patterns of differences between model types. Two results do stand out: the probability correct, given a case was predicted "Violent" was .66 for D17, and given a case was actually Nonviolent was .95 for D10; both results were greater than two standard deviations above the means for the respective conditions.

Insert Table 5 about here

False Positive and False Negative Ratios

Figure 6 shows the ratios of false positive to correct positive predictions (FP/CP) and false negative to correct negative (FN/CN) predictions across each of the nine models.

The minimum false positive ratio across all models was .53 for D17 (approximately one FP for every two CPs, or 1:2), and the maximum false positive ratio was 1.13 (approximately 1:1). The minimum false negative ratio was .20 for N17 (1:5), and the maximum false negative ratio was .32 (1:3).

Insert Figure 6 about here

Increasing Decision Thresholds

The performance of each of the models was further compared at decision thresholds of increasing stringency. For the purposes of this comparison, the values of the two output nodes were combined according to the following formulas:

$$\textit{Combined Output} = \frac{o_V + (1 - o_N)}{2}, \quad (6)$$

where o_V = Violent output, and o_N = Nonviolent output, for **positive** predictions ($o_V > o_N$); and

$$\textit{Combined Output} = \frac{o_N + (1 - o_V)}{2}, \quad (7)$$

for **negative** predictions ($o_N > o_V$).

Positive predictions were counted as correct for a

particular threshold level if: (a) the case was actually violent and (b) the Combined Output was greater than or equal to the decision threshold. All other positive predictions were counted as incorrect. Negative predictions were classified as correct or incorrect in the same fashion.

Figures 7 and 8 display the proportions of correct positive predictions $[CP/(CP + FN)]$ and correct negative predictions $[CN/(CN + FP)]$, respectively, at decision thresholds of .50, .60, .70, .80, and .90.

Insert Figure 7 about here

Insert Figure 8 about here

Notice the general trend toward superior performance of the neural network models, over the discriminant analysis models, as the decision threshold increases. This trend held for both positive and negative predictions. The proportions correct at decision thresholds of .70 or greater for positive predictions, and .80 or greater for negative predictions, were higher for every neural net model than for any discriminant model. Also observe that the proportion of correct positives generated by the best neural net model (N17-12) decreased by only .13 from the least stringent threshold (.50) to the most stringent (.90)

threshold. The corresponding decreases for the best (at .50 threshold) discriminant analysis models, D17 and D53, were .43 and .36, respectively. For correct negative predictions, the degree of separation between the best neural net models and the best discriminant analysis models also increased with increasing decision thresholds of .70, .80, and .90, with separations .03, .12, and .24, respectively.

DISCUSSION

"Best-One-Wins"

The performance of two classes of prediction models, neural network (NN) and discriminant analysis (DA), was first compared, using a "Best-One-Wins" (BOW) decision rule. Both classes of models significantly exceeded the accuracy expected on the basis of the *proportional* criterion. The results for the models using the entire set of predictor variables (60 inputs) showed substantially better classification by the NN model (16% higher hit rate) over the accuracy of the DA model for the construction sample (as was true for all NN models), but no statistically significant advantage in classification accuracy for the NN model over the DA model on the cross-validation sample. The entire set of predictor variables accounted for 38% of the variance in the criterion as calculated on three cross-validation samples--a level 3% greater than that obtained by Jones and colleagues (1981) on their **construction** sample, using the 22 MMPI scales and demographic data to predict intra-prison violent behavior.

Although it would be highly desirable to be able to directly compare the hit rates obtained in this study to those of previous studies, the gross differences in terms of factors discussed previously (e.g., definitions of violence, base rates) preclude any direct comparisons. One prospective study (Wilbanks, 1985), however, did produce comparable classification

accuracy in predicting parole violation for 427 parolees. Wilbanks achieved a 77% total group hit rate, quite comparable to the 76% (NN) and 79% (DA) achieved in this study. It should be noted, however, that in the Wilbank's study, the base rate of failure on parole was 67% (half of which were rearrests for felony charges, and half were reincarcerations for technical parole violations), nearly the complement of the 27% violent base rate in the present study.

When the best subset of variables (17) selected by stepwise discriminant analysis was used to develop both classes of models, the NN performed as well as it had using the full set of predictor variables. This lends some support to the claim (e.g., Hartzberg et al., 1990) that using a large set of intercorrelated predictor variables does not detract from the goodness-of-fit of a NN model. The DA model, however, showed a significant 6% gain in accuracy on the cross-validation sample over the 60 variable model, providing a total group hit rate statistically equivalent to the NN hit rate. This DA model (D17) was the only model that was significantly better in accuracy than expected on the basis of the base rate (*maximum*) criterion. When three additional subsets of predictor variables (53, 29, 10) were selected on the basis of Garson's (1991) method of analyzing the NN weight matrix, all models (NN and DA) performed equally well, but no better than the accuracy obtained with the entire set of predictor variables.

Although the performance of the neural net models did not significantly differ from the performance of the discriminant analysis models overall, one capability of the neural network models is important to note concerning the issue of predicting a low base rate criterion. All of the discriminant analysis models had a built-in advantage, in that the prior probabilities for each class, Violent and Nonviolent, were given as *a priori* information to the models. Thus the discriminant analysis model could use this base rate information in computing Bayesian posterior probabilities. The neural net models, in contrast, had to "learn" this information, simply by processing a given number of examples of each class. This would seem to document an attribute of neural nets that is not possessed by the linear discriminant models. Had the discriminant model been given equal prior probabilities for the two criterion classes, it is unlikely that it would have achieved such high classification accuracy. In most prediction applications, however, the base rate is known *a priori*, at least to the extent it can be estimated from the sample at hand; thus although this attribute of neural networks is interesting, it may not result in any practical advantage for neural networks over the discriminant analysis models in application.

Three conclusions may be drawn from this first set of results using a BOW decision rule. First, in terms of cross-validation total group hit rates, the linear (DA) approach

worked about as well as the nonlinear (NN) approach. Second, stepwise discriminant analysis proved to be a useful strategy for selecting a subset of predictor variables for use with either class of models. Third, using a large set of intercorrelated predictor variables did not detract from the goodness-of-fit of the NN model.

The results of conditionalizing the probability of a correct classification showed that both classes of models significantly exceeded the expected separate group hit rates [$p(\text{corr}|V)$ and $p(\text{corr}|NV)$]. These two conditions may be considered to be of greater consequence, in terms of implementing a prediction model, than the remaining two conditions [$p(\text{corr|"V"})$ and $p(\text{corr|"NV"})$]. The separate group hit rates indicate the types of errors that were actually committed by the models, by quantifying the probability that the model detected those cases that were actually violent and actually nonviolent. The remaining two conditional probabilities measured the accuracy of the model in terms of its internal predictive validity, that is, the probability that a "violent" prediction was correct, or a "nonviolent" prediction was correct. Overall, the 17-input models produced the highest accuracy with respect to all four conditions (with one exception), but there were no distinguishable patterns of differences between the two classes of models. The exception was for model D10, which detected 95% of the Nonviolent cases,

4% higher than the rate for D17. Further inspection, however, reveals that this result was at the expense of the Violent group hit rate which was only 17% for D10, detecting 30% fewer Violent cases than D17. In other words, model D10 approached the results one would obtain by classifying all cases as "Nonviolent," therefore, the high Nonviolent group hit rate should be disregarded as a measure of superiority for model D10.

In terms of false positive rates, once again, model D17 produced the best results, with a rate of .56, or approximately one false positive for every two correct positive predictions. This result, compared to those reported throughout the violence prediction literature (e.g. Steadman, 1980) is quite good, reflecting the quality of the subset of 17 predictor variables. Indeed, nearly all of the models produced less than one false positive for every one correct positive. The results do not, however, demonstrate consistent superiority of one class of models over the other, in terms of false positive rates. False negative rates were comparatively very low, with D17 and N17 producing approximately equal rates of about one false negative prediction for every five correct negative predictions.

Increasing Decision Thresholds

When the performance of NN and DA models were compared at decision thresholds of .50, .60, .70, .80, and .90, a different pattern of results was observed. As the decision threshold increased, the NN models retained their accuracy in terms of

both correct positives and correct negatives, whereas the DA models dropped in accuracy at each successively higher threshold level. This result was consistent with previous results obtained by Gordon (1991a). N17-12, the best NN model in terms of correct positive predictions, was only 13% less accurate in classifying violent offenders with a decision threshold of .90 (41% CP), than with a threshold of .50 (54% CP). Model D17, on the other hand, fell in accuracy from 48% CP predictions at a threshold of .50 to only 5% at a threshold of .90. This general trend held for both positive and negative predictions. Correct negative predictions produced by D17 fell from 91% correct at .50 threshold, to only 41% correct at the most stringent threshold. The corresponding decrease for N17-12 was from 84% to 74%. Performance of all of the NN models exceeded the accuracy of all of the DA models at or above decision thresholds of .70 and .80 for positive and negative predictions, respectively. Thus, when performance of the prediction models was analyzed at increasing levels of decision thresholds, the neural networks showed substantially more accurate prediction than the linear discriminant models, providing strong support for the hypothesis that neural networks can offer more relative efficacy in prediction, given that there is reason to produce predictions with higher degrees of certainty.

The advantage for NN prediction accuracy stems from the capacity of a NN model to be trained to any arbitrary degree of

accuracy (White, 1989). DA models, on the other hand, construct the best model possible in a single procedure of simultaneous processing of the data, minimizing the error in one "epoch." Moreover, these results suggest that prediction accuracy with high certainty may require, for some behavioral prediction problems, a nonlinear model. The linear discriminant model apparently didn't fit the underlying patterns in the data to a degree comparable to that which was obtained with a nonlinear fit.

The practical implications of this set of results would seem substantial. Any agency that would want to implement a prediction device for violent behavior, would surely want to have the highest degree of confidence possible in its predictions. A NN model offers this advantage, and does so in such a way as to allow a given agency the capacity to set its desired decision threshold, and make decisions to release or retain an individual based, in part, on outputs that meet, or fail to meet this threshold, respectively.

Predictor Variables

The second research question posed for this project was to determine if the weight matrix from a NN could provide useful information concerning the relative contribution of each of the individual predictor variables. The Garson (1991) method of weight matrix analysis produced information in the form of a rank ordering of the relative contributions of the predictor

variables, which was used to select subsets of variables. However, the subsets selected from this analysis did not produce better classification results than the entire set of variables. Nor were the results of the weight matrix subsets of predictor variables as good as the subset produced by stepwise DA. Thus, although the weight matrix analysis produced information regarding each of the predictor variables, no support was found for the hypothesis that this information was "useful." The best subset was selected by stepwise DA; and the best DA model, as well as the best NN model, were developed with this subset of 17 variables. This would suggest that stepwise DA is a profitable method for selecting a subset of the total set of predictor variables, to be followed with the development of a NN model on the basis of that subset of variables. When one compares the ranked list generated by the weight matrix analysis to the subset of 17 variables selected by stepwise discriminant analysis, however, one gets a somewhat different picture of which variables were most predictive of violence. The Spearman correlation between the rankings of the top 17 variables from stepwise DA, and rankings for the same variables produced by the weight matrix analysis was .56. This discrepancy in rankings was expected, since stepwise DA was limited to selecting variables on the basis of their linear relationships to the criterion; whereas the NN model had the capacity to select variables in terms of whatever nonlinear relationships produced

the best results.

Seven of the best 17 predictor variables selected by the stepwise DA procedure, also ranked in the top 17 of the variables ranked by the weight matrix analysis. These seven predictors included four positive predictors of violence: income, married, living with a mate, and benzodiazepine use; and three negative predictors of violence: Native American race, probation and parole status. Apparently, these seven variables were of relatively high predictive power for both types of models. Further interpretation of the results of the weight matrix analysis would be premature, since the Garson method has yet to rigorously analyzed, but the results of this study suggest that further evaluation of the Garson method is warranted. The conclusion to be drawn, however, is that the two classes of models were each relying on different variables, yet produced very similar results in terms of overall accuracy (BOW) of classification.

Further inspection of the 17 variables selected by the stepwise analysis reveals that violent offenders in this sample were less likely than nonviolent felons to have been on probation for a previous offense at the time of arrest for the instant offense; more likely to be younger, to have more previous arrests for violent acts, to score higher on MMPI clinical scale 0, and to be married than nonviolent offenders. These five variables were ranked as the top five, respectively,

in terms of partial correlations, and accounted for 20% of the variance combined. Four additional variables, ranked six through nine accounted for an additional 5% of the variance; violent offenders more frequently used benzodiazepines, "lived with" a mate, and were less likely to be Native American, or to have performed unskilled labor than nonviolent offenders. The remaining eight variables in this subset accounted for another 6% of the variance. Those variables positively associated with the criterion were income, irregular work history, MMPI clinical scales 3 and 8, and supported family; parole status, Beta IQ, and MMPI scale 1 were negatively associated with violent offenders. A total of 32% of the criterion variance in the entire data set was accounted for with this subset of 17 predictor variables.

It is interesting to note that all categories of predictor variables were represented in the best 17: demographic (income, age, Native American, unskilled labor), criminal history (probation and parole status, previous violent arrests), psychometric (Beta IQ, MMPI clinical scales 1, 3, 8, and 0), situational factors (being married, or "living with" a mate, irregular work history, and supporting a family), and substance abuse (benzodiazepines). It was surprising, however that substance abuse was not more discriminating, given the extensive support in the violence literature (e.g., Mednick et al., 1982; Pritchard, 1977; Smith, 1980). This result may be due to a

confound present in the only available measure of substance abuse. Although the information came from a medical interview at the time of a physical examination (which should lend credence to the information), the subjects were asked if they had *used* each substance, recently or at anytime in their past. Thus, the measure did not distinguish between use and *abuse*, which, in the case of alcohol, for example, would clearly diminish its discriminatory power. It is also conceivable that benzodiazepine use, the one substance measure that was discriminating, was confounded by the *prescribed* use of such a drug for individuals after they had committed a violent act. This possibility, although speculative, should detract attention away from benzodiazepine use as a predictor of violence, although there was support for its role as such in previous literature (Moyer, 1976); perhaps, at least in this study, it was an "aftereffect" of violent behavior.

Other discrepancies between the variables found predictive in this study and previous literature include the lack of a white-nonwhite racial difference (e.g., Silberman, 1978), child abuse (Klassen & O'Connor, 1988), and age of first arrest (Hamparian et al., 1978; State of Michigan, 1978; Wolfgang, 1972). The direction of association of marital status was opposite of that reported by the State of Michigan study (1978), whose risk assessment scale used single or divorced status as a positive correlate of parole violation (not necessarily by

violence). The negative relationship of being married or living with a mate, to violent behavior, might be explained if one views the responsibility level involved in such a commitment as increasing the general stress level an individual would experience in providing for another person. This would seem even more likely if the individual were of low intelligence level. Another opposite effect was obtained for income level; previous literature (Pritchard, 1977; Wolfgang et al., 1972) indicated low socioeconomic status was a correlate of violent behavior--this study found income to be positively associated with violent offenses. The "highest" level of income for this study was relatively low, however--\$30,000 and above constituted the highest category.

MMPI scales 1, 3, 8, and 0 ranked in the top 17 variables of the stepwise DA subset. Scale 1 was negatively associated with violent behavior in this study. Low scores on scale 1 are generally indicative of a lack of somatic preoccupation, or concern with physical problems or health (Graham, 1977; Dahlstrom et al., 1972). Scale 3, a positive predictor in this study, is indicative of psychological immaturity, desire for attention and affection, and a tendency to deny these and other troubling feelings (Graham, 1977; Dahlstrom et al., 1972). Dahlstrom and colleagues further suggest that scale 3 measures social facility; that high scorers frequently claim that others are untrustworthy, irresponsible, and unlikable. Perhaps more

directly relevant to violent behavior, Graham suggests that high scorers occasionally act out in a sexual or aggressive manner, and furthermore tend to have problems with authority figures, often including a "rejecting father to whom males reacted with rebellion or overt hostility" (p. 40). Scales 1 and 3 are highly correlated; 20 of the 33 items comprising scale 1 are also included in scoring scale 3 ($r = .46$; Dahlstrom et al., 1972). The opposite direction of the relationship of scales 1 (-) and 3 (+) to violent behavior, therefore, is somewhat puzzling. One might consider, however, that the lack of admission to physical problems (low scores on scale 1) and a tendency to deny, or "repress" problems (high scores on scale 3) is consistent.

High scores on MMPI scale 8 seem more logically consistent with violent behavior in that the scale represents a factor of general maladjustment, anxiety, distress, and thought disturbance (Dahlstrom et al., 1972). High scorers tend to be socially isolated and alienated from peers, impulsive, and lacking in problem solving skills (Graham, 1977). Scale 8 was the only one of these scales that had previously been associated with violent behavior, and then usually in combination with scale 4 (Jones et al., 1981), which did not appear in the selected subset of predictor variables. Furthermore, elevated scores on scale 8 were characteristic of the two types of criminal offenders found by Megargee and Bohn (1979) to have the

highest rates of recidivism (reconviction or reincarceration after release from custody), and of the four types ranking the highest (of 10 types) in intra-institutional violent disciplinary infractions. The absence of scale 4 as a significant predictor of violence in this study may be partially explained by its degree of item overlap with scale 8 (10 of 50 items comprising scale 4, $r = .16$; Dahlstrom et al., 1972).

Finally, high scores on scale 0 indicate social introversion, general maladjustment and self-depreciation (Graham, 1977). Such individuals tend to withdraw from social contacts and responsibilities (Dahlstrom et al., 1972). These descriptors, as well, seem to fit logically with the other MMPI indicators of general maladjustment as predictive of individuals who might resort to violence.

Limitations of a Neural Network Approach

The very empirical nature of applying neural networks to behavioral prediction problems limits the confidence one can achieve in terms of knowing whether better results could have been achieved if the network had been designed differently. This limitation results from the lack of consistent rules for designing a network in terms of its architecture, complexity fit, learning rule, and activation function. One must rely either on rules-of-thumb suggested by other investigators as a result of their experience with different sorts of problems, or trends that can be discerned in the course of obtaining one's

own results.

Backpropagation is only one of many existing neural network architectures. The conclusions drawn from this study are limited in generalization to this particular architecture. One cannot, on the basis of this study, make any statement about the efficacy of neural networks in general.

Future Research

The results of this study suggest that a prospective study of the prediction of violence, using a neural network approach, would be worthwhile. Further research may identify variations in neural network architectures, learning rules, or activation functions, that would improve the efficacy of a neural network approach to predicting violence, or other low base rate human behaviors.

References

- American Psychiatric Association (1974). Clinical aspects of the violent individual. Washington, D.C.: author.
- Arnoldo, C. M., Miller, W. D., & Gonzalez, L. P. (1990, November). The geometry of backpropagation training: Visualization, heuristics, and theory. Paper presented at the 4th Oklahoma Symposium on Artificial Intelligence, Stillwater, OK.
- Bailey, D. L., & Thompson, D. (1990, September). Developing neural network applications, AI Expert, 34-41.
- Bandura, A. (1969). Principles of behavior modification. New York: Holt, Rinehart & Winston.
- Becker, S., & Hinton, G. E. (1991). Learning to make coherent predictions in domains with discontinuities [Abstract]. Proceedings of the IEEE Conference on Neural Information Processing Systems--Natural and Synthetic, 9.
- Bem, D., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. Psychological Review, 81, 506-520.
- Beverly, R. F. (1964). Base expectancies and the initial home visit research schedule (Research Report 37). Sacramento, CA: California Youth Authority.
- Black, T., & Spinks, P. (1985). Predicting outcomes of mentally disordered and dangerous offenders. In D. P.

- Farrington & R. Tarling (Eds.), Prediction in Criminology (pp. 174-192). Albany: State University of New York Press.
- Bohn, M. J., Jr. (1978, April). Classification of offenders in institution for young adults. Paper presented at the Nineteenth International Congress of Applied Psychology, Munich.
- Borden, H. G. (1928). Factors for predicting parole success. Journal of American Institute of Criminal Law and Criminology, 19, 328-336.
- Brown, F. G. (1976). Principles of educational and psychological testing (2nd edition). New York: Holt, Rinehart and Winston.
- Burgess, E. W. (1928). Factors determining success or failure on parole. In A. A. Bruce, E. W. Burgess, & A. J. Harno (Eds.), The workings of the indeterminate-sentence law and the parole system in Illinois (pp. 205-249). Springfield, IL: Illinois State Board of Parole.
- Carpenter, G. A. (1989). Neural network models for pattern recognition and associative memory. Neural Networks, 2, 243-257.
- Caudill, M. (1987, December). Neural networks primer, Part I. AI Expert, 46-52.
- Caudill, M. (1988, June). Neural networks primer, Part III. AI Expert, 46-52.
- Cohen, J. (1968). Multiple regression as a general

- data-analytic system. Psychological Bulletin, 70, 426-443.
- Cohen, M. L., Groth, A. N., & Siegel, R. (1978). The clinical prediction of dangerousness. Crime and Delinquency, 24, 28-39.
- Copas, J. B. (1985). Prediction equations, statistical analysis, and shrinkage. In D. P. Farrington & R. Tarling (Eds.), Prediction in Criminology (pp. 232-255). Albany: State University of New York Press.
- Copas, J., & Tarling, R. (1984). Some methodological issues in making predictions. Paper prepared for the National Academy of Sciences Panel on Research on Criminal Careers.
- Cottrell, G., Munro, P., & Zipser, D. (1987). Image Compression by Backpropagation (Tech. Rep. No. TR-154), La Jolla: University of California at San Diego, Institute for Cognitive Science.
- Crick, F. H. C., & Asanuma, C. (1986). Certain aspects of the anatomy and physiology of the cerebral cortex. In J. L. McClelland, D. E. Rumelhart, & the PDP research group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2. (pp. 333-371) Cambridge, MA: Bradford Books.
- Crowder, R. S., III, (1991). Predicting the Mackey-Glass timeseries with cascade-correlation learning. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski, & G. E. Hinton (Eds.). Connectionist Models: Proceedings of the 1990 Summer

- School (pp. 117-123). San Mateo, CA: Morgan Kaufmann.
- Dahlstrom, W., Welsh, G., & Dahlstrom, L. (1972). An MMPI Handbook (Vol. 1). Minneapolis: University of Minnesota Press.
- Davis, K. R., & Sines, J. O. (1971). An antisocial behavior pattern associated with a specific MMPI profile. Journal of Consulting and Clinical Psychology, 36, 229-234.
- Ellinwood, E. H., Jr. (1971). Assault and homicide associated with amphetamine abuse. American Journal of Psychiatry, 127, 1170-1175.
- Elman, J. (1988). Finding the Structure in Time (Tech. Rep. No. 8801), La Jolla: University of California at San Diego, Center for Research in Language.
- Elman, J., & Zipser, D. (1987). Learning the hidden structure of speech (Tech. Rep. No. 8701), La Jolla: University of California at San Diego, Department of Linguistics and Institute for Cognitive Science.
- Elstein, A. (1976). Clinical judgment: Psychological research and medical practice. Science, 194, 696-700.
- Farrington, D. P. (1982). Longitudinal analyses of criminal violence. In M. E. Wolfgang & N. A. Weiner (Eds.). Criminal Violence (pp. 81-170). Beverly Hills: Sage.
- Farrington, D. P. (1985). Predicting self-reported and official delinquency. In D. P. Farrington & R. Tarling (Eds.), Prediction in Criminology (pp. 150-173). Albany:

- State University of New York Press.
- Farrington, D. P., & Tarling, R. (1985). Prediction in criminology. Albany: State University of New York Press.
- Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. Cognitive Science, 6, 205-254.
- Fisher, R. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.
- Flanagan, T. J., & Maguire, K. (1990). Sourcebook of Criminal Justice Statistics--1989. Washington, D.C.: U.S. Department of Justice.
- Fozzard, R., Bradshaw, G., & Ceci, L. (1989). A connectionist expert system for solar flare forecasting. In D. S. Touretzky (Ed.), Advances in neural information processing systems I, (pp. 264-271). San Mateo, CA: Morgan Kaufmann.
- Gallinari, P., Thiria, S., Badran, F., & Fogelman-Soulie, F. (1991). On the relations between discriminant analysis and multilayer perceptrons. Neural Networks, 4, 349-360.
- Garson, G. D. (1991, April). Interpreting neural-network connection weights. AI Expert, pp. 47-51.
- Glueck, S., & Glueck, E. T. (1950). Unraveling juvenile delinquency. Cambridge, MA: Harvard University Press.
- Gordon, J. S. (1991a). [Probability correct classification as a function of increasing decision thresholds]. Unpublished raw data.
- Gordon, J. S. (1991b). [Backpropagation neural networks vs

discriminant analysis: Classification accuracy with low base rate data]. Unpublished raw data.

- Gottfredson, S. D., & Gottfredson, D. M. (1985). Screening for risk among parolees: Policy, practice, and method. In D. P. Farrington & R. Tarling (Eds.), Prediction in Criminology (pp. 54-77). Albany: State University of New York Press.
- Gottfredson, M. R., Hindelang, M. J., & Parisi, N. (1978). Sourcebook of criminal justice statistics--1977. Washington, D.C.: National Criminal Justice Information and Statistics Service.
- Graham, J. R. (1977). The MMPI: A practical guide. New York: Oxford University Press.
- Hair, J. F., Jr., Anderson, R. E., & Tatham, R. L. (1987). Multivariate data analysis with readings (2nd ed.). New York: Macmillan.
- Hamparian, D. M., Schuster, R., Dinitz, S., & Conrad, J. P. (1978). The violent few. Lexington, MA: D. C. Heath.
- Hartzberg, J., Stanley, J., & Lawrence, M. (1990). BrainMaker User's Guide and Reference Manual [Computer program manual]. Sierra Madre, CA: California Scientific Software.
- Hathaway, S. R., & McKinley, J. C. (1989). MMPI-2: Manual for administration and scoring. Minneapolis: University of Minnesota Press.
- Hecht-Nielsen, R. (1988, March). Neurocomputing: picking the human brain. IEEE Spectrum, 36-41.

- Hinton, G. (1986). Learning distributed representations of concepts, Proceedings of the Eighth Annual Conference of the Cognitive Science Society, Hillsdale, NJ: Lawrence Erlbaum Associates, 1-12.
- Hirschi, T., & Hindelang, M. (1977). Intelligence and delinquency: A revisionist review. American Sociological Review, 42, 571-587.
- Hoffman, P. B., & Beck, J. L. (1985). Recidivism among released federal prisoners: Salient factor score and five-year follow-up. Criminal Justice and Behavior, 12, 501-507.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2, 359-366.
- Howell, J. (1990, November). Inside a neural network. AI Expert, pp. 29-33.
- Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis. Psychological Bulletin, 95, 156-171.
- Jones, T., Beidleman, W. B., & Fowler, R. D. (1981). Differentiating violent and nonviolent prison inmates by use of selected MMPI scales. Journal of Clinical Psychology, 37, 673-678.
- Kinoshita, J., & Palevsky, N. G. (1987, May). Computing with neural networks. High Technology, 24-31.
- Klassen, D., & O'Connor, W. A. (1988). Predicting violence in

- schizophrenic and non-schizophrenic patients: A prospective study, Journal of Community Psychology, 16, 217-227.
- Klimasauskas, C. C. (1991a, Jan./Feb.). Applying neural networks: An overview of the series (Part 1 of a series). PC AI, pp. 30-33.
- Klimasauskas, C. C. (1991b, April). Neural nets tell why: A technique for explaining a neural network's decision-making process. Dr. Dobb's Journal, pp. 16-24.
- Klimasauskas, C. C. (1991c, May/June). Applying neural networks: Training a Neural Network (Part 3 of a series). PC AI, pp. 20-24.
- Kruschke, J. K. (1991). Dimensional Attention Learning in Connectionist Models of Human Categorization (Tech. Rep. No. 50), Bloomington, IN: Indiana University Cognitive Science.
- Langan, P. A. (1991). America's soaring prison population. Science, 251, 1568-1573.
- Lapedes, A., & Farber, R. (1987). Nonlinear signal processing using neural networks: Prediction and system modelling (Technical Report LA-UR-87-2662), Los Alamos: Los Alamos National Laboratory.
- Lippmann, R. P. (1987, April). An introduction to computing with neural nets. IEEE ASSP Magazine, pp. 4-22.
- Loeber, R., & Dishion, T. (1983). Early predictors of male delinquency: A review. Psychological Bulletin, 94, 68-99.
- McClelland, J. L. (1991). Stochastic interactive processes and

- the effect of context on perception. Cognitive Psychology, 23, 1-44.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18, 1-86.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception, Part I: An account of basic findings. Psychological Review, 88, 375-407.
- McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. Journal of Experimental Psychology: General, 114, 159-188.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153-157.
- Mednick, S. A., Pollock, V., Volavka, J., & Gabrielli, W. F. (1982). Biology and violence. In M. E. Wolfgang & N. A. Weiner (Eds.), Criminal Violence (pp. 81-170). Beverly Hills: Sage.
- Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychological Bulletin, 52, 194-216.

- Megargee., E. I. (1973). Recent research on overcontrolled and undercontrolled personality patterns among violent offenders. Sociological Symposium, 9, 37-50.
- Megargee, E. I. (1982). Psychological determinants and correlates of criminal violence. In M. E. Wolfgang & N. A. Weiner (Eds.), Criminal Violence (pp. 81-170). Beverly Hills: Sage.
- Megargee, E. I., & Bohn, J. J. (1979). Classifying criminal offenders. Beverly Hills: Sage.
- Megargee, E. I., Cook, P. E., & Mendelsohn, G. A. (1967). Development and validation of an MMPI scale of assaultiveness in overcontrolled individuals. Journal of Abnormal Psychology, 25, 519-528.
- Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. Psychological Review, 80, 252-283.
- Monahan, J. (1981). Predicting violent behavior: An assessment of clinical techniques. Beverly Hills: Sage.
- Moody, J., & Utans, J. (1991). Selecting neural network architectures via the prediction risk: Application to corporate bond rating prediction [Abstract]. Proceedings of the IEEE Conference on Neural Information Processing Systems-Natural and Synthetic, 26.
- Moyer, K. E. (1976). The psychobiology of aggression. New York: Harper & Row.

- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 9, 127-147.
- Munro, P. W. (1986). State-Dependent factors influencing neural plasticity: A partial account of the critical period. In J. L. McClelland, D. E. Rumelhart, & the PDP research group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2. (pp. 471-502). Cambridge, MA: Bradford Books.
- Nelson, M. M., & Illingworth, W. T. (1991). A practical guide to neural nets. Reading, MA: Addison-Wesley.
- Neter, J., Wasserman, W., & Kutner, M. H. (1989). Applied linear regression models (2nd ed.). Homewood, IL: Irwin.
- Newman, G. (1979). Understanding violence. Philadelphia: Lippincott.
- Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. In Proceedings of the International Joint Conference on Neural Networks (pp. II-163-168). San Diego.
- Palmer, J., & Carlson, P. (1976). Problems with the use of regression analysis in prediction studies. Journal of Research in Crime and Delinquency, 13, 64-81.
- Parker, D. (1982). Learning logic. (Invention report No. S81-64) File 1, Office of Technology Licensing, Stanford University.

- Persons, R. W., & Marks, P. A. (1971). The violent 4-3 personality type. Journal of Consulting and Clinical Psychology, 36, 189-196.
- Petersilia, J., Greenwood, P., & Lavin, M. (1977). Criminal careers of habitual felons. Santa Monica, CA: Rand.
- Poggio, T. & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. Science, 247, 978-982.
- Pollack, J. (1988). Recursive auto-associative memory. Neural Networks Supplement: INNS Abstracts, 1, 122.
- Pritchard, D. (1977). Stable predictors of recidivism. Journal Supplement Abstract Service, 7, 72.
- Quay, H. C. (1984). Managing adult inmates. College Park, MD: American Correctional Association.
- Rector, M. (1973). Who are the dangerous? Bulletin of the American Academy of Psychiatry and the Law, 1, 186-188.
- Redl, F., & Wineman, D. (1957). The aggressive child. New York: Free Press.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65, 386-408.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), Parallel distributed processing: Explorations in the

- microstructure of cognition. Volume I. (pp. 318-362).
Cambridge, MA: Bradford Books.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986b).
Learning representations by back-propagating errors. Nature,
323, 533-536.
- Rumelhart, D. E., & McClelland, J. L., & the PDP Research Group.
(1986). Parallel distributed processing: Explorations in the
microstructure of cognition. Volume I. Cambridge, MA:
Bradford Books.
- Sanger, T. D. (1991). Basis-function trees for approximation
in high-dimensional spaces. In D. S. Touretzky, J. L. Elman,
T. J. Sejnowski, & G. E. Hinton (Eds.), Connectionist models:
Proceedings of the 1990 summer school (pp. 117-123). San
Mateo, CA: Morgan Kaufmann.
- Sarbin, T. R. (1942). A contribution to the study of actuarial
and individual methods of prediction. American Journal of
Sociology, 48, 593-602.
- Sejnowski, T. J. (1986). Open questions about computation in
cerebral cortex. In J. L. McClelland, D. E. Rumelhart, & the
PDP research group (Eds.), Parallel distributed processing:
Explorations in the microstructure of cognition. Volume 2.
(pp. 372-389). Cambridge, MA: Bradford Books.
- Sejnowski, T. J., & Rosenberg, C. (1987). Parallel networks
that learn to pronounce English text. Complex Systems, 1,
145-168.

- Shah, S. A. (1978). Dangerousness: A paradigm for exploring some issues in law and psychology. American Psychologist, 33, 224-238.
- Sharda, R., & Patil, R. B. (in press). Connectionist approach to time series prediction: An empirical test. Journal of Intelligent Manufacturing.
- Silberman, C. (1978). Criminal violence, criminal justice. New York: Random House.
- Simon, F. H. (1971). Prediction methods in criminology. London: Her Majesty's Stationery Office.
- Simpson, P. (1990). Artificial neural systems: Foundations, paradigms, applications, and implementations. New York: Pergamon Press.
- Smith, D. E. (1980). A clinical approach to the treatment of phencyclidine (PCP) abuse. Psychopharmacology Bulletin, 16, 67-70.
- State of Michigan. (1978). Summary of parole risk study. Unpublished manuscript, Department of Corrections.
- Steadman, H. J. (1980). The right not to be a false positive: Problems in the application of the dangerousness standard. Psychiatric Quarterly, 52, 84-99.
- Steadman, H., Vanderwyst, D., & Ribner, S. (1978). Comparing arrest rates of mental patients and criminal offenders. American Journal of Psychiatry, 135, 1218-1220.
- Tarling, R., & Perry, J. A. (1985). Statistical methods in

- criminological prediction. In D. P. Farrington & R. Tarling (Eds.), Prediction in Criminology (pp. 210-231). Albany: State University of New York Press.
- Thorndike, E. L. (1918). Fundamental theorems in judging men. Journal of Applied Psychology, 2, 67-76.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.
- Waller, I. (1974). Men released from prison. Toronto: University of Toronto Press.
- Walker, N., Hammond, W., & Steer, D. (1967). Repeated violence. Criminal Law Review, 465-472.
- Webster, W. (1978). Crime in the United States--1977. F.B.I. Washington, D.C.: Supt. of Docs., U.S. Government Printing Office.
- Weigend, A. S., Rumelhart, D. E., & Huberman, B. A. (1991). Back-propagation, weight-elimination, and time series prediction. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski, & G. E. Hinton (Eds.). Connectionist models: Proceedings of the 1990 summer school (pp. 105-116). San Mateo, CA: Morgan Kaufmann.
- Weiss, S. M., & Kulikowski, C. A. (1991). Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems. San Mateo, CA: Morgan Kaufmann.
- Wenk, E. A., Robison, J. O., & Smith, G. W. (1972). Can

- violence be predicted? Crime and Delinquency, 18, 393-402.
- Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences, unpublished doctoral dissertation, Harvard University, Boston.
- Werbos, P., & Titus, J. (1978). An empirical test of new forecasting methods derived from a theory of intelligence: The Prediction of Conflict in Latin America. IEEE Transactions on systems, man, and cybernetics, SMC-8, 657-666.
- Wheeler, L., & Caggiula, A. R. (1966). The contagion of aggression. Journal of Experimental Social Psychology, 2, 1-10.
- White, H. (1981). Consequences and detection of misspecified nonlinear regression models. Journal of the American Statistical Association, 76, 419-433.
- White, H. (1989). Learning in artificial neural networks: A statistical perspective. Neural Computation, 1, 425-464.
- White, H. (1989b). Some asymptotic results for learning in single hidden-layer feedforward network models. Journal of the American Statistical Society, 84, 1003-1013.
- Wilbanks, W. L. (1985). Predicting failure on parole. In D. P. Farrington & R. Tarling (Eds.), Prediction in Criminology (pp. 78-94). Albany: State University of New York Press.
- Williams, R. J. (1986). The logic of activation functions. In D. E. Rumelhart, J. L. McClelland, & the PDP research group

- (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1. (pp. 423-443).
Cambridge, MA: Bradford Books.
- Wolfgang, M. E. (1958). Patterns in criminal homicide.
Philadelphia: University of Pennsylvania Press.
- Wolfgang, M. E., & Ferracuti, F. (1967). The subculture of violence. London: Tavistock.
- Wolfgang, M. E., Figlio, R. M., & Sellin, T. (1972).
Delinquency in a birth cohort. Chicago: University of Chicago Press.
- Zipser, D. (1986). Biologically plausible models of place recognition and goal location. In J. L. McClelland, D. E. Rumelhart, & the PDP research group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2. (pp. 432-470). Cambridge, MA: Bradford Books.

Appendix A

Offenses Considered Violent from NCIC Uniform Offense Codes

(modified from Megargee, 1982, p. 168-170)

	<u>NCIC Code</u>
<u>Homicide</u> (0900)	
Homicide--willful killing--family--gun	0901
Homicide--willful killing--family--(other weapon)	0902
Homicide--willful killing--nonfamily--gun	0903
Homicide--willful killing--nonfamily--(other weapon)	0904
Homicide--willful killing--public official--gun	0905
Homicide--willful killing--public official--(other weapon)	0906
Homicide--willful killing--police officer--gun	0907
Homicide--willful killing--police officer--(other weapon)	0908
Homicide--willful killing--gun	0911
Homicide--willful killing--(other weapon)	0912
<u>Kidnapping</u> (1000)	
Kidnap minor for ransom	1001
Kidnap adult for ransom	1002
Kidnap minor to sexually assault	1003
Kidnap adult to sexually assault	1004
Kidnap minor	1005
Kidnap adult	1006
Kidnap hostage for escape	1007
Kidnap--hijack aircraft	1009

Sexual Assault (1100)

Rape--gun	1101
Rape--(other weapon)	1102
Rape--strong arm	1103
Sex assault--sodomy--boy--gun	1104
Sex assault--sodomy--man--gun	1105
Sex assault--sodomy--girl--gun	1106
Sex assault--sodomy--woman--gun	1107
Sex assault--sodomy--boy--(other weapon)	1108
Sex assault--sodomy--man--(other weapon)	1109
Sex assault--sodomy--girl--(other weapon)	1110
Sex assault--sodomy--woman--(other weapon)	1111
Sex assault--sodomy--boy--strong-arm	1112
Sex assault--sodomy--man--strong-arm	1113
Sex assault--sodomy--girl--strong-arm	1114
Sex assault--sodomy--woman--strong-arm	1115
Sex assault--carnal abuse	1117

Robbery (1200)

Robbery--business--gun	1201
Robbery--business--(other weapon)	1202
Robbery--business--strong-arm	1203
Robbery--street--gun	1204
Robbery--street--(other weapon)	1205
Robbery--street--strong-arm	1206
Robbery--residence--gun	1207

	118
Robbery--residence--(other weapon)	1208
Robbery--residence--strong-arm	1209
Forcible purse-snatching	1210
Robbery--banking-type institution	1211
<u>Assault</u> (1300)	
Aggravated assault--family--gun	1301
Aggravated assault--family--(other weapon)	1302
Aggravated assault--family--strong-arm	1303
Aggravated assault--nonfamily--gun	1304
Aggravated assault--nonfamily--(other weapon)	1305
Aggravated assault--nonfamily--strong-arm	1306
Aggravated assault--public officials--gun	1307
Aggravated assault--public officials--(other weapon)	1308
Aggravated assault--public officials--strong-arm	1309
Aggravated assault--police officer--gun	1310
Aggravated assault--police officer--(other weapon)	1311
Aggravated assault--police officer--strong-arm	1312
Aggravated assault--gun	1314
Aggravated assault--(other weapon)	1315
<u>Arson</u> (2000)	
Arson--business--endangered life	2001
Arson--residence--endangered life	2002
Arson--public building--endangered life	2008

Appendix B

Predictor Variables

<u>Variable</u>	<u>Values/Levels</u>	<u># cases with missing value</u>
Demographic		
1. Age	Years	0
2. Education	Years	3
3. Income	1 - < \$10,000	12
	2 - \$10,000 to 19,999	
	3 - \$20,000 to 29,999	
	4 - \geq \$30,000	
4. Population	1 - < 4,000	0
	2 - 4,000 to 15,999	
	3 - 16,000 to 49,999	
	4 - 50,000 to 300,000	
	5 - > 300,000	
5. Occupation	Unemployed	8
	Unskilled Labor	
	Skilled Labor	
	Professional	
6. Race	Caucasian	0
	Black	
	Native American	

Psychometric

7. Beta IQ	Score	7
8-25. MMPI	3 Validity scales	0
	10 Clinical scales	
	5 Auxiliary scales	

Criminal/Violence History

26. Age First Arrest	Years	1
27. Arrests	# Violent	0
28.	# Nonviolent	0
29. Convictions	# Violent	0
30.	# Nonviolent	0
31. Commitments	#	0
32. Status	Free	0
	Probation	
	Parole	
	Escape	
33. Time Free	Months	3
34. Child Abuse	Present/Absent	54

Situational

35. Marital Status	Single	0
	Live With	
	Common Law	
	Married	
	Divorced	
36. Single Marriage	Present/Absent	26

37.	Supported Family	Present/Absent	31
38.	Irregular Work	Present/Absent	130

Substance Abuse History

39.	Alcohol	Present/Absent	0
40.	Amphetamines	Present/Absent	0
41.	Barbiturates	Present/Absent	0
42.	Heroin	Present/Absent	0
43.	Other Opiates	Present/Absent	0
44.	Marijuana	Present/Absent	0
45.	Mescaline	Present/Absent	0
46.	Benzodiazepines	Present/Absent	0
47.	LSD	Present/Absent	0
48.	Inhaling Vapors	Present/Absent	0

Appendix C

Predictor Variables (17) Selected by
Stepwise Discriminant Analysis

<u>Variable</u>	<u>Partial R²</u>	<u>Prob > F</u>
Probation	0.0597	0.0001
Age	0.0411	0.0001
Violent Arrests	0.0405	0.0001
MMPI-0	0.0397	0.0001
Married	0.0214	0.0039
Benzodiazepines	0.0149	0.0163
Live With	0.0136	0.0217
Native American	0.0122	0.0304
Unskilled Labor	0.0103	0.0466
Parole	0.0097	0.0538
Income	0.0094	0.0585
Irregular Work History	0.0088	0.0668
Supported Family	0.0078	0.0865
Beta IQ	0.0077	0.0873
MMPI-3	0.0077	0.0885
MMPI-1	0.0068	0.1082
MMPI-8	0.0061	0.1305

Appendix D

Predictor Variables Ranked by Neural NetworkWeight Matrix Analysis

<u>Rank</u>	<u>Variable</u>	<u>Share of Violent Output</u>
1	Professional	0.0442
2	Escape	0.0361
3	Heroin	0.0360
4	Live With	0.0342
5	Native American	0.0312
6	Probation	0.0253
7	Violent Convictions	0.0251
8	Inhaling Vapors	0.0224
9	Unemployed	0.0213
10	Alcohol	0.0206
-----Subset of 10-----		
11	Married	0.0196
12	Skilled Labor	0.0188
13	Parole	0.0185
14	Nonviolent Convictions	0.0184
15	Benzodiazepines	0.0184
16	Common Law Marriage	0.0180
17	Income	0.0178
18	Unskilled Labor	0.0177
19	LSD	0.0176

20	Single	0.0173
21	Divorced	0.0173
22	Violent Arrests	0.0169
23	Mescaline	0.0167
24	Barbiturates	0.0166
25	Child Abuse	0.0163
26	MMPI-K	0.0162
27	Single Marriage	0.0162
28	Other Opiates	0.0161
29	Time Free	0.0152

-----Subset of 29-----

30	Amphetamines	0.0148
31	Irregular Work History	0.0148
32	Education	0.0147
33	Population	0.0145
34	Prior Commitments	0.0145
35	MMPI-FAM	0.0141
36	Caucasian	0.0140
37	Black	0.0137
38	Nonviolent Arrests	0.0136
39	Supported Family	0.0135
40	Marijuana	0.0134
41	MMPI-L	0.0132
42	Free	0.0131
43	MMPI-9	0.0127
44	MMPI-HC	0.0123
45	MMPI-OH	0.0123

46	MMPI-4	0.0120
47	Age	0.0119
48	MMPI-0	0.0119
49	MMPI-8	0.0110
50	MMPI-7	0.0108
51	MMPI-3	0.0107
52	MMPI-F	0.0105
53	MMPI-HOS	0.0104

-----Subset of 53-----

54	Age at First Arrest	0.0099
55	MMPI-1	0.0097
56	MMPI-PV	0.0096
57	MMPI-5	0.0094
58	MMPI-2	0.0093
59	MMPI-6	0.0091
60	Beta IQ	0.0089

Appendix E

Oklahoma Department of Corrections Approval Forms

Attachment A

REQUEST FOR ACCESS FORM

Request for access to correctional client criminal case history information maintained by Oklahoma Department of Corrections, from _____

Jolene R. Scully Gordon authorized and duly representing _____ (individual)

Oklahoma State University, hereinafter called requestor. (agency)

1. Information requested:

Demographic information, psychometric data, criminal history, substance abuse history; strictly archival data

2. Requestor request this information

- (X) on a continuing basis - until completion of my research (expected May 1992)
(O) on a one-time basis

3. The purpose for which information is requested:

- () To implement a statute, ordinance, or executive order. (Submit copy or give citation)
(O) To provide services required for the administration of criminal justice pursuant to an agreement with a criminal justice agency. (Attach agreement)
(X) Research, evaluative or statistical activities
(O) Such purposes as authorized by court rule, decision, or order. (Attach or cite)
(O) Other purpose Explain:

8-14-91 Date

Jolene R. Scully Gordon Signature of Requestor Representative

Request Granted X

Request Denied

If denied, reason denied:

8-14-91 Date

Jack ... Signature of Oklahoma Department of Corrections Representative

NON-DISCLOSURE AGREEMENT

This agreement is made and entered into by and between (Oklahoma Department of Corrections) hereinafter called Agency; and (Jolene R. Scully Gordon) hereinafter called Recipient.

- A. This agreement is to provide administratively created correctional client criminal case history information for research, evaluative, or statistical activities. The recipient agrees that the information will not be used to the detriment of the record subject nor for any purpose other than those stated in the research plan. The recipient agrees further to abide by the confidentiality and security provisions of Section 524 (a) of the Omnibus Crime Control and Safe Streets Act of 1973 and any regulations issued pursuant to that section.
- B. Agency agrees to provide Recipient with the correctional client case history information requested in the attached access request.
- C. Recipient agrees to limit the use of this information to the purpose for which it was provided and to destroy the source documents when they are no longer needed for the purposes for which it was provided.
- D. Recipient agrees that the only persons allowed access to this information are: recipient and advisor, and agrees not to disseminate or re-disclose the information to any other agency or person.
- E. Recipient agrees to implement reasonable procedures to protect this information from unauthorized access, alteration, or destruction.
- F. Recipient agrees to abide by the laws or regulations of this state and the federal government, any present or future rules, policies, or procedures adopted by Agency or adopted by NCIC after approval by the NCIC Policy Board to the extent that they are applicable to the information provided under this agreement.
- G. Recipient agrees to indemnify and save harmless this state, Agency, other criminal justice agencies as defined by the Code of Federal Regulations, Title 28, Chapter I, Part 20, the electronic data processing agencies with whom this state has contract to process correctional client criminal case information and the employees of any of the above entities (1) from and against any and all causes of action, demands, suits and other proceedings of whatsoever nature, (2) against all liability to other including any liabilities or damages by reason of or arising out of any arrest, or imprisonment or any cause of action, whatsoever, and (3) against any loss, cost expense and damage resulting therefrom, arising out of or involving any negligence on the part of the recipient in the exercise or enjoyment of this agreement.
- H. If the agreement is to provide correctional client criminal case history information on a continuing basis, Agency reserves the right to immediately suspend furnishing information under this agreement when any rule, policy, procedure, regulation, or law described in Section F is violated or appears to be violated.
- I. If this agreement is to provide correctional client criminal case history information on a continuing basis, then either Agency or Recipient may, upon 30 days notice in writing, terminate this agreement.

8-14-91

Date

Jolene R. Scully Gordon

Signature of Recipient Representative

8-14-91

Date

Jack Scully, Warden Joseph Hays

Signature of Oklahoma Department of Corrections Representative

Table 1

Parole Violation by Salient Factor Score

SFS	Risk category	Percent non-recidivists	Percent recidivists	N
10 - 6	Very Good/Good	86	14	766
5 - 4	Fair	71	29	423
3 - 0	Poor	60	40	614
All cases		74	26	1806

Note. Adapted from "Recidivism Among Released Federal Prisoners: Salient Factor Score and Five-Year Follow-Up" by P. B. Hoffman and J. L. Beck, 1985, Criminal Justice and Behavior, 12, p. 505.

Table 2

Performance Comparison of Neural Net and Discriminant Analysis
with two different base rates, on 10 test files of random input

Construction Sample Base Rate ("Violent")	Proportion Classified "Violent"			
	Neural Net		Discrim Anal	
	<u>Mean</u>	<u>Range</u>	<u>Mean</u>	<u>Range</u>
.50	.60	(.56 - .67)	.48	(.38 - .54)
.20	.15	(.11 - .18)	0	(0)

Table 3

Frequencies of Four Classification Outcomes for Neural Network
(N) and Discriminant Analysis (D) Models

Model	<u>Construction Sample</u>			<u>Validation Sample</u>		
	Predicted Viol	NonV	Correct	Predicted Viol	NonV	Correct
N60-25						
Actual Viol	70	0		15	20	
Actual NonV	0	191		11	84	
			261			99
D60						
Actual Viol	41	29		16	19	
Actual NonV	12	179		16	80	
			220			96
N17-12						
Actual Viol	70	0		19	16	
Actual NonV	0	191		15	81	
			261			100
D17						
Actual Viol	33	37		17	18	
Actual NonV	11	180		9	87	
			213			104
N53-26						
Actual Viol	70	0		15	20	
Actual NonV	0	191		17	79	
			261			94
D53						
Actual Viol	40	30		17	18	
Actual NonV	13	178		15	81	
			218			98
N29-15						
Actual Viol	70	0		17	18	
Actual NonV	0	191		16	80	
			261			97
D29						
Actual Viol	30	40		12	23	
Actual NonV	12	179		11	85	
			209			97
D10						
Actual Viol	17	53		6	29	
Actual NonV	10	181		5	91	
			198			97

Table 4

Total Group Hit Rates

Inputs	Training Data		Test Data	
	Neural Net	Discrim Anal	Neural Net	Discrim Anal
60	1.000	0.843	0.756 ^P	0.730 ^P
17	1.000	0.813	0.761 ^P	0.791 ^{P,m}
53	1.000	0.836	0.720 ^P	0.743 ^P
29	1.000	0.798	0.743 ^P	0.740 ^P
10	---	0.759	---	0.740 ^P

^m = $p_{\max} < .005$, one-tailed

^P = $p_{\text{proport}} < .00001$, one-tailed

Table 5

Conditional Probabilities by Model Type: Neural Network (NN)
and Discriminant Analysis (DA)

	p(corr V)		p(corr NV)		p(corr "V")		p(corr "NV")	
	.27 ^P		.73 ^P					
Inputs	NN	DA	NN	DA	NN	DA	NN	DA
60	.42**	.46**	.88**	.83*	.54	.49	.81	.81
17	.54**	.47**	.84**	.91**	.56	.66	.83	.83
53	.43**	.48**	.83*	.84**	.47	.53	.80	.81
29	.50**	.33**	.83*	.89**	.52	.51	.82	.79
10	--	.17	--	.95**	--	.55	--	.76

^P Separate group hit rates expected by proportional criterion

* $p < .05$, one-tailed

** $p < .01$, one-tailed

Figure 1

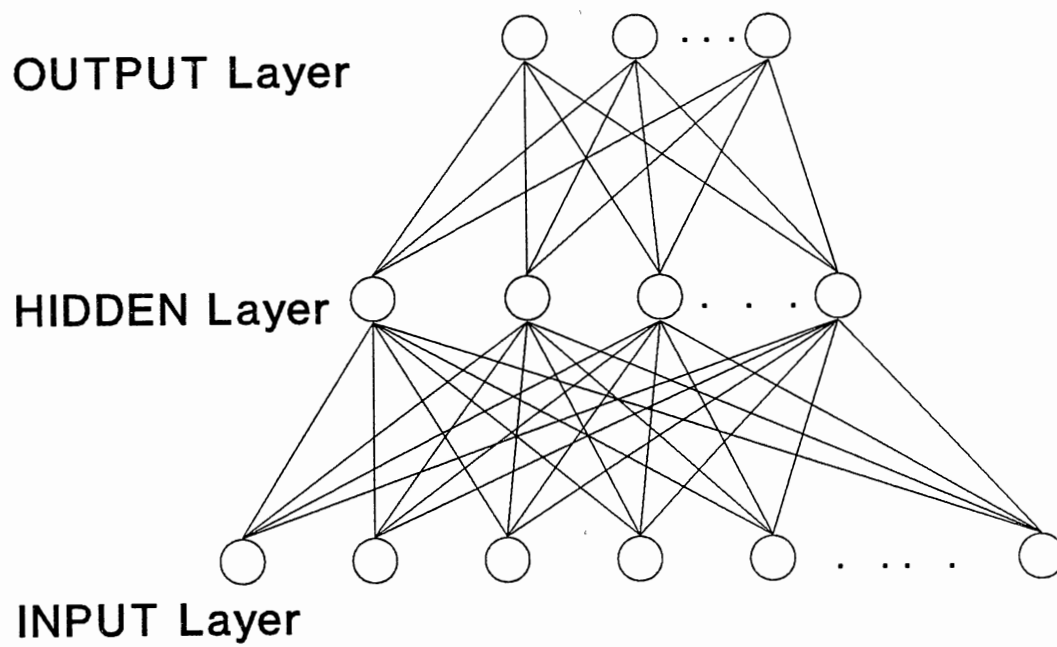
Layers of a Backpropagation Neural Network

Figure 2

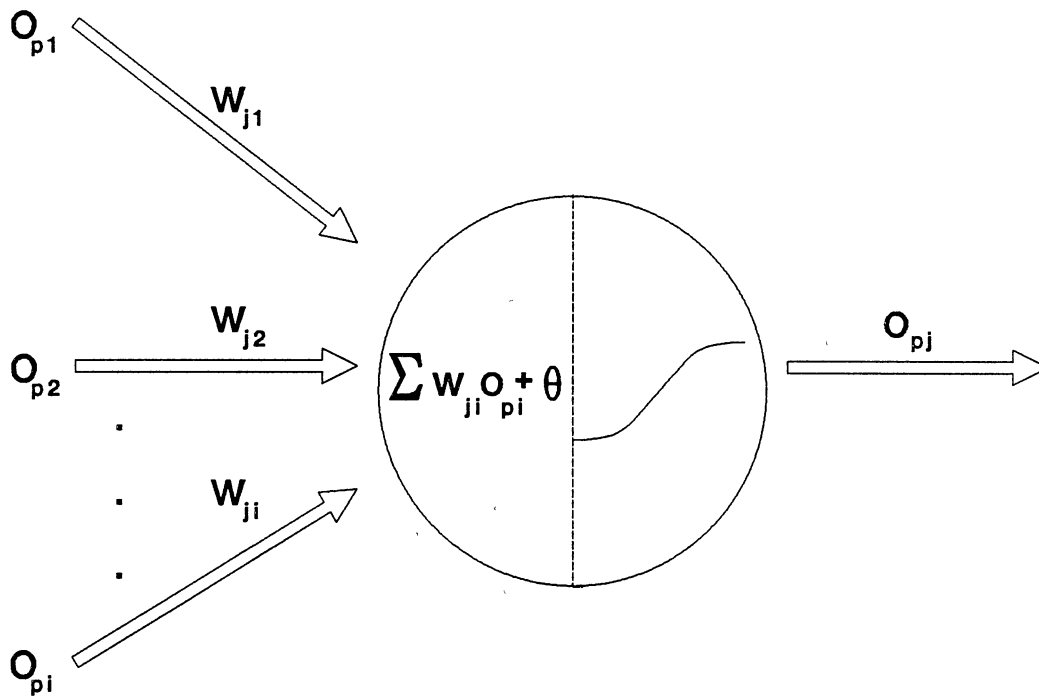
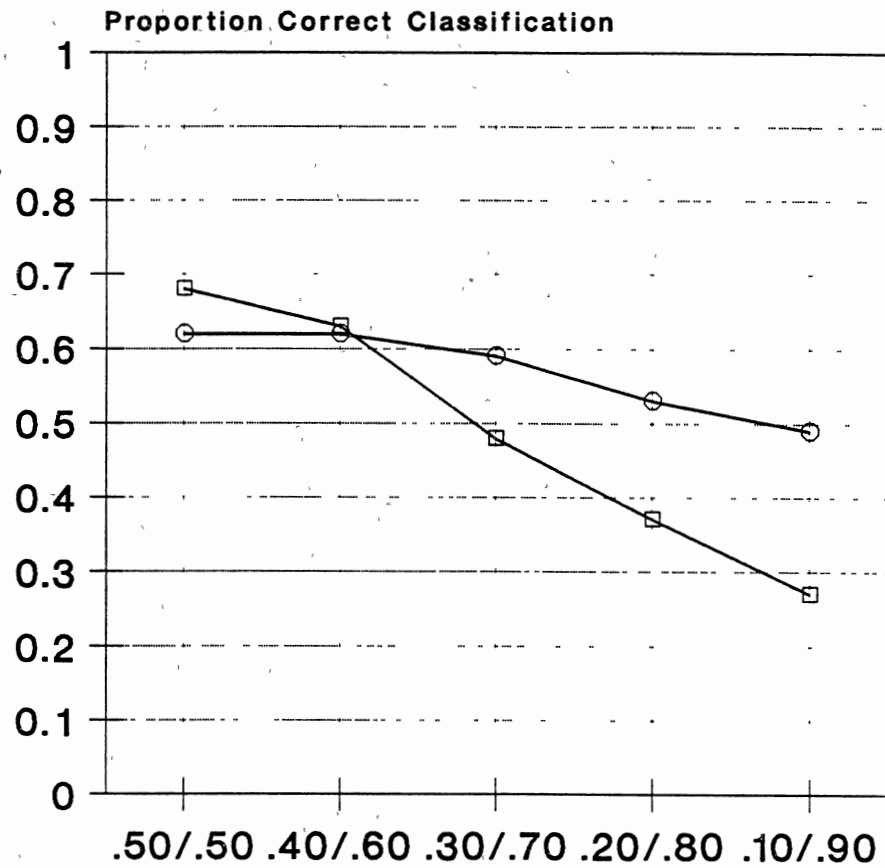
Activation Function

Figure 3

Proportion Correct Classification at Increasing Decision
Thresholds



Neural Network	0.62	0.62	0.59	0.53	0.49
Discrim Analysis	0.68	0.63	0.48	0.37	0.27

Decision Thresholds

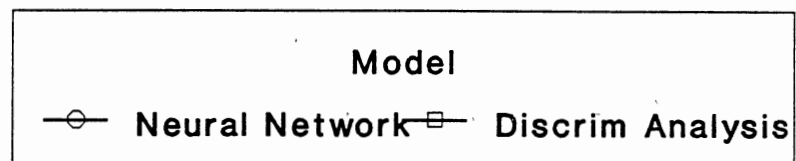


Figure 4

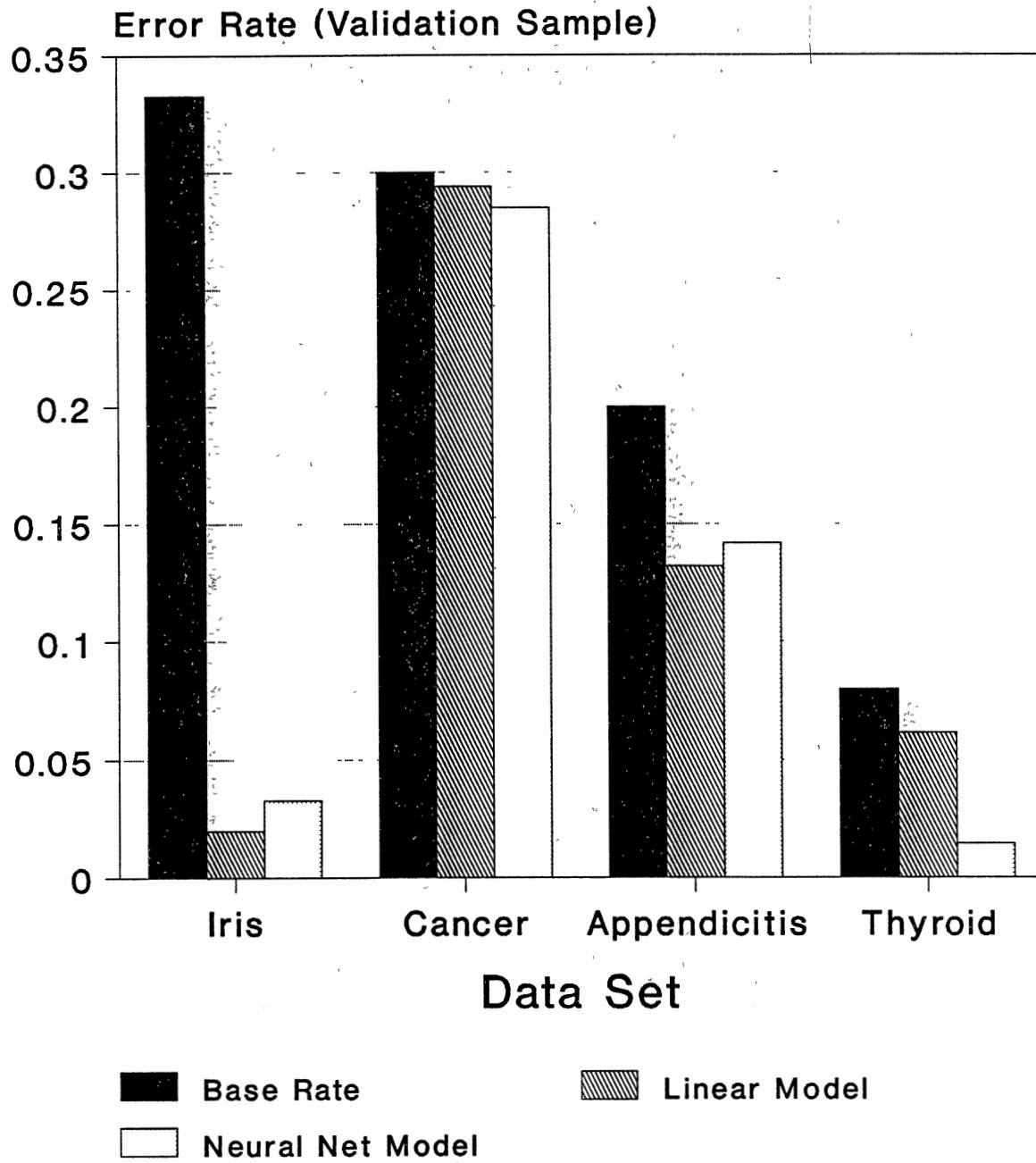
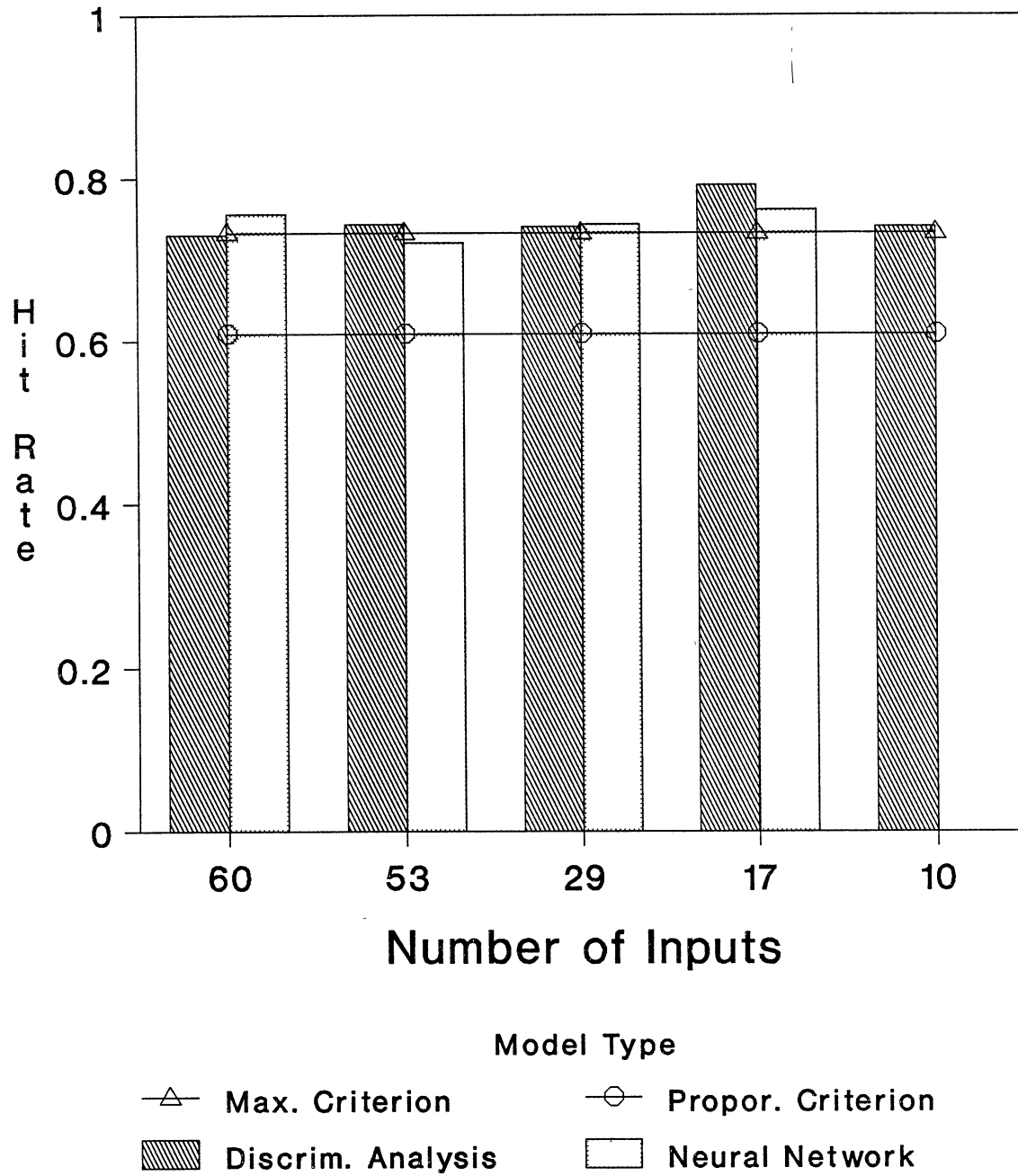
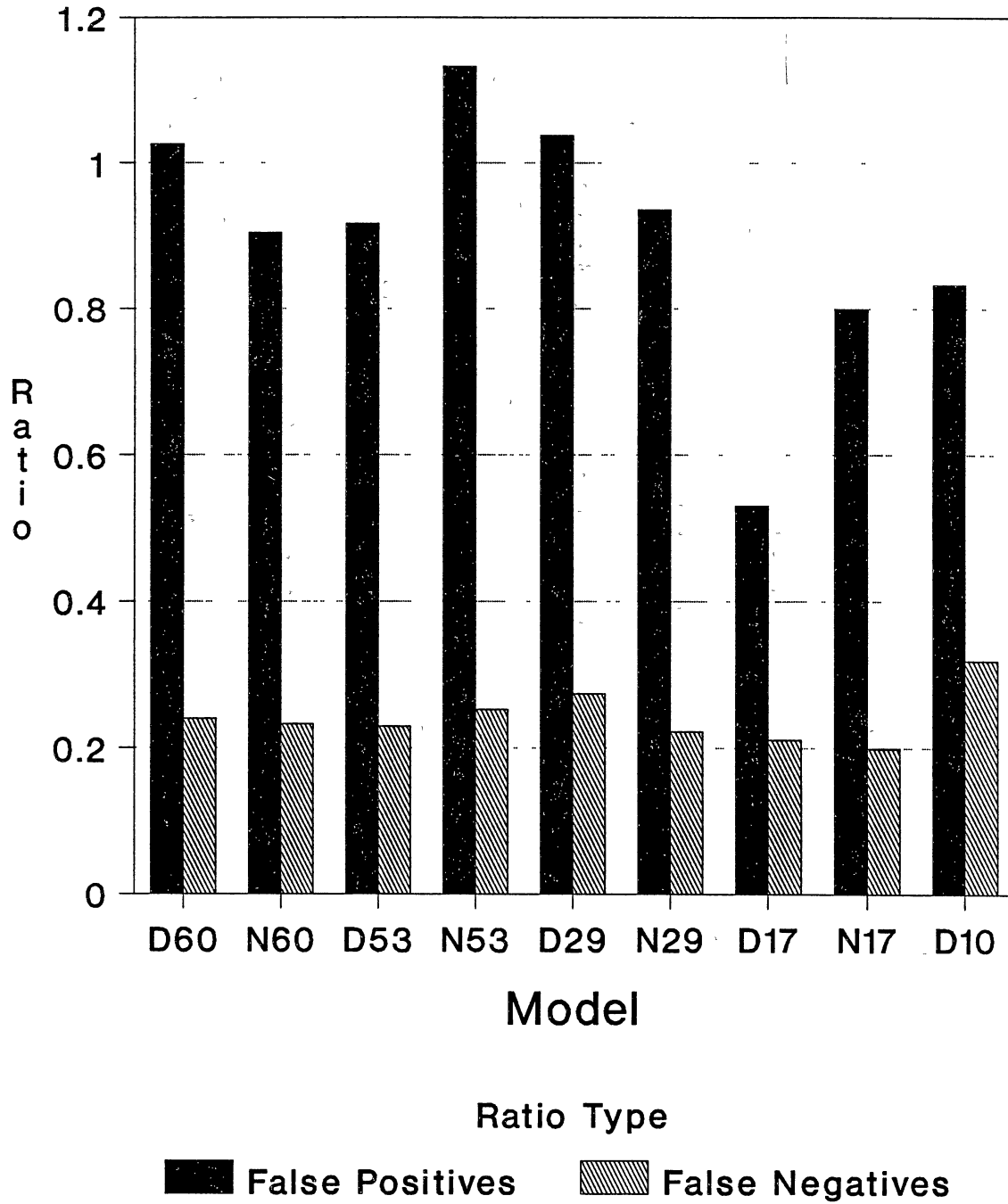
Comparative Performance of Two Prediction Methods vs. Base Rate

Figure 5

Total Group Hit Rates by Model Type

$$TGHR = (CP + CN)/N$$

Figure 6

False Positive and False Negative Ratios

Ratio = # False : 1 Correct

Figure 7

Proportion Correct Positive Predictions at Increasing Decision
Thresholds

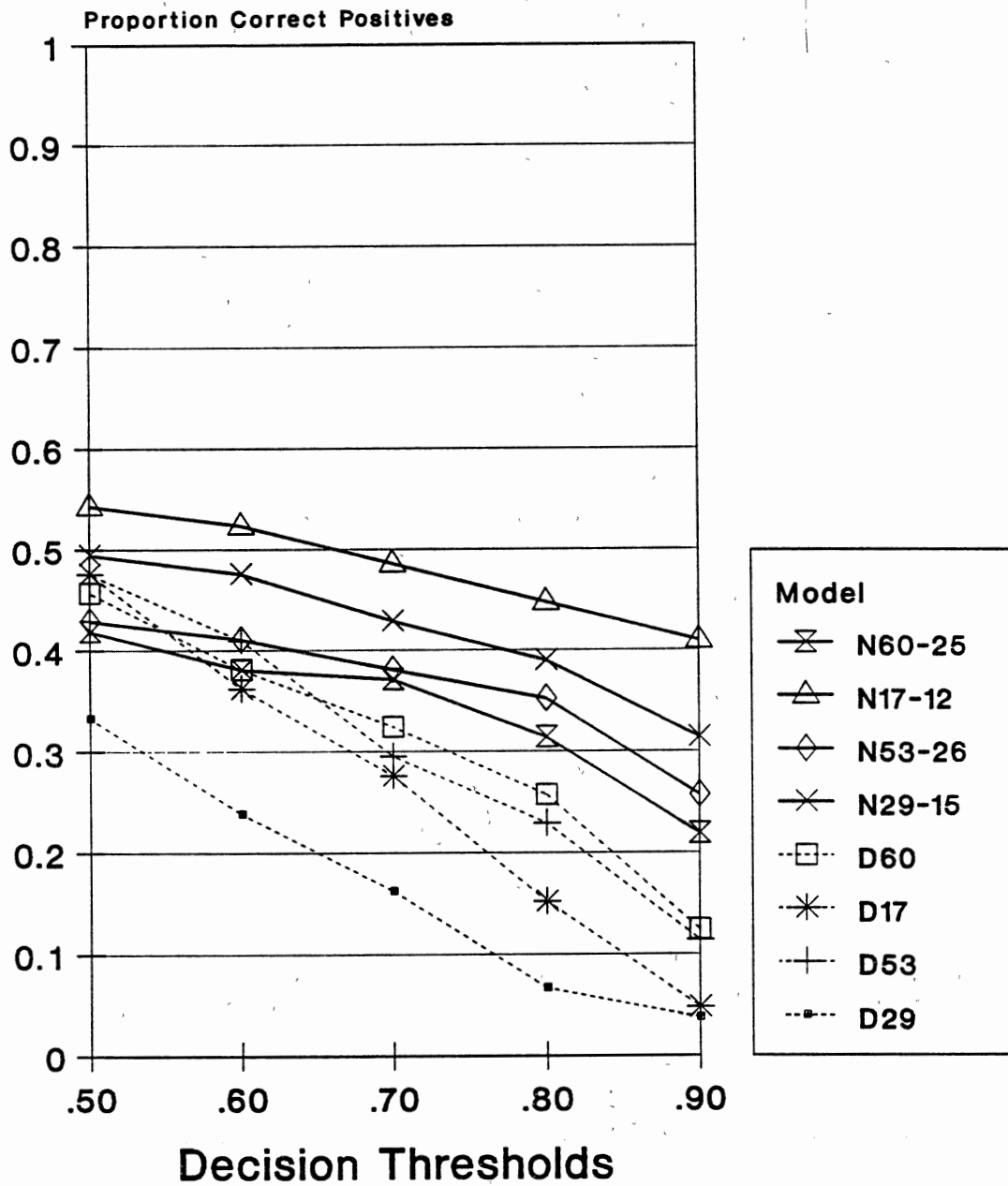
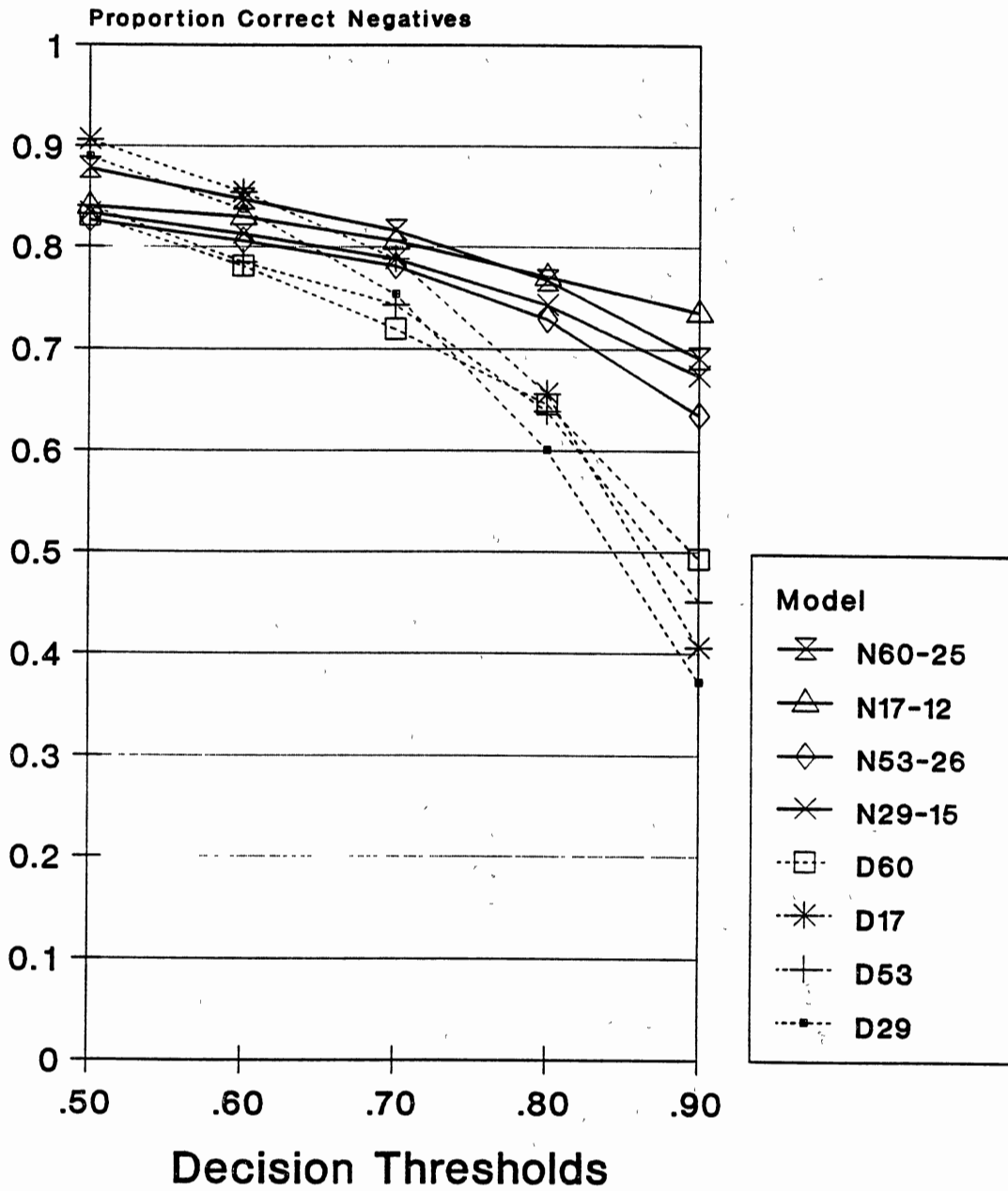


Figure 8

Proportion Correct Negative Predictions at Increasing Decision Thresholds



$p(\text{CN}) = \text{CN}/(\text{CN}+\text{FP})$

VITA

Jolene Scully Gordon

Candidate for the Degree of

Doctor of Philosophy

Thesis: A NEURAL NETWORK APPROACH TO THE PREDICTION OF VIOLENCE

Major Field: Psychology

Biographical:

Personal Data: Born in Independence, Missouri, September 5, 1957, the daughter of R. Harold and Lois V. Scully; married Richard T. Gordon, August 11, 1979; mother of Ashley D. Gordon, born September 26, 1985.

Education: Graduated from Van Horn High School, Independence, Missouri, in May, 1974; received Bachelor of Arts Degree in Education from the University of Missouri-Kansas City in May, 1978; received a Master of Science in Education Degree in Educational Psychology and Research from the University of Kansas in May, 1983; received a Master of Science Degree in Psychology in July, 1989; completed requirements for the Doctor of Philosophy Degree from Oklahoma State University in July, 1992.

Professional Experience:

Research: Research Assistant, Department of Psychology, Oklahoma State University, 1989-1990, partial funding from Martin-Marietta Electronic Systems, Inc.; Research Assistant, School of Civil Engineering, Oklahoma State University, 1991.

Teaching: Teacher, regular classroom, Fort Osage School District, Independence, Missouri, 1978-1981; Master Teacher, gifted/talented students, North Kansas City School District, Kansas City, Missouri, 1981-1988; Teaching Assistant, Department of Psychology, Oklahoma State University, 1988-1992.

Professional Organizations: American Psychological Society, Cognitive Science Society, International Neural Network Society, Southwestern Psychological Association, Oklahoma Psychological Society.