A COMPUTATIONAL INVESTIGATION INTO MALADAPTIVE AGGRESSION

A dissertation presented

by

Glen Anthony Coppersmith

to
The Department of Psychology

In partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in the field of
Psychology

Northeastern University
Boston, Massachusetts
July, 2008
A COMPUTATIONAL INVESTIGATION INTO MALADAPTIVE AGGRESSION

by

Glen Anthony Coppersmith

ABSTRACT OF DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology in the Graduate School of Arts and Sciences of Northeastern University July, 2008
Chapter 1

Abstract

Maladaptive aggression is a serious, growing, and ill-understood problem for today’s society. This is due, at least in part, to a lack of knowledge regarding how economic, social, environmental, and/or psychiatric factors influence the incidence of maladaptive aggression at the individual patient level. Standard statistics have teased out the etiological factors that correlate with the incidence of maladaptive aggression in the population as a whole, but have proven ineffective at predicting which patients will display maladaptive aggression and which will not. This failing is likely due to the high number of interactions implicated in the development of maladaptive aggression, the heterogeneous nature of maladaptive aggression, a distinct lack of adequate data sets, or some combination thereof. Thus, the most comprehensive data set on maladaptive aggression available to date was examined with a variety of techniques to overcome some of the difficulties inherent in predicting maladaptive aggression. The techniques employed were: adapted standard statistics, statistical pattern recognition, machine learning, and a suite of novel predictive analysis tools developed during the process of this dissertation. The results of this investigation provide a method capable of illuminating the complex causes and correlates of maladaptive aggression with both expected and unexpected factors implicated by the current data set. Notably, this method is easily adapted for use with other data sets and a broad range of predictive problems, especially the investigation of mental illnesses.
Acknowledgements

There are too many people who have helped along this journey to list them all here, so I won’t even attempt it.

Thank you to my family, friends, colleagues, and especially to those that fall into more than one category. You each provided your own kind of support in your own special way, and for that you have my eternal gratitude.

For all those better angels who shouted down the demons throughout this endeavor, I thank you.


## Contents

1 Abstract

2 Background and Significance 5
   2.1 Introduction .......................................................... 5
      2.1.1 Subtyping aggression ........................................... 11
   2.2 Risk and Protective Factors ......................................... 14
   2.3 Shortcoming of Standard Methods ................................. 15
   2.4 Purpose .................................................................. 17
   2.5 Machine Learning ...................................................... 18

3 Data Set 23
   3.1 Devereux Institute: Inpatient Data ................................. 23
   3.2 Novel and Training Data .............................................. 26

4 Methods 27
   4.1 The Ψ Metric ............................................................ 27
   4.2 Diagnosis from Binary Outcomes .................................... 28
   4.3 Techniques .............................................................. 29
      4.3.1 Normalization ....................................................... 29
      4.3.2 Regression .......................................................... 32
      4.3.3 K-Nearest-Neighbor .............................................. 34
      4.3.4 Neural Networks .................................................. 42
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.5 Evolutionary Algorithms</td>
<td>62</td>
</tr>
<tr>
<td>4.3.6 Collaborative Classifiers: Forests</td>
<td>64</td>
</tr>
<tr>
<td>4.3.7 Human-created Model</td>
<td>67</td>
</tr>
<tr>
<td>4.3.8 Synergy [1]: Optimizing Neural Network Structure</td>
<td>67</td>
</tr>
<tr>
<td>4.3.9 Synergy [2]: Collaborative Classifiers: Forest of Neural Networks</td>
<td>68</td>
</tr>
<tr>
<td>4.3.10 Synergy [3]: Hybridized Neural Network</td>
<td>69</td>
</tr>
<tr>
<td>4.3.11 Synergy [4]: Collaborative Disparate Classifiers</td>
<td>69</td>
</tr>
<tr>
<td>4.3.12 Controls</td>
<td>72</td>
</tr>
<tr>
<td>4.3.13 Expectations</td>
<td>77</td>
</tr>
<tr>
<td>4.3.14 Model Selection</td>
<td>79</td>
</tr>
<tr>
<td>4.3.15 Decomposing Classifiers</td>
<td>83</td>
</tr>
<tr>
<td>5 Results</td>
<td>84</td>
</tr>
<tr>
<td>5.1 Chance</td>
<td>86</td>
</tr>
<tr>
<td>5.2 Standard Statistics</td>
<td>86</td>
</tr>
<tr>
<td>5.3 Neural Networks</td>
<td>89</td>
</tr>
<tr>
<td>5.4 Generalization</td>
<td>94</td>
</tr>
<tr>
<td>5.5 Streamlining</td>
<td>94</td>
</tr>
<tr>
<td>5.6 Decomposing Classifiers: Distilling Analysis</td>
<td>94</td>
</tr>
<tr>
<td>6 Discussion</td>
<td>97</td>
</tr>
<tr>
<td>7 Conclusions and Future Directions</td>
<td>105</td>
</tr>
<tr>
<td>8 Glossary</td>
<td>108</td>
</tr>
<tr>
<td>8.0.1 General Mathematics</td>
<td>108</td>
</tr>
<tr>
<td>8.0.2 Neural Networks</td>
<td>108</td>
</tr>
<tr>
<td>8.0.3 Evolutionary Algorithms</td>
<td>108</td>
</tr>
</tbody>
</table>
Chapter 2

Background and Significance

2.1 Introduction

The prevalence of maladaptive aggression in today’s society is a serious and growing problem. About 15 percent of violent crime arrests in 2001 were of perpetrators under the age of 18 (Federal Bureau of Investigation, 2002). Moreover, for every teen arrested, it is estimated that ten teens are engaged in violence with the potential for serious bodily injury or death (U.S. Department of Health and Human Services, 2001). Additionally, roughly 9 percent of all homicides in the United States in 2000 were perpetrated by those under the age of 18 (Fox and Zawitz, 2002). Finally, between one and four percent of all children meet the diagnostic criteria for conduct disorder, the diagnosis used in the mental health field for disproportionate levels of violence and aggression (U.S. Department of Health and Human Services, 2001). As these statistics suggest, youth violence is a serious problem society is facing. For reasons that will become apparent later, it is difficult to ascertain accurate statistics of aggression itself, but aggressive acts cost society monetarily in a number of ways: psychiatric staff, medications, and damages caused by aggression among them. In fact, maladaptive aggression takes up most of the resources of child mental health systems, juvenile justice systems, and special education systems (Borduin, 1999). Moreover, it is one of the most common reasons for psychiatric referral in youth (Steiner et al., 2003).
2.1. **INTRODUCTION**

The ability to diagnose and predict the likelihood of an individual displaying maladaptive aggressive tendencies is needed to alleviate this economic burden. In order to do this, the precursors and treatments of aggression need to be better understood and the ability to target those in need of treatment needs to be improved. Ideally, research will lead to an improved ability to predict the kind of treatment(s) that will be beneficial to a patient as well. Unfortunately, there is no straightforward way to make these improvements. This task is even more complicated considering that aggression is not uniform.

Adaptive aggression is commonly linked with an organism’s drive and motivation, often considered necessary for survival. In contrast, maladaptive aggression is aggression detrimental to the organism’s success and survival (Connor, 2002). It is often true that only the context in which an aggressive act is exhibited differentiates between adaptive and maladaptive aggression. In other words, the same aggressive act may be adaptive in some circumstances and maladaptive in others. For the purposes of this document, maladaptive aggression refers to violent acts committed in a society where these acts incur negative consequences.

Studies into the roots of maladaptive aggression point to the fact that the adage 'violence begets violence' holds true, highlighting the severity and complexity of identifying and treating maladaptive aggression. Children exposed to aggression during their early development are more likely to become maladaptively aggressive later in life and those who exhibit persistent maladaptive aggression show poorer quality of life in almost all areas examined (Moffitt, 1993). In fact, some measures of maladaptive aggression at age 8 significantly predict legal convictions and self-reported aggression at age 30 (Huesmann et al., 1984). There is additional evidence supporting the need for early intervention and prevention: studies investigating mid-to-late adolescent aggressors indicate that treatment (both psychosocial therapy and pharmacologic therapy) was effective while administered, but post-treatment long-term effects (even at six months) were minimal (Frick, 1998). Conversely, psychosocial therapy and pharmacologic interventions during early life were often the only treatments with lasting post-treatment effects (Kazdin, 1997; McMahon and Wells, 1998). Generally,
treatments target specific maladaptive behaviors that the patient performs, however all such behaviors are rarely targeted by therapy. Thus, the desistance of non-targeted behaviors has been examined as a measure of treatment generalization. The interventions with consistent generalization target pre-adolescence and focus on parental behavior modification, rather than treating the child alone. These treatments have been shown to have both generalizing and persisting positive effects on maladaptive aggression (McMahon and Wells, 1998; Forehand and Long, 1988; Wells et al., 1980). Beyond this, relatively few strategies have evolved highlighting causal relationships and centering treatments around the causes of maladaptive aggression rather than the symptoms. Collectively this evidence suggests that some of the roots of maladaptive aggression lie in early childhood.

The early childhood years provide numerous ways in which aggression cycles can be started and perpetuated. When a parent uses aggression to control an unruly child, this sends two signals to strengthen the aggression cycle. First, if the child was brought under control, this success would reinforce the parent to use similar methods during future episodes of misbehavior. Second, the child was shown that aggression is an acceptable tactic to mediate social conflicts. Additionally, when resistance by a child leads the parent to back down, this also perpetuates the cycle in two fashions. First, the child has learned a successful method to avoid punishment. Second, the child is again shown that aggression is an acceptable tactic by which to mediate conflicts (Connor, 2002). There is evidence to suggest that even inadvertently, maladaptive aggressive behaviors may be developed and maintained by parent child interactions (Kazdin, 1997; Patterson, 1982, 1986). Furthermore, there is evidence to indicate that lacking normative influences (such as the positive reinforcement necessary to break either of the cycles detailed above) during the early childhood years allows deviant streaks to grow unchecked. These effects have been measured as early as six years of age (Loeber et al., 1993; Connor, 2002). The parents contribute more than direct effects, as their choice of environment, to an extent, dictates the rearing environment of the child as well. In essence, the parental influences go beyond genetics and the home environment.
2.1. INTRODUCTION

This evidence examined together, suggests a cycle forming between parents and children: a child growing up in an aggressive environment (both at home and in the environment chosen by the parents) is more likely to be maladaptively aggressive. Further, when this maladaptively aggressive child grows to adulthood and has children of his or her own, these offspring are also likely to develop maladaptive aggressive tendencies for the same reason their parents did. The perpetual reinforcement of this behavior suggests that the problem will only grow with each passing generation, provided that aggressors are indeed breeding. Given the incidence of maladaptive aggression across psychiatric diagnosis, identification and characterization of the psychiatric, psychosocial, environmental, and familial factors that influence maladaptive aggression is of particular importance.

Indeed, many of the illnesses described in the Diagnostic Statistical Manual (IV) have aggression or an aggression-like behavior listed as a symptom, some samples quoting up to 80% comorbidity (Steiner and Remsing, 2007). Psychiatric conditions generally come about as some abnormality of brain function. Thus, it is not surprising to see certain symptoms clustering together across illnesses, assuming that these symptoms are associated with some similar deficit in the brain. However, maladaptive aggressive acts are routinely co-occurring with not only more illnesses but more distinct classes (clusters) of illnesses than any other symptom (American Psychiatric Association, 1994). Given this broad range, it seems unlikely that aggression is a part of each of these illnesses, but rather a separate and related phenomenon. This sort of relationship is common in the medical world as evidenced by the co-occurrence of a fever with a wide range of physical ailments.

In the same manner that a fever is a component of a physical illness, it has been theorized that aggression is a component of a psychiatric illness (Connor, 2002; Connor and McLaughlin, 2006). Along these lines, aggression may be a byproduct of a mind trying to mend itself or perhaps an indicator that something in the brain is malfunctioning (as a fever is of a bodily infection). Given either scenario, aggression is a byproduct rather than a disease all its own. Though this perspective may not immediately impact the treatment of maladaptive aggression, it is informative for future research directions. Further,
2.1. INTRODUCTION

the concept of aggression as a fever directs not only the study of aggression, but also the study of any psychiatric illness that it co-occurs with. Currently the research and treatment of psychiatric illness with co-occurring maladaptive aggression is to approach the two as a single problem, which has been met with little success. Another potential strategy is to deal with aggression separately (i.e. as the physician deals with the fever), devising treatment strategies to mitigate the symptoms of maladaptive aggression specifically.

Assuming aggression can be treated separately (even temporarily), this may expose the underlying problem to be studied by peeling away the complications associated with treating an aggressive or violent patient. Further, this will enable psychiatric treatments to be more targeted to the underlying cause, without having to additionally be concerned with relieving the aggressive tendencies co-occurring with the illness right away (of course, the implication is that by treating the underlying cause, the maladaptive aggression will subside as well). This is important considering that it is not the mental illness, but the manifestation of maladaptive aggressive tendencies that is the primary reason for most institutionalization (Connor, 2002). Institutionalization, by definition, precludes the dangers of maladaptive aggressive behaviors for the patient and others, but a recent study shows that those patients who are institutionalized take longer to recover from psychiatric illnesses and take longer to return to normal functional roles in society (Connor and McLaughlin, 2006). This finding can be interpreted two ways: either those who are institutionalized have more severe illnesses than those who are not (and that is why institutionalization is correlated with longer recovery times), or it may be that being in an institution, removed from normal society and normal function, hampers a patient’s ability to cope, respond to treatment, and/or rebound. Taken together with evidence that implies that aggression, not illness severity, is the primary reason for institutionalization and that patients with a supportive network do better than those without (Leavy, 1983; Uchino et al., 1996), it is likely that being able to shift a patient from inpatient to outpatient status would expedite their recovery. So if a treatment for maladaptive aggression can be found, it would reduce the cost of operating an institution by limiting the dangerous population, drastically reduce
2.1. INTRODUCTION

the number of inpatients, and allow for more efficient research and treatment strategies for psychiatric illnesses. Lastly, the financial burden removed from society would be significant (Borduin, 1999).

Understanding the underlying phenomenon of maladaptive aggression would be the first step towards such a treatment. Evidence suggests that a set of underlying constructs are responsible for aggression and that treatments (both psychosocial and psychopharmacological) specifically targeted for aggression should be investigated (Connor and McLaughlin, 2006). The notion that similarities in the occurrence of maladaptive aggression across different co-occurring psychiatric illnesses is supported by this finding as well. Furthermore, this highlights the possibility of diagnosing aggression in its own right, separate from those co-occurring illnesses. In situations like these, researchers often develop models of the illness’s occurrence, specifically its causes and symptomology. Such a model, capable of analyzing the nature of aggression, would be a useful step in diagnosing and treating maladaptive aggression. To be effective, this model would separate adaptive aggressors from maladaptive aggressors in need of treatment as well as indicating an adequate treatment. This would be an aid to researchers and clinicians alike.

The information gathered from this model is necessary given the complex nature of aggression. Research to date implicates interaction effects in the development of the maladaptive aggressive phenotype (Connor, 2002). In other words, the maladaptive aggressive phenotype is not produced by simple relationships between risk factors in isolation. In actuality it appears to be a complex relationship between different aspects of a child’s biological makeup and development that moderate and mediate the appearance of the aggressive phenotype. To take an example from the preclinical literature, zebrafish require a three-way interaction between age, place in a social hierarchy, and experience aggressing to address their aggression levels (Ricci et al., 2008), as illustrated in Figure 2.1.

As is suggested above, genetics are heavily implicated in the development of maladaptive aggression, for a review, see (Nelson, 2006). Moreover, genetic effects appear to contribute

---

1Ideally, the most effective, but even identification of an adequate treatment would be significant.
2.1. INTRODUCTION

Figure 2.1: Evidence of a three-way interaction to describe the aggression levels for zebra fish: experience with aggressing, age, and place in a social dominance hierarchy (yellow plane indicates submissive fish, blue plane indicates dominant fish) interact to adequately explain the aggression level (height in the graph).

most-significantly as part of an interaction with familial and environmental factors (Nelson, 2006). Additionally, the number of interactions implicated by the literature prohibits an examination here, but for a review see (Connor, 2002; Nelson, 2006).

2.1.1 Subtyping aggression

Often, it appears that maladaptive aggressive behavior is a single phenotype characterized by unacceptable violence. However, there are at least two disparate underlying subtypes that can lead to the appearance of the single phenotype.

There are many ways that aggression has been subtyped: some classifications rely on the
actions taken, some on the target of the actions, and some the motivation behind the actions. However, only one classification system implicates separate biological underpinnings.

One way of classifying the aggressive acts has been by their furtiveness. Overt aggressive acts are those openly directed at a target, such as fighting, verbal aggression, self-directed aggression, mugging, and assault. In contrast, covert aggressive acts are those not directed at a target or those indirectly aimed at a target, such as property destruction and rumor spreading. Again, knowing if a patient is going to exhibit overt aggression would be helpful for the clinician or institution staff, but there is little underlying biology to indicate whether an aggressor’s tendencies will manifest in an overt or covert manner (Quay, 1986a,b; Loeber and Schmaling, 1985). By virtue of its nature, covert aggression is more difficult to detect, giving rise to a skew in the existing aggression data towards overt aggression. Taken together, complications arise in generalizing from a data set (with a strong selection bias towards overt aggressors) for the population of maladaptive aggressors.

There are a number of other classification systems used when attempting to subtype human aggression: offensive / defensive (Blanchard and Blanchard, 1984), relational (Crick and Grotpeter, 1995; Crick and Werner, 1998) or indirect (Hood, 1996), but most have the same failing: a paucity of neurobiological evidence. However, the maladaptive aggressive phenotype has also been subtyped by taking into account the motivation behind the aggressive acts. The resulting categories are reactive and proactive aggression classifications (Dollard et al., 1939; Hartup, 1974; Bandura, 1973). Unlike any other human classification, this way of subtyping aggression implicates separate underlying neurobiological pathways underlying the aggressive phenotypes.

The reactive aggression phenotype (also referred to as affective, defensive, or impulsive aggression) is derived from the Frustration-Aggression model (Berkowitz, 1989). Reactively aggressive acts are violent acts in response to a stressor, such as a threat to oneself or a loved one. As such, reactive aggression is classically seen as a lack of proper inhibition. This inhibition is not only manifest in uncontrolled (or undercontrolled) thoughts, but also in a high autonomic nervous system (ANS) response to stress and to aggressive episodes.
2.1. INTRODUCTION

In response to a stressor, reactive aggressors show increased heart rate, increased sweating (indicated by skin-conductance) and greater pupil dilation when compared to non-aggressive control subjects or proactive aggressors (Dodge, 1991; Crick and Dodge, 1996). The causes and risk factors of reactive aggression most often occur early in the aggressor’s life, often in the early childhood years (Dodge et al., 1997). Generally reactive aggression is displayed as a lack of inhibition and proper impulse control, both regarding thoughts and actions (Connor, 2002). Subsequent aggressive acts are typically disproportionately high in their intensity with respect to the stimulus eliciting them.

In contrast, the proactive aggression phenotype (also referred to as instrumental or predatory aggression) is derived from Bandura’s Social Learning Theory and is characterized by aggression used in pursuit of a goal (Bandura, 1977; Berkowitz, 1993). Thus, proactive aggression manifests in acts such as mugging, bullying, and carjacking, often-times co-occurring (Roland and Isdoe, 2001). Since proactive aggressors use aggression as a tool rather than displaying a lack of inhibition, it seems that this phenotype has risk and protective factors associated with a period in development when social strategies are learned. Thus, the risk factors for proactive aggression occur later in childhood than those for reactive aggression. Proactive aggressors also, in general, lack the characteristic high ANS response to stress seen by reactive aggressors, though this is not universally true. It has been theorized that proactive aggressors have learned how to control their aggressive tendencies, putting them to work in pursuit of their goals (Dodge et al., 1997). For them, aggression is simply a means to an end.

These two subtypes of aggression appear to have their antecedents diametrically opposed, with reactive aggression characterized by a lack of control and proper inhibition, while proactive aggression is the ability to control aggressive tendencies so as to use them in the pursuit of a goal. However, these phenotypes are sometimes both manifest in the same subjects, and often characterized by identical actions: violent acts. In order to further explore the nature of this seemingly contradictory relationship, the underlying neurobiology must be examined. All violent actions observed within either of these subtypes must in
some part be the product of the complex interplay of sensory, motor, and integrative areas of the brain, generating overtly similar behavioral responses. However it appears that aside from this overlap, the two subtypes of aggression show at least partially divergent neural pathways (Coccaro, 1989; Brunner et al., 1993; Grimes and Melloni Jr., 2006; Ricci et al., 2007; Nelson and Trainor, 2007; Barratt et al., 1999; Crowe and Blair, 2008).

Given all this evidence, it seems this problem is not one of a single underlying illness, but of two disparate deficits with similar symptoms (namely violent tendencies). Thus, instead of attempting to design a treatment (pharmacologic or otherwise) to treat aggression, one is needed for proactive aggression while another is needed for reactive aggression. At present, there are no pharmacologic treatments targeted for these subtypes, nor is there an easy way to differentiate between the two subtypes in a given patient, especially in the absence of contextual cues.

### 2.2 Risk and Protective Factors

The risk factors investigated are defined and typed by Raine, Brennan, and Farrington (Raine et al., 1997). Their division includes: familial, environmental, heritable, biological, pharmacological, cultural, and personality variables. As an example of the importance of the interaction between these variables, Cadoret and colleagues studied the interaction between environmental and biological factors. Twin studies indicate that environment alone accounts for the occurrence of 5% of aggressive tendencies, while genetic factors alone accounts for 10% of the same tendencies. However, when both are taken together, they serve to explain 40% of the variability (Cadoret et al., 1995). For a review, see (Connor, 2002). Clearly this is not an additive effect, and there is an interaction of some sort influencing the generation of the aggressive phenotype. Given the complex nature of this data set and the number of underlying individual factors and classes of factors, it appears likely that there are many undiscovered interactions that drive the formation of these aggressive tendencies. Thus, an analysis was undertaken which integrates all the different risk and protective factors in an
2.3. **SHORTCOMING OF STANDARD METHODS**

attempt to put together a comprehensive model, rather than only being concerned with a small subset of the known factors.

Such an investigation has not been performed before for a number of reasons. The sheer amount of data needed to adequately perform such an analysis is large by any standard. This poses two problems: attaining a large enough sample, and then undertaking a computational analysis beyond what most in this field are accustomed to. Additionally, the risk and protective factors are too numerous to be reasoned about efficiently, so the creation of a model through typical means is a difficult undertaking. Furthermore, until such a model can be created, the statistical techniques typical to the field are not going to be of much use, since they tend to center around confirmatory analyses rather than large-scale exploratory analyses. As such, machine learning techniques were used to create a model instead.

2.3 Shortcoming of Standard Methods

In order to adequately diagnose and effectively treat maladaptive aggression, clinicians and researchers require an understanding of what causes maladaptive aggressive tendencies. That is to say that we seek the factors which account for the difference between a healthy individual and an individual expressing maladaptive aggression. Since there are no comprehensive models or hypotheses that incorporate the contribution of all the known risk and protective factors (both individually and in concert with one another), an exploratory analysis is required.

There are statistics adept at confirming non-linear relationships, but they require a hypothesis of how all of the risk and protective factors interact to form this relationship. Moreover, they only provide information as to how well the proposed relationship explains the data, not what can be done to improve it. In theory, data sets with small numbers of variables can have non-linear relationships tested through linear regression (by creating a variable for each interaction and multiplying the contributing factors to represent the unique contribution of the presence of all the contributing factors), however applying this
work-around to larger data sets is problematic.

Ideally, all of the possible interactions need to be tested for their unique contribution to the developmental course of aggression. Unfortunately, this undertaking via the regression work-around is computationally unfeasible with current technology. The data from the Devereux Institute contained 63 items, for which the total combination of possible interactions is $2^{63} - 1$, approximately $9.2 \times 10^{18}$. If interactions were limited to a maximum of 5 items, the number of values would still be over $7.7 \times 10^7$. Even this reduced data set creates a problem of computational explosion, thus making the computation cumbersome (if even possible) and the interpretation complex and unrealistic. This is referred to in some fields as the *Curse of Dimensionality*, in that for every additional variable to be investigated, the computation grows exponentially and the interpretation of the results likewise (Bishop, 1995).

Aside from it being computationally daunting to explore such options, most techniques require more subjects than variables to provide an accurate result, in which case there simply have not been enough humans in the history of the race to satisfy the former condition, with the latter condition requiring far more subjects than are available to us. That’s not to say that we can ignore this concern entirely, but to point out the complete infeasibility of approaching this problem with standard methods.

To simplify the problem and make it more tractable, it would seem natural to attempt some form of data reduction. However, in doing so, we would run the risk of eliminating variables that individually contribute little to the overall variance explained, but are a part of an interaction that provides worthwhile information. Without testing each interaction the variable could be a part of, we would be unable to distinguish a variable that contributes little to the explained variance from a variable that contributes little on its own, but contributes significantly as part of an interaction.
2.4 Purpose

Though methods for examining large data sets exist for exploratory linear regression, confirmatory linear regression, and confirmatory non-linear regression, none exist in the standard repertoire for exploratory non-linear regression. Thus, we derived methodologies from fields where problems of this size and nature have been explored with some success.

There is an important balance to be struck between performance and interpretability. A perfectly fitting classifier that cannot be broken down to its component parts does little good in furthering research. Further, such a classifier is likely to be met with skepticism if we are unable to explain precisely why and how it works. Conversely, an easily decomposable classifier that has high error rates is as likely to contribute incorrect information as correct information and thus is of little value either as a diagnostic tool or as a research-driving technique. In truth, this pursuit is not of a single classifier, but of two classifiers: one to aid in diagnosis (diagnostic) and one to aid in research (exploratory). Ideally, of course, these two purposes could be served by a single well-fitting and easily decomposable classifier. Unfortunately, most emphasis placed on research in this area is on best-performing classifiers, not on interpretable ones. This results in the addition of mathematical transforms and computational tricks, which improve performance, but obfuscate the underlying relationships. Many of the techniques we will employ, therefore, are adaptations to these high-powered techniques, designed to increase the chance of interpretability, while potentially impacting the classification performance (i.e. misclassifying more patients).

Standard methods have fallen short in elucidating the occurrence of maladaptive aggression, while the need to understand and treat aggression (prophylactically or otherwise) has never been greater. Assessing the relative importance of the numerous factors implicated by the literature provides a clearer view of maladaptive aggression. Narrowing the field of possibilities that lead to maladaptive aggression allows for more targeted future research and the first step towards new, better-informed, treatment strategies.
2.5. MACHINE LEARNING

Figure 2.2: Pixelized representations of numbers, as if read from an LCD display. Each box representing a single pixel, blue for present, white for absent. These are simplified versions of what optical character recognition software uses when categorizing letters.

2.5 Machine Learning

The problem of deciphering the way in which risk and protective factors interact to produce the aggressive phenotype seems intractable, as all methods typically used to investigate these types of problems are inadequate (with respect to power or to compute time). However, there are techniques from other fields suited to similar identification problems. In essence, this problem is one of classification: separating those subjects that exhibited signs of maladaptive aggression (proactive, reactive, or both subtypes) from those that do not. The underlying assumption is that there is some element (more to the point, some combination of elements) in the collected data that can reliably differentiate between these classifications.

This task is similar to a number of classification problems examined in machine learning (a subfield of computer science), notably classification in computer vision. In essence, object classification based on visual cues is a problem of finding distinctive qualities in the visual field that differentiate between classes of objects. This problem can be seen in many different applications and at many levels of complexity: object recognition and classification in a scene (Serre et al., 2007), visual inspection of the machines in an assembly line (Stomski and Elmaghraby, 1989), handwriting recognition (Plamondon and Srihari, 2000), or optical character recognition (OCR) from printed text (Amit and Geman, 1997), for example. This
2.5. MACHINE LEARNING

Digital images of letters (or objects) are typically used to represent the visual field (pixelized numbers are shown in Figure 2.2). The image is quantified by moving a 'window' over the image and examining the substructures present in that window for each position, as shown in Figure 2.3. By limiting the size and complexity of the analyzed window, shape extraction is more easily accomplished when compared to analyzing the entire image (both in terms of the complexity of the shapes found and in the complexity of differentiating between those shapes). This allows the image to be quantified by the number of and location of these shapes present (as shown in Figure 2.4 for the digits 0 and 3). There are various methods for putting together these subimages and shape statistics to identify the character to be detected, the specifics of which are more cumbersome than useful. For these examples, the number of 'horizontal' bars, 'corners', and 'vertical' bars are counted, a combination of which gives an adequate differentiation between the two images (also note how large this set of 'features' would get if an entire image is analyzed). For the example in Figure 2.4, 3s and 0s can be differentiated by examining the 'horizontal' value alone. Greater than 3 'horizontal' windows indicates a 3, while less than 3 indicates a 0. If there

Figure 2.3: a. The moving window. b. The window is moved across the image in steps, with data collected at each step, each image above representing a single step. At each step the structure of the pixels in the window is recorded and compiled as show in Figure 2.4.
2.5. MACHINE LEARNING

Figure 2.4: a. The shape of the digit 3 in pixels, and the corresponding vector computed by the moving window method. The bottom row is a sum of each of the shapes, while the top row is the sum of all the sub-images (windows) possessing that shape (as indicated by the lines). The same is done for the digit 0 in b, while c is the digit 0 with a missing (erroneous) pixel (indicated in red). Despite the error, the resulting horizontal value for b and c match, while not matching the value for a.
Figure 2.5: This is the representative analog between optical character recognition and the application to maladaptive aggression. Instead of examining the presence of pixels, we examined the presence of risk and protective factors in a subject’s history.
2.5. MACHINE LEARNING

are 3 'horizontal' bars (indicating an error somewhere), we could further extend the rule to incorporate 'vertical' bars as well. Thus, this technique is at least relatively resistant to errors, as can be seen in Figure 2.4. Figure 2.4.c is the digit 0 missing a pixel (indicated in red), and though that changes the totals computed over the image, the differentiation by the shape count (i.e. horizontal bars) does not change. The sets of values drawn from the moving window identification (collectively referred to as a vector) are very similar to the data we used to represent a subject’s history, in that they represent data in which a pattern relevant to the classification is present. For notation, let \( V \) be the set of vectors, \( v_i \) be the \( i \)th vector in the set, and \( v_{ij} \) be the \( j \)th value of the \( i \)th vector. These vectors of pixels are not directly informative to the classification sought, but, simple shapes can be recognized by examining windows (sub-images) containing a number of pixels in close proximity. Furthermore, when many of these sub-images are examined together, their relative arrangement can indicate the shape of the letter (or object) in the visual field. In sum, though the pixels themselves are uninformative, an interaction between the pixels in close proximity can provide enough shape-information to correctly classify letters and numbers most of the time (Amit and Geman, 1997). Likewise, the information collected from a subject’s history is largely uninformative when taken alone, but may be useful when examined together for identifying the underlying construct(s) leading to aggression. The analog between OCR and maladaptive aggression is illustrated in Figure 2.5. Following this logic, we employed and adapted some techniques common in optical character recognition to the classification of aggressive individuals, substituting history variables for pixels and aggressive outcome for the letter detected.
Chapter 3

Data Set

3.1 Devereux Institute: Inpatient Data

Since the most effective treatments have been interventions into early life, the need to examine the incidence of maladaptive aggression in youth is clear: early detection and effective treatment. The Devereux Institute is an inpatient facility where some of the most violent youths are sent for care. The database we used is a snapshot database collected upon patient admission to the facility, compiled by Daniel Connor during his tenure as the director of the facility. Controls (primarily non-aggressors) were collected from a range of New England towns.

The data was collected from each patient upon entrance to the facility, providing a cross-sectional (snapshot) database. This data set covers a number of different categories of risk and protective factors collected (biological, environmental, parental, psychiatric, neurological, and medical), indicative of what a typical clinician can attain from examining a patient’s chart and conducting a diagnostic psychiatric interview. Thus, we examined the data available to a typical clinician when assessing the aggressive risk of a new patient.

As can be seen from Figure 3.1, the incidence of proactive aggression in this data set is very sparse. In the general populace, after an incident of proactive aggression, the ag-

---

1Due to legal restrictions placed on the facility, further data cannot be collected, thus eliminating the possibility of longitudinal data.
gessor is far more likely to end up in the correction system than the mental health system (unpublished observations, Daniel Connor). This might be explained by the perception of the motivation behind an aggressor's action: proactive aggression can be more easily rationalized than reactive aggression, as it is disproportionately intense in pursuit of a goal (and thus often seen as premeditated). Conversely, reactive aggression appears to be less rational, and thus the aggressors are more likely to be deemed mentally ill. Since this data set was collected upon entrance to a mental health facility, proactive aggressors are necessarily going to be undersampled. In this case, this undersampling is dramatic: only one proactive only aggressor out of 390 subjects. Additionally, an argument could be made that the proactive aggressor present in this data set is an atypical proactive aggressor, since they ended up in the mental health system rather than the correction system. Given that the proactive-only data was undersampled and likely misrepresented, we did not use it in our analysis. Further, attempting to classify the occurrence of both proactive and reactive aggression when there is no proactive-only data to examine (i.e. the appropriate negative side) seemed irrational at best. Thus, our analysis from here forward is limited to the differentiation of those subjects exhibiting reactive aggression from those that do not.

Even considering all the flaws mentioned above, this data set is the most comprehensive and largest available to date. So while it lacks significant biological measures implicated in many known interactions, it provides a look at what a typical clinician would see when admitting a patient. Thus, drawing direct conclusions regarding the nature of aggression is difficult and should be replicated in further studies (incorporating more biology, for example) before being accepted. However, direct conclusions can be drawn regarding the sorts and direction of questions that clinicians should focus on when attempting to establish the risk posed by a patient.

2Importantly, this does not impact the generalizability of the methods developed, as discussed in more detail later. Though the conclusions that can be drawn regarding aggression can only be applied to reactive aggression.
3.1. DEVEREUX INSTITUTE: INPATIENT DATA

Figure 3.1: a. The data set, broken down by subtype of aggression (Non, Proactive, Reactive, or Both). b. The data set, with proactive aggression excluded, broken down by subtype of aggression (Non or Reactive).
3.2 Novel and Training Data

It is not good enough to have a model capable of correctly classifying data it has already seen. The only way any model will be useful is if it is capable of correctly classifying data that was not used in its creation (as would be the case with a new patient whose aggressive phenotype is unknown). In order to test the ability of any model to generalize to novel data, the data set is divided into a training set and a novel set. The model will be trained only using data from the training set, and then evaluated on the novel data presented in the novel set. In order to attain a fair metric, and to avoid testing on a single random novel set, we divided the data set into three randomly selected blocks. Instead of training and evaluating the models once, the models were created from scratch three times, each time using a different one of the blocks for the novel data and the other two blocks for the training data. Thus, the metric of fit will be averaged over the three trials to give the final fit. In doing so, each data point is used once and only once for testing, in an unbiased manner.
Chapter 4

Methods

From this point forward, the mathematics of the techniques will be specified and discussed. Every effort has been made to explain the general principles in english directly prior to the formulas, so as to make this accessible to those unfamiliar with statistical pattern recognition and machine learning. Generally, the theoretical basis for each technique is spelled out before the equations used to compute it. Comprehension of the equations themselves is not necessary for understanding the endeavor (and to an extent) the approach taken, nor is it necessary to understand the significance of the findings. However, to aid those unfamiliar with the symbols used here, a glossary is included in Section 8.

An additional note, everything will be discussed only in terms of differentiating non-aggressors from reactive aggressors, but that does not limit these techniques for use in a single classification task. These techniques can be expanded (explicitly or implicitly) to encompass more complex classification tasks.

4.1 The $\Psi$ Metric

In order to give an accurate estimation of how well the techniques performed on all of the potential diagnoses, the patients were split into two groups based on their 'correct' diagnosis, as assessed by clinicians: (Non-Aggressive [Non], or Reactively Aggressive [Re]).
This diagnosis will be referred to as the patient’s correct diagnosis. The percent of correct classifications for each of these groups was computed as a metric of how well the technique predicted that specific diagnosis ($C_{Diagnosis}$). The average of the diagnoses was taken as a metric of the overall fit for a technique ($\Psi$). This was done to remove the bias created by the unequal numbers of the diagnoses.

$$C_{Re} = \frac{\# \text{Correctly classified 'Reactive Aggressive Only' Patients}}{\# \text{Total 'Reactive Aggressive Only' Patients}}$$

$$\Psi = \frac{\sum_{n \in D} C_n}{|D|}$$

where $D = \{\text{Non,Re}\}$ for our purposes, but can be expanded to encompass any set of subtypes. In keeping with the statistical pattern recognition field,

For the sake of clarity, the techniques will be analyzed and discussed in terms of performance (or loss) on differentiating Non-Reactive from Reactive patients ($D = \{\text{Non,Re}\}$), without loss of generality.

### 4.2 Diagnosis from Binary Outcomes

As previously stated, these techniques can be expanded to work on more complex classification tasks. Here is one approach to doing so, applicable to any pair of classifiers differentiating between binary (or boolean) inputs. If we were to examine the incidence of both Proactive and Reactive aggression, instead of the two outcome states possible ($D = \{\text{Non,Re}\}$), we have four possible outcome states Non Aggressive, Reactively Aggressive Only, Proactively Aggressive Only, or Both Proactively and Reactively Aggressive (more formally: $D = \{\text{Non,Re,Pro,Both}\}$). Either a single classifier can be created to determine the outcome of both Proactive and Reactive aggression (as seen in Figure 4.1), or two separate classifiers can be created, each to classify a single subtype (as seen in Figure 4.2). These two classifiers each produce a single binary decision regarding a subtype of aggression: one differentiates $Pro$ from $\neg Pro$ while the other differentiates $Re$ from $\neg$
4.3. TECHNIQUES

A composite of these binary outcomes can create a decision for one of the four desired diagnosis (Non, Pro Only, Re Only, or Both Pro and Re) in the following manner (as is also shown in Figures 4.1 and 4.2):

\[-Pro \land \neg Re \rightarrow \text{Non Aggressive}\]
\[Pro \land \neg Re \rightarrow \text{Proactively Aggressive}\]
\[-Pro \land Re \rightarrow \text{Reactively Aggressive}\]
\[Pro \land Re \rightarrow \text{Both Proactively and Reactively Aggressive}\]

4.3 Techniques

4.3.1 Normalization

All of the raw data was transformed before use in any of the techniques described. Typically in the social sciences, data normalization is accomplished by centering the data’s mean around 0 and constraining the range between −1 and 1 (as illustrated in Figure 4.3.a). Machine learning takes a different approach, constraining the range from 0 to 1, without regard for where the mean lies (as illustrated in Figure 4.3.b).

Rather than following either of those conventions we opted to explicitly encode the protective and risk values of the input factors, using the literature and Daniel Connor’s extensive clinical experience to differentiate between risk and protective values. Unless otherwise stated, each variable was normalized to the range \([-1, 1]\), with 0 reserved for missing data. Further, values implicated as 'protective' were relegated to the range \([-1, 0)\) and values implicated as 'risk' were assigned to the range \((0, 1]\). Variables with binary values were similarly assigned: the 'protective' value to \(-1\) and the 'risk' value to 1, with 0 reserved for missing data. This can be seen in Figure 4.3.c for continuous values, and 4.3.d for binary values. Though other techniques for filling in missing data could have been employed, this tactic was employed with specific regard towards creating predictive models, such that any data that was missing would have no effect (positive or negative), rather than
Figure 4.1: A classifier (everything below the dotted line) with two separate binary outcomes, Pro and Re. We can combine these two binary outcomes to produce one of four states expected for a diagnosis: Non Aggressive, Proactively Aggressive Only, Reactively Aggressive Only, Both Proactively and Reactively Aggressive. Since this is a modification of a classifier to produce two outputs instead of one, this is the implicit expansion described above.
Figure 4.2: A pair classifiers (everything below the dotted line) with two separate binary outcomes: one classifier for Pro and a second for Re. We can combine these two binary outcomes to produce one of four states expected for a diagnosis: Non Aggressive, Proactively Aggressive Only, Reactively Aggressive Only, Both Proactively and Reactively Aggressive. In this manner, no modification is required for the techniques to produce these classifiers, but different output variables are investigated (the specifics of this will be discussed in later sections).
any effect dictated by the rest of the training data. This transformation will be noted as $X'$ for the vector and $x_i'$ for individual values.

### 4.3.2 Regression

The approach generally applied to such a problem (at least in psychology) is multiple linear regression or logit regression (Aldrich and Nelson, 1984). This, however, only addresses the individual contributions of each factor (in linear combination), ignoring the possible interactions between two or more factors. Using multiple linear regression, the $r^2 = 0.37$ for the presence of reactive aggression in the Devereux data set. So 37% of the variance in the presence of reactive aggression is able to be explained by a linear model. However, this figure is misleading: multiple linear regression was designed to determine the relationship between dichotomous and continuous input variables to continuous output variables. Classification requires a dichotomous decision to be made, thus the output variable is not continuous, so the performance of multiple linear regression ($R(X)$) is not a fair metric, at least at face value.

$R'(X)$ is an adaptation of multiple linear regression that builds a classifier from a regression equation. While this does not compensate for the flaw (and the likely reduced-fit because of it), it will allow a fair evaluation of the performance of multiple linear regression. The regression equation provides information useful for classification, in the constant ($c$) and weights ($\beta_i$ for $i = 1...d$ where there are $d$ input variables). Given that the value on any outcome is either 0 or 1 (absent or present respectively), the point halfway between them (0.5) can be used as a threshold ($\tau$). Values from the regression equation ($R(X)$) greater than $\tau$ are classified as present (1) while those resulting values less than $\tau$ are classified as absent (0). This relationship is graphically represented in Figure 4.4 and formally defined as:

$$R(X) = \sum_{i=1}^{d} \beta_i(x_i') + c$$
4.3. TECHNIQUES

Figure 4.3: a. The spectrum of values generally used in psychology and psychiatry: range from $-1$ to 1 with the mean ($\bar{x}$) at 0. b. The spectrum of values generally used in machine learning: range from 0 to 1. c. and d. are the spectrums used here, a range from $-1$ to 1, corresponding to Protective and Risk values respectively. c. is the continuous version, where the output can be any number between $-1$ and 1. d. is the step-version, where the possible values are constrained to three values: $-1$, 0, or 1, as was used for dichotomous variables.
4.3. TECHNIQUES

Figure 4.4: Graphical representation of creating a classifier from a regression equation \( R(X) \) plotted on the \( y \)-axis, that takes a vector \( X \) as input, plotted on the \( x \)-axis. The shaded region indicates a classification of Reactive, whereas the unshaded region indicates a classification of Non.

\[
R'(X) = \begin{cases} 
0 & \text{for } R(X) < \tau \\ 
1 & \text{for } R(X) \geq \tau 
\end{cases}
\]

A similar approach (selecting the midpoint) can be used with logit to create a classifier, as indicated by Figure 4.5 (Aldrich and Nelson, 1984). Unfortunately, SPSS only allows a maximum of 9 input variables to be used as independent variables in a Logit Analysis, so in order to gain an adequate metric of performance, we would have to create our own implementation of logit regression.

4.3.3 K-Nearest-Neighbor

One common approach to pattern recognition is the \textit{K-Nearest-Neighbor} analysis (Cover and Hart, 1967). Generally, the idea is to find the subject in the training set that is most-similar a novel subject whose classification is unknown, and assume that the class label is
Figure 4.5: Graphical representation of creating a classifier from a Logit equation \( L(X) \) plotted on the y-axis, that takes a vector \( X \) as input, plotted on the x-axis. The shaded region indicates a classification of \textit{Reactive}, whereas the unshaded region indicates a classification of \textit{Non}. 
the same. This would be referred to as a 1-Nearest-Neighbor classifier. However, taking more than the closest single subject into account reduces error, providing more accurate classification. Allowing the $k$ nearest subjects in the training data set to ‘vote’ for the classification of the novel one allows a more accurate predictor.

A graphical representation of the problem is shown in Figure 4.6 with the $K$-Nearest-Neighbor approach shown in Figure 4.7. The $x$ and $y$ axes represent two hypothetical history variables ($n_x$ and $n_y$). A series of subjects from the training set are plotted with their expressed phenotype of aggression (in this case, Reactive [plus] or Non [circle]). A novel subject is also plotted (square), whose phenotype we wish to determine.

For the 1-Nearest-Neighbor case described above, Figure 4.7.a where $k = 1$ indicates how the single training data point in closest proximity determines the predicted outcome. As shown, as $k$ increases, the predicted value may change, given the relative distribution of the known data points. As seen in $k = 9$, the majority of the nearest points are of the Reactive class, rather than the Non-Aggressive class for $k \in \{1, 3, 5\}$, thus the prediction would be Non-aggressive for $k \in \{1, 3, 5\}$ and Reactive for $k = 9$.

The distance between vectors $l$ and $m$ (plotted as singular points in the figures) in these examples are merely the squared Euclidean distance:

$$\Delta_{lm} = \sum_{n=1}^{d} w_n (l_n - m_n)^2$$

where there are $d$ history variables examined ($d = 2$ for the examples and figures) and $w_n \in W$ is the $n$th entry in a weight vector. In Euclidean space, $w_n = 1$ for all $n$, so all dimensions carry equal weight. This naïve assumption is a direct result of the paucity of a priori hypotheses. There are a few methods to replace the use of Euclidean space with something more sensible.

First, we can use Mahalanobis Distance instead of Euclidean distance. Mahalanobis Distance takes the covariation of the history variables into account when calculating the difference scores (this approach is analogous to multidimensional scaling (Ripley, 1996)).
An example will demonstrate the usefulness of this: in the simplest case, let there be one underlying cause (a) that is represented by two variables in the data (say through two different questions: 1 and 2) when all the other causes (b and c) are represented by only one question each (b is represented by 3 and c is represented by 4). Since each question is given a single dimension in the computation of Euclidean distance, underlying cause a will be weighted twice that of b or c because a has two dimensions devoted to it, where b and c have only one apiece. However, when computing the Mahalanobis Distance, the fact that questions 1 and 2 always covary is taken into account. Thus the weight that 1 and 2 carry is reduced so that when taken together they carry weight equal to that which 3 or 4 carries individually. Mahalanobis distance, in this sense, can be used to more appropriately weight the variables such that these weights are not influenced by frequent covariation, and thus are no longer quite as constrained by the number of questions reflecting a single underlying cause (as is the case with Euclidean distance). A graphical representation can be found in Figure 4.9. Specifically, let $\Sigma$ be the covariance matrix and $\Sigma^{-1}$ be the inverse covariance matrix. Also, let $v^T$ be the transpose of any vector $v$. The Mahalanobis Distance between points $l$ and $m$ ($\Delta_{lm}^M$) is:

$$
\Delta_{lm}^M = (l - m)^T \Sigma^{-1} (l - m)
$$

Secondly, we can use the $\beta$ weights from a multiple linear regression (from the history variables onto the aggressive outcome) to determine how important each history variable is to the the overall outcome. Since we are computing distance, the use of negative weights would be counter productive, so the the dimensions are weighted by the absolute value of the $\beta$-weights or the $\beta$-weight squared instead. In this way, the importance of the relationship is reflected when computing the distance, regardless of the direction. This weighted Euclidean distance ($\Delta^\beta$) is:

$$
\Delta_{lm}^\beta = \sum_{n=1}^{d} |\beta_n| (l_n - m_n)^2
$$

Thirdly, we can combine these two to account for both how much the history variables covary and the weight they carry with regards the the aggressive outcome. Since the inverse
Figure 4.6: The $x$ and $y$ axis represent two hypothetical variables. Known phenotypes are plotted, along with an unclassified point.

covariance matrix $(\Sigma^{-1})$ is multiplied by the distance vector between $l$ and $m$ twice (once in the vector’s original form, and once as a transpose of the vector), we transformed the the weight vector to compensate. When $W$ is made up from $|\beta|$: 

$$W' = \sqrt{|\beta|}$$

and when $W$ is made up from $\beta^2$: 

$$W' = |\beta|$$

Thus, the distances computed for the $\beta$-weighted Mahalanobis distance between points $l$ and $m$ is:

$$\Delta^M_{lm} = (w(l-m))^T \Sigma^{-1} (w(l-m))$$

Though $K$-Nearest-Neighbor offers a simple classifier, this classifier does not provide any information regarding what aspects of a history vector made the classification correct or incorrect. While this may provide adequate empirical results, there is nothing about the
Figure 4.7: Representative Euclidean $K$-Nearest-Neighbor for $k = 1, 3, 5, 9$ with the resulting outcome. The black square is the subject to be identified and the grey circles represent the $k$ closest neighbors being taken into account.
4.3. TECHNIQUES

Figure 4.8: Representative $\beta$-weighted Euclidean K-Nearest-Neighbor for $k = 1, 3, 5, 9$ with the resulting outcome. In this example, the history variable plotted on the x axis is given more weight than the history variable on the y axis, resulting in an oval-shaped grouping of the $k$ nearest neighbors.
4.3. TECHNIQUES

Reactive  Non  Unknown

a. $k = 1$, Result:  
b. $k = 3$, Result: 
c. $k = 5$, Result:  
d. $k = 9$, Result:

Figure 4.9: Representative Mahalanobis distance $K$-Nearest-Neighbor for $k = 1, 3, 5, 9$ with the resulting outcome. In this example, the history variables have some covariance to them, so the grouping of the $k$ nearest neighbors is oval-shaped and on an angle inversely proportional to the covariance.
underlying relationship that can be inferred from the method itself. Some of the adaptations mentioned above could be interpreted to glean further information, but would be nothing in addition to what the statistics used to generate $W$ (i.e. covariance matrix, $\beta$ weights, etc.) would provide. Additionally, the history vector is only examined by linear combination, so the important and unimportant interactions are indistinguishable.

4.3.4 Neural Networks

Classification via neural networks is prevalent in machine learning, utilizing an abstraction of the framework for a set of neurons as a method to model interactions. It is important to note that despite the technique’s name, it is not being used to model actual neurons in brain. Instead, the operation of the network is modeled after a group of neurons. In order to prevent the confusion created by the unfortunate misnomer, any reference to a set of actual neurons in the brain will be referred to as a pathway, thus reserving the term neural network or network for the machine learning technique. These networks have previously been shown effective in many pattern recognition situations including the computer vision problem outlined above (Serre et al., 2007; Amit and Geman, 1997; Stomski and Elmaghraby, 1989; Plamondon and Srihari, 2000). They also scale with the dimensionality of the problem, making their use ideal for large data sets, both in terms of number of subjects and in terms of variables per subject (Bishop, 1995).

The general application of a neural network is to uncover hidden relationships between variables. Moreover, the network serves to separate the few interactions of inputs meaningful to the classification of the output from the numerous non-meaningful interactions. In other words, this relationship is a mapping of how the precursors lead to the outcome or classification\(^1\). There are a number of techniques for developing this mapping, which will be explored in more detail later. For now, let us assume that the mapping provides clues as to how the inputs interact to influence the outputs. In our case, the inputs are the patient histories and the outputs are their aggressive phenotype (Non-Aggressive or

\(^1\)While some of the relationships may be causal, the network does not imply causality in and of itself; it is a correlative technique.
4.3. TECHNIQUES

Figure 4.10: The general structure of a neural network

Reactively Aggressive). In order to understand how a neural network can learn and test all these possible interactions, let us examine its components.

A neural network (represented in Figure 4.10) is composed of a series of nodes \( (n_i \in N) \) representing the various aspects of the relationship between input and output. The nodes at the bottom of the figure (green rectangles) represent the inputs. In this case, these are variables used to represent a patient’s history. Sometimes these nodes will be labeled by whatever real-world state they correspond to. \( n_{Re} \) represents reactive aggression, for example. The nodes at the top of the figure (purple diamonds) represent the aggressive phenotype or outcome. Finally, there are the unnamed nodes in the middle (blue circles) which represent some form of interaction between the nodes that feed into them, rather than history variables in and of themselves \( (n_i \text{ for } i = 1, ..., m) \). These unnamed nodes are referred to as the hidden layer or hidden nodes since they represent states not directly observed in the patient. Each of these nodes (input, hidden, and output) can be activated or inhibited, indicating the presence or absence of the state they represent. Let \( a_i \) be the activation of \( n_i \).

\(^{2}\)These states may or may not be directly observable.
4.3. TECHNIQUES

While nodes represent different states within a patient (be it history, interactions, or phenotype), connections between the nodes (referred to as edges) represent the real-world associations between those states. More to the point, these edges have weights associated with them to represent the strength and polarity of the connection between two states. This allows for a representation of how one state \((n_i)\) can influence another state \((n_j)\). \(\omega_{ij}\) represents the weight of the edge between \(n_i\) and \(n_j\) (and thus the strength of influence that \(n_i\) has on \(n_j\)). This influence is represented in the network by changing the activation of \(n_j\) in some manner dependent upon both the activation of \(n_i\) and the weight of the edge connecting \(n_i\) to \(n_j\). In turn, \(n_j\) will influence a node that it is connected to \((n_k)\) via edge \(\omega_{jk}\) and so on up the entire network. In the networks used here, activation can only propagate from input to output (this is referred to as a feed-forward network). Thus any \(n_i\) can have an influence on \(n_j\) but not vice versa (represented by \(n_i \prec n_j\)). For simplicity, the nodes are arranged in layers such that a node can only receive input from the layers below it and project output to the layers above it. Importantly, nodes in the same layer cannot influence each other, thus maintaining a feed-forward network. It is this feed-forward activation that will be used to determine the outcome (phenotype) predicted by the model given the input (history).

Up until this point, this system still seems linear (in fact approximating linear regression), thus lacking any clear advantage over the existing methodologies. In order to accommodate nonlinearity there are a few changes to be made: the inclusion of a transform function, a thresholding function, and/or multiplicative edges.

The first method, often employed in neural network applications, is the transformation of activation levels in a logistic manner (Bishop, 1995; Ripley, 1996; Rumelhart and McClelland, 1986; Devroye et al., 1997). After a node’s incoming activation has been summed up, the following is generally used to compute the activation to be propagated by that node \((a_i)\):

\[
a_i = \frac{e^x}{1 + e^x}
\]
4.3. TECHNIQUES

Figure 4.11: The standard sigmoidal transform: \( y = \frac{e^x}{1 + e^x} \) applied to the incoming activation of a node (\( x \) axis) to get the outgoing activation (\( y \) axis).

This log transform (graphically represented in Figure 3.11) constrains the range of \( a_i \) from 0 to 1, regardless of the summed incoming activation. In this manner, each node adds a non-linearity to the model, rather than merely passing on the summed activations. Additionally, the output of each node is constrained to identical maximum and minimum activations, despite the magnitude of the incoming activation to those nodes. For our purposes, however, the inclusion of the negative activation is important (otherwise there is less of a mechanism to propagate the difference between the protective spectrum of a value and a missing or neither protective nor risk area of the spectrum). So when referring to a sigmoidal transform, we mean the following adaptation, represented graphically in Figure 3.12:

\[
a_i = \frac{2e^x}{1 + e^x} - \frac{1}{2}
\]

One drawback to the use of the sigmoidal transformation, however, is the lack of a clear indication when a risk or protective factor is present. Sigmoidal transforms are useful for increasing the performance of a classifier, but since they (almost) always pass on activation of some magnitude, the interpretation of that classifier is made more complex. At what level do we place the transition from risk to neither risk nor protective and then again from neither risk nor protective to protective? A more interpretable, though less elegant transform is a linear step function. If we define a threshold value to separate the three potential
Figure 4.12: Our modification to the sigmoidal transform \( y = \frac{2e^x}{1+e^x} - \frac{1}{2} \) incoming activation on the \( x \) axis and outgoing activation on the \( y \) axis. This modification allows for outgoing activation to range from \(-1\) to \(1\) instead of the standard \(0\) to \(1\).

outcomes, we can propagate a different activation for each outcome of a state (e.g. \(1\) for \textit{risk}, \(0\) for \textit{neither risk nor protective} and \(-1\) for \textit{protective}), while immensely simplifying the eventual interpretation of the model. In essence, we are trading the continuous sigmoidal transform for a non-continuous (binning) transform, allowing conclusions regarding the presence or absence of a given underlying state to be clear-cut.

Let \( \tau_i \) be the threshold at \( n_i \).\(^3\) The activation of \( n_i \) only propagates to the nodes downstream of \( n_i \) (\( \forall n_j : n_i < n_j \)) if the activation exceeds the threshold (\( |a_i| \geq \tau_i \)). There are a few methods for how thresholding can control the flow of activation (and thus the way in which nodes influence each other). \( \mathcal{X} \) will be used in the equations to denote the following linear step function:

\[
\mathcal{X}_i(a_i) = \begin{cases} 
-1 & \text{for } a_i \leq -\tau_i \\
0 & \text{for } -\tau_i < a_i < \tau_i \\
1 & \text{for } a_i \geq \tau_i 
\end{cases}
\]

\( \mathcal{X}_i \) is a linear step function, only propagating a signal if the activation of \( n_i \) is larger than the threshold at \( n_i \) (\(|a_i| > \tau_i \)). Additionally, it propagates the same signal regardless of the extent to which the activation exceeds the threshold. Thus, the output of \( \mathcal{X}_i(a_i) \) is a good

\(^3\)Where sigmoidal transforms are used there are no thresholds, so \( \tau_i = 0 \), necessarily.
4.3. TECHNIQUES

\[ X^i(a_i) = \begin{cases} 
-1 & \text{for } a_i \leq -\tau_i \\
0 & \text{for } -\tau_i < a_i < \tau_i \\
1 & \text{for } a_i \geq \tau_i 
\end{cases} \]

\[ X^{s}_i(a_i) = 2(0.5 - e^x / 1 + e^x) \]

\[ X^{1}_i(a_i) = \begin{cases} 
0 & \text{for } a_i < \tau_i \\
0 & \text{for } a_i \geq \tau_i 
\end{cases} \]

\[ X^{-1}_i(a_i) = \begin{cases} 
-1 & \text{for } a_i \leq -\tau_i \\
0 & \text{for } -\tau_i < a_i < \tau_i \\
0 & \text{for } a_i \geq \tau_i 
\end{cases} \]

Table 4.1: Various methods of determining activation to be passed on by a node (y axis) from the sum of the incoming activation to the node (x axis).
representation of the presence or absence of the state represented by \( n_i \) in the patient. We implemented and evaluated a number of alternate approaches to this step function, shown in Table 3.1.

The third manner in which we can introduce nonlinearity to the model is via more complicated edge types: multiplicative edges. Like additive edges (\( \omega_{ij} \)), multiplicative edges represent a relationship between the nodes they connect. Unlike the additive edges, however, these multiplicative edges take the activation of two nodes as an input. In order to determine the change in the activation of the downstream node, the activations of both precursor nodes must be present (more specifically, \( \mathcal{X}_i(a_i) \neq 0 \) for both nodes). Given \( n_i \), \( n_h \), and \( n_j \) such that \( n_i \prec n_j \) and \( n_h \prec n_j \), let \( \omega_{hij} \) be the multiplicative edge that allows \( n_h \) and \( n_i \) to influence \( n_j \) together.

In implementation, these multiplicative edges have two distinct behaviors, separated by how the input activations are combined. The first we will refer to as gating edges (though they are also referred to as \( \Sigma \Pi \) edges in (Rumelhart and McClelland, 1986)), and are governed by the following:

\[
\Delta a_j = \omega_{hij}^G \times \begin{cases} 
  a_h a_i & \text{if } |a_h| > \tau_h \text{ and } |a_j| > \tau_i \\
  0 & \text{otherwise}
\end{cases}
\]

Gating edge behavior is graphically represented in 4.13. These edges are less directly interpretable, but have at least one notable property: it has previously been shown that any interaction between more than two variables (as you would expect for a three-way interaction, say) can be effectively modeled with these two-input multiplicative edges, avoiding the need for three-input multiplicative edges (Rumelhart and McClelland, 1986). However, this is contingent upon the range potential outputs being from 0 to 1, not from \(-1 \) to 1 like we have here.
Figure 4.13: The behavior of a *gating edge*; activation (indicated by + or − on the edge) is only passed along if both input nodes (blue rectangles) are activated (indicated by + or − on the node). The polarity of the activation passed along is determined by multiplying the input activations.
Figure 4.14: The behavior of our modification to the Rumelhart and McClelland *gating edge*: activation (indicated by + or − on the edge) is only passed along if both input nodes (blue rectangles) are activated (indicated by + or − on the node) and have opposite polarities. The polarity of the activation passed along is determined by the first node (the ordering of which is arbitrary, but consistency is important).
Figure 4.15: The behavior of a interaction edge; activation (indicated by + or − on the edge) is only passed along if both input nodes (blue rectangles) are activated in the same direction (indicated by + or − on the nodes). The polarity of the activation passed along is determined by taking the polarity of a single input node (thus matching the direction).
4.3. TECHNIQUES

The second we will refer to as interaction edges, and are governed by the following:

\[
\Delta_{a_j} = \omega_{hij}^T \times \begin{cases} 
  a_h a_i & \text{if } a_h > \tau_h \text{ and } a_i > \tau_i \\
  -a_h a_i & \text{if } a_h < \tau_h \text{ and } a_i < \tau_i \\
  0 & \text{otherwise}
\end{cases}
\]

Interaction edge behavior is graphically represented in Figure 4.15. These edges are designed to model the effect of having two factors of like polarity (protective and risk for negative and positive respectively) with an effect beyond what each contributes alone, and only when both are present. This is not typical in the neural network field, but it is likely to provide a level of interpretability that the gating \((\omega^G)\) edge does not. The interaction edges are contingent upon our correct assessment regarding which values of a given factor are protective and which are risk. This can be compensated for by incorporating another variation of the multiplicative edges, simply propagating activation when the polarity of activation of both input nodes does not match (but also are non-zero), as is illustrated in Figure 4.14.

In these multiplicative edge types, the absence of activation in either of the precursor nodes \((n_i \text{ or } n_h)\) will mean that this edge has no effect on the activation of the downstream node \(n_j\). With this property in mind, the multiplicative edge becomes a good component to represent the interactions examined in the network.

Now that all components have been examined and explained, the activation of a single node is as follows:

\[
\Omega_j = \sum_{i=0}^{n} \omega_{ij} x_i(a_i)
\]

\[
\Omega_j^G = \sum_{h,i=0}^{n} \omega_{hij}^G x_i(a_i)x_h(a_h)
\]

\[
\Omega_j^T = \sum_{h,i=0}^{n} \omega_{hij}^T I\{x_i(a_i) > 0\} = (x_h(a_h) > 0) x_i(a_i)x_h(a_h) - \sum_{h,i=0}^{n} \omega_{hij}^T I\{x_i(a_i) < 0\} = (x_h(a_h) < 0) x_i(a_i)x_h(a_h)
\]
Figure 4.16: Components of a node with a threshold transform ($n_i$): $\omega$ are the edges, with arrows indicating direction of activation flow. Activation from previous nodes feed into $n_i$ from the edges on the bottom. This activation is summed to produce the raw activation of the node ($a_i$). This activation is transformed via a step function to produce the outgoing activation. This outgoing activation is propagated via the top edges to the nodes $n_i$ influences.
Figure 4.17: Components of a node with a sigmoidal transform \((n_i)\): \(\omega\) are the edges, with arrows indicating direction of activation flow. Activation from previous nodes feed into \(n_i\) from the edges on the bottom. This activation is summed to produce the raw activation of the node \((a_i)\). This activation is transformed via a sigmoidal function to produce the outgoing activation. This outgoing activation is propagated via the top edges to the nodes \(n_i\) influences.
4.3. TECHNIQUES

\[ a_j = \Omega_j + \Omega^G_j + \Omega^I_j \]

\[ \forall n_j \in N \text{ where } \exists \omega_{ij} \land n_i < n_j \]

where \( \Omega_j \) is the contribution from additive edges, \( \Omega^G_j \) is the contribution from gating edges, and \( \Omega^I_j \) represents the contribution from interactive edges. For a graphical representation of an individual node and its components, see Figure 3.16 for one with a threshold transform, and see Figure 3.17 for one with a sigmoidal transform. For the overall schematic of the construction of a neural networks, see Figure 4.19.

The states and conditions present in the patient will be activated (for state \( i, X_i > 0 \)), those not present will not be activated (\( X_i = 0 \)), and those inhibited or protected against will be negatively activated (\( X_i \leq 0 \)). This holds for all states: input, hidden/interaction, and output. Thus, after propagation, if \( n_{Re} \) is activated, the network predicts that this patient will display reactive aggression, whereas if \( n_{Re} \) is not activated or inhibited, the network predicts that this patient will not display reactive aggression.

In sum, a neural network that accurately models the relationship between a patient’s history and their aggressive phenotype will activate the correct pattern of output nodes when given the patient’s history as an input vector. The next section will address how to develop this accurate model using the data available.

Training

There are a number of parameters that specify how this model was created, and the optimum values were picked by an evolutionary algorithm (described more fully in Section 4.3.5). So for the moment, let us assume that we have optimum or near-optimum values for these parameters.

Before any model is created, we have a few things held constant: the set of input nodes (each representing a value in our data set) and the set of output nodes (each corresponding to an aggressive phenotype, and in our running example, only a single output node \( n_{Re} \)). The network manipulates the nodes and edges between the input and output (collectively the...
Figure 4.18: The propagation of activation (indicated in orange) and inhibition (indicated in blue) through a neural network is shown here in steps. The first image shows the activation of the input layer for a single subject who has a protective value in the IQ variable and a risk factor in the Drug variable, with no information available on the Gang variable. The second image shows the activation propagating from these input nodes to the nodes they are connected to. A node is activated (\(n_1\)) or inhibited (\(n_2\)) by summing the input activation. The third image shows the activation reaching the output node (Reactive). The output node is inhibited in this case, so the model predicts this person does not display maladaptive aggression.
4.3. TECHNIQUES

Figure 4.19: Graphical representation of an entire neural network and its functioning mathematical components. $\sum$'s are points where the input from the connecting edges (additive and multiplicative alike) are summed. This summed activation is then entered into $\mathcal{X}$ (the step-function transform) on the $x$-axis of the step-function graph (which represents the transform in the diagram). The resulting value ($y$-axis) is propagated on to all the connected edges. There is a slight difference in the transform on the output node, as indicated. The output can only be activated or non-activated, so inhibited values were incorporated into non-activated.
hidden nodes) to improve the accuracy of classification. Both the number of hidden nodes and the number of layers of hidden nodes are parameters to be tuned by the procedure explained in Section 4.3.5. In order to allow for all possible interactions, the network is maximally connected. In other words, every node of the graph has an edge connecting it to every other node in the network (save those in the same layer). At the outset, each of these edges is assigned a random weight where \(-1 < \omega < -0.5\) or \(0.5 < \omega < 1\) and thus, represents only random connections, with no ability to detect aggression beyond chance. Pre-training performance tests support this premise. In order to train the network to recognize the underlying relationship, we used data for which we can connect a history vector with the corresponding aggressive phenotype: the Devereux Institute admission data.

Some of these initial chaotic connections are meaningful to the detection of the underlying relationship while others are not. We wanted to strengthen the influence of the former while weakening the influence of the latter. We separated the two by presenting the model with a patient’s history vector and propagating the activation through the network. The resulting activation of the output layer indicated what the network, in its current state (and more specifically, this configuration of weights), predicted for this patient. Since we have information about what aggressive phenotype this patient actually exhibited, we then checked the output of the model to evaluate its accuracy. When the model correctly classified the patient, we changed nothing. When the model incorrectly classified the patient, we weakened the edges that led to the incorrect conclusions and strengthened the edges that would have lead to the correct conclusion (as is shown in Figure 4.20). We repeated this procedure for each patient in the training set, and then repeated the entire training set a number of times (ranging from 1-10, tuned by the procedure described in Section 4.3.5). At the end, we expected to have decayed away the non-meaningful links while strengthening the meaningful ones.

For example, when \(n_{Re}\) provided a correct outcome (\(X_{Re} > \tau\) for a phenotype that was present, and \(X_{Re} \leq \tau\) for a phenotype that was absent; \(\tau = 0\) if sigmoidal transforms are used), there is also some merit to non-maximally connected networks, but the resulting interpretation regarding the relative usefulness of variables is less straightforward than the maximally-connected approach.
are used) then we did nothing. However, when \( n_{Re} \) provided the incorrect outcome, we recursively applied the following procedure (assuming that \( n_i \) is the current node, \( a_i^* \) is the correct activation, and we start with a \textit{strengthen} flag):

1. Examine the veracity of each edge feeding in to the node \( V_{ij} = a_i^* \times \omega_{ij} \times a_j \).
2. If \( V_{ij} > 0 \) set the same flag as coming in [in the initial one, \textit{strengthen}] \(^5\)
3. If \( V_{ij} < 0 \) set the opposite flag as the one coming in [in the initial one, \textit{weaken}]
4. If the flag we try to set is already set, we can exit the procedure
5. Recur for every edge feeding in to \( n_i \), passing along the type of flag set.

This is shown graphically in Figure 4.20. Once the recursion reached the input nodes (and thus has no further to go) we went through each edge and increased the weight of those edges with \textit{strengthen} flags and decreased the weight of those with \textit{weaken} flags. Further, if both flags are set, we do not change the weight.

The amount the weight is changed is called the \textit{delta rule}. This can be derived from the \textit{Generalized Delta Rule} (Rumelhart and McClelland, 1986):

\[
\Delta_{\omega_{ji}} = \eta(t_{pj} - o_{pj})i_p
\]

where \( \eta \) is a constant (optimized by the algorithm described in Section 4.3.5), \( t_{pj} \) is the target (correct) activation, \( o_{pj} \) is the obtained (actual) activation, and \( i_p \) is the activation at node \( p \). Since the target and actual activations are binary values, \( t_{pj} - o_{pj} \in \{-1, 0, 1\} \). Further, \( i \in \{-1, 0, 1\} \) since the linear step function is used. So, \( \Delta_{\omega_{ji}} = 0 \) when the output is correct \( (t_{pj} - o_{pj} = 0) \) or if the node \( i \) is not activated (and thus does not contribute

\(^5\)If the flag we try to set is already set, we can exit the procedure since the path we are going to strengthen or weaken has already been appropriately flagged.
Figure 4.20: Progression of training for the neural network. The top row indicates when \textit{Re} patient is presented but the network labels them as \textit{Non}. The edges that contribute positive activation were strengthened (indicated by green edge outlines) and those that contributed negative activation were weakened (indicated by blue edge outlines). The final image in the row indicates that after training, when the same patient is presented, now labeled (correctly) as \textit{Re}. The bottom row indicates when a \textit{Non} patient is labeled by the network as \textit{Re}. The edges that contribute positive activation are weakened and those that contribute negative activation are strengthened. The final image in the row indicates that after training, when the same patient is presented, now labeled (correctly) as \textit{Non}. Additionally, when the \textit{Re} patient is presented again, they too are labeled correctly as \textit{Re}. 
anything to the outcome. The simplified version of the equation is thus:

\[ \Delta_{\omega_{ji}} = \begin{cases} 
\eta & \text{if } i = 1 \\
0 & \text{if } i = 0 \\
-\eta & \text{if } i = -1 
\end{cases} \]

Up until this point, there is no mechanism for edges of the wrong polarity (say a positive weight where it should be a negative weight) to switch. In order to accommodate this situation, a polarity switch point was established (and tuned via the algorithm in Section 4.3.5). When an edge was flagged to be weakened beyond this point, it instead had its sign inverted. Thus, any edge, regardless of its initial weight and polarity could achieve any (reasonable) weight. Additionally, if an edge represented a relationship that did not exist (and thus theoretically was strengthened and weakened more or less randomly), we would expect the edge to flip-flop between positive and negative polarity a number of times over the training period. In order to remove the noise that such a edges contributed, a time to live parameter was also instituted (and tuned via the algorithm in Section 4.3.5). Each edge was given a number of times it was allowed to switch polarity before it was deemed noise and deleted from the system (the tuned value of that parameter was 3).

**Interpreting a Neural Network**

One major drawback of using a standard neural network comes in the lack of interpretability, despite potential improved performance over other techniques. We have outlined procedures to reduce this complexity as much as possible, but even with these measures in place, the number of relationships examined (by expanding from linear combinations of the input vectors to nonlinear combinations) is increased exponentially, thus decreasing the interpretability at least geometrically. In the trained model, relationships (individual and interaction alike) have weights that should be proportional to their importance (at least in relation to those feeding in to the same node). This can be extended to allow weight-based comparisons across the entire model by normalizing the thresholds of each node (and like-
4.3. TECHNIQUES

wise normalizing each weight \( \omega'_{ijk} = \omega_{ijk}/\tau_k \). In essence, after this normalization each \( \tau = 1 \). This, however, still leaves a great number of relationships to be examined. Since many of these weights are working against one another, it is possible that two or more relationships only exist to balance each other out. In light of this, after training is completed, a streamlining process was employed to cull the unnecessary relationships (both nodes and edges).

For each component (node or edge), the model’s performance was assessed on the entirety of the training set, with and without the component in question. If removing the component was not detrimental\(^6\) to the performance of the model, the component was eliminated. To remove any effects of order, this procedure was applied once progressing forward (from the beginning to the end of the nodes and from the beginning to the end of the edges) and then again progressing backwards (from the end to the beginning for each). Differences in order can be attributed to covariation in the input nodes, and thus the forward-streamlining and backward streamlining were combined to reach any conclusions. Additionally, since the elimination of a component has the ability to remove the balance of a component previously examined, this procedure is recursively applied until one pass of the entire model warrants no changes (i.e. there are no more component to remove without negatively impacting the performance).

4.3.5 Evolutionary Algorithms

Another machine learning technique employed is from the class of techniques commonly referred to as Evolutionary Algorithms (EAs). EAs draw from Darwinian theory (Darwin, 1859), using techniques based on genetics and natural selection to solve optimization problems. In this case, we are interested in optimizing the values for the parameters of other techniques (for example, the number of layers and number of hidden nodes in the neural network). Not all values for the parameters are optimum, and if the parameters were in-

---

\(^6\)If using the sigmoidal transform, there is no need to normalize the weight of the edges, since the thresholds are necessarily the same: \( \tau = 0 \).  
\(^7\)Specifically, if the fit within each possible diagnosis (D) did not decrease the component was eliminated. Increases in \( \Psi \) via deletion were permitted.
dependent, they could be methodically examined and maximized. However, many of the
parameters are interdependent, so finding the optimum value for parameters \( p_i \) and \( p_j \) is
non-trivial. If we seek to optimize \( p_i \) while holding \( p_j \) constant, we will get the optimum
value for \( p_i \) for that particular value of \( p_j \). If we then seek to optimize \( p_j \) and we change
the value of \( p_j \) in the process, \( p_i \) may no longer be the optimum value, and thus we would
continue ad infinitum. Instead, if we vary all of the values together, letting them all settle
together to an optimized set of parameters, we can optimize the system (the combination
of all of the parameters) regardless of the interdependent nature of the parameters.

The exact employment of these EAs in the current research is detailed more fully in the
Synergy sections. Here we will only focus on how EAs operate and the function of the EA
used here.

In order to use an EA, a fitness function is needed to evaluate how close to an optimum
solution any set of parameters is. Let \( v_i \in V \) represent the vector for a set of parameters.
Further, let \( v_{ij} \) be the \( j \)-th factor of the \( i \)-th vector such that \( 0 \leq v_{ij} \leq 1 \). Let \( F(v_i) \)
be the fitness of vector \( v_i \in V \). The exact equation used for \( F(x) \) will change depending
upon the problem it is used for, but for our purposes we will always seek to minimize \( F(x) \),
(\( \min(F(x)) \) or \( F^* \)).

Generally, the EA works in terms of generations (\( G^{(t)} \) for the set of all individuals in
the generation at time \( t \), and \( G^{(t)}_i \) for individual \( i \) at time \( t \)). Each of these generations is
made of parents (\( P^{(t)} \) for the parents at time \( t \), \( P^{(t)}_i \) for an individual parent \( i \) at time \( t \))
and offspring (\( Q^{(t)} \) for the offspring at time \( t \), and \( Q^{(t)}_i \) for an individual offspring \( i \) at time
\( t \)). Each individual in a generation is composed of genes, represented by a vector (\( v \in V \)).

EAs are primarily classified by the makeup of each of their generations of offspring.
We implement a \((\mu, \lambda)\)EA, which means that we use \( \mu \) of the best solutions from \( G^{(t-1)} \)
to make up \( P^{(t)} \). The entire \( G^{(t)} \) is made from \( \lambda \) offspring (\( Q^{(t)} \)) bred from the \( \mu \) parents
(\( P^{(t)} \)) (notably excluding \( P^{(t)} \) from \( G^{(t)} \)). Another common type of EA is the \((\mu + \lambda)\)EA.
This is exactly the same as \((\mu, \lambda)\)EA but generation \( n \) consists of \( \lambda \) offspring derived from
\( \mu \) parents, including the \( \mu \) parents. To be specific, this work was done with a \((3,10)\)EA.
4.3. TECHNIQUES

Initially we assign random values \(0 \leq v_{ij} \leq 1\) to each gene for each individual in \(G^{(0)}\). For every subsequent generation,

\[
P^{(t)} = \min_\lambda (F(v_i)) \forall i \in G^{(t-1)}
\]

In other words, the parents for generation \(t\) are the \(\lambda\) individuals with the best fit from generation \(t-1\). Each individual in \(Q^{(t)}\) is then generated by randomly selecting each of the possible values in \(v_i\) with the following probability (\(h\) is a random individual selected from the group listed, selected separately for each gene):

\[
\forall Q^{(t)}\text{in the following manner: } Q^{(t)}_{ij} = \begin{cases} 
0.4 & P^{(t)}_h \\
0.5 & P^{(t)}_h + r \quad \text{where } -0.18 < r < 0.18 \\
0.1 & r \quad \text{where } 0 \leq r \leq 1 
\end{cases}
\]

Once all \(Q^{(t)}\) are established, all of \(G^{(t)}\) is evaluated by \(F\). From this set, \(P^{(t+1)}\) are selected, \(Q^{(t+1)}\) are created from \(P^{(t+1)}\), and the process repeats itself until some predetermined value of \(F^{(t)}\) is reached.

4.3.6 Collaborative Classifiers: Forests

In the techniques described above there is an amount of variation between the models created (this is especially true for the neural networks), contributed to by both random starting conditions and the use of only a subset of the data as training data. Though the latter is a difficult problem to address, the former can be countered, to some extent, via a technique commonly referred to as forests or collaborative classifiers. The name comes from placing a number of decision trees together, thus cleverly making a forest (Breiman, 2001).

This technique relies upon the fact that a number of models may have similar performance, but do so by misclassifying different subjects (as illustrated in Figure 4.21. If each misclassification is due either to something inherent about the subject (an outlier) or some-
4.3. TECHNIQUES

Figure 4.21: Depiction of how combining a few models that misclassify different patients can be advantageous. The first three circles in a and b each represent a model, the yellow is patients that they classify correctly, the black ones that they classify incorrectly. The single circle on the right is representative of using all three together to make a classification, each providing a vote for the classification. The patients encompassed by one of three black circles are correctly classified, and by two or three circles are misclassified. a. indicates then the misclassification rate is smaller and less overlapping, where the amount of misclassifications is drastically reduced from what a single model could do. b. indicates the result when larger errors and greater overlap occurs.

thing inherent about the classifier (a non-outlier that the model did not classify correctly for any number of reasons), then we assume that the latter class of subjects are likely to be randomly distributed as errors among all of the models generated. Following that logic, it would be possible to correctly classify more subjects by allowing a number of models to vote for the outcome, where the majority vote is the classification to be accepted by this collaborative classifier. In this scheme, the outliers would likely still be misclassified, but the non-outliers will be correctly classified by most classifiers, while erroneously classified by the rest. In this way, the small variations and noise present in the individual classifiers can complement each other to provide a more robust and better-performing classifier.\(^8\)

\(^8\)Importantly, by increasing the performance of the classifier, we do not necessarily add anything to the interpretability of the underlying relationship, so the usefulness of this technique is more or less empirical in nature, rather than theoretical.
Figure 4.22: Using 5 decision trees, each casting a single vote for the outcome of a subject.
4.3.7 Human-created Model

We also assembled a model from the collection of literature on aggression, to compare performance to what a well-read human can create. The majority of the literature focuses on the interaction between biology (in the form of genetics, levels of neurochemicals in the blood, or pharmacology) and the environment, we have little to no information on the former in our data set. Thus, the human classifier should not be taken as an indication of the performance of the collective knowledge of the field in diagnosing or predicting the incidence of maladaptive aggressive tendencies. In order to ascertain a patient’s value for most of these biological measures, costly, invasive, or unethical steps would be required. Instead, since the Devereux data set is representative of the information available upon admission to a residential treatment facility, this should be an indication of how well a typical clinician could predict aggression, given the available published knowledge and the information available about their patient.

This model was based in the framework of the neural network, and was created using the same components (nodes, additive edges, gating edges, and interaction edges). However, rather than being maximally connected, the edges and hidden layer nodes were selectively connected. These edges were placed where the literature has implicated some form of relationship (or where we have inferred a potential or implied relationship from the literature). Rather than attempting to assign weights to these relationships (as there is no a priori evidence as to the comparative strength of these relationships), the edges were assigned equal weights (1.0 for positive associations and −1.0 for negative associations). This model then had its weights adjusted by the same training procedure used in the neural network training (see 4.3.4). Thus, Ψ was computed in the same manner as for the neural networks.

4.3.8 Synergy [1]: Optimizing Neural Network Structure

The manner in which neural networks operate can be described by a number of parameters. If these parameters were independent of one another, the optimum values could be selected easily and individually for each parameter. However, this is not the case: many of the
parameters are dependent upon one another, and thus need to be optimized together. In order to accomplish this, we employed an EA.

The parameters that were tuned with the EA are: number of hidden nodes in each layer, number of hidden layers, the amount of activation required to exceed the threshold ($\tau_i$), the amount by which an edge was strengthened or weakened ($\eta$), the maximum amount that an edge can be, at what point the edge’s polarity was switched, how many times an edge was allowed to switch polarity before it was deemed unimportant and removed from calculation, and the number of times the entire training set was used to train the system. In addition to the continuous parameters, there were a number of discrete parameters to be optimized (usually in a binary fashion): whether or not edges were allowed to span multiple layers, whether edges were decayed away or were simply removed after a number of weaken calls, whether nodes were connected with singular (either a positive or a negative link) or dual links (both a positive and a negative edge of differing strengths), whether edge weights were altered once per patient or for the number of times the edge contributed to an incorrect solution, whether links were allowed to propagate negative activation, and whether a strengthen and a weaken flag on the same edge cancelled each other out.

Models were generated using these tuned parameters for further analysis.

4.3.9 Synergy [2]: Collaborative Classifiers: Forest of Neural Networks

Associated with each neural network model (much like every other classifier) is a portion of the population which it incorrectly classifies. Some of these patients are outliers, and there is little chance of correctly classifying them given the data available. The rest of the misclassifications represent a set of subjects where the model failed to recognize the relationships leading to the correct outcome (or assigned them the wrong weight). We expect this to happen because each neural network is somewhat determined by the initial chaotic connections, resulting in different relationships being captured. Thus, it is possible that of the set of patients misclassified, some will be misclassified by almost every model (the outliers) while some will only occasionally be misclassified (the non-outliers). So, by placing
a few classifiers together and having them each vote for the outcome, the non-outliers can be correctly classified by the synergy, assuming that they are correctly classified by more than half of the individual networks.

Given the nature of neural networks and their dependence upon initial edge weights to discovering classifiers, it is unlikely that a single neural network will capture all of the relationships present in any given data set, unless the initial weights are optimized as well. Instead of attempting to optimize starting conditions, we opted to assume that each neural network will only contain a subset of nodes and edges that represent the relationships. Thus, by combining a number of neural networks we expect a higher performance since the relationships covered will differ between networks, as illustrated in Figure 4.23.

### 4.3.10 Synergy [3]: Hybridized Neural Network

Along similar lines as Section 4.3.9, perhaps this error would be best compensated for by allowing the models to interact with one another at the output level, rather than allowing them autonomy at the voting stage. If sureness is conveyed in distance from the threshold of an outcome node (which is not altogether implausible) then it may be beneficial to allow models to collaborate by combining activation rather than only by binary votes. In a scenario where many models in the Forest are unsure of the outcome, they carry the same weight in their vote as a model that is very sure. Under the Forest of Neural Networks (Section 4.3.9) methodology, a single correct and sure model will be eclipsed by many incorrect but unsure models. Here, however, the sureness of the model can be taken into account, and perhaps create a more accurate classifier. This is illustrated in Figure 4.24.

### 4.3.11 Synergy [4]: Collaborative Disparate Classifiers

From a purely performance standpoint, it seems logical that combining diverse types of classifiers would be beneficial, since they all capture slightly different types of the underlying relationships: nearest neighbor classifiers may capture things not directly measured and other similarities, human-created models can allow subtle implications to come to light,
Figure 4.23: A Forest of Neural networks uses multiple neural networks, each casting a single vote for the outcome of a subject.
Figure 4.24: The grayed out portions of the neural network are components shown to be ineffective by the streamlining process. By putting together all these functional components of the neural network, we created a hybrid neural network (below the dotted line).
4.3. TECHNIQUES

despite those implications not being found in the current set of data, and neural networks can model complex relationships that are arguably less than obvious to the human model creator. In a similar fashion to the forests, we can allow disparate classifiers to vote for the classification of a novel subject, as illustrated in Figure 4.25.

4.3.12 Controls

Since this research is entering uncharted waters, several positive and negative controls were used to validate the claims made about the method used. Additionally, this allows any negative result to be attributed to a failing in the data collected, rather than a failing of the modeling process. We wanted to be certain that if a complex pattern exists within the data leading from a patient’s history to a patient’s aggressive phenotype, that we would be capable of detecting it. So in order to benchmark the modeling process’s capability, we created test data sets in which relationships between input and output existed.

If we had a variable in the history vector that adequately predicted, on its own, the aggressive outcome, we would expect the neural network to be able to detect this and exploit it in the classifier. An example of this, even from the current data set, is that some of the Devereux variations of the DSM narrow band rating scales (DSM-D), when normalized and centered, perfectly predict the incidence of aggressive outcomes: Impulsivity (DSM-D:Impulsivity) for reactive aggression and Conduct Problems (DSM-D:Conduct) for proactive aggression (Coppersmith et al., 2008). Thus, the first test data set (Figure 4.26.a) encoded the binary outcome for each phenotype in the input variables. So the expected value of $n_{P_{re}}$ was the value assigned to the first input variable while the expected value of $n_{R_e}$ was assigned to the second. Three other variables whose values were random numbers $-1 \leq v_{ij} \leq -0.5$ or $0.5 \leq v_{ij} \leq 1$.

Secondly, if we had a pair of variables in the history that adequately predicted the aggressive outcome when taken together, the network would be have to be able to detect that. Thus, the second test data set (Figure 4.26.b) encoded each phenotype over a pair of variables instead of one. So only when both of the input variables were at 1 was the
Based on the *forest* methodology, we can combine a number of disparate techniques, having each vote for the outcome. Here, we have a decision based on three techniques, Random Forest, K-Nearest-Neighbor, and a Forest of Neural Networks.
Figure 4.26: Graphical representations of the positive controls: Diamonds are aggressive outcomes, circles are hidden nodes, orange ovals are random inputs, green ovals are portions of an interaction dictated by the aggressive outcomes (the dotted line indicates their exclusion from the input nodes, but their signal was encoded as an interaction over the nodes they connect to). 

a shows a model capable of predicting the outcomes if the outcomes were given as inputs and b shows the same, with distractor randomness added as well. c shows the output encoded as an interaction in the input nodes, while d shows the same with distractor randomness.
4.3. TECHNIQUES

Figure 4.27: Graphical representation of the negative control: Diamonds are aggressive outcomes, circles are hidden nodes, and orange ovals are random inputs. Classification is only performed by this model at chance.

phenotype present. If the phenotype is not present, one of the variables was 1 and the other was a random number between -1 and .5.

To simulate more complicated interactions, the third test data set (Figure 4.26.c) was the same as the second, but encoded each phenotype over three variables instead of two.

We also assume that when putting so many variables into such an analysis, some of them will inevitably only be adding noise. In order to be certain that we can differentiate between noise and signal, the fourth test data set (Figure 4.26.d) is the same as the third, but includes seven variables who vary freely ($-1 \leq v_{ij} \leq 1$). This is to simulate the potentially useless variables present in our data set.

The fifth test data set, rather than encoding a simple interaction, encoded an interaction analogous to an "exclusive or" relationship (XOR). A relationship was only present when the two nodes did not match in polarity (i.e. $n_1 = -1, n_2 = 1$ or $n_1 = 1, n_2 = -1$).

These tests benchmark the model’s capability for detecting complex patterns leading from input data to output classifications. Thus, if we can find no model capable of prediction using these methods, the data we are using must not have sufficient information regarding the patterns sought.

As a negative control (Figure 4.27), completely random data was entered as inputs, and the resulting fit (shown in Figure 3.28) shows how no relationship could be found.
Figure 4.28: The performance ($\Psi$) of the neural network classifier on random data it was trained on ($x$ axis) and novel random data ($y$ axis). As can be seen, the performance on the training data set has no bearing on the performance on the novel training set.
4.3. TECHNIQUES

4.3.13 Expectations

Here we lay out expectations for the potential performance of the models and what conclusions can be drawn from that regarding the nature of the underlying data. One assumption being made right away is that subsequent applications of these techniques (the neural networks in particular) will not produce equivalent models, due to the randomness involved in their initial conditions. With that in mind, we expect to see a range of fits across even the training data, as the models will capture different underlying relationships and different amounts of the underlying relationships. We must also assume that the data being used for this analysis is made of up signal and/or noise. Either something helps explain the relationship we are investigating (signal) or it does not (noise). Further, the assumption is that the signal remains constant across the data set (in that the relationships that make up the signal are universal relationships, thus applying to data regardless of whether they are in the training data or in the novel data). Additionally, it is assumed that the noise does not remain constant across the data set, in that the noise in the novel data set does not match the noise of the training data set.\(^9\)

Let us start with the simplest of the potential relationships: if the data used as input has no relevance to the outcome sought, we would expect to see no relationship as to how well the model performed on the data it was trained on and how it did on the novel data, represented in Figure 3.29.a. Additionally, the fit of the negative control, which speaks directly to this scenario, can be seen in Figure 3.28.

Conversely, if the information in the training data is sufficient to determine the outcome sought, we would expect to see the performance increase on both scales equally, as the model learns more of the underlying relationships, the performance increases on both sets of data. This is represented in Figure 3.29.b. The positive controls fit this description, but in reality the plots are uninteresting, as only in rare cases do the models misclassify any of the data.

However, there is a lot of ground between those two extremes. If we have data to capture

\(^9\)If the noise \textit{does} match across the novel and the training set, than it is necessarily not noise, but signal.
Figure 4.29: Theoretical models have their performance ($\Psi$) on their training data plotted on the $x$-axis with performance ($\Psi$) on novel data on the $y$-axis. 

- **a.** represents the scatter plot we would expect if there was no relationship to be found in the data.
- **b.** shows the scatter plot we would expect if the outcome can be perfectly explained by the data.
- **c.** shows the expected scatter plot for a data set that does not completely capture the data, but also does not introduce misleading information into the equation.
- **d.** shows the scatter plot expected for a data set in which there is both signal and noise.
part of the underlying relationship, but not enough to fully explain the relationship, we would expect to see a parabolic curve (as illustrated in Figure 3.29.c). While the fit is low on the training data, it is also expected to be low on the novel data, as most of the relationship is not being captured. As the fit on the training data increases, we expect the model to learn both the noise and signal of the training data. So when the model is applied to the novel data, we expect to see the amount of signal captured lead to increased performance, while the noise that was also captured not increase the performance (as illustrated in Figure 3.29.c) or in some cases, decrease the performance (as illustrated in Figure 3.29.d). This decrease in performance is commonly called overfitting or overtraining, since it indicates that the model has learned too much of the patterns in the training data. In this scenario the noise in the training data has been incorporated into the classifier, and thus negatively impacts the performance when applied to novel data. Figure 3.29.c is expected if the noise is homogeneous, and thus has its weight decreased and is essentially eliminated from the model. Figure 3.29.d is expected if the noise is heterogeneous, having some structure in the training set that does not match that of the testing data.

Realistically, we expect something in between Figures 3.29.c and 3.29.d, as most of the noise is homogenous (and thus discounted) while some is heterogeneous (by chance or otherwise) and thus decreases the fit on the novel data as the training data reaches the highest point. Additionally, variance in the data set will create a cloud of fits, rather than the lines shown here. In order to determine the underlying relationship, a parabolic regression was performed on the train and novel performances.

4.3.14 Model Selection

In order to select the best-performing model, we need to examine how well it fits both the data it was trained on ($\Psi(Train)$) and the data it was not trained on ($\Psi(Novel)$). As illustrated in Figure 3.30, the selection of models is best understood from examining a scatterplot of each model plotted with it’s $\Psi(Train)$ on the $x$-axis and $\Psi(Novel)$ plotted on the $y$-axis. We can regress a parabolic line to best fit the data, as discussed in Section
4.3. TECHNIQUES

Figure 4.30: In order to select models that generalize and capture the maximal amount of variance, we want to select models from the shaded region above, where the parabolic regression line peaks. This indicates that the maximal amount of the relationship has been captured, without overtraining.

4.3.13. The ideal models have a high fit for both the training and the novel data, so we would assume that the best models are in the shaded region indicated in Figure 3.30, at the top of the parabolic curve. As can be seen in Figure 3.31, merely selecting on the $\Psi(Train)$ will provide a wide range of fits, and thus is not advised. This is the primary reason for holding back some of the data from the training set, as it allows a metric to be established as to both the fit and the likely generalization of any given model. Thus, the model (or models) can be selected by their ability to fit both the training and the novel data. We selected models for further analysis by finding the model for which the $\Psi(Novel)$ was highest, but not higher than $\Psi(Train)$ (as shown in Figure 3.32). The best fit on the novel data indicates that the model has captured the largest amount of the underlying relationship possible, and is able to apply this knowledge to novel cases successfully. By eliminating the models for which $\Psi(Novel) > \Psi(Train)$ for the training data, we eliminate the models that, by chance, capture the noise in the novel data (which necessarily does not match the noise in the training data). Thus, the models examined further are ones with a good $\Psi$ for the entire data set (training and novel), indicating that as much of the underlying relationship as possible has been captured.
Figure 4.31: In order to select models that generalize and capture the maximal amount of variance, we want to select models from the shaded region above, where the parabolic regression line peaks. This indicates that the maximal amount of the relationship has been captured, without overtraining. To do this by only examining training data proves unfeasible, as is indicated above.
4.3. TECHNIQUES

Figure 4.32: The alternative used for model selection (indicated by the black box). The red line indicates where $\Psi(\text{Train}) = \Psi(\text{Novel})$, so any model above that line fits the novel data better than the training data ($\Psi(\text{Novel}) > \Psi(\text{Train})$). All of the models shown in this graph are drawn from a single parameter set and a single split of the data. The best model was selected for each of six parameter sets and each of three data splits, yielding 18 models for use in the *Distilling Analysis* (Section 4.3.15).
4.3. TECHNIQUES

4.3.15 Decomposing Classifiers

Understanding the relationships underlying the incidence of maladaptive aggression is the ultimate goal. Ideally, we would like to find the interactions that lead to successful classification, but only if they are succinct enough to reason about and sufficient enough to explain the relationship. However, even without identifying the individual interactions that lead to successful classification, we can identify which input variables are not important to successful classification. We can examine classifiers that show good performance on both training and novel data (as an indication of how well that model will generalize), and examine the unique contribution of each variable from the patient’s history, without having to understand the entirety of the interactions it is involved in. This gives a rough estimate of the contribution of the history variables, but taking into consideration the interactions that they are a part of, rather than merely their individual contribution (as is the case with Multiple Linear Regression, Logit, and, to an extent, K-Nearest-Neighbor).

We calculated the percent of the fit that can be attributed to a given history variable ($\Xi_i$ for the contribution of history variable $n_i$) by subtracting the model’s performance without the variable ($\Psi^{-n_i}(X)$) from the performance with the variable included ($\Psi(X)$). Additionally, we will take the average over a number of well-performing models ($m$), to account for the differing initial conditions of the model while also lessening the cruciality of selecting a single (or small subset) of models to perform more time-consuming analysis on. More formally:

$$\Xi_i = \sum_{j=0}^{m} \frac{\Psi(X) - \Psi^{-n_i}(X)}{m}$$

We performed this procedure for 18 models ($m = 18$). These models were selected as the best model (as shown in Figure 3.32) from each of six parameter sets and each of three data splits. This will allow us to order the importance of each variable in the patient’s history in terms of importance to the performance of a classifier (on average).
Chapter 5

Results

The techniques, models, and synergies will be compared on their $\Psi$ based on a resampling procedure. The patients from the Devereux data set were resampled, with replacement, 10000 times, calculating $\Psi$ each time. More formally, new data sets were created via bootstrap resampling ($\text{Bootstrap}_1, \text{Bootstrap}_2, \text{Bootstrap}_3, ... \text{Bootstrap}_b; b = 10000$). The $\Psi$ presented on the graphs is the average $\Psi$ for the bootstrapped data sets.

$$\Psi(\text{Bootstrap}) = \frac{1}{b} \sum_{i=1}^{b} \Psi(\text{Bootstrap}_i)$$

In some cases, Pearson’s $r$ (calculated between $\Psi(\text{Train})$ and $\Psi(\text{Novel})$) will be used to compute a metric for generalization.

For the sake of clarity, the techniques were analyzed and discussed in terms of performance (or loss) on differentiating Non-Reactive from Reactive patients ($D = \{\text{Non}, \text{Reactive}\}$), without loss of generality.

For significance tests between methods McNemar’s Test will be used, $\alpha < 0.05$ (Fleiss, 1981). In a most general sense, McNemar compares the performance of two models by only taking into account the patients where one model provides a correct classification and the other does not (as can be seen in Figure 5). McNemar’s Test was computed as follows: let $n_A$ be the number of errors made by method A and not method B and $n_B$ be the number
Figure 5.1: Visual representation of McNemar’s test. Each box represents the patient population, white indicates patients both models correctly classified, orange shading indicates the patients model A misclassified, blue shading indicates the patients that model B misclassified, and green shading indicates patients that both models misclassified. As the bar on the left shows, when both models misclassify different patients, $n_A$ is just the errors made by model A, and $n_B$ the errors made by B. Conversely, the bar on the right shows how this changes when the same number of misclassifications are made by each model, but some of the patients are misclassified by both models.
5.1 Chance

First, let us establish chance: since the outcome is a binary decision, chance of getting any patient correct is 0.5. Moreover, since \( \Psi \) averages the two diagnoses (Non and Re), it is not advantageous to guess one outcome over the other, as guessing one outcome for every patient will end up with \( \Psi = 0.5 \), regardless of the outcome chosen.

\[
\Psi = \frac{\sum_{n \in D} C_n}{|D|} \quad \text{where} \quad D = \{\text{Non, Re}\}
\]

\[
\therefore \quad \Psi = \frac{0 + 1.0}{2}
\]

5.2 Standard Statistics

The conditions used to represent the standard statistics are multiple linear regression (MLR; \( \Psi(\text{Novel}) = 0.66 \)) and the human-created model based on the literature (Human(\text{Novel}; \Psi = 0.69). MLR and Human are not significantly different from one another, but are both significantly different from chance.

For each of the following algorithms, the highest-performing instance of the algorithm was selected for discussion and testing, rather than trying to justify selecting a single model \textit{a priori}. In all cases, the model was selected that best performs on the novel data set, while \( \Psi(\text{Novel}) \leq \Psi(\text{Train}) \). Section 4.3.14 explains this process in more depth.
Figure 5.2: Performance ($\Psi$) of the K-Nearest-Neighbor techniques along with the discriminator based on multiple linear regression (MLR) and the human-created discriminator (Human). The two groupings indicate the distance metric that K-Nearest-Neighbor was based on, Euclidean distance or Mahalanobis distance. These groups are subdivided by the weighting given to each dimension: Unweighted means all are equally weighted, $|\beta|$ indicates the absolute value of the MLR $\beta$-weights, $\beta^2$ is the squared value of the $\beta$ weights. * indicates significance from MLR ($\alpha < 0.05$). Error bars indicate two standard deviations, as derived from the resampling procedure. The y axis starts at chance (0.50).

K-Nearest-Neighbor (KNN), in its simplest forms (Unweighted Euclidean Distance and Unweighted Mahalanobis Distance), performed no differently than MLR. When the dimensions are weighted by the $\beta$ weights from MLR ($\beta^2$), the differences do reach significance for Euclidean distance ($p < 0.05$), but not for $|\beta|$ nor for either weight set used with Mahalanobis Distance ($\beta^2$ or $|\beta|$). For a graph comparing the various KNN implementations to MLR and Human, see Figure 5.2. For a graph of the various KNN implementations by the size of $K$ (the number of nearest-neighbors used), see Figure 4.3.
Figure 5.3: Performance ($\Psi$) of the various K-Nearest-Neighbor techniques by the size of $K$ (the number of neighbors used to determine the diagnosis).
5.3 Neural Networks

Neural Network classifiers did perform significantly better than MLR and Human models (as shown in Figure 4.4). Neural networks created with sigmoidal transforms on the nodes did not perform significantly differently from those with step transforms at the nodes (but both performed significantly better than MLR, so only models with step transforms were further analyzed, due to their ease of interpretation.

Additionally, the number of layers of hidden nodes in the network showed no significant changes in Ψ (as can be seen in Figure 4.4). All edge sets used to create the neural networks (A represents additive edges, I represents interactive edges and G represents gating edges) performed significantly better than MLR ($p < 0.05$ [One-A], $p < 0.01$ [Multiple-A, Multiple AI, Multiple AIG] or $p < 0.001$ [One-AI]) except for One-AIG. One-AI is also significantly higher than One-AIG ($p < 0.05$), but no other significance can be found between the neural network models. Thus, we selected models with a single hidden layer for further analysis, as they are composed of fewer edges and nodes.

When compared to Human (see Figure 4.4) the networks using additive and interaction edges performed significantly better ($p < 0.05$ [Multiple-AI], $p < 0.01$ [One-AI]), while those composed of additive only or additive, interaction, and gating edges did not.

Examining the performance of the streamlined models, there is no significant difference between streamlined and unstreamlined versions of the same model, as indicated by Figure 4.5. All models created from neural network classifiers (a single neural network, a forest of neural networks, and a hybrid model created from a number of neural networks) perform significantly better than MLR ($p < 0.01$ [Forest-Unstreamlined, Forest-Streamlined, Hybrid-Unstreamlined] or $p < 0.001$ [Single-Unstreamlined, Single-Streamlined, Hybrid-Streamlined]) and Human($p < 0.05$ [Forest-Unstreamlined, Forest-Streamlined, Hybrid-Unstreamlined] or $p < 0.01$ [Single-Unstreamlined, Single-Streamlined], or $p < 0.001$ [Hybrid-Streamlined]).

Finally, the best of each category of algorithm is compared in Figure 4.6. A disparate
Figure 5.4: Performance ($\Psi$) of single neural network models, grouped by numbers of hidden layers (One has 1 hidden layer and Multiple has either 2 or 3). These groups are subdivided by the components used in each model (A indicates additive edges only, AI indicates additive and interaction edges, and AIG indicates additive, interaction, and gating edges). Asterisks indicate performance significantly higher than MLR (**: $p < 0.01$, ***: $p < 0.001$). Daggers indicate performance significantly higher than Human (**: $p < 0.05$, †: $p < 0.01$). Double daggers indicate performance significantly higher than One-AIG (**: $p < 0.05$). Importantly, there is no significant difference among the neural networks (aside from One-AIG and One-AI). Error bars indicate two standard deviations, derived from the resampling procedure. The $y$ axis starts at chance (0.50).
Figure 5.5: Performance ($\Psi(\text{Bootstrap})$) of the best models before and after streamlining, as well as the MLR and Human controls. Asterisks indicate performance significantly higher than MLR (**: $p < 0.01$, ***: $p < 0.001$). Daggers indicate performance significantly higher than Human (†: $p < 0.01$, ††: $p < 0.001$). There is no significant difference between Single, Forest, or Hybrid of these models, in either their unstreamlined (blue) or their streamlined (white) versions. Error bars indicate two standard deviations, derived from the resampling procedure. The $y$ axis starts at chance (0.50).
classifier was also created by minimizing the number of patients that would be misclassified by combining three different techniques, thus comprised of a forest of neural networks, a multiple layer neural network, and the \textit{KNN} based on Euclidean distance and $\beta^2$ weights. Neural networks (in this case represented by forest of neural networks created from the top performer for each of the data blocks) performs significantly better than all of the other algorithms, except the disparate classifier (which uses the same forest of neural networks as one of its classifiers). The disparate classifier also performs significantly better than the rest of the algorithms, except the neural network classifier. K-Nearest-Neighbor performs significantly better than \textit{MLR} as previously discussed ($p < 0.05$).
### Figure 5.7: Matrix of the results of McNemar’s Test, comparing the best of each algorithm to each other.
5.4 Generalization

Correlations between the performance of neural networks on training ($\Psi(\text{Train})$) and novel ($\Psi(\text{Novel})$) data indicate that the models generalize well better than $MLR$. $r^2 = 0.19$ for Neural Networks and $r^2 = 0.04$ for classifiers based on $MLR$.

5.5 Streamlining

The streamlining procedure, though it eliminates around 90% of the edges for most of the models, does not produce an easily-interpretable model. The edges are too numerous and the number of edges loading onto each node in the hidden layer (which represent the interactions) are too high to draw clear conclusions from. Additionally, ordering effects are present in the edges selected for removal. Frequent covariation in the input variables may be contributing to both these difficulties, so application of this technique to different data may produce more interpretable models.

5.6 Decomposing Classifiers: Distilling Analysis

For the distilling procedure, we used the top single neural network performers for each parameter set (single or multiple hidden node layers; additive edges only, additive and interaction edges, and additive, interaction, and gating edges). One network was selected from each training block, providing three networks from each parameter set, and eighteen in total ($m = 18$). The results can be found in Figure 5.6.

The most important factor, according to this technique is the patient’s rating on the Clinical Global Impairment scale, indicative of how debilitating a mental illness is (5.3% reduction in $\Psi$). This was followed by physical abuse and the number out-of-home placements the patient had in his or her lifetime. Also ranked highly was the presence of a nervous system disorder.

The least important factor, according to this technique was the age of the first out-of-
5.6. DECOMPOSING CLASSIFIERS: DISTILLING ANALYSIS

Devereux - Known to Generalize
Neural Network A, A1, A1G
Unstreamlined
n=18 (3 Best from each class)
home placement. This is followed by the presence of a few pharmacotherapies the patient was prescribed prior to admission to the facility (clonidine and lithium). These are followed by performance and verbal IQ (with full scale IQ coming only a bit further up in the ordering), and then by gender.
Chapter 6

Discussion

We investigated the use of standard statistics, standard statistical pattern recognition tools, and novel approaches to elucidate the underlying relationships between a patient’s history and their reactively aggressive phenotype. In doing so, we found relationships both expected and unexpected, given previous literature.

Given the low number of proactive aggressors in our data set, we excluded examination of proactive aggression, and focused wholly on differentiating reactive aggressors from non-aggressors. These methods were designed such that it is a trivial matter to incorporate other types of classifications. Thus, should an adequate data set be compiled on proactive aggressors, it would be near-trivial to examine it with this methodology as well.

Though a large portion of this paper is devoted to discussion of the methodology and techniques used in this endeavor, the significance of those evaluations of performance is not of primary concern, other than the fact that these techniques capture a significant amount of the underlying relationship, where standard methods capture only trivial amounts.

First, the performance of multiple linear regression (MLR) is the typical choice for the field of psychology and psychiatry when attempting this sort of a problem. While the technique is not designed for mapping relationships between continuous and dichotomous input variables to dichotomous outcome variables, the technique was adapted to perform the classification task (Aldrich and Nelson, 1984). MLR performed poorly on the classification
task (correctly classifying 66% of the patients, when chance is 50%). Further, it was shown to be unable to generalize, explaining only 4% of the variance between performance on the data it was trained on ($\Psi(Train)$) and novel data ($\Psi(Novel)$). In fact, if the regression is performed on the entire data set $\Psi(Train + Novel) = 0.58$, indicating that the result used as a control comparison was actually a spuriously high performance for this technique. Thus, this technique surely should not be applied to such data sets in the future, and the previous findings based on application of MLR should be reexamined with a more appropriate technique.

The next technique investigated was $K$-Nearest-Neighbor ($KNN$), a technique routinely applied to such problems in statistical pattern recognition. In its most simplistic forms, $KNN$ performed on par with MLR. When the dimensions were weighted according to the $\beta$ weights derived from an MLR equation, $KNN$ classified the data significantly better than MLR. This technique, like MLR, only examines linear combinations of the input variables in the patient’s history. Given the literature’s sweeping implications of interaction effects in the development and incidence of maladaptive aggression, we sought a technique that would combine these variables in a nonlinear fashion.

The first technique we employed to examine nonlinear combinations of variables was a model created by examining the literature. We creating a classifier (referred to as Human) based on what has been found in the literature to date. Human performed better than MLR, but not significantly so\footnote{However, Human does perform significantly better than the performance of MLR on the entire data set ($\Psi(Train + Novel)$).}. This was used as a comparison for how a classifier designed by a well-read human would compare to those models based on standard and novel techniques. In the creation of Human, it came to light that a good portion of the literature examines the interactions between biology and environment (since the collection of such data is often cost-prohibitive, invasive, and/or unethical). Since the Devereux data set did not record most of the biological measures examined in the literature, the Human classifier only represents the portion of the literature for which we had all of the interaction terms (with linear relationships between the input and outcome variables to substitute for the interactions
between variables collected and those not).

We then applied a series of novel techniques, based on neural network methodology, and specifically tailored to this problem (though these techniques were carefully designed to be applicable to other similar classification problems, specifically in the field of psychology and psychiatry). These techniques, drawn from machine learning and statistical pattern recognition, would be most accurately classified as a modification of existing neural network techniques. They were shown to work significantly better than the standard techniques \textit{(MLR, Human, and K-Nearest-Neighbor)}.

In order to decipher what relationships and variables were important to the development of maladaptive reactive aggression, we sought to decompose the best-performing of these neural network models. In streamlining the classifier we hoped to find a small number of relationships which are responsible for the improved performance of the classifier over standard techniques. In order to streamline the classifier by separating the 'contributing' components from the 'non-contributing' components, we examined nodes and edges individually to assess their importance to the performance of the classifier. The application of this streamlining procedure lead to ordering effects, with the variables examined first more likely to be removed as 'non-contributing' variables. Moreover, despite the streamlining procedure removing a significant portion (in most cases, more than 90\% of the components were removed as 'non-contributing'), the number of components was still too large to reason effectively about. These two issues likely come about because of frequent co-occurrence of the history variables. In many cases, this will cloud the interactions, since differentiating the individual effects of the variables is not possible in the current data set. Thus, this procedure has merit in future applications, but is not feasible for application to the present research.

Given the above performance, it appears this methodology is fatally flawed in its complexity. However, the ordering effect can be used as a boon in the selection of variables for future comprehensive studies in aggression. Instead of allowing the ordering to be arbitrary, as is currently the case, if the variables are ordered by the difficulty, reliability, or cost to
collect, with the most expensive (least reliable, etc.) ones being first and the least expensive ones being last (the least reliable being first, the most reliable being last, and so on), the ordering effects will work in our favor. The ordering effect will force the removal of the most expensive variables first if they can be eliminated without negatively impacting the performance of the classifier.

Regardless of the complexity of the interactions leading to maladaptive reactive aggression, we can certainly examine the opposite side: what variables are not important to the incidence of maladaptive reactive aggression. The distillery process examined the unique contribution of each history variable to the performance of the best-fitting neural network classifiers. This provides an analog to the $\beta$-weights derived from a MLR equation. The $\beta$-weights provide a rank ordering of the variables in terms of contribution to the outcome variables. Specifically, the $\beta$-weights provide a polarity as well as a magnitude, indicating whether the relationship between that input variable and the outcome variable is a positive or negative relationship. The distillery rank ordering does not indicate that (as it is likely that many variables have both positive and negative relationships to the outcome variable). Thus, it is more appropriate to compare the distillery rank ordering to the magnitude of the $\beta$-weights (i.e. the distance from zero with no regard for the direction, or $|\beta|$). The difference between the distillery rank-ordering and the $\beta$-weight rank-ordering is that the former is based on the magnitude of unique contribution of the variables in linear and nonlinear combination, when the latter is ranked only by the magnitude of the linear contribution. In practice, when examining these two rank-orderings, they are highly dissimilar, again calling into question any results derived from the application of MLR in this way.

Thus, let us explore the results of distillery process in the context of the existing literature on maladaptive reactive aggression. In the Devereux data set, the severity of impairment caused by mental illness ($CGISeverity$) was the single variable with the highest contribution. This makes intuitive sense, especially when taken in context of the Frustration-Aggression model, the root for reactive aggression (Bandura, 1977). The severity of mental illness is likely to contribute to the frustration a patient feels, increasing the potential for
explosive outbursts.

This is followed in importance by a number of measures of physical abuse in the patient’s history. Exposure to violence has long been associated with the development of reactive aggression (Neller et al., 2006; Frazzetto et al., 2007) and both reactive and proactive aggression (May, 1986; Connor, 2002). Thus it is unsurprising to find its magnitude so high. Conversely, sexual abuse ranks in the bottom half of the variables, indicating that not all abuse has equal effect. This indicates that it may not be the stress induced by abuse, but perhaps the violence exposure that has the debilitating effects. Interestingly, it does not appear that the relationship of the abuser to the patient seems to be as important as the existence of abuse: all of the roles abuse-perpetrators have been classified into rank together, save parental perpetrating abuse, which ranks well below the rest.

Following this in terms of importance is the number of out-of-home placements in the patients history. Interestingly, the age of the first out-of-home placement is ranked as having the least magnitude of importance. A caveat have to be made here: first, patients only make it to the Devereux institute after many other attempts have failed. In this way, out-of-home placements may be measuring the number of failed treatments attempted for the patient in addition to anything actually tied to the number of out-of-home placements. Since both of these variables were separated into scales based on protective and risk values for these variables, (−1 for those patients who were not place out of their home at all, and values ranging from (0 − 1] for either age or number of out-of-home placements), it is unlikely that we get differing results for the measurement of the same variable. In fact, what’s more likely is that age and number of out-of-home placements do not matter. The number of out-of-home placements may be acting as a surrogate for the patient’s resistance to treatment. Thus, it is unsurprising that a patient’s resistance to treatment is indicative of their likelihood to express maladaptive reactive aggression. Surprisingly, this would also mean that out-of-home placements and out-of-home placement age themselves are not implicated in the development of the maladaptive reactive aggressive phenotype, contrary to conventional thought (Lamb, 1996). Note, however, that this has no bearing on the relation-
ship between out-of-home placements (number and age of first) with proactive aggression. The studies implicating the importance of out-of-home placements with aggression may be linking it with proactive aggression.

Also ranking highly is the presence of a disorder of the nervous system (both central nervous system and non-central nervous system). It has been suggested such a link would be present for early-onset aggression (Moffitt, 1993).

Aside from out-of-home placement age, the next few least-important variables in the Devereux data set are the use of clonidine and lithium. This is unsurprising since they are not typically used to treat aggressive tendencies, so their contribution to the incidence of maladaptive aggression is understandably low.

Following those are measures of IQ. IQ, specifically verbal IQ, has long been implicated in the incidence of aggression (Huesmann et al., 1987). However, this data set indicates that these variables may not be as important as once thought. There are two possibilities created by this analysis: IQ in many other studies was acting as a surrogate for another variable or IQ is associated with the incidence of proactive aggression, rather than reactive aggression. There is existing support for the relation between IQ and proactive aggression (Silverthorn and Frick, 1999). IQ has also been found to correlate with a number of other aspects of patient’s history, such as socioeconomic status (SES) (Turkheimer, 2003). Given this, other studies may have implicated IQ instead of some combination of its covariates in the incidence of maladaptive aggression. A few possibilities are apparent for the link between IQ to proactive aggressors, ranging from violence exposure (associated with SES), shortsighted planning (which would impact both learning and the use of violence to accomplish a goal), or the inability to talk out differences, more often leading to violence escalation. The possibilities are numerous, and without further study interpretation of this result is unwise.

Following IQ is gender. Though females were undersampled in our population: 77 of the patients, 36 Reactive Aggressors and 39 Non. This represents 19-20% of the population for each diagnosis. it appears than gender itself has a very small magnitude of effect, especially when compared with what the literature has classically implicated as gender’s role. Gender
could have a role very similar to IQ, standing as a surrogate for the collective effects of
different societal norms or some other macro-level factor. Like IQ, this warrants much
further investigation before conclusions are drawn from it.

Beyond conclusions about the specific history variables included in the Devereux data
set, information can also be gleaned about the data set itself. First, the maximum amount
of variability captured by the models should be discussed. For MLR-based classifiers, the
variance explained between the $\Psi(\text{Train})$ and $\Psi(\text{Novel})$ is only 4% ($r^2 = 0.04$). The same
correlation for the neural network models is 19% ($r^2 = 0.19$). Thus, while the neural net-
works are able to capture significantly more of the underlying relationship between the
history variables and the incidence of maladaptive aggression (as inferred from the greater
ability to generalize), they are still incapable of capturing the entire relationship. What’s
more, the positive and negative controls (see Section 4.3.12) show that if some combination
of the history variables was sufficient to explain the incidence of maladaptive aggression,
the neural networks would be able to capture it. Thus, we are left to conclude that the
Devereux data set alone does not have adequate information to predict the incidence of
maladaptive aggression. Additionally, this data set lacks one crucial factor for creating use-
ful diagnostic tools: information regarding desistance. Equally important to understanding
the underlying causes and correlates of maladaptive aggression is an understanding of the
efficacy of treatments. This is simply not possible with a snapshot database (since all of the
collection is necessarily at the same time point; admission to the facility in this case), but
these techniques can be adapted for use with a longitudinal database. This will be further
discussed in Section 7.

This data set also lacks significant biological measures (e.g. hormone levels, genetics,
and stress-response information). A significant portion of the preclinical literature has
examined the relationships between these factors, see (Nelson, 2006) for a review. Moreover,
not knowing how the environment may act to alter the internal biology of a patient, the
biological evidence from preclinical literature cannot be directly or indirectly applied to
humans. Thus, the construction of a new data set is warranted. The exact nature of
constructing this data set is beyond the scope of this project.
Chapter 7

Conclusions and Future Directions

The application of neural networks and the distilling analysis has provided some interesting findings: both replications of previous work and novel discoveries. The lack of predictive power displayed by the classifiers based on standard practices in the field (multiple linear regression $[MLR]$ and the human-synethsized-classifier based on the current literature $[Human]$) is a clear indicator that more advanced analysis must be applied to this and similar problems. This first attempt has not only provided clues regarding the incidence of maladaptive aggression, but also a proof of concept for the application of machine learning and statistical pattern recognition techniques to the problems in the fields of psychiatry and psychology.

This methodology has demonstrated the ability to make better-fitting models than standard statistics and techniques can. In fact, it has created generalizable models where the standard statistics did not. Additionally, this methodology allows us to ascertain the relative importance of each of these variables alone, in linear combinations, and in nonlinear combinations.

As previously stated, this pursuit was seeking two classifiers: one to aid in diagnostics and one to aid in exploratory research. The combination of neural networks and the distilling analysis has fulfilled the exploratory need, providing useful information for the direction of aggression research. The diagnostic need is more complicated and warrants application of
more advanced mathematical analysis.

If we were to pursue the development of a diagnostic tool, a number of changes could be made to improve the performance of this algorithm. Notably, incorporating Bayesian priors corresponding to the incidence of aggression subtypes in the population we are investigating would allow us to dispense with the naïve notion that all subtypes are equally likely to present themselves to the model. Moreover, depending upon how this model would be used, the Bayesian priors would reflect the incidence in that subpopulation (i.e. the priors for inpatient populations ≠ the priors for outpatient referrals ≠ the priors for the general population, since the percent of patients expressing a given subtype of aggression are different depending upon the subpopulation investigated) (Bishop, 1995). Additionally, as alluded to previously, applying less-transparent mathematical and computational tricks may boost the performance of the models created (additional parameter tuning, sigmoidal rather than step activation, etc.), making them closer to an acceptable diagnostic level. Furthermore, since maladaptive aggression is typically viewed on a spectrum, rather than a binary classification, replacing the binary classification with a spectrum would be useful for those purposes. Also, incorporation of more patients (including proactive aggressors) would be necessary.

Turning back to the exploratory research, there is one additional component to the *distilling* analysis that can further elucidate the contributions of the variables. By separating the importance of the factors by their contribution when the factor is present and when it is absent may allow additional results to emerge. If a factor has a contribution when it is present, but only to a few patients, then we would expect that factor to rank low on the importance scale (since excluding it only negatively impacts the few patients who had a *risk* value in that factor, but the rest of the population was left unchanged by it). If, instead, we separated the contribution when the factor is present in a *risk* value from the contribution when the factor is present in a *protective* value, we can get a clearer picture as to how the factor contributes. For example, if sexual abuse contributes to the incidence of aggression, but only for a few patients, the change in performance by the exclusion of sexual
abuse would only affect those few patients, probably excluding it from being important to aggression. However, when we split the contribution by present and absent factors, we can see that all of those with sexual abuse were misclassified after the exclusion of sexual abuse, while the rest of the population was unchanged.

Beyond increasing the knowledge that can be gained from this technique, this technique can also be applied to other similar psychiatric problems (and to longitudinal databases). Part of the beauty of this technique is its versatility. Any set of inputs can be coupled with any output, and the technique will function the same. Thus, by taking out aggression and putting in depression, anxiety, post traumatic stress disorder, or any other psychiatric disease is relatively simple. Moreover, this could be extended beyond psychiatric diseases to any form of disease or behavior, transcending the realm of psychology into medicine or economics.

Further, when longitudinal information is introduced, models can be designed that deal with the efficacy of treatment, rather than the incidence of disease. With history variables as inputs and the relative efficacy of treatments as outcomes, this method can be applied to rank-order treatments in terms of likely efficacy, rather than the relatively-arbitrary ranking system currently in use. Taking this a step further, we can transcend the categorical DSM criteria by combining information from ranges of patients for ranges of mental illnesses. Regardless of the diagnosis and symptomology, the effective treatments can be assessed based on what has previously been shown to work for patients of similar histories.

In sum, we have demonstrated a proof of concept for applying novel techniques to problems in psychology, psychiatry, and beyond. We have taken the first step towards a new series of techniques and analysis based on computational methods. The full ramifications for computational psychiatry won’t be known for years to come, but the potential is self-evident.
Chapter 8

Glossary

8.0.1 General Mathematics

\(\neg\) = Not, as in: \(\neg True = False\).

\(\land\) = And, as in: \(True \land False = False\).

\(\sum_{m=a}^{b} x_m\) = Sum of all the \(x\)'s from \(a\) to \(b\).

\(\in\) = In, as in: \(1 \in \{1,2,3\}\)

8.0.2 Neural Networks

\(n_i\) = a Node.

\(A(n_i) = a_i\) = Activation of \(n_i\).

\(\omega_{ij}\) = Weight of edge from \(n_i \rightarrow n_j\).

\(n_i \prec n_j = n_i\) is upstream of \(n_j\). \(n_i\) can influence \(n_j\) but not vice versa.

\(\omega_{hij}\) = Weight of a multiplicative edge (Gating/\(\Sigma\Pi\) or Interaction) with \((n_h \land n_i) \prec n_j\).

\(\tau_i\) = Threshold at \(n_i\).

\(n_{Pro}\) and \(n_{Re}\) = Output nodes.

\(n_{VIQ}\) and \(n_{Abuse}\) = Input nodes.

8.0.3 Evolutionary Algorithms

\(G^{(t)} = \) Generation at time \(t\)
\(\mathcal{P}(t) = \text{Parents/Elites for } \mathcal{G}(t)\)

\(\mathcal{Q}(t) = \text{Offspring for } \mathcal{G}(t)\)

Let \(x_{ij}, i = 1, \ldots, \lambda \text{ and } j = 1, \ldots, n\) be the \(j\)-th factor of the \(i\)-th offspring in \(\mathcal{Q}(t)\).
Bibliography


