A NEURAL NETWORK ANALYSIS OF MILITARIZED DISPUTES, 1885–1992

Temporal Stability and Causal Complexity

Monica Lagazio and Bruce Russett

Great progress has been made in predicting and explaining interstate conflict. Improved data, theory, and methods all deserve credit. Yet much remains to be done. First, whereas many variables (e.g., geographical proximity, relative power, alliances, political regime type, economic interdependence) have important effects, even the most successful multivariate analyses leave much of the variance in conflict behavior unaccounted for, due to inadequate data, specification, or theory, or simply random variation. Consequently, questions arise about the predictive power of such analyses. Can we identify, with enough accuracy for policy purposes, those relationships very likely or very unlikely to experience militarized disputes? Can we reduce the number of false negatives and false positives? Second, interstate conflicts are complex phenomena often displaying nonlinear and nonmonotonic patterns of interaction. Those complexities are hard to model. Finally, there are questions about whether causal or predictive relationships are stable across time and space. One such question is whether democracy reduced the risk of interstate conflict throughout the twentieth century (Thompson and Tucker 1997; Maoz 1998; Russett and Oneal 2001) or its effect was limited to the Cold War era (Gowa 1999) due to particular conditions like ideological rivalry, bipolarity, or nuclear weapons. Some early COW analyses (e.g., Singer and Small 1968a) also emphasized nineteenth- and twentieth-century systemic differences.

Recent innovations have employed neural network analysis, a mathematical technique especially suitable to the interactive, nonlinear, and contingent relations across the variables that may trigger mil-

itarized interstate disputes (Schrodt 1991; Beck, King, and Zeng 2000). As a descriptively predictive rather than overtly theoretical tool, neural network analysis does not require rigid a priori assumptions on the mathematical nature of such complex relationships as do commonly used multivariate statistical techniques (Garson 1991; Zeng 1999). Moreover, it provides a clear answer to questions about predictive accuracy, with measures of the percentage of correct predictions both for dyads that actually experienced disputes and those that did not. And it readily lends itself to analyses whereby one can inductively establish a pattern of regularities in a data set for one time period (e.g., the Cold War era) and then measure how accurately that empirically derived pattern of regularities postdicts to disputes, and their correlates or causes, in a data set for another period (e.g., the decades preceding the Cold War).

To assess the possibility of uncovering durable conflict dynamics with a complex model we develop and test a neural network model of Cold War interstate conflicts, then test its performance on data covering more than a century (1885–1992). Since predictive accuracy is a major criterion by which models are assessed, our exercise can reveal the extent to which the Cold War causal structure is representative of earlier historical contexts. Thus the first of our three goals for adding to existing neural network analyses of international conflict is to discover whether the process at work in determining interstate conflicts during the Cold War was a consequence of specific systemic conditions, such as East–West confrontation or U.S. hegemony, or whether some complex regularities at the dyadic level were characterizing conflict outcomes. We find, rather, that much the same regularities exist in the pre–Cold War era. Using many data sets originating in the COW Project, we can correctly predict 82 percent of militarized disputes and 72 percent of nondisputes in the Cold War era and nearly 65 percent of disputes and nondisputes in the pre–Cold War decades, with economic interdependence, democracy, and international organizations providing strong input to the predictions in both periods.

Another extension of previous work is to further develop the methods of neural network analysis for international conflict. Schrodt's (1991) and Garson's (1991) first efforts to use this technique in political science exposed two major drawbacks. One is that neural networks are not efficient classifiers of rare events, because they are biased toward the modal value (the most common value in the output). This can be a serious problem in large-*N* conflict analysis as militarized disputes are indeed rare events, with 95 percent of observations usually coded as zero. In addition, it is hard to comprehend the causal model that the trained

neural networks have internally constructed. To address the issue of rare events' prediction in neural models, we propose a sampling technique called *balanced training with cross-validation strategy.* It makes use of the advantages of selecting on the dependent variable, while avoiding selection bias.

Our third goal is to offer three measures as guides to interpreting the relative influence of different variables on conflict. The different information provided by each generates insights into the complex regularities discovered by the network. Since our analysis uses variables from both realist and liberal theories, interpreting the network model also provides a test for hypotheses from the two theoretical perspectives.

The first two sections of this chapter consider why a neural network model is suitable for studying international conflict and then discuss the model. Next we focus on the analytical issues of rare event prediction in neural networks and use of the balanced training with cross-validation strategy as a solution for the rare event bias in neural networks. The fourth section briefly discusses the data utilized, and the final sections summarize the results. Much of the discussion of the methodological innovations must be technical. We nonetheless think these innovations are important, and that the results are substantively and theoretically interesting.

WHY NEURAL NETWORKS IN CONFLICT ANALYSIS?

The COW Project has ranged over several levels of analysis. J. David Singer (1961b) initially endorsed the systemic level as most promising and expressed skepticism about the power of the nation-state level of analysis—a position shared, ironically, by a very different kind of scholar (Waltz 1979). Subsequently, Small and Singer (1976) expressed doubt about the value of a middle level of analysis between the systemic and state levels; that is, on pairs of states, or dyads. Notably they questioned the democratic peace hypothesis, a position Henderson too quickly endorses in this volume.[1] (Geller and Singer 1998, 85–96, might suggest some subsequent mellowing on this point.) Nevertheless, Singer was centrally involved in the origin and development of the militarized international dispute data set, which has proven one of the great achievements of the COW project. His early influence is evident in conceptualizing (Leng and Singer 1977) MIDs as bilateral interactions and in a more recent report on use (Jones, Bremer, and Singer 1996). Much of COW use of MIDs in the early years was predominantly descriptive and inductive in mapping characteristics of the international system and had an implicit assumption that the important relationships would be more or less

linear. Recently, however, other COW associates (Maoz and Abdolali 1989; Bremer 1992) have made major innovations in theoretically driven use of MIDs at the dyadic level. This chapter extends that development, working further in what is now called the Kantian peace research program that extends the scope of dyadic influences from democracy to economic interdependence and international organizations. In so doing it looks intently at the implications of using neural network analysis to relax the assumption that key relationships are fundamentally linear.

In light of the dominance of statistical methods in conflict research, we need to consider the benefits that can result when neural network methodologies are applied to conflict data. Statistically trained political scientists may ask, why neural networks? Wouldn't simpler and more established multivariate statistical techniques do better? Answers to these questions can be articulated both from the methodological and theoretical levels.

First, neural networks can provide a powerful method to develop nonlinear and interactive models of militarized disputes, redressing the restrictive linear and fixed effect assumptions that have dominated the field.[2] Recent development in the liberal peace literature seems to indicate that the causal processes at work in interstate conflicts result from complex interactions. In recognition of these complex dynamics, Russett and Oneal (2001, 39) express doubt that individual causal relationships can be considered well in isolation. Peace may result from multiple and overlapping liberal behaviors, shaped by democracy and interdependence, which interact with the opportunities offered by the realist variables. A synthesis of Kantian and realist effects emphasizes an interpretation of constraints on states' willingness and ability to resort to violence. Beck, King, and Zeng (2000) similarly interpret the realist variables as creating a pre-scenario of low or high ex ante probability of military conflict from which the influence of the liberal variables is plucked. Or the effects of relative power on dispute outcomes—strong between nondemocracies—may be much weaker when democracies settle their disputes (Gelpi and Griesdorf 2001). Other studies assume that a reciprocal relationship, or feedback loop, runs between democracy and interdependence. Democratic institutions may indirectly increase the weight of the economic constraints on militarized behavior by empowering economic interest groups in the state. Another link may run as interdependence in turn increases the number of international political constraints. High levels of dyadic trade often create a need for new institutions to manage and stabilize the existing commercial relations. These new institutions add more restraints on militarized behavior.

This interactive and nonlinear perspective can be fully embraced by neural network models. By superposing multiple nonlinear functions and avoiding a priori constraints on the functional nature of the data examined, multilayer networks can construct different causal structures in the same model and combine them together in a systematic way. Indeed, a wide variability of the inputs' effect is allowed, while avoiding the independence assumption of the random effect model (Beck, King, and Zeng 2000, 25).

Second, neural networks do not require independent observations, and thus they deal better with the suspected influences that militarized events exercise on each other (Sarle 1994). As Beck, Katz, and Tucker (1998) argue, the conflict history of a state can either positively or negatively affect the state's willingness to become involved in future conflict. To overcome this problem, they suggest that a control for the number of years that have elapsed from the most recent occurrence of a conflict should be used. However, that solution is problematic. While solving the independence problem, the year correction opens new issues. It rests on the assumption that the effect of the other explanatory variables and time can be separated. This seems very unlikely for some of the important variables in conflict analysis. For instance, liberal theory expects interdependence to fall with the outbreak of a conflict, but to rise over time after a conflict ends.

Finally, model formulation in neural networks is shaped not only by theoretical but also empirical considerations, making the neural methodology a middle-range approach between deductive and inductive model building. This characteristic of neural networks should not be regarded as a negative aspect. The increasing number of factors and reciprocal interactions that may characterize the causal structure of international conflicts makes the development of fully specified theories more difficult.[3] As result, some aspects of the causal interaction may remain undefined. What is left unexplained by the theory can be "discovered" by the neural networks themselves since their flexible methodology enables them to learn from the empirical data. Without deemphasizing model building based on first principles, neural modeling can strengthen theory building by supporting a constant interplay between theory and data.

THE NEURAL NETWORK MODEL

Backpropagation multilayer networks implement a nonlinear mapping, or a function approximation, from a set of n input, $x_1, x_2, \ldots, x_n$, to a set of m outputs, $y_1, y_2, \ldots, y_m$. Although mapping is not new in

quantitative studies, the way in which neural networks perform this input-output transformation represents an important development for mathematical models of complex input-output relations. In neural networks the complex relation between the inputs and outputs is modeled using a superposition of multiple nonlinear functions, represented by neurons. By increasing the number of functional transformations, the model can approximate any hypothetical relations between the selected explanatory and dependent variables. The key point is that the number of nonlinear functions, or neurons, need grow only as the complexity of the mapping itself grows (Hornik, Stinchcombe, and White 1990; Hornik and Stinchcombe 1992).

The major implication of using the flexible functional form provided by the superposition method is that the network functions become nonlinear functions of the network adaptive parameters, the weights (w_I). Because of the complexity involved in the mathematical structure of the network, the procedure for determining the value of the parameters becomes a problem in nonlinear optimization, which requires finding efficient learning algorithms to reduce the network overall error function (Bishop 1996). Our network model utilizes the backpropagation algorithm to calculate the weight values. The backpropagation process is relatively simple in concept. The objective is to compare the *actual output* calculated by the network with the *target output,* given in the training set, each time the network is presented with a sample case. This comparison produces an error value in the output layer that is calculated as a function of the weights. Then the error is propagated backward to the previous layer and used to adjust the weight values in each nonlinear function in order to reduce the difference between the actual output and the target output.[4] Each time a new comparison is made, the error is further reduced.[5]

Besides selecting the learning algorithm, another issue in neural network modeling is how to determine the optimum network architecture. This mainly involves deciding the number and type of functional transformations needed. Selecting the appropriate network architecture is an important part of model building. A higher number of functions (neurons) will produce highly flexible networks, which may learn not only the data structure but also the underlying noise in the data. Too few neurons will produce networks that are unable to model complex relationships. In addition to deciding the number and type of activation function for the network, other parameters need also to be selected. Because of the number of parameters involved in the selection process, choosing the appropriate network architecture is a multicriterion search problem. To perform a global search through the space of possible combinations

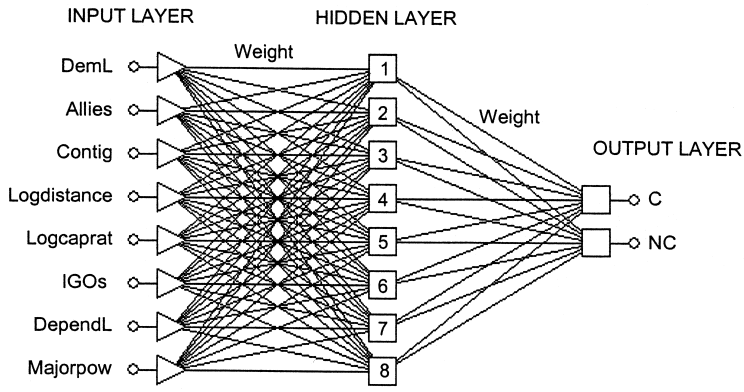

Fig. 1. Diagram of the best neural network configuration

we adapt a genetic algorithm that uses a Darwinian model of evolution (Holland 1992; Davis 1991). The genetic algorithm was used to calculate the number of hidden layers and hidden neurons, best activation functions, and values of the learning rate, η, and momentum, α.[6] The training patterns were entered in a shuffled order, and the training was repeated ten times to avoid bias.[7] The genetic algorithm found that the optimal configuration is a multilayer neural network with one input layer, one hidden layer, and one output layer. The input and hidden layers contain eight processing units each, with two in the output layer. The hidden units utilize the *tanh* function, whereas the output units adopt the logistic function.[8] Finally, the networks perform better using a learning rate and momentum equal to 0.7. Figure 1 provides a schematic representation of the optimal configuration. The appendix presents the genetic algorithm optimization process and its results.

Neural Network Models for Rare Events: A Balanced Training Set with Cross-Validation

Interstate conflict data are often coded as binary dependent variables, with zero as by far the most common value. This implies that a data set of such events will be characterized by an unbalanced dependent variable, with important consequences for the analysis and prediction of interstate disputes with neural network models as with other multivariate models. In such rare event domains the estimated event probability may be so small as to make efficient event prediction very difficult (King and Zeng 2000). Unbalanced data sets also affect the perform-

ance of neural network models. The learning process that neurally based models use to update their weight estimates is biased toward commonly encountered (modal) values in the training sample (Garson 1998, 88). Consequently, practical strategies need to be developed, during the training phase, to improve the neural networks' prediction ability for rare events.

The neural network literature has given little attention to the rare event problem. Few attempts have been made to address it, and the results have not been very successful. Schrodt's (1991) experimental work on conflict data utilizes a replication strategy to increase the number of conflict cases in the training set. It does increase prediction in the training set, but at the cost of reducing the model's ability to correctly classify new cases. What is missing is a correction strategy to reduce the bias produced by intentionally selecting training cases on the dependent variable.

To develop a practical procedure that increases accuracy in the neural network classification of rare events, we extend a strategy suggested by King and Zeng (2000) for logistic regression models to the neural network approach. This solution focuses on selecting data on the basis of the dependent variable (endogenous stratified sampling) while at the same time using a statistical correction (prior correction) for the logistic estimates to avoid selection bias. In the neural network analysis, instead of directly correcting the estimates, as Beck, King, and Zeng (2000) suggest, the correction mechanism is provided indirectly by the cross-validation set. However, both corrections rely on the same principle, which is based on prior information about the incidence of the rare event in the population, τ.

As with any form of data analysis, the meaningfulness of a neural network prediction depends heavily on the extent to which the relevant explanatory variables are selected and included among the network input. If important explanatory variables are omitted, the neural network models cannot produce meaningful predictions. Assuming that the important explanatory variables for the classification have been selected by the researcher and used as the network inputs, we can think of two reasons for poor performance by neural network classifiers that utilize unbalanced training sets: (1) the inadequate type of information that the unbalanced training set provides and (2) the way in which the learning algorithm, which is used during the backpropagation process, minimizes the prediction error and changes the weight values.

Regarding the first point, many political scientists (e.g., Maoz and Russett 1993, 627) suggest that most of the nonconflict cases provide the model with little information. The data on these dyads is often similar

and repetitive, as many of the conditions allowing stability show little variation. More information lies where the action (conflict) is, since disputes are often preceded by changes in other patterns of international interaction. Moreover, since researchers commonly believe that conditions causing international conflicts are highly nonlinear and interactive, the effect of the explanatory variables (the neural network's inputs) may vary widely over the observations. Whereas the effect of many explanatory variables may be undetectable for most dyads—the nonconflict ones—it may be very substantial for the conflict cases.

Since the conflict observations display sensitive input-input and input-output effects, this sensitivity becomes key to understanding and predicting the likelihood of conflicts in the international context. Feeding the neural network with an unbalanced training set, heavily loaded with nonconflict dyads, runs the risk of overemphasizing a single output pattern (nonconflict). The network, in this case, is not exposed to all the possible input-output effects stored in the databases. As a result, the internal model constructed by the neural network during the training will be only partially representative of the much more complex "causal" model embedded in the data and so will be unable to generalize (Garson 1998, 87).

Another important contributor to poor classification performance in neural network models with unbalanced training sets derives from how the backpropagation algorithm works. As explained earlier, backpropagation is achieved in neural networks by an iterative process, as the network repeatedly tries to learn the correct output for each training pattern. During this learning phase, the weights are modified on the basis of error signals generated from the output learned by the network. Thus if the majority of the outputs used in the learning process belong to one class (nonconflict), the error minimization process will concentrate overwhelmingly on that class. The less frequent value of the output (dispute) will account for only minor changes in the network's weights. This negative process would be further strengthened by the fact that neural networks do not have a linear response to the input and are less sensitive to outliers (Schrodt 1991, 372). Indeed, the nonlinear functions used by the network to model the input-output relation have the effect of "squeezing" the values of the data in the training set, especially at the high and low ends of the data range, thus reducing the effect of outliers. Consequently, incidences of dispute, which are the outliers in the conflict data, will not produce dramatic change in the network's internal model even if the input-output effect that they reproduce is quite large.[9] Because few changes in the weight values are determined by the dispute output, and because of the nonlinear impact of the dispute dyads on the

neural network's weights, the learning process in the backpropagation algorithm will be biased toward the modal values of the training set. Unless care is taken to construct training sets with a more balanced representation of the two output classes, the neural network models will be less useful in predicting rare events.

One obvious way to avoid an inadequate information flow to the network and an error minimization process that is biased toward the modal value is to adopt an endogenous stratified sampling for the training set, which is also known as choice-based or case-control design.[10] This sampling strategy focuses on selection within the range of the dependent variable. A predefined number of observations, for which the dependent variable is equal to one, is randomly selected. The same number of cases is also randomly chosen from all the observations with output equal to zero (the control), perfectly equalizing the two outputs in the training set. A perfectly balanced training set alone does not, however, provide the optimal solution. Selecting on the dependent variable is widely recognized as a possible source of biased conclusions.[11] A correction mechanism should be used, together with the balanced approach, so as to avoid biases that will produce the opposite effect of the unbalanced training set.

A way to integrate the equally important needs for a balanced training set and to correct against possible selection bias is to provide the network with prior knowledge of the actual distribution of classes' occurrence. This prior knowledge should be able to correct the weight estimates produced by the balanced training set without affecting the network's ability to learn equally from both classes. Unfortunately, it is hard to incorporate prior knowledge into neurally based classifiers (Barnard and Botha 1993). In neural networks, information about input-output relations is distributed across multiple weight values, making the direct correction of the weight with a priori probabilities of class membership nearly impossible.[12] Since it is so difficult to manipulate the weight values directly, we need new types of correction mechanisms that are both efficient and easy to apply. An alternative approach to incorporate prior knowledge is to use a cross-validation set reproducing the frequency of the rare event in the population, τ.[13] This makes it possible to "correct" the value of all the weights in the network in a distributed way, while allowing the network to produce weight values that focus on both classes.

How does this cross-validation correction strategy work to intervene on the weight values calculated by the training set? The function of cross-validation is to stop the training when the prediction error in the cross-validation set, *MSE,* reaches a predefined local minimum.[14] Thus,

when a balanced training set is used together with an unbalanced cross-validation set, a double divergent process on the error minimization calculation occurs. On the one hand, the error minimization process in the training will be shaped by both classes, since the training set is balanced. The network's search for the optimal weights would be executed on the basis of a minimal error that takes into consideration the correct prediction in both classes. On the other hand, an early stop in the training process is determined by the reduction of the error in the cross-validation sample. Since the more dominant class in the sample is the non-conflict class, training ends when the prediction in this class reaches its minimum error. The goal is to select the combination of weights that equally improves the prediction result on both classes but, at the same time, can offer the best prediction on the more dominant class of the population.[15]

By experimenting on the double levels of the error minimization process (one level supervising the weight change and the other determining the end of the training) the balanced training set with cross-validation correction strategy can solve some of the problems that rare event prediction, specifically conflict prediction, presents to neural network models.[16]

VARIABLES AND ANALYSIS

The network inputs include MIDs as the dependent variable and eight dyadic independent variables. Our theoretical perspective is that of the Kantian research program, addressed to the system of directed and reciprocal relations among democracy, economic interdependence, international organizations, and militarized conflict or the lack thereof, as laid out in Russett and Oneal (2001). Consistent with the view that liberal states carry on relations with each other differently from their power-oriented relations with other states, the analysis includes five variables usually associated with realist analysis, and three Kantian variables. The realist variables include *Allies,* a binary measure coded 1 if the members of a dyad are linked by any form of military alliance. *Contig* is also binary, coded 1 if both states are geographically contiguous, and *Logdistance* is an interval measure of the distance between the two states' capitals. *Majorpow* is a binary variable coded 1 if either or both states in the dyad is a major power, and *Logcaprat* measures the dyadic balance of power on an interval scale. The first Kantian variable, *DemL,* is a 21-point scale for the level of democracy in the less democratic state in each dyad. *DependL* is a continuous variable measuring the level of economic interdependence (dyadic trade as a portion of a state's gross

domestic product) of the less economically dependent state in the dyad. *IGO* measures the number of international organizations in which the two states share membership. Most of these measures (e.g., MIDs, alliances, contiguity, major power, capability, and IGOs) derive from conceptualizations of the COW project and are measured by COW. We lag all independent variables one year to make any inference of causation temporally plausible.[17]

Our data set is the population of politically relevant dyads for the pre–Cold War period (PCW), from 1885 to 1945, and the Cold War and immediate post–Cold War period (CW), from 1946 to 1992, as described extensively and used by Russett and Oneal (2001). For the first population, PCW, only the initial year of the two world wars, 1914 and 1939, is included in the data set. This restriction ensures that the analysis is not unduly influenced by World Wars I and II and by the absence of adequate trade data for the wartime and immediate postwar years.

We chose the politically relevant population (contiguous dyads plus all dyads containing a major power) because it sets a hard test for prediction. Omitting all distant dyads composed of weak states means we omit much of the influence that variables not very amenable to policy intervention (distance and national power) would exert in the full data set; by that omission we make our job harder by reducing the predictive power of such variables, but also make it more interesting. By applying the training and cross-validation sampling technique we show that a strong performance is achieved even when the analysis is restricted to the politically relevant group. By focusing only on dyads that either involve major powers or are contiguous, we test the discriminative power of the neural network on a difficult set of cases.[18] The neural network system is fed with only highly informative data since every dyad can be deemed to be at risk of incurring a dispute, yet it is harder for the network to discriminate between the two classes (dyad-years with disputes and those without disputes) because the politically relevant group is more homogeneous (e.g., closer, more interdependent) than the all-dyad data set. If the balanced training with cross-validation correction strategy outperforms the other techniques with these data, by providing models that can successfully generalize in different time frames, it should also be successful for researchers who wish to consider the entire population of dyads.[19]

The unit of analysis is the dyad-year. There are a total of 27,737 cases in the Cold War population, with 26,845 nondispute dyad-years and 892 dispute dyad-years. The pre–Cold War population comprises 11,686 cases, with 11,271 nondispute dyads and 415 dispute dyads. The dependent variable (*Dispute*), or network output, is 1 if a militarized

interstate dispute (MID) was begun and 0 otherwise. Only dyads with no dispute or with only the initial year of the militarized conflict are included, since our concern is to predict the onset of a conflict rather than its continuation. Other investigations (e.g., Bennett and Stam 2000b; Russett and Oneal 2001) find substantial commonality between the influences on dispute initiation and dispute continuation, but it is best not to assume their similarity, so we limit ourselves to the former.

The CW data are used to generate three training sets (*Balanced Training, Balanced-Replicated Training,* and *Unbalanced Training*), five cross-validation sets (*CV prior correction I, CV prior correction II, CV no correction III, CV no correction IV,* and *CV no correction V*) and five testing sets (*CW Test I, CW Test II, CW Test III, CW Test IV,* and *CW Test V*) according to the different training strategies and their relative sampling rules. The size of these sets varies slightly for each training strategy, as more or fewer dyads are needed to satisfy sampling requirements. Another testing set, *PCW Test,* comprises the complete population of the pre–Cold War period. The training tests are used to fit the model through the backpropagation learning algorithm, the cross-validation tests determine the end of the training (fitting) process, and the error matrices of the testing sets measure the accuracy of each training strategy. Since one objective is to determine whether the pattern discovered by the neural networks for the Cold War period can also explain the pre–Cold War period, only CW data are used in the training and cross-validation set, while the PCW cases are used only as a testing set. The difference in accuracy between the CW and PCW testing sets, achieved by the different training strategies, gives us a measure of stability for the CW model. Similar accuracy for the two periods means that the relationships at work during the Cold War were already in place during the previous era.

The *Balanced Training* set contains a randomly predefined equal number of conflict and nonconflict cases. In the *Balanced-Replicated Training* set we replicated the number of conflict cases of the *Balanced Training* set once and then randomly selected an equal number of nonconflict cases to match the duplicated conflict observations.[20] Finally, the *Unbalanced Training* set comprises an uneven number of conflict and nonconflict cases randomly sampled from the CW population. This last training set is almost half of the entire CW population. By selecting such a large training sample we can show empirically that, quite differently from multivariate statistical techniques, in neural networks the rare event bias does not decrease in large samples.[21]

For the cross-validation phase we generated two sets (*CV prior correction I* and *CV prior correction II*) reproducing the class frequency in the CW population for use as prior correction for the two balanced

training sets (*Balanced Training* and *Balanced-Replicated Training*). In this case, the nonconflict class represents 97 percent of the population, while the conflict class is only 3 percent. The remaining cross-validation sets (*CV no correction III, CV no correction IV,* and *CV no correction V*) provide no correction to the training since they equalize the two classes. These balanced cross-validation sets can test the performance of our correction strategy.

The CW testing sets contain the remaining dyads after the training and cross-validation sets were selected. There are no common cases in the training set, cross-validation set, and testing set, which are used together for each training strategy. Table 1 shows all the totals. To evaluate the performance of the three different training strategies, with and without the cross-validation correction, we compute the kappa and conditional kappa coefficients, K_{hat} *statistic* and K_{hatk} respectively (an estimate of kappa). Kappa analysis, a discrete multivariate technique, offers a comprehensive accuracy measurement for neural classifiers applied to

TABLE 1. Summary of the Data Sets Used for the Neural Network Simulations

Data Set	C	NC	Total Cases
Balanced Training (Bal)	564	564	1,128
Balanced-Replicated Training (Rep)	1,128	1,128	2,256
Unbalanced Training (Unb)	382	10,712	11,094
CV prior correction I (used with the Balanced Training)	10	312	322
CV prior correction II (used with the Balanced-Replicated Training)	19	625	644
CV no correction III (used with the Unbalanced Training)	92	2,682	2,774
CV no correction IV (used with the Balanced Training)	161	161	322
CV no correction V (used with the Balanced-Replicated Training)	322	322	644
CW Test I (used with the Balanced Training with correction)	318	25,969	26,287
CW Test II (used with the Balanced Training without correction)	167	26,120	26,287
CW Test III (used with the Balanced-Replicated Training with correction)	309	25,092	25,401
CW Test IV (used with the Balanced-Replicated Training without correction)	167	25,395	25,562
CW Test V (used with the Unbalanced Training)	418	13,451	13,869
PCW Test	415	11,271	11,686

rare event domains. While overall accuracy stresses the overall result of the classification by focusing only on the main diagonal of the classifier's error matrix, kappa analysis calculates how the accuracy is distributed across the individual classes. By considering both individual class accuracy and overall accuracy, kappa analysis does not bias accuracy evaluation toward the dominant class in the testing set as overall accuracy does.[22] Moreover, it is especially appropriate here because dispute data are not continuous and normally distributed (Jensen 1996, 250). Once the kappa and conditional kappa coefficient have been calculated, a pairwise test Z statistic is used to determine whether the prediction results of an error matrix are significantly better than a random result, as well as whether similar error matrices, which consist of identical classes but are the product of different classifiers, are significantly different.[23]

RESULTS AND DISCUSSION

We initially discuss the results of the kappa analysis for all the training strategies, with and without the cross-validation correction, on the CW data set. We also show individual error matrices for each strategy to further support the kappa results.[24] Finally, we focus on the ability of the Cold War models, selected from the best training strategy, to postdict the pre–Cold War dyads. By comparing the accuracy between the CW and PCW testing sets, we test the hypothesis that the causal relationships triggering interstate conflicts have been stable over time.

Table 2 summarizes the results of the kappa analysis for the Cold War period, in the form of significance matrices.[25] Those matrices show all the Z values from comparing the kappa coefficients of the different

TABLE 2. Significance Matrix for Comparing the Training Strategies Using Kappa Analysis and the Cold War Testing Sets

CW	CVBal	Bal	CVRep	Rep	Unb
KAPPA	0.045	0.022	−0.016	−0.009	0
VAR	0.000009	0.000004	0.000002	0.000001	0
CVBal	15				
Bal	**6.38**	11			
CVRep	**18.39**	**15.51**	**−11.31**		
Rep	**17.08**	**13.86**	4.04	**−9**	
Unb	**15**	**11**	11.31	9	**0**

Note: The table also presents the Kappa coefficient and the variance for each training strategy. Bold Z values indicate a significant improvement in the performance of the training strategies at 95% confidence level ($Z > 1.95$). The bold values in the main diagonal indicate that the classification of the training strategy is worse than a random one at 95% confidence level ($Z > 1.95$).

training strategies (off-diagonal elements) two at a time, as well as the Z statistic that measures the significance of each individual classification (main diagonal elements). The tables also present the kappa coefficient and variance for each training strategy in the first two rows. If the Z value exceeds the critical value, $Z_{\alpha/2}$, then the classifications are significantly different, or, as with the Z values in the main diagonal, are worse than a random one. A better performance is given by the training strategy with the higher kappa coefficient.

The first two rows following the kappa coefficient and variance row show the result of the two balanced training techniques starting with the balanced training using the cross-validation prior correction (*CVBal*) and then the one without it (*Bal*). The balanced-replicated training follows, again initially utilizing the cross-validation prior correction (*CVRep*) and then without it (*Rep*). Finally, the result of the unbalanced training is shown (*Unb*). We use the acronyms *C* (conflict) and *NC* (nonconflict) in the conditional kappa matrices.

The importance of constructing a meaningful training set emerges clearly from the table 2 results. Three classifications, *CVRep, Rep,* and *Unb,* are statistically insignificant, since the Z values in the main diagonal are smaller than 1.95 (the critical value at $p < .05$), $Z = -11.31$, $Z = -9$, and $Z = 0$, respectively (significant values are in boldface). Only the two balanced training strategies (*CVBal, Bal*) are significantly better than a random classification. This underlines three important factors. First, balanced training is a key to produce robust classifications in backpropagation networks. Second, in backpropagation classifiers, rare event bias remains significant in large unbalanced samples. Finally, replication strategies are not efficient since they largely reduce the network's generalization ability. This is because the duplication of cases in the training sets leads the network to learn too well the training cases, so fitting data noise rather than data structure. And table 2 stresses another important result: *CVBal* performs statistically better than all the other training sets. Indeed *CVBal* also achieves better accuracy than *Bal* ($Z = 6.38$). This is empirical evidence that the cross-validation prior correction is effective in significantly reducing selection bias in balanced samples.

In table 3, the error matrix of the *CVBal* technique in the Cold War analysis shows that this training strategy achieves the highest individual class accuracy. It correctly predicts 82.4 percent of the militarized outcomes and 72.2 percent of the nonmilitarized outcomes during the Cold War period. Because it far less often fails to identify the politically costly and dangerous dispute dyads as nondisputes while still providing a high accuracy on the nondispute class, *CVBal* emerges as the best training strategy to adopt with rare event data in international relations.[26] Since

TABLE 3. Error Matrices of the Three Training Strategies with and without the Cross-Validation Prior Correction for the Cold War Testing Set

CW/Method	CVBal		Bal		CVRep		Rep		Unb	
Output/Desired	C	NC	C	NC	C	NC	C	NC	C	NC
C	262	7229	137	7778	73	18109	35	18172	0	0
NC	56	18740	30	18342	236	6983	132	7223	418	13451
Class Accuracy	82.39	72.16	82.03	70.22	23.62	27.83	20.96	28.44	0.00	100
Overall Accuracy	72.29		70.30		27.79		28.39		96.97	

these predictions are out-of-sample forecasting, they indicate that the interactive model developed by the neural network and the balanced with cross-validation correction strategy can extract much of the influences embedded in conflict data. Consequently, we believe that neural networks together with the balanced training with cross-validation correction constitutes a viable and efficient method to improve model forecasting ability in conflict analysis, especially analyses of pooled annual dyadic time-series data.

Finally, we address a key theoretical and substantive question: Are the patterns stable? The error matrices of *CVBal* for the PCW data set in table 4 show that the CW model provides a high level of accuracy for the pre–Cold War period too. As table 4 shows, postdiction of the pre–Cold War dyads is similar to the Cold War result (64.8 and 65.5 percent accuracy for the PCW dispute dyads compared with 82.4 and 72.2 percent for the CW ones). However, these results are deceptive since the PCW and CW testing sets are different in size. To prevent differences in sample size from influencing the result, we performed kappa and conditional kappa analysis together with a pairwise Z test statistic on the *CVBal* result as a measure of normalized accuracy, thus making the PCW and CW error matrices directly comparable.

TABLE 4. Error Matrices of CVBal Strategy for the Pre–Cold War Testing Set

PCW/Method	CVBal	
Output/Desired	C	NC
C	269	3888
NC	146	7383
Class Accuracy	64.81	65.50
Overall Accuracy	65.48	

Table 5 summarizes the kappa and conditional kappa analysis results in the form of a significance matrix. Whereas the levels of accuracy for the two periods are similar, the overall performance of the CW model is *significantly better* on the pre–Cold War dyads ($Z = 2$), with the kappa coefficient for the PCW accuracy larger than for the CW ($0.056 > 0.045$). This is mainly because the class accuracy for the conflict dyads in the pre–Cold War years is substantially better than in the Cold War era ($Z = 2.21$) with the conditional kappa for the PCW being larger than the CW ($0.03 > 0.023$). However, while the model does substantially better in predicting disputes (avoiding false positives) in the pre–Cold War period, it loses some predictive ability regarding nondisputes (more false negatives). The accuracy for the CW nonconflict class is significantly better than for the PCW one, though the difference in performance is not as large as for the conflict dyads since, as mentioned before, PCW accuracy is better overall. In the case of the nondispute class $Z = 8.12$, this time the conditional kappa for CW nondisputes exceeds the PCW coefficient ($0.454 < 0.754$).

These findings lead to three important inferences. First, not only is the pattern of dyadic influence discovered by the networks for the Cold War disputes reasonably representative of the pre–Cold War context, those influences enabling conflicts were even stronger in the earlier period. Second, in relation to peace, the structure of influence was already in place in the pre–Cold War years, although showing slightly less strength. This slight difference in strength can be explained by the maturation of democratic institutions and the transformation of the economy from national to global in the twentieth century. Most likely, these two factors have increased the positive influence of democracy and economic interdependence, which was already coming into place late in the

TABLE 5. Significance Matrix for Comparing the CW and PCW Accuracy of the CVBal Training Strategy Using Kappa and Conditional Kappa Analysis

CW/PCW	CW	PCW	CW_(C)	PCW_(C)	CW_(NC)	PCW_(NC)
KAPPA	0.045	0.056	0.023	0.03	0.754	0.454
VAR	0.000009	0.000027	0.000002	0.000008	0.000888	0.001287
CW	15					
PCW	**2**	10.78				
CW_(C)	•	•	14.67			
PCW_(C)	•	•	**2.21**	10.83		
CW_(NC)	•	•	•	•	25.29	
PCW_(NC)	•	•	•	•	**8.12**	12.65

Note: The table also presents the Kappa coefficient and the variance for each training strategy. Bold Z values indicate a significant improvement in the accuracy at 95% confidence level ($Z > 1.95$).

nineteenth century, in the subsequent years (Russett and Oneal 2001; Sachs 1998).

Finally, since the reduction in influence on peace does not match the increase in influence on disputes, the overall pattern fits the pre–Cold War years even better than it does the Cold War ones. As some differences do emerge in the strength of the effect, we next turn to the relative performance of individual predictor variables and whether they underline stability or difference.

MODEL INTERPRETATION

We now interpret the causal model that the neural networks using the balanced with correction strategy developed during the training. Although valid indications about the causal structure can be offered by the following analysis, our task is difficult and, at this stage, should be regarded as tentative. Interpretation of the causal hypotheses represented by a trained neural network is a complex exercise for several reasons. First, neural network models encode their knowledge across hundreds or thousands of parameters (weights) in a distributed manner. These parameters embed the relationships between the input variables and the dependent output. The sheer number of parameters and their distributed structure make the task of extracting the network knowledge not an easy one. Second, the weight parameters of a multilayer network usually represent nonlinear and nonmonotonic relationships across the variables, making it difficult to understand both the relative contributions of each single variable and their dependencies. Thus, to extract a causal model developed by the trained network we utilize three different approaches. By doing so we can interpret the network model from single variable and dependency perspectives. We calculated three measures of input influence: a relative evaluation of the general influence of each input variable, the specific influence of the individual input variables on the network output, and the input relation factor of each input.

General Influence of Inputs

The general influence measure, *GI,* provides an estimate of the relative overall influence exerted by each input variable on the output. It relies on the absolute value of the weight of the trained network. Inputs connected to the hidden and output units by weights of large absolute magnitude will have more relative influence than those inputs with smaller magnitudes. Since *GI* focuses on single input variables in isolation

without taking their dependencies into account, it should be regarded as an approximate measurement. Following Howes and Crook (1999):

$$GI(x_i, net) = \frac{\sum_{j=1}^{h} \left| \left(w_{ji} \div \sum_{k=0}^{n} \left| w_{jk} \right| \right) v_j \right|}{\sum_{j=0}^{h} \left| v_j \right|} \quad (1)$$

with x_i being the i input variable, *net* referring to the network architecture, w_{ji} being the weight from the ith input node to the jth hidden node, and v_j giving the weight from the jth hidden node to the output node.[27] Because our network presents two output nodes, we compute two separate *GI* values for each input node.

Table 6 reports the general influence results for the balanced with cross-validation correction network. Economic interdependence and democracy exert the greatest general influence on both the conflict and nonconflict outcomes (*DependL* = 0.173 for dispute and 0.213 for nondispute, *DemL* = 0.160 and 0.197, respectively). This supports the liberal thesis that the state with the lower level of interdependence and democracy in the dyad has the major impact on dyadic relationships. Another Kantian variable, *IGOs,* also has a high general influence value (0.154 for the conflict output and 0.188 for the nonconflict one), indicating that international organizations can constrain interstate behavior. Other variables that matter are *Logdistance* (0.142 for *C* and 0.174 for *NC*), *Allies* (0.136 for *C* and 0.167 for *NC*), and *Logcaprat* (0.131 for *C* and 0.161 for *NC*). Thus proximity, alliance, and power also play a part in providing opportunities and incentives for interstate action. Overall, the result supports theories first of the democratic peace, then the liberal peace of both democracy and economic interdependence, and finally the Kantian peace of democracy, trade, and IGOs. However, three realist variables, *Logdistance, Allies,* and *Logcaprat,* cannot be ignored. The relationship of democracy and interdependence and interstate conflicts is to some extent mediated by both the dyadic balance of power and geographical proximity. This supports Russett and Oneal's (2001) and Beck, King and Zeng's (2000) syntheses of liberal and realist in-

TABLE 6. General Influence of the Network's Input Variables

GI	DemL	Allies	Contig	Logdistance	Logcaprat	IGOs	DependL	Majorpower
C	0.160	0.136	0.066	0.142	0.131	0.154	0.173	0.038
NC	0.197	0.167	0.081	0.174	0.161	0.188	0.213	0.047

fluences. For example, proximity is positively related to both trade and the probability of disputes, so a failure to control for distance can readily produce the erroneous impression that trade causes conflict.

Specific Influence of Inputs

The specific influence of inputs, *SI,* measures the degree to which each input variable contributes to the dependent output. Instead of relying on the weight value, *SI* compares the output of the network with the network's new output produced by a modified form of the input pattern. Using an approach similar to Saito and Nakano's (1998), we iteratively increase the value of one input variable in the training set by a small amount (initially 0.1 of a standard deviation from the mean, to keep the increases comparable across variables, then 0.3, and finally 0.5), while keeping all the other inputs unchanged.[28] Then we reinterrogate the network, record the difference in the output values as percentages, and then present the overall result as the average. Those input variables producing a large percentage change on the dependent output contribute significantly to the network's prediction. The measurement nevertheless should be regarded as an estimate, since the interdependence across variables means that no scheme of single ratings per input can reflect all the subtleties of the full situation.

Table 7 shows the *SI* of the eight input variables for the conflict and nonconflict outcome. In both cases, the input variables identified by the *GI* as the most significant are still the ones having the greatest general influence on the output. In addition, this time, *DependL* (*SI* = 23.77 and 25.65) and *DemL* (*SI* = 21.31 and 20.55) have the strongest *SI* value, for both the conflict and nonconflict outcomes. This means that small increases in values for the state with the lower economic dependence or democratic score move the conflict output toward the nonconflict outcome and make the nonconflict value more evident. Again this underlines that the degree of democracy and economic interdependence in the less constrained state in the dyad has a strong influence on the probability of conflict. *Logdistance* (15.55 and 15.50) and *IGOs* (12.83 and 12.57) follow, both for conflict and nonconflict dyadic outcomes. Greater geographical distance between states or a larger num-

TABLE 7. Specific Influence of the Network's Input Variables

SI	DemL	Allies	Contig	Logdistance	Logcaprat	IGOs	DependL	Majorpower
C	21.31	8.58	6.94	15.55	12.42	12.83	23.77	3.79
NC	20.55	8.46	6.84	15.50	11.21	12.57	25.65	3.70

ber of shared memberships in international organizations cuts the probability of conflict by nearly 16 percent and almost 13 percent, respectively. Conversely, the nonconflict probability increases almost exactly the same amount. Finally, another variable, *Logcaprat,* is also important from the *SI* perspective (12.41 and 11.21). Increases in the dyadic power ratio reduce the probability of conflict while increasing the chance of a peaceful outcome.

The *SI* measurements once more stress the influence of increasing economic interdependence and democracy on reducing the incidence of interstate disputes. The results also show the importance of two key realist variables: geographical proximity and power ratio. And as in the case of *GI,* the *SI* values indicate the need to hypothesize complex causal patterns of interaction across the variables deemed to trigger interstate disputes.

Input Relation Factor

The input relation factor (*RF*) tries to uncover the dependencies across the input variables. An input variable may have a low *GI* and *SI* but a high *RF.* This means that input variable would not likely trigger the outcome, but it is important in enabling the other explanatory variables to do so. In other words, *RF* measures the degree to which an input variable is *necessary* for producing the network output, although it alone may not be *sufficient* to determine it. To calculate the *RF* of our eight input variables we developed a heuristic procedure, switching off one variable at a time in the Cold War and pre–Cold War testing sets by replacing its values with zero. We then calculated the deterioration in modeling performance by comparing the change in class accuracy between the test with all the active input variables and the ones with one input variable switched off.[29] Since the network learned the causal structure taking in consideration all the relationships across all the variables, the deterioration in accuracy indicates the enabling power of the switched-off input variable.

The *RF* values for the eight variables produced by the CW and PCW tests in tables 8 and 9 identify the same variables as the main ones, although the model deterioration for the PCW period is higher. This

TABLE 8. Input Relation Factors for the Cold War Years

CW	All	DemL	Allies	Contig	Logdistance	Logcaprat	IGOs	DependL	Majorpower
C	82.39	6.50	62.58	88.05	41.82	52.51	66.99	0	73.90
NC	72.16	94.66	80.88	54.56	86.42	86.81	75.66	100	75.84

shows that the *interactive* pattern in the two periods is similar, but with stronger effects for several variables in the pre–Cold War period. The liberal variables make the most difference on the conflict outcome. *DependL* has the greatest *RF* (low *RF* values indicate high impact). When this input variable is switched off, the model's performance drops hugely, from the original 82.4 percent and 64.8 percent for Cold War and pre–Cold War years respectively, to *0 percent* in both periods. *DemL* follows with a very significant *RF* value for the conflict outcome (6.5 percent for the CW set and 4 percent for the PCW data). As previously indicated, *Logdistance* (41.82 percent *RF* value for the Cold War dispute cases and 25.54 percent for the pre–Cold War disputes) and *Logcaprat* (52.51 percent for CW conflicts and 40.72 percent for the PCW conflicts) also have substantial enabling power. Finally, *Allies* (62.58 percent for CW and 42.62 percent for PCW) and *IGOs* (62.58 percent for CW and 43.61 percent for PCW) to a lesser degree influence the effect of the other variables on disputes in both periods. In addition, for *IGOs,* the model deterioration is higher in the PCW years. From these results, not only do we reject the hypothesis that low democracy and shared participation in international organizations had weaker dispute-enabling effects in the pre–Cold War era, we support the opposite hypothesis: that they had *even stronger effects earlier.* Furthermore, low economic interdependence appears to be the most important necessary condition for conflict in both periods.

These results once again show the power of interdependence, democracy, and distance. Since these variables have high values by each test—*GI, SI,* and *RF*—they emerge as key variables. They affect war directly and enhance other variables' influence on dispute initiation. This supports Beck, King, and Zeng's (2000) conclusion that the effect of the input variables varies significantly across dyads as a consequence of input-to-input interactions as well as a feedback loop between democracy and economic interdependence (Papayoanou 1997; Burkhart and Lewis-Beck 1994; Weede 1996; Przeworski and Limongi 1997).

In predicting nondispute outcomes, the only input variable with a significant *RF,* both for the CW and PCW test, is *Contig* (54.56 percent for the Cold War data and 40.28 percent for the pre–Cold War years).[30] (Not surprisingly, contiguity was more important in the pre–

TABLE 9. Input Relation Factors for the Pre–Cold War Years

PCW	All	DemL	Allies	Contig	Logdistance	Logcaprat	IGOs	DependL	Majorpower
C	64.81	4.00	42.63	82.89	25.54	40.72	43.61	0	59.52
NC	66.50	98.00	80.97	40.28	90.83	80.37	78.54	100	62.89

Cold War years of less effective military technology to exert force at a distance.) Contiguity makes *DemL, DependL, IGOs, Logdistance,* and *Logcaprat*—the variables with significant *GI* or *SI* values—more important in reducing disputes. Again the close *RF* value for the Cold War and pre–Cold War period stresses a stable structure of influence over time at the interaction effect level.

The difference between the predictions of disputes and nondisputes means that the interaction between the explanatory variables is also nonlinear. Although low levels of economic interdependence, democracy, distance, power imbalance, and shared membership in international organizations and alliances interact to create multiplicative effects that enhance the likelihood of a dispute, high levels of those variables do not have the same multiplicative effect on peace. Low values produce strong interaction effects, while high values display more of an additive relationship in which they complement each other more than they interact with each other. If two states are geographically close, low levels of interdependence and democracy, a relatively equal balance of power, and low level of participation in international organizations and alliances interact with each other and with proximity to substantially raise the risk that a dispute will occur. But for more distant states (not contiguous), each variable makes a substantial contribution to keeping the peace even in the absence of much help from the others.

Disputes can be quite effectively explained as deriving from low levels of democracy and interdependence, geographical proximity, a relatively equal balance of power, and low shared membership in international organizations and alliances within the dyad. That is, "unhappy" relationships stem from the lack of one or more of the constraints that interdependence, democracy, distance, an imbalance of power, and international organizations could provide. These are the conditions under which the anarchic Hobbesian world of each against all identified by the realists applies. By contrast, "happy" relationships are a mixture of those dyads where one or more of the constraints do apply, and of those states that are not contiguous and so lack the immediate set of opportunities and capabilities which contiguity provides as inducements to disputes among states not otherwise constrained. For the former, "happiness" derives more from the liberal attributes that suppress violent conflict, whereas for the latter it derives more from their separation (Kinsella and Russett 2002).

The causal interpretation provided by the three measures of input influence does not uncover the full model developed by the neural network. However, it does offer interesting findings with which to refine

theories on peace and war. Relationships across the variables do appear to be nonlinear, contingent, and nonmonotonic. Also, the liberal variables—economic interdependence, democracy, and international organizations—play important direct as well as indirect roles in producing war and maintaining peace. Their influence is strengthened both by some interaction between them and by their interactions with the realist variables of geographical proximity, contiguity, balance of power, and alliances.

CONCLUSION

The interstate dispute model we developed, using backpropagation multilayer neural networks, a balanced training with cross-validation strategy, and Cold War data, improves on the dispute prediction capability of the initial pioneering efforts utilizing neural network methodologies.[31] Our preferred model correctly categorizes 82.4 percent of the Cold War dispute dyads and 64.8 percent of the pre–Cold War ones. For the nondispute cases the accuracy is high: 72.2 percent for the Cold War years and 65.5 percent for the pre–Cold War. But when we compare these postdiction results using kappa and conditional kappa analysis—which is necessary because of the different sizes between the pre–Cold War and Cold War testing sets—we find that the overall accuracy of the model for the pre–Cold War period is significantly better than for the Cold War years. These results over both periods indicate an underlying stability of the network structure, both overall and in the consistent strong effect of economic interdependence, democracy, and to some extent international organizations, on the conflict outcome. Indeed, the somewhat better accuracy on disputes for the pre–Cold War period underlines that the pattern of interactions leading to conflict remains highly representative across time and space, not one vulnerable to changes in systemic or state-level characteristics. Although nonlinear and nonmonotonic relationships often characterize the interaction across the variables, our findings indicate that this complex interaction was fully in place during the pre–Cold War era.

In relation to peace (nondisputes), the results are different. Although the Cold War model for nondisputes does well in the pre–Cold War period, it has less predictive power. While interdependence, democracy, distance, difference in power ratio, and shared participation in international organizations and alliances all were important variables for maintaining peace in the previous period, their influence was slightly weaker than in the Cold War context.

The final analysis, of dependencies across the eight input variables

for the Cold War and pre–Cold War period, provides additional evidence that the relationships producing interstate disputes are structurally similar but stronger for the pre–Cold War context. Furthermore, analysis of the peaceful interactions shows how similar the dependence structure was in the two periods. While the variables display a slightly weaker influence on peace in the pre–Cold War years, the interactive effect leading to peace is stable over time. Interdependence and democracy consistently emerge as key variables, together with proximity, power ratio, international organizations, alliances, and contiguity. All this extends and deepens earlier indications (Russett and Oneal 2001) that the same fundamental pattern of influences applied for more than a century. In doing so it gives a stronger basis to believe that it will continue to apply in the twenty-first century—under conditions when democracy, interdependence, and international organizations are deeper and more widespread than ever before.

Our analysis indicates, moreover, that the pattern of relationships affecting disputes often is not linear, and that interactions are common. For example, instead of exerting a constant effect, economic interdependence and democracy may vary their influence as they are either enabled or not by interaction effects between themselves and with the realist influences. Russett and Oneal (2001) suggested some of these interactions, but they are more apparent with the new neural network model used here. This analysis, however, only begins to understand what those relationships may be. It represents a challenge to theorists and methodologists to carry on the task of understanding the complexity, even in terms of the limited number of variables employed here, of the international system in which we try to live in security and peace.

APPENDIX

A genetic algorithm solves optimization problems by creating a population or group of possible solutions to the problem at hand. In our case this method starts when a large random population of network configurations is constructed following the number of input and output variables and the general neural structure required. Each configuration is then expressed as a string of values, a "chromosome," in which each value, a "gene," represents a network parameter. Then each network in the population is trained and a fitness score assigned to it on the basis of a fitness function. The fitness function may incorporate many criteria in evaluating the network quality. Here, the accuracy of the network, the complexity of the configuration, and the ability to learn rapidly are of importance. Indeed, the genetic algorithm process aims to

TABLE 10. Output of the Genetic Algorithm Optimizing the Network Configuration

Rating	Fitness	Genetic String	1 Hidden Layer Neurons	2 Hidden Layer	Learning Rate	Momentum	Output Neurons
1	0.8320	11111111000000001100000011010000011010	8 Tanh	none	0.7	0.7	2 Logistic
2	0.8314	11111111000000000001000001101111000111	17 Tanh	none	0.7	0.7	2 Logistic
3	0.8297	11111111000000001101000011000101111111	2 Logistic 31 Tanh 2 Linear	none	0.7	0.7	2 Logistic
4	0.8292	11111111000000001000000010000111111111	3 Logistic 58 Tanh 3 Linear	none	0.7	0.7	2 Logistic
5	0.8291	11111111000000001011000010100111111010	2 Logistic 45 Tanh 2 Linear	none	0.7	0.7	2 Logistic
6	0.8289	11111111000000001100000011010000011010	2 Logistic 36 Tanh 2 Linear	none	0.7	0.7	2 Logistic
7	0.8289	11111111000000001001000010010001001011	3 Logistic 51 Tanh 3 Linear	none	0.7	0.7	2 Logistic
8	0.8283	11111111000000000101000011111001111011	2 Logistic 39 Tanh 3 Linear	none	0.7	0.7	2 Logistic
9	0.8282	11111111000000000001000001101111001010	17 Tanh 2 Linear	none	0.7	0.7	2 Logistic
10	0.8280	11111111000000000010000001001000111111	27 Tanh 2 Logistic	none	0.7	0.7	2 Logistic

minimize the network's training and cross-validation errors, complexity, and learning time. Furthermore, each criterion is normalized and weighted according to its importance.

On the basis of the assigned fitness score, the best network configurations are selected. These networks go through a process of genetic manipulation, which involves crossover of genetic material (mating of genes) and mutation (randomizing of genes). The crossover mechanism allows us to recombine the fittest network parameters while narrowing down the genetic algorithm's search space. Instead, mutation encourages extension of the search space and is useful if the population has converged on a local suboptimum solution.

The genetically modified network configurations are then retrained, their fitness calculated and compared with previous values. If the resulting networks offer higher fitness scores than previous attempts, then they are used as parents for the next generation. In this way, the genes, which represent the network parameters of the fittest networks, are maintained during the evolution process. This sequence of training, replication, crossover, and mutation continues either until a prespecified number of generations has been reached or until a desired fitness value has been achieved.

Table 10 shows the final result after twenty generations of optimization of the neural network configuration. Only the top ten performing networks with their calculated fitness values and genetic string are shown. The combination of neural activation functions within the hidden and output layers is also illustrated, as well as the network's learning rate and momentum. As can be seen from the table, the optimal configuration selected by the genetic algorithm is the one described previously in figure 1.

NOTES

Authors' Note: We thank the Carnegie Corporation of New York, the Ford Foundation, the National Science Foundation, the Weatherhead Initiative on Military Conflict as a Public Health Problem, the Economic and Social Research Council, and the Foreign and Commonwealth office for financial support, and Richard Aldrich, Neal Beck, Scott Boorman, Evan Govender, Gary King, John Oneal, Carlos Vieira, and Langche Zeng for helpful comments. We gave an earlier version of this chapter at the annual meeting of the Peace Science Society (International), New Haven, October 2000. Our data, from Russett and Oneal (2001), are at www.yale.edu/unsy/democ/democ1.htm.

1. Henderson is provocative but mistaken. His major effort is to run many additive and multiplicative combinations of regime scores to show the insignificance of the democratic peace in the presence of political distance. But there

are better ways to test this. The simplest is in Oneal and Russett (1997), which is the source of Henderson's data: Instead of combining the lower and higher regime scores of a dyad in any way, just include each in the regression model. This shows clearly that political distance does matter, in what we call the cats and dogs effect—but there is also a democratic peace. Democratic dyads are most peaceful, autocratic dyads less so, and mixed dyads least peaceful. Also see Peceny and Beer (2002), who show that even when autocracies are divided into dyads of similar types, democratic dyads still are more peaceful.

Alternatively, create an indicator that identifies truly democratic pairs (both states above +6), and one that gives political distance (*DemH* minus *DemL*). Enter both. Both are significant. This test also shows that the effect of regimes is best captured by truly democratic dyads (above +6) and truly autocratic dyads (below −6). There—with coherent regimes—the democratic peace is clearest. Neural networks analysis is well suited to find such unanticipated nonlinearities.

2. Related concerns drove the debate over whether fixed-effects models are useful in analyses of disputes, with the consensus in the negative. See Green, Kim, and Yoon (2001) with rebuttals by Beck and Katz (2001), Oneal and Russett (2001), and King (2001); also see Bennett and Stam (2000b). Neural networks analysis addresses some of the problems Green et al. identify.

3. As Weinberg (1975, 18–25) underlines in his conceptualization of scientific inquiry, social behavior belongs to the realm of organized complexity and, as such, may be too complex for analytical treatment.

4. The weight change, ΔW, at the time t for all the weights' value in the network is

$$\Delta W_{(t)} = \eta\delta X + M\Delta W_{(t-1)} \tag{2}$$

where η is a small positive constant called learning rate, usually between 0 and 1, δ is the local error gradient for the neuron considered, X is the input of the neuron, M is a constant called the momentum coefficient ranging between 0 and 1, and $\Delta W_{(t-1)}$ is the change in error in the weight value in the previous time period, $t - 1$. Ripley (1994) describes the backpropagation algorithm.

5. However, the backpropagation algorithm is not guaranteed to find the global minimum. The error surface that is the geometrical representation of the error function is multidimensional, displaying not only a global minimum but also multiple local minima. In its search for the optimal solution, the backpropagation algorithm can easily get trapped in these local minima. To avoid this, a momentum term, α, is added to the backpropagation formula. Another strategy in avoiding local minima is to run the training process numerous times starting from randomly selected initial values—that is, from different ordering of the training data and/or different initial random weights (Garson 1998, 50–54). Early stopping, which adopts a cross-validation set during the training, may also reduce the danger of settling in a local minimum (Sarle 1995).

6. For genetic algorithms applied to neural network optimization see Miller, Todd, and Hegde (1989), Yao (1999), and Blanco, Delgato, and Pegalajar (2000).

7. The order in which the training patterns are presented to the network can also affect performance. If the data are grouped in the same manner, rather than being randomly organized, the system may "remember" the last group better than the previous ones, with the obvious consequence that the system's predictive result would be biased toward the information contained in this last grouping. To avoid this problem, the neural analysis should be repeated multiple times using differently ordered data and different initial weight values, both randomly chosen. Moreover, in order to further reduce the possibility of local mimima, the network training was repeated ten times, and the lowest accuracy (training and cross-validation errors) was selected to be used in the fitness function (Bengio 1996, 31).

8. On the *tanh* function see Abramowitz and Stegun (1966, 83).

9. As mentioned before, because of the unusual character of dispute events, the dispute dyads often contain information on large input-output effects.

10. The term *choice-based sampling* is used in econometrics, while *case-control design* is more common in epidemiology. For further discussion see Breslow (1996).

11. As King and Zeng (2000, 7) stress, "Designs that select on Y [the dependent variable] can be consistent and efficient but only with the appropriate statistical correction." This is because a sample selected on the values of the dependent variable can increase the effect of the input on the output and the estimated probability of conflict events for all dyads (note that this bias is exactly opposite to the one produced by an unbalanced training set).

12. Prior correction in conventional statistical classifiers involves estimating the coefficients using the sample selected on the dependent variable and then correcting the estimates with a priori probability of class membership. This a priori probability takes into account the ratio of the classes in the population and in the sample. For examples of a priori correction see King and Zeng (2000, 5–6), McKay and Campbell (1982), and Strahler (1980).

13. The importance of class size as a means to provide the network with prior knowledge of class allocation in the population is discussed in the neural network literature. On incorporating prior knowledge into neural network classifiers see Foody (1995) and Foody, McCulloch, and Yates (1995).

14. The MSE for the cross-validation set is

$$MSE = \frac{1}{n}\sum_{p=1}^{n}(t_p - y_p)^2 \tag{3}$$

where t_p is the target output of each cross-validation sample, y_p is the actual output calculated by the network, and n is the number of cases in the cross-validation set.

15. In the literature, cross-validation has mainly been used to improve the generalization ability of the network model. Instead of stopping the training process when the MSE on the training set reaches the minimum, the MSE of the cross-validation set is used for early stopping. By doing this we avoid possible

overfitting of the data, which refers to the extent to which the network has gone beyond learning the optimal pattern from the training data to also learning the idiosyncratic noise particular to that specific training set (Mosteller and Tukey 1977, 36–41). However, while performing this function, the cross-validation can also be used to provide additional information to the network. Without moving away from the traditional use of cross-validation, we suggest extending the cross-validation method to encompass prior knowledge functionality. Consequently, we use cross-validation both for early stopping during training and to provide the network with prior knowledge on class distribution.

16. Some analytical solutions suggested in the literature directly correct the probabilistic output produced by the neural network or, in the case of a network with a logit output function, the constant term in the hidden neuron to output layer (see King and Zeng 2000, 24). Though these efforts are valid and theoretically well grounded, any estimation errors in the network parameters could be aggravated by the analytical correction, making the final corrected result less accurate than the uncorrected one. Thus the cross-validation correction strategy we employ still provides, in this case, a better solution for producing more accurate results.

17. When attributing cause in reducing conflict we follow the theoretical reasoning of Russett and Oneal (2001), supported by the lag. Oneal, Russett, and Berbaum (2003) use distributed lag models to validate that reasoning more persuasively, finding that trade and peace constitute a feedback loop of mutual reinforcement, and that IGOs increase trade. Other known causal links (Russett and Oneal 2001, chap. 6) include from peace, democracy, and trade to IGOs, and democracy to trade. Pevehouse (2002) reports a link from IGOs to democratization. Kant saw such influences as creating what is now called a dynamic feedback system.

18. This restriction may provide another theoretically relevant advantage. By dropping the non-PRDs, characterized by great distance and weakness, we eliminate many dyads for which such constraints on dispute initiation as democracy and trade may have a lesser role to play in preventing disputes that are highly unlikely to arise anyway. Of the relatively few disputes falling outside of the politically relevant dyads, many are multistate disputes with small powers being drawn into disputes between major powers (Lemke and Reed 2001b). Expected utility calculations seem less informative with nonrelevant dyads; see Bennett and Stam (2000c).

19. Like logistic regression and many other multivariate models, network analysis does not readily identify such historical dynamics as contagion, diffusion, and imitation.

20. We designed the replication strategy so as to increase the size of the training set. The need to keep a balanced ratio of the two classes in the training set reduces the size of the training sets, since the number of conflict cases that can be utilized is limited. To assess whether bigger balanced training sets produce better results we replicate the conflict cases in the training set.

21. King and Zeng (2000, 17–19) show that in logit analysis the rare event

bias becomes minimal in large samples, since increase in sample size improves efficiency. In backpropagation networks the opposite seems to happen. The larger the sample (so the larger the class unbalance) the less robust the network. Indeed, large unbalanced training samples dramatically reduce the ability of the network to discriminate between the modal and rare class, to the extent that no conflict case is correctly predicted. Again, we believe this situation is caused by the error minimization process adopted by the backpropagation algorithm. With large samples, the weight parameters are mainly determined by the nonconflict class, since the large difference between the two classes significantly increases the proportion of changes in error in the weight value, Δw, calculated on the basis of the nonconflict class. In conclusion, it appears that in backpropagation neural networks a significant loss of efficiency is associated with increase in class unbalance (rare event bias), which large samples imply. This plays a bigger role in comparison to the parallel increase in efficiency that the larger sample size provides.

22. Schrodt (1991, 370), one of the first political scientists to suggest an alternative accuracy method to overall accuracy, stresses the inadequacy of overall accuracy as the measure for conflict prediction models. His solution is to use an entropy ratio (ER), which is equal to model entropy (ME) divided by the dependent variable entropy (DE).

23. Comprehensive reviews of the calculation involved in kappa analysis and the test Z statistics can be found in Goodman and Kruskal (1963) and Congalton and Green (1999, 43–57).

24. Kappa analysis was implemented by the software FUNCPOW.C, authored by Carlos Vieira of the School of Geography, University of Nottingham. The software was developed in standard C language and on the Unix platform.

25. The term *significance matrix* is relatively new. It has been adopted in the remote sensing literature by researchers dealing with classifiers' performance (Vieira and Mather 1999).

26. The difference in accuracy across training strategies becomes more evident when focusing on the other results. *Unb* cannot discriminate between the dispute and the nondispute class. Indeed, the unbalanced training set correctly predicts no dispute dyad, although it predicts 100 percent of nonconflict outcomes. However, as stressed by Beck, King, and Zeng (2000, 29), "This is not great success, of course, since the optimistic claim that conflict will never occur is correct 96 percent [in our case 97 percent] of the time." This result again shows how rare-event bias can preserve and even increase its negative influence in large heavily unbalanced samples. Also, the two balanced-replicated training strategies, *CVRep* and *Rep,* do not offer high accuracy on either class. Their accurate prediction on both dispute and nondispute dyads is less than 30 percent.

27. The denominator in (1) operates as a normalizing factor, which avoids the negative effect of the network activation function squeezing the weight value into a smaller range.

28. In order to increase variable comparability we also normalized the network's inputs so as to achieve means of zero for all input variables.

29. Here we can directly compare class accuracy since the testing sets have the same size.

30. Conditional kappa analysis and a pairwise test *Z* statistic were performed between the model with all the active inputs and those with one input switched off. The result indicates a significant deterioration, at least in one class accuracy, when *DependL, DemL, Logdistance, Logcaprat* are switched off.

31. Beck, King, and Zeng (2000, 29) say that their model, developed on the 1947–85 period, predicts 99.4 percent of the nondispute cases in 1986–89, but only 16.7 percent of disputes.