1. INTRODUCTION
The National Severe Storms Laboratory (NSSL) has developed numerous algorithms for the detection of various atmospheric phenomena. These algorithms reside collectively within an “umbrella” program called the Severe Storm Analysis Package (SSAP). Two such algorithms designed for the detection of tornados and/or damaging wind have recently been supplemented with neural networks (Marzban and Stumpf 1995, 1998) and their performance has been compared with conventional statistical methods in Marzban, Paik, Stumpf (1997). The NSSL has also developed a Hail Detection Algorithm (HDA) that computes a number of attributes believed to be relevant for the detection/prediction of severe hail (Witt et al. 1998a). As such, it is natural to develop a similar neural network complementing the HDA.

In this article, data produced by the National Severe Storms Laboratory’s Hail Detection Algorithm (HDA) will be employed to illustrate the importance of attaining a sufficiently deep local minimum in determining the optimal architecture of several neural networks developed for the prediction of maximum hail size. The performance of the networks is assessed in terms of both categorical and probabilistic measures.

2. DATA
The input variables provided to the neural network include a mix of Doppler-radar derived parameters along with several parameters representing the near-storm environment. The radar parameters include two based on reflectivity data, cell-based VIL (Johnson et al. 1998) and the severe hail index (SHI; Witt et al. 1998a), as well as two based on velocity data, storm-top divergence (Witt and Nelson 1991) and midaltitude rotational velocity (Witt 1998). The near-storm environment parameters include three based on thermodynamic data and two based on kinematic data (Table 1). The vertically-integrated wet-bulb temperature parameter is computed by integrating the wet-bulb temperature profile from the surface to the height of the wet-bulb zero. These near-storm environment parameters are either calculated within the SSAP using numerical model data, or they can be calculated from sounding data and manually entered into an adaptable parameter file. For this study, all the near-storm environment parameters were calculated from sounding data.

The verification data on hail size comes from Storm Data. Because Storm Data is a collection of severe weather reports, the minimum hail size in this study is 19 mm (0.75 inch). There are numerous problems associated with using Storm Data for verifying radar-based algorithm predictions (Witt et al. 1998a; Witt et al. 1998b). For predictions of maximum hail size, the primary concern involves the possibility that any given hail report is not representative of the maximum size being produced by the storm (at the time of the report). To minimize the impact of this possibility, the analysis was restricted to the maximum size observed per hailstorm (Witt 1998). For each severe hailstorm, a 20 minute time window (Witt et al. 1998a) was used to relate the predictor variables to the maximum reported hail size. For the radar-based parameters, the maximum value within the time window was used, whereas for the near-storm environment parameters, an average value was used.

Algorithm output was generated for 53 storm days from across the U.S. (Table 2). The distribution of reported hail sizes for these 53 days is shown in Fig. 1. The common practice of reporting hail size using familiar circular or spherical objects (e.g., various coins or balls) is clearly evident, as reports tend to be clustered along discrete sizes. The highest frequency corresponds to dime, nickel and quarter (coin) size hail (19 - 25 mm), golfball size hail (44 mm) and baseball size hail (70 mm). It would appear that few hail reports are actually measured to obtain a precise reading of their size, and that most hail sizes are estimated. Hence, one must assume that a certain amount of “rounding-off” error exists in the observations, and this error appears to increase as hail size increases. Be-
cause of all this, we have chosen to approach the prediction of maximum hail size from two distinct directions.

3. OVERFITTING

All nonlinear regression and classification models can overfit data; overfitting occurs when the flexibility/complexity of a model allows it to fit a data set or a decision boundary to such high accuracy that the fit is driven by the statistical fluctuations in the data. Consequently, the predictive capability of such a model is compromised. This phenomenon occurs when the model order greatly exceeds the sample size. To restrain overfitting, one traditional approach calls for splitting the data into several sets - a training set, employed to estimate the parameters of the model, a validation set for determining the complexity of the model, and a test set for estimating the unbiased performance of the model. The complexity (nonlinearity) of a neural network - a Multilayered Perceptron (MLP), in particular - is determined by two quantities: The number of hidden nodes, and the magnitude of the weights in the MLP, and the task of determining their optimal values falls into the realm of model-selection.

Within the split-sample method, resampling methods can be employed to determine the optimal number of hidden nodes. In fact, only two sets - a training and a validation set - are sufficient to that end. An example of such a resampling method for model selection is Bootstrapping (Efron and Tibshirani 1993). In its simplest form, one repeatedly trains with subsamples of the data; then, the average of the performance of the MLP on the unused subsamples is a measure of generalization performance. As such, the optimal network is the one with the lowest average (over the subsamples) validation error.

An implicit assumption in this method is that the network converges to a global minimum. In practice, however, not only is reaching a global minimum an unrealizable task, but also most learning algorithms can get trapped in local minima of the error function. Therefore, a sufficiently deep local minimum is often all that is attainable. The question then arises as to what precisely is a “sufficiently deep” local minimum. The (partial) answer lies in the range over which the local minima vary. Specifically, the distribution of the values of the error function at the local minima can aid in assessing the likelihood of obtaining a deeper local minimum than a given one. Thus, a local minimum that is highly unlikely to be won over by a deeper one is considered to be sufficiently deep.

In the remainder of the article, data produced by the HDA will be employed to illustrate the importance of attaining a sufficiently deep local minimum in determining the optimal number of hidden nodes in an MLP. Indeed, it will be shown that for the data at hand this procedure for model selection leads to an MLP whose optimal number of hidden nodes is zero. As such, the underlying function mapping the various attributes to hail size is linear. However, the high quality of the probabilistic forecasts produced by such an MLP justify its development, in spite of the linearity of the underlying function.

4. METHOD

The problem of predicting maximum hail size can be approached either from the point of view of regression or classification. This can be illustrated by examining the distribution of the hail size data set (Fig. 1). The number of distinct values (i.e., 14) is sufficiently large to warrant a regression analysis wherein the predictand is actual hail size. On the other hand, the existence of clusters in that distribution suggests three distinct classes for hail size, corresponding to “coin size”, “golfball size”, and “baseball size.” Both approaches are fruitful in that the former can provide precise estimates for hail size, and the latter can provide probabilities for a given class of hail size.

In the regression approach, an appropriate measure of error is the Mean Square Error (MSE),

<table>
<thead>
<tr>
<th>Region</th>
<th>No. days</th>
<th>No. hailstorms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western U.S.</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>High Plains</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Southern Plains</td>
<td>21</td>
<td>113</td>
</tr>
<tr>
<td>Midwest</td>
<td>6</td>
<td>31</td>
</tr>
<tr>
<td>S.E. U.S.</td>
<td>13</td>
<td>47</td>
</tr>
<tr>
<td>N.E. U.S.</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>253</td>
</tr>
</tbody>
</table>
defined as

\[ MSE = \frac{1}{N} \sum (t - y(x, \omega))^2, \tag{1} \]

where \( x \) is the vector of inputs, \( \omega \) is the vector of the weights, and \( t \) is the target value that is to be produced by the output \( y(x, \omega) \). The summation is over the number of cases, \( N \), in the relevant data set (training or validation). The activation function for the hidden nodes is taken to be the logistic function, \( f(x) = 1/(1 + \exp(-x)) \), while that of the output nodes is the linear function. The former introduces the necessary nonlinearity into the MLP, and the latter allows for the output node to take the full range of values taken by the target.

In the classification approach, the appropriate error function is cross-entropy, defined as:

\[ S = -\frac{1}{N} \sum [t \log y(x, \omega) + (1 - t) \log(1 - y(x, \omega))], \tag{2} \]

In the 2-class case, with a single output node representing class membership, if the activation functions are the logistic function, then the output can be interpreted as the posterior probability of class-membership, given the inputs (Bishop 1996). Such a probability is precisely what is required for forecasting purposes.

Given the inadequacies of hail size measurements discussed in Section 2, henceforth most of the emphasis will be on the classification networks. This emphasis is further justified since the classification networks produce probabilities - a feature that is indispensable in a forecasting context.

For the current application, the existence of three classes somewhat complicates the problem. It can be shown (Bishop 1996) that if the activation function is the so-called softmax function, then the minimization of cross-entropy does yield posterior probabilities, but only if 1-of-c coding is employed for coding the classes among the output nodes. This means that 3 output nodes are required to represent the 3 classes, one output node for each class. This is undesirable when dealing with small samples because the increase in the number of output nodes causes a corresponding increase in the number of parameters in the MLP, and this in turn, renders the network more likely to overfit the data.

To preclude this problem, for the classification analysis, the problem of predicting the 3 classes is decomposed into three 2-class problems. In other words, three MLPs are developed - one for discriminating between coin-size hail and otherwise, a second for discriminating between golfball-size hail and otherwise, and a third MLP for discriminating between baseball-size hail and otherwise. As such, each network requires only one output node, and a simple application of Bayes’ theorem shows that the output nodes represent posterior probabilities for the corresponding classes.

In both the regression and the classification schemes the number of input nodes is 9, namely the total number of hail attributes computed by HDA and SSAP. Experimentation with smaller number of input nodes suggests that nothing is gained by employing subsets of the nine attributes as inputs. Each attribute is centralized (i.e., converted to z-scores) by subtracting off the mean and dividing by the respective standard deviation. This can stabilize and even expedite the training procedure. As mentioned previously, the number of hidden nodes (on one hidden layer) was determined via bootstrapping. It was found unnecessary to go beyond one hidden layer.

The performance of the regression network must be assessed in terms of a different measure than that of the classification networks. The former is best assessed in terms of the scatterplot of the actual versus predicted hail size. The \( R^2 \) of the model is a measure of the goodness of the fit, and it reflects the amount of variance explained by the model. Residual plots (Draper and Smith 1981) display a wide range of qualitative features of the model, as well.

The performance of the classification networks can be assessed in terms of both categorical and probabilistic measures. There exist a large number of categorical measures each of which captures a different facet of performance. However, a (nonexhaustive) examination has suggested that almost all categorical measures have certain pathologies (Marzban 1998). One measure that is somewhat “healthy” is the so-called
Heidke Skill Statistic (HSS). It is defined as

$$HSS = \frac{H - E}{N - E},$$  \hspace{1cm} (3)

where $H$ is the number of hits, $N$ is the sample size, and $E$ is the expected number of hits based on climatology.

A commonly employed probabilistic measure is a skill score based on the Brier Score. Since the outputs of the networks are the forecast probabilities, the Brier Score is equal to MSE in equation (1). The corresponding skill score - which we shall call the Brier Skill Score (BSS) - is defined as

$$BSS = 1 - \frac{\text{Brier Score}}{p(1 - p)},$$  \hspace{1cm} (4)

where $p$ is the climatological probability of hail of a given size (Wilks 1995).

Due to its scalar nature, BSS fails to capture all the facets of probabilistic forecasts. Consequently, the quality of the probabilities will also be assessed within a probabilistic scheme (Murphy and Winkler, 1987, 1992; Wilks, 1995). One multidimensional measure of the quality of probabilistic forecasts is their reliability, expressed in reliability plots. There exist other facets that are neglected herein due to space limitations.

5. RESULTS

To better understand the relation between the individual attributes (predictors) and hail size, it is useful to compute the corresponding linear correlation coefficients, $r$. This can aid in anticipating the performance and/or the complexity of the MLP. The $r$ between each attribute and hail size is reported in Table 3. It can be seen that the four “radar attributes” (1-4) all have approximately equal correlation with hail size, while the “environmental attributes” (5-9) have linear correlations that are zero within the standard errors, with the possible exception of variable 8. Of course, the true underlying relationship may be nonlinear.

Furthermore, the $r$ between the various attributes themselves is important in ascertaining the amount of collinearity among the inputs. Identifying collinear inputs and their exclusion as inputs to the MLP can reduce the likelihood of overfitting the data. The most collinear pair is (5,6) with $r = 0.83$, followed by the pair (6,7) with $r = 0.68$. Neither of these $r$’s is sufficiently large to justify the exclusion of either member of either pair. As a result, all 9 attributes were employed as input nodes.

As mentioned previously, a number of MLPs were developed - a regression MLP designed to estimate actual hail size, and three classification MLPs for modeling the probability of the 3 hail-size classes. The optimal number of hidden nodes for all four networks was found via bootstrapping, as described in the Introduction. For a given number of hidden nodes, 100 local minima were visited, and the one with the lowest training error was considered to be the “global minimum.”

To illustrate the effect of the local minima on the optimal number of hidden nodes, only one of the classification networks (discriminating between coin-size hail and otherwise) will be considered; the general features found are common to all the networks. Fig. 2 (top) shows the values of the training and validation errors for the local minima of networks with 0, 2, 4, and 8 hidden nodes, for one bootstrap trial. Such “tv diagrams” display numerous useful features. For instance, it can be readily seen that the 100 training errors generally decrease with increasing number of hidden nodes. This is as expected, since the larger networks have more parameters for fitting the data. It can also be seen that the validation errors generally increase, indicating that the larger networks overfit the data. For the single bootstrap trial shown in Fig. 2 (top), the lowest validation error occurs for a network with 2 hidden nodes.

This result is, in fact, misleading. It is commonly assumed that the network with the lowest validation error is the optimal one. However, as seen from the tv diagram, the 2-hidden node network with the lowest validation error does not have the lowest training error, and is therefore trapped in a local minimum. The lowest training error of the 2-hidden node network is accompanied by a validation error that is higher than that of the 0-hidden node network. According to this tv diagram, the optimal number of hidden nodes is zero.

The distributions of the training and validation errors for the various networks are shown in Fig. 2 (middle). Evidently, all of the distributions are generally bell-shaped. This implies that it is unlikely to find a local minimum deeper than the deepest local minimum found. As such, the deepest
local minimum can be considered to be a global minimum.

According to the bootstrap approach, in order to assure that the results are not jeopardized due to sampling effects, one must repeat the entire training/validating process with different training and validation data. For the current analysis, 4 such trials were made and the results were averaged over the trials. Fig. 2 (bottom) shows the resulting averages and the standard errors due to sampling. This plot confirms the initial expectation that the optimal number of hidden nodes is zero, because it is that network which has the lowest average validation error.

Fig. 3 shows the scatter plot of predicted versus actual hail size for one of the bootstrap trials. The corresponding network has zero hidden nodes, for that was found to be the optimal number (via bootstrapping). The circles and the squares correspond to the training and validation data, respectively. The $R^2$ of the model is 0.398, and so it is capable of accounting for 40% of the variance in the data. Also shown is the diagonal line representing a perfect model. An examination of the residue-plot (not shown) indicates that the performance of the network deteriorates with increasing hail size. That, of course, is partially a consequence of the skewed nature of the distribution of hail size (Fig. 1).

Continuing with performance issues, that of the classification networks is assessed in terms of HSS, probability of detection (POD), the False Alarm Ratio (FAR), and BSS, and the quality of the probabilities is displayed in reliability plots. Fig. 4 shows the HSS, POD and FAR of the “baseball-size network” as a function of a threshold placed on its output node. It can be seen that a binary (warning/no-warning) forecast based on a probability threshold of approximately 50% yields the highest skill of HSS=65%. The HSS, POD and FAR of all three classification networks are shown in Table 4. The maximum HSS for the “golfball-size network” is obtained at a threshold of approximately 30%, and that of the “coin-size network” appears at about 45%. Naturally, since the networks are trained on the training data, the HSS values as obtained from the training data are positively biased. Given the sampling standard error of approximately 5% the three networks have approximately equal skill in terms of the validation HSS.

The BSS values for the three classification networks are also given in Table 4. The standard error due to sampling is approximately 5%. Therefore, as seen from the validation values, the “coin-size network” and the “baseball-size network” have comparable skill, but the “golfball-size network” has no significant skill in terms of BSS. This is easy to understand, for discriminating be-

Figure 2: The scatterplot (top), the distribution (middle), and the bootstrap-averaged (bottom) of validation vs. training errors for 100 local minima reached by networks of 0, 2, 4, and 8 hidden nodes. For clarity, the 8-hidden-node distributions are not shown in the middle figure.
between extreme events is a relatively simple task. In contrast, disambiguating intermediate events from the extreme ones is more difficult.

Finally, the quality of the outputs can be assessed in terms of reliability plots. These plots are shown in Fig. 5; they also display the 95% confidence intervals due to sampling. It can be seen that all three networks have "near perfect" reliability to within the error bars. Evidently, the "golfball-size network" does not produce probabilities exceeding 60%; this reflects the difficulty in predicting golfball-size hail with a high degree of confidence. It is interesting that the "baseball-size network" does produce high probabilities, but these high probability forecasts are relatively rare.

6. SUMMARY

Several neural networks (MLPs) have been developed for hail-size prediction. It is shown that their optimal architecture can be inferred via bootstrapping, but only if proper attention is given to local minima. It is found that the optimal networks are linear. Although the linearity of the problem might have obviated the development of MLPs, it is shown that the high quality of the produced probabilities justifies their development. The performance of the networks is also assessed in terms of scalar and probabilistic measures; the results suggest that predicting "coin-size" and "baseball-size" hail is relatively simpler than that of "golfball-size" hail.

7. ACKNOWLEDGEMENTS

The authors are grateful to V. Lakshmanan and Conrad Ziegler for reviewing the manuscript. This study was partially supported by the WSR-88D Operational Support Facility.

8. REFERENCES


---

Table 4: The HSS, POD, FAR, and BSS for the three classification networks.

<table>
<thead>
<tr>
<th></th>
<th>coin-size</th>
<th>golfball-size</th>
<th>baseball-size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSS</td>
<td>43</td>
<td>18</td>
<td>65</td>
</tr>
<tr>
<td>POD</td>
<td>73</td>
<td>62</td>
<td>75</td>
</tr>
<tr>
<td>FAR</td>
<td>27</td>
<td>62</td>
<td>24</td>
</tr>
<tr>
<td>BSS</td>
<td>27</td>
<td>6</td>
<td>46</td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSS</td>
<td>22</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>POD</td>
<td>78</td>
<td>50</td>
<td>33</td>
</tr>
<tr>
<td>FAR</td>
<td>34</td>
<td>68</td>
<td>62</td>
</tr>
<tr>
<td>BSS</td>
<td>10</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>
Figure 5: The reliability plots for the “coin-size network” (top), “golfball-size network” (middle), and the “baseball-size network” (bottom). Also shown are the 95% confidence intervals due to sampling.