CERN-PPE/92-98

9 June 1992

Identification of Tau Decays Using a Neural Network

V. Innocente¹⁾, Y.F. Wang²⁾ and Z.P. Zhang³⁾

- 1) CERN and INFN-Sezione di Napoli, Italy
- 2) INFN-Sezione di Firenze and University of Florence, Italy
- 3) World Laboratory, FBLJA Project, Geneva, Switzerland and
University of Science and Technology of China, Hefei, China

Abstract

A neural network is constructed to identify the decays $\tau \rightarrow \rho\nu$ in the L3 detector at LEP. The same network is used to identify $\tau \rightarrow \pi(K)\nu$ and $\tau \rightarrow e\nu\bar{\nu}$ as a cross check. High efficiency in the rho channel is obtained. A major effort has been made to study the systematic errors introduced by the use of a neural network and no obvious bias has been found.

(Submitted to Nucl. Instr. and Meth.)

1. Introduction

The tau particle, the heaviest known lepton, can decay both leptonically and hadronically via the exchange of a W boson. The exclusive identification of tau decays is important, both in understanding the charged current couplings by measuring the tau decay branching ratios, and for precision tests of the Standard Model by measuring the longitudinal polarization of taus in e^+e^- annihilations. The difficulties in identifying tau decays at the Z^0 resonance are mainly in the hadronic channels, where the opening angles among decay products ($\pi^\pm$ and π^0 's) are too small to be reconstructed precisely due to the large Lorentz boost of taus. It is very hard to distinguish among decays like $\tau \rightarrow \pi\nu, \tau \rightarrow \rho\nu \rightarrow \pi\pi^0\nu, \tau \rightarrow a_1\nu \rightarrow \pi\pi^0\pi^0\nu$ and $\tau \rightarrow \pi\pi^0\pi^0\pi^0\nu$.

Recently it has been reported that the Neural Network technique (NN) [1] is a good tool for identifying gluon jets from quark jets and/or for tagging b quarks [2]. In this paper we describe our attempts to identify tau decays with NN techniques using a full simulation of the detector. The results are then checked with data from the L3 experiment at LEP. A major effort has been made to study the systematic errors introduced by the use of a neural network.

2. The L3 Detector and the Event Sample

The L3 detector consists of a central tracking chamber, a high-resolution electromagnetic calorimeter composed of BGO crystals, a ring of scintillation counters, a uranium and brass hadron calorimeter with proportional wire chambers readout, and a precision muon chamber system. These detectors are installed in a 12 m diameter magnet which provides a uniform field of 0.5 T along the beam direction. A detailed description of the detector can be found in reference [3].

An event is reconstructed from tracks in the central tracking chamber TEC, bumps (local maxima of energy deposit) in the BGO calorimeter and clusters in the hadron calorimeter. Combinations of these measurements give the smallest resolvable "physical objects" (SRCs) which correspond to individual particles or to unresolved groups of particles. SRCs are then grouped into jets by an algorithm described in reference [4]. Each jet corresponds in general to a hadronic jet from a quark or a gluon in hadronic decays of Z^0 , and/or a lepton in leptonic decays of Z^0 .

A set of preselection cuts which are similar to those described in reference [5], is used to identify $Z^0 \rightarrow \tau^+\tau^-$ events. It gives an efficiency of 84.9% within a fiducial volume ($|\cos\theta| < 0.7$) with 1.8% background. A total of 10411 events from 1990 and 1991 runs are selected. An additional cut is then applied to select only the one prong decay: each jet should have one and only one good track¹ with momentum greater than 0.3 GeV/c.

¹A good track means: Number of Hits > 30 out of 62; length in the R/ϕ plane > 216 mm; distance of closest approach to average beam position < 10 mm. For the rho channel, two additional cuts are applied in order to have good momentum resolution: χ^2 probability of the circle fit > 0.001; tracks away from wire planes > 10 mrad.

3. The Network Structure

A feed-forward neural network [1] is constructed with one input layer, one hidden layer and one output layer.

The output layer has only one node which is designed to have an output value of 1 for the exclusive tau decay channel assumed to be the signal and a value of 0 for all other decays which constitute the background.

The hidden layer consists of 18 neurons. However our final result is not sensitive to their exact number.

The input layer contains 18 neurons, each of them corresponds to a physical variable of a tau, namely a jet, measured by the detector. The following variables are chosen:

- 1) E_{BGO} . The energy deposited in the BGO electromagnetic calorimeter.
- 2) E_{had} . The energy deposited in the hadron calorimeter.
- 3) P_{TEC} . The momentum of the track measured in the central tracking chamber.
- 4) Σ_{BGO}^2 . The energy sum of the two most energetic BGO crystals in the most energetic bump.
- 5-6) Σ_{BGO}^9 (Σ_{BGO}^{25}). The energy sum of 9 (25) crystals surrounding the most energetic crystal in the most energetic bump.
- 7-8) $E_{src}^{1^\circ}$ ($E_{src}^{2^\circ}$). The energy of the most (second most) energetic SRC in the calorimeters.
- 9-10) $E_b^{1^\circ}$ ($E_b^{2^\circ}$). The energy of the most (second most) energetic bump in the BGO calorimeter.
- 11-12) $E_{nb}^{1^\circ}$ ($E_{nb}^{2^\circ}$). The energy of the most (second most) energetic neutral bump in the BGO calorimeter, where neutral means that there is no track pointing to this bump within 15 mrad in the transverse plane.
- 13-14) W_θ (W_ϕ). The width of the most energetic bump in the direction along (transverse to) the beam.
- 15) N_{nb} . The number of neutral bumps in the jet.
- 16) N_{src} . The number of SRCs in the jet.
- 17) R_a . The ratio between the two axes of the lateral profile of the shower in the BGO.
- 18) Δ_{tb} . The ϕ -angle difference between the track in the TEC and the center-of-gravity of the most energetic bump in the BGO.

Fig. 1 shows some of these variables for different tau decay channels from Monte Carlo simulation. All these variables are then normalized to be within the range $[0, 1]$.

4. The Network Training

The network is trained by Monte Carlo $e^+e^- \rightarrow \tau^+\tau^-$ events which are generated by the program KORALZ 3.18 [6]. It takes into account the radiative corrections including tau decays. The detector simulation is based on GEANT 3.14 [7] which uses GHEISHA [8] for hadronic interactions.

We used 40K Monte Carlo $e^+e^- \rightarrow \tau^+\tau^-$ events with full simulation of the L3 detector for training the network using the back-propagation procedure as described in reference [1], where the “temperature” T is set to be 1, the “momentum” α is chosen to be 0.7 and the learning strength parameter η decreases from 0.05 to 0.001 during the training process. The initial values of weight W and threshold C are set randomly between $[-0.1, 0.1]$. During each loop, one signal event and one background event are randomly chosen from the training sample and fed into the network. The weights and thresholds are adjusted for each loop and updated for every five loops. The training procedure uses typically three million loops and spends about one hour on an Apollo DN10000 CPU.

We have applied this method to identify ρ 's from tau decay which is the most sensitive channel to the tau polarization. By taking $\tau \rightarrow \rho\nu$ as the signal and all others as background, we obtain the NN output (y) distribution as shown in Fig. 2a. The $\tau \rightarrow \rho\nu$ candidates are then identified as those which have a NN output value $y > Y_{cut}$ where the cutoff Y_{cut} is determined by optimising both the selection efficiency and the purity of the sample. The selection efficiency is defined as the number of selected signal events divided by the total number of signal events within the fiducial volume. The purity (background) is defined as the number of selected signal (background) events divided by the total number of selected candidates with $y > Y_{cut}$. Fig. 2b shows the selection efficiency and the purity as a function of the cutoff Y_{cut} .

The same procedure is repeated for $\tau \rightarrow \pi(K)\nu$ and $\tau \rightarrow e\nu\bar{\nu}$ decays with different definition of signals and backgrounds. As our intention is to make a cross check, we use the same network as for ρ 's so that the input variables are not optimized and consequently the efficiencies and backgrounds are not the best one can obtain. The corresponding NN outputs and efficiencies are shown in fig. 3 and fig. 4. Table 1 shows the typical selection efficiency and background contamination for various tau decay channels.

Decay Channel	Cutoff Y_{cut}	Efficiency(%)	Background (%)
$\tau \rightarrow \rho\nu$	0.8	55.2	16.1
$\tau \rightarrow \pi(K)\nu$	0.9	62.5	14.9
$\tau \rightarrow e\nu\bar{\nu}$	0.9	75.5	1.9

Table 1. Typical selection efficiency and background contamination from training for different tau decay channels.

5. The Test of Trained Results

During the training procedure, there are several parameters which are tuned by hand. For example, the number of training loops is determined such that the error function does not decrease further with more training loops. It is not proved that we arrived at the absolute minimum; there is no way to prevent the training from finding local minima. The same problem exists with a simple cut method: one can not prove that a set of cuts is the best one absolutely. However the efficiency and purity obtained are adequate for the analysis. Once the network is trained, it becomes a fixed “function box” which gives a certain answer for a given input. For each input event, its output is related to the probability that such an event is a signal (or a background) event. The performance of the network can be understood by comparing its output value with the desired value, using Monte Carlo events and/or known data events. This also provides tests against systematic bias.

We have done two kind of checks described in the following to test the NN. The self-consistency check uses Monte Carlo events to test the performance of NN and compares the agreement with data. The checks on weights provide further understanding of the behaviour of the network.

5.1 The Self-Consistency Check

We first test the network using 140K M.C. $\tau^+\tau^-$ events which are statistically independent from those we used for training. By passing them through the network, we get the selection efficiency for each tau decay channel and the corresponding background contamination as shown in Table 2. The selected number of events for each decay channel from real data are also listed in table 2. Comparing table 1 and table 2, one can see how the efficiencies obtained from the training sample are slightly higher than those from the test sample. Also the estimation for the background contamination from the training is slightly lower than those from the test. This is due to the limited statistics of M.C. events used for training. During the learning procedure the network improperly takes into account some statistical differences between signal and background. This is a common problem to any discriminant analysis and just confirms that the training sample should only be used to compute the network parameters while the performance of the network should be evaluated using a different test sample.

Decay Channel	Cutoff y_{cut}	Efficiency (%)	Background (%)	Data Events
$\tau \rightarrow \rho\nu$	0.8	53.8	17.2	3533
$\tau \rightarrow \pi(K)\nu$	0.9	62.9	15.6	1941
$\tau \rightarrow e\nu\nu$	0.9	76.4	1.8	3274

Table 2. Typical selection efficiency and background contamination from test for different tau decay channels. The selected number of events for each decay channel from real data are also listed.

We have checked that there are no biases on the input variables and on the output

values by comparing data and Monte Carlo. Fig. 5a shows some of input variables for all taus before NN selection, i.e., the whole input of network. Fig. 5b-5d shows all of them after NN selection for rhos as signals (with an output value $y > 0.8$), for backgrounds ($y < 0.1$) and for intermediate region $0.1 < y < 0.8$. A good agreement between data and Monte Carlo can be seen. More importantly, fig. 6 shows the distribution of the output value for all taus. Again the data and Monte Carlo agrees well.

For the selected rho candidates we can construct the invariant mass of $\pi^0\pi^\pm$ system which is shown in figure 12. It peaks around $800 MeV/c^2$. This is compatible with the mass of rho $M_\rho = 768.3 \pm 0.5 MeV/c^2$ as in reference [9]. The width is dominated by the charged track momentum resolution.

We performed another test to understand how the network output is sensitive to any small bias on the input data, i.e., to small discrepancies between data and Monte Carlo for certain input variables. Keeping the network parameters fixed, using the test sample as the reference, we scaled one of the input variables by 10% for each event of the test sample and then compared the resulting efficiency and purity with the reference values. Fig. 7 and 8 show the deviation of efficiency and background by changing all 18 input variables one by one and all together. It can be seen from the plot that almost all changes are within 5%. We then shifted each variables by 0.05 (remember that all variables are normalized to be within $[0, 1]$). The variations of efficiency and background are also within 5% as shown in fig. 7 and 8. We also smeared the input variables using a gaussian function with a sigma of 10% of its value. The deviation of efficiency and purity are smaller than those from scaling, i.e., within 1% as shown in fig. 7 and 8. In the previous test we did not consider the correlations/anticorrelations between input variables. To take these effects into account, we varied some of known correlated/anticorrelated variables by $\pm 10\%/\mp 10\%$ together. For example, we increased all energies related to the BGO calorimeter by 10%, or we increased energies measured in BGO by 10% and decreased energies measured in hadron calorimeter by 10%; we got the efficiencies and purities as shown in fig. 9. It can be seen from the plot that their changes are within 3%. It should be noticed that 10% scaling/smeared and 0.05 shifting is much larger than realistic case, especially for quantities measured by the BGO calorimeter which has an energy resolution of about 1%. If we take a reasonable amount to scale/shift/smeared, the total efficiency change (quadratic sum of scale/shift/smeared) is only 1.3% and background 0.7%. All of these tests demonstrate that any small bias of the input variables would have little effect on the output of the network. This is in fact one of the advantages of NN which takes into account the input pattern in a global fashion and hence it is not sensitive to small biases of input variables.

Since the efficiency and the purity of the selection of $\tau \rightarrow \rho\nu$ is dependent on the branching ratios of tau decays, we varied the branching ratios of tau decays in the Monte Carlo simulation by twice the errors quoted by the particle data group [9]. Fig. 10 shows that efficiencies and purities change less than 3% after these variations.

In order to understand the stability of the network, we chose a subset of input nodes to form a new network with the same number of hidden nodes. Table 3 lists four such kinds of networks. Using the same training and testing procedures as before, we obtain

the efficiency and purity curves as shown in Fig. 11 compared with the reference. It can be seen that the efficiency and purity decrease slowly as input nodes are removed. With only one third of the input nodes, we can still maintain 90% of the original efficiency. This proves that the network is quite stable. On the other hand, if we look at the common events selected by these networks both from data and Monte Carlo, we can obtain a matrix as shown in table 4. The diagonal elements are the total number of events selected by the corresponding network normalized to the number of events selected by network 18 (reference). The other elements give the common events selected by the two networks denoted in the corresponding column and line normalized as before. The lowest line denoted as “.and.” gives the common events of several networks. For example, in column “network 13” it gives the common events of networks 18, 17 and 13; in column “network 6” it gives the common events of all networks. From this table we can make the following observations: (1) The efficiency decreases slowly with a reduction of the number of input nodes. (2) The common events of any two networks are quite high (75% – 85%) and also the common events of all networks is nearly 70% which is quite satisfactory. (3) The data and Monte Carlo agree generally well in spite of a small (< 3%) systematic discrepancy. This gives the order of systematic errors on the Monte Carlo prediction and suggests that Monte Carlo takes into account correlations to within 3%.

Network	Input Variables
18	$E_{BGO}, E_{had}, P_{TEC}, \Sigma_{BGO}^2, \Sigma_{BGO}^9, \Sigma_{BGO}^{25}, \Delta_{tb}, E_{nb}^{1^\circ}, E_{nb}^{2^\circ}, E_{src}^{1^\circ}, E_{src}^{2^\circ}, E_b^{1^\circ}, E_b^{2^\circ}, W_\theta, W_\phi, N_{src}, R_a, N_{nb}$
17	$E_{BGO}, E_{had}, P_{TEC}, \Sigma_{BGO}^2, \Sigma_{BGO}^9, \Sigma_{BGO}^{25}, \Delta_{tb}, E_{nb}^{1^\circ}, E_{nb}^{2^\circ}, E_{src}^{1^\circ}, E_{src}^{2^\circ}, E_b^{1^\circ}, E_b^{2^\circ}, W_\theta, W_\phi, N_{src}, R_a$
13	$E_{BGO}, E_{had}, P_{TEC}, \Sigma_{BGO}^2, \Sigma_{BGO}^9, \Sigma_{BGO}^{25}, \Delta_{tb}, E_{nb}^{1^\circ}, E_{nb}^{2^\circ}, E_{src}^{1^\circ}, E_{src}^{2^\circ}, E_b^{1^\circ}, E_b^{2^\circ}$
9	$E_{BGO}, E_{had}, P_{TEC}, \Sigma_{BGO}^2, \Sigma_{BGO}^9, \Sigma_{BGO}^{25}, \Delta_{tb}, E_{nb}^{1^\circ}, E_{nb}^{2^\circ}$
6	$E_{BGO}, E_{had}, P_{TEC}, \Sigma_{BGO}^2, \Sigma_{BGO}^{25}, \Delta_{tb}$

Table 3. Networks constructed by subsets of standard input variables. The reference network denoted as “18” is also listed for comparison.

Input Nodes	18		17		13		9		6	
	Data	M.C.	Data	M.C.	Data	M.C.	Data	M.C.	Data	M.C.
18	1	1								
17	0.86	0.88	0.95	0.97						
13	0.83	0.86	0.81	0.84	0.93	0.94				
9	0.81	0.84	0.80	0.83	0.83	0.84	0.92	0.93		
6	0.77	0.79	0.76	0.77	0.76	0.78	0.78	0.79	0.89	0.89
.and.	1	1	0.86	0.88	0.77	0.80	0.72	0.75	0.66	0.69

Table 4. Percentage of common events for networks constructed by subsets of standard input variables.

5.2 Checks on Weights

Since the network is constructed by a set of weights, it is natural to ask what the weights look like. Although it is almost impossible to understand the exact meanings of them because the input variables have different units and their weighted sums have no physical meanings, the order of magnitude of the weights shows their relative importance. Fig. 13a shows the weights between hidden nodes and output node. Since all nodes in the hidden layer should have nearly the same role, the corresponding weights should be more or less the same. The “total weights” between input nodes and output are evaluated by summing the input-hidden weights $\times$ hidden-output weights over all hidden nodes. Fig. 13b shows these calculated weights which reflect roughly the relative importance of input variables. It can be seen that some variables play a more important role but all of them have weights of the same order of magnitude. This shows that all variables are used for producing the output and there are no unusual weights which dominate the selection based on some unphysical features of the input.

As a further test on network stability, we add one input node with random values between $[0, 1]$. It is found that the network remains stable as we expect. The efficiency and purity change by less than 1% as shown in fig. 14a and the corresponding weight for this random node is the smallest one as shown in fig. 14b.

6. Discussion

Although we show that data and Monte Carlo agree for each input variable and output, we have not fully proven that the correlations among input variables are correctly simulated in Monte Carlo and hence produce no bias in the result. This is not only a problem for NN selection, but also for any kind of selection based on Monte Carlo simulation. For any function constructed from measurable quantities, the verification of its performance is obtained by comparing its input (measurable quantities) and output (function value) between data and Monte Carlo. If both agree well, we can get a confidence in the function. Furthermore, if one can check the correlations between input and output in detail, say for different tau decays or different bins of output value, or on an event by event basis, one can increase the level of confidence. Both tests have been done and are quite satisfactory as shown above. On the other hand, it is very difficult to imagine that the distributions of input variables are in good agreement between data and Monte Carlo but their correlation differs significantly. This would mean that the Monte Carlo simulation was wrong on an event by event basis. Even if this happens, the function value for each event would be different and hence the output distribution would be different. Therefore we believe that there is no unknown correlation between input variables which could bias our results significantly.

Another important issue is how to estimate the systematic errors on the NN selection. The conventional way for simple cut selection is to make changes in the cuts and to see their influence on the final results. Since there are no cuts in the NN, we can do it in a contrary way: instead of shifting the cuts, we can shift and/or smear and/or

rescale the data or MC distribution, to see their influences on the physics results such as polarization or branching ratios. Using this method we can get a reasonable estimation on the systematic errors due to the NN selection. It has been shown in the previous section that this error is reasonably small.

7. Conclusions

A feed-forward perceptron type neural network with one input layer consisting of 18 nodes, one hidden layer consisting of 18 nodes and one output layer with only one node is used to identify tau decays with the L3 detector at LEP. A selection efficiency of 54% with a purity of 83% is obtained for decays $\tau \rightarrow \rho\nu$. Using the same network we obtain an efficiency of 63% with a purity of 84% for decays $\tau \rightarrow \pi(K)\nu$ and a efficiency of 76% with a purity of 98% for decays $\tau \rightarrow e\nu\bar{\nu}$.

The performance of the network is checked using Monte Carlo and real data events. It is found that the result of the network is predictable as conventional method and its systematic errors can be well controlled.

Acknowledgments

The authors would like to thank all their colleagues of the L3 collaboration for the continuous support, stimulating discussions, constructive criticism and valuable suggestions, especially H.S. Chen, R. Mount, M. Pohl and B. Monteleoni.

FIGURE CAPTIONS

- Figure 1. The NN input variables for some tau decay channels. The distributions are arbitrarily normalized in order to be seen more clearly.
- Figure 2. The NN trained results for $\tau \rightarrow \rho\nu$ decays. (a) The NN output value for signals and backgrounds; (b) the selection efficiency and purity as a function of y_{cut} .
- Figure 3. The NN trained results for $\tau \rightarrow \pi(K)\nu$ decays. (a) The NN output value for signals and backgrounds; (b) the selection efficiency and purity as a function of y_{cut} .
- Figure 4. The NN trained results for $\tau \rightarrow e\nu\nu$ decays. (a) The NN output value for signals and backgrounds; (b) the selection efficiency and purity as a function of y_{cut} .
- Figure 5. Some NN input variables for all tau decays from data and Monte Carlo. The histograms are Monte Carlo events from all tau decays; the dots are data events and the shadowed area are $\tau \rightarrow \rho\nu$ from Monte Carlo simulation. (a) all events; (b) events with NN output value for ρ selection $y > 0.8$; (c) events with NN output value for ρ selection $y < 0.1$; (d) events with NN output value for ρ selection $0.1 < y < 0.8$.
- Figure 6. The NN output for data and MC.
- Figure 7. The deviations of efficiency of modified samples from reference sample.
- Figure 8. The deviations of purity of modified samples from reference sample.
- Figure 9. The efficiencies and purities from modified samples whose correlations (anti-correlations) are taken into account.
- Figure 10. The efficiencies and purities after modifying the branching ratios of tau decays in the Monte Carlo simulation.
- Figure 11. The efficiencies and purities from networks constructed by subset of input variables compared with the reference “network 18”. The efficiencies and purities decrease with the reduction of the number of input nodes.
- Figure 12. The invariant mass of $\pi^0\pi^\pm$ system of selected rho candidates.
- Figure 13. (a) The weights between hidden and output layer for each hidden nodes; (b) the “total weights” between input and output for each input variables.
- Figure 14. (a) The efficiency and purity curve for the network with a random input node compared with the reference ones; (b) the “total weights” between input and output for each input variable with a random input node.

REFERENCES

1. For an introduction, see P. Wasserman, "Neural computing: Theory and Practice", Van Nostrand Reinhold, N.Y. (1989).
2. C. Bortolotto *et al.*, "Neural Networks in Experimental High Energy Physics", Udine Report 91/09/CB-Rev. and references therein, to be published in International Journal of Modern Physics, C.
3. L3 Collaboration, B. Adeva *et al.*, Nucl. Instr. and Meth. A 289 (1990)35.
4. O. Adriani *et al.*, Nucl. Instr. and Meth. A 302 (1991) 53.
5. L3 Collaboration, B. Adeva *et al.*, Z. Phys. C 51 (1991) 179.
6. S. Jadach *et al.*, "Z Physics at LEP I", CERN Report CERN-89-08, Vol. III, (1989) 69.
7. R. Brun *et al.*, "GEANT 3", CERN DD/EE/84-1 (Revised), September 1987.
8. H. Fesefeldt, RWTH Aachen PITHA 85/02 (1985).
9. Particle data group, Phys. Lett. B 239 (1990).

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5a

Figure 5b

$Y < 0.1$

Figure 5c

$$0.1 < Y < 0.8$$

Figure 5d

Figure 6

Figure 7

Figure 8

Figure 9

Figure 10

Figure 11

Figure 12

Figure 13

Figure 14