ALEPH 93-125
PHYSIC 93-106
G. Stimpfl et P. Yepes
28.7.1993

Higgs Search and Neural-Net Analysis

Georg Stimpfl-Abele[(a)] and **Pablo Yepes**[(b)]

(a) Laboratoire de Physique Corpusculaire, Université Blaise Pascal
Clermont-Ferrand, F-63177 Aubiere, France
e-mail address: stimpfl@cernvm.cern.ch

(b) Rice University, Houston, TX 77251-1892, USA
e-mail address: yepes@physics.rice.edu

May 30th, 1993

Abstract

Two aspects of neural-net analysis are addressed: the application of neural nets to physics analysis and the analysis of neural nets. Feed-forward nets with error back-propagation are applied to the search for the Standard Higgs Boson at LEP 200. New methods to select the most efficient variables in such a classification task and to analyse the nets are presented. The sensitivity of the nets for systematic effects is studied extensively. The efficiencies of the neural nets are found to be significantly better than those of standard methods.

(Submitted to Computer Physics Communications)

1 Introduction

The search for new elementary particles is among the most important tasks in High Energy Physics. In general events containing new particles are produced along with a much larger number of conventional events. Hence an analysis looking for new phenomena needs a filtering process intended to separate signal and background events.

In this study a traditional filtering method, using standard one-dimensional cuts, is compared with a Neural Net (NN) approach in the search for the Higgs boson at LEP 200. The following two mass hypotheses are chosen: 70 and 90 GeV/c^2. The lower mass represents the easier case because the signal is higher and the backgrounds can be better discriminated. The higher mass is rather challenging since it is just below the Z mass.

Standard feed-forward nets with one hidden layer and error back-propagation are used. Emphasis is given to the selection of the best input-variables by analyzing their utility inside the net. Systematic effects are studied in detail.

The physics case is discussed in section 2, followed by a description of the generation and the preselection of the input data. Section 4 is dedicated to the standard analysis based on one-dimensional cuts. Section 5 contains the technical details of the net generation, like architecture and learning procedure, and a description of the methods developed to analyse neural nets and to select the best input-variables. The performance of the NNs in the Higgs search is demonstrated in section 6. Systematic effects are studied extensively in section 7 in order to test the reliability of the methods. Finally conclusions are given.

2 Higgs production and backgrounds at LEP 200

The Standard Model of Particle Physics [1] is the commonly accepted theory to explain the interactions among elementary particles. This model predicts the existence of the Higgs boson, H, responsible of the so-called Symmetry Breaking mechanism [2]. During the last years the Large Electron Positron collider (LEP) at CERN, operating with a center-of-mass energy (E_{cm}) around 91 GeV (LEP 100), has performed a very intensive Higgs search in the mass range $0 < m_H < 60$ GeV/c^2 [3]. Unfortunately at present no evidence of the Higgs boson has been found. However the second phase of LEP (LEP 200), running at $E_{cm} = 170 - 200$ GeV, will extend the search up to Higgs masses around 90 GeV/c^2. In this study a center-of-mass energy of 190 GeV is assumed.

At LEP 200 the neutral Higgs boson could be produced through the reaction:

$$e^+e^- \to H\ Z\ .$$

Its cross section at $E_{cm} = 190$ GeV is 0.82 pb for a 70 GeV/c^2 Higgs (H_{70}) and 0.36 pb for a Higgs with a mass of 90 GeV/c^2 (H_{90}) [3]. A heavy Higgs ($m_H > 15$ GeV/c^2), as in the case considered here, decays predominantly into a quark-antiquark ($q\bar{q}$) pair. The Z decays into hadrons ($\approx 70\%$), charged leptons ($\approx 10\%$) or into neutrinos ($\approx 20\%$). In this analysis only the most abundant decay channel will be considered, that is:

$$e^+e^- \to H\ Z\ , \quad H \to q\bar{q}\ , \quad Z \to q\bar{q}\ .$$

The physical backgrounds for this reaction along with their production cross sections are given in Table 1 [4].

3 Simulation and preselection

A fast Monte Carlo program of a LEP detector was implemented. This allows to produce very high statistics event-samples for signal and background processes. The simulation includes realistic numbers for momentum resolution and reconstruction efficiencies of charged tracks. The detector response for neutral hadronic and electromagnetic particles is also implemented. In addition, the detector ability to identify electrons and muons is simulated, as well as the capability to detect secondary vertices from heavy flavor decays. The parameters are tuned to achieve Higgs detection efficiencies and background rejection with standard cuts similar to those obtained in previous studies [4, 5].

Reaction	Cross Section [pb]
Signal:	
$e^+e^- \rightarrow H_{70}\ Z$	0.82
$e^+e^- \rightarrow H_{90}\ Z$	0.36
Backgrounds:	
$e^+e^- \rightarrow Z\ Z$	0.6
$e^+e^- \rightarrow W^+W^-$	18.0
$e^+e^- \rightarrow q\bar{q}gg$	90.0

Table 1: The cross sections for Higgs production at 70 and 90 GeV/c^2 and its main backgrounds in the hadronic channel at E_{cm} = 190 GeV.

As already mentioned this work is focused on the hadronic channel, that is when the Higgs and the accompanying Z decay each into a quark-antiquark pair. Subsequently each of those quarks will generate a jet. In this analysis jets are reconstructed using the LUCLUS algorithm [6]. Then only events with 4 or more jets are accepted. In order to avoid contamination from channels with leptons in their final state, events are required not to have electrons nor muons with more than 20 GeV energy. In addition events with energetic (> 20 GeV) and isolated particles are not considered for further analysis. A particle is considered isolated when its energy is more than 90% of the energy of the jet it belongs to. Finally only events with a fitted Higgs mass (m_H^f) in a ±5 GeV/c^2 window around the searched Higgs mass are selected (the fit algorithm is explained in the next section). The number of events for the different samples left after the preliminary cuts is listed in Table 2, normalized to an integrated luminosity of 1000 pb^{-1}. All results presented in this paper are based on this luminosity. It is higher than what is expected but only one Higgs channel is considered and the results can easily be scaled to other luminosities.

Reaction	H_{70}	H_{90}
$e^+e^- \rightarrow H_{70}\ Z$	227	-
$e^+e^- \rightarrow H_{90}\ Z$	-	115
$e^+e^- \rightarrow Z\ Z$	28	135
$e^+e^- \rightarrow W^+W^-$	2774	844
$e^+e^- \rightarrow q\bar{q}gg$	2746	1010

Table 2: Number of events left for m_H = 70 and 90 GeV/c^2 after the preselection for an integrated luminosity of 1000 pb^{-1}.

Since the number of events for a process is given by the cross section times the integrated luminosity the total number of events for each background (ZZ, WW, $q\bar{q}$) is 600, 18000 and 90000, respectively (see Table 1). Many more events are used for the analyses in order to lower statistical fluctuations. The results obtained in this study are based on a production of about 80k ZZ events, 240k WW events, 800k $q\bar{q}$ events, and 50k signal events.

4 Higgs search with standard cuts

The standard analysis is tuned for both Higgs masses (70 and 90 GeV/c^2). The selection criteria used are very similar but not identical as it is explained in the following.

The background with the highest cross section is $e^+e^- \rightarrow q\bar{q}gg$. At this energy the quark pair is often produced with a radiated energetic gamma. This kind of events can be rejected by demanding no photon with more than 30 GeV energy (E^{γ}_{max}). The second cut against this background is based on the so-called thrust (T). This quantity reflects the event isotropy: T $\approx$ 0.5 means an isotropic event, while T $\approx$ 1 indicates a narrow back-to-back event. Higgs candidates are required to have a thrust smaller than 0.9. Some of the background events left are found to have a large multiplicity. They are rejected by applying the cut $N_{trk} < 40$ (47) for H_{70} (H_{90}).

Additional cuts are based on secondary vertices from particles containing the b quark. The Higgs boson decays predominantly into b quarks while backgrounds tend to disintegrate into all quarks. In order to take advantage of this characteristic the number of secondary particles (N_{off}) is calculated. A particle is considered to be a secondary, if the distance between its origin and the main vertex is larger than three times the spatial resolution of the detector. The total invariant mass of all secondaries (m_{off}) is also calculated. A large fraction of the background is then rejected by demanding $N_{off} > 4$ and $m_{off} > 10$ GeV/c^2.

Since a topology with four jets is searched for, all events are forced into four jets. Then a mass fit procedure [7] is applied with two hypotheses: HZ, with the Higgs mass as a free parameter, and WW. From the first hypothesis the fit quality, χ^2_{HZ}, and m^f_H are obtained. While the second only provides the fit quality χ^2_{WW}. The WW background is further suppressed by demanding $\chi^2_{WW} - \chi^2_{HZ} < 0.5$ (3.0) for the H_{70} (H_{90}) analysis.

When the event is fitted with the Higgs hypothesis, two of the four jets are assumed to come from the Higgs decay. The number of secondaries, N^H_{off}, should be high if those jets come from a Higgs decay, as explained above. Therefore the requirement $N^H_{off} > 5$ is applied to further reduce the background.

The number of events left for the signal and the different background sources are shown in Table 3 for the two Higgs masses considered.

The efficiency of a given analysis can be measured by the statistical significance defined as: $N_\sigma = N_s/\sqrt{N_b}$, where N_s (N_b) is the number of expected signal (background) events. The minimum luminosity for discovery is defined as the luminosity for which the statistical significance is five. These two quantities are also shown in Table 3.

Reaction	H_{70}	H_{90}
$e^+e^- \to H_{70}$ Z	60.5	–
$e^+e^- \to H_{90}$ Z	–	38.0
$e^+e^- \to q\bar{q}gg$	22.3	11.9
$e^+e^- \to W^+W^-$	25.0	11.8
$e^+e^- \to$ Z Z	1.4	11.4
Total background	48.7	35.1
Statistical significance	8.7	6.4
Min. luminosity [pb^{-1}]	330.	610.

Table 3: Signal and background events left after applying the standard cuts, statistical significance, and minimum luminosity for both masses. The statistical significance is the signal divided by the square root of the background. The minimum luminosity is defined as the luminosity for which the statistical significance is 5.

5 Generation and analysis of NNs

5.1 Network layout

Throughout the whole study simple nets with one hidden layer and one output neuron are used. The neurons have a sigmoid activation function. The inputs to the nets are the values of the selected variables normalized to the interval [0,1]. The number of hidden neurons (N_{hid}) is about half the number of input neurons (N_{in}) in case of more than 15 input neurons and $N_{in} - 1$ else. A cut is applied to the value of the output neuron: events below the cut are regarded as background, the other ones as signal. The cut is chosen such that the statistical significance is maximal.

5.2 Learning procedure

The nets are trained with a mixture of events from the three different background samples and from the signal. The output values 0 and 1 are demanded for background and for signal events, respectively. The weights are adjusted with the well-known error back-propagation algorithm. A description of such NN training can be found in [8].

The learning data are selected applying the preliminary cuts described in section 3 except from the cut on the fitted mass m_H^f. The elimination of this cut does not diminish the learning capability and it allows to train mass-independent nets. Like in a previous NN study [8] the best performance is found using twice as many background as signal events. As shown in Table 2 almost the same number of $q\bar{q}$ and WW events is left after the preselection, whereas the number of ZZ events left is much smaller. In order to train all three background classes reasonably well the backgrounds $q\bar{q}$, WW, and ZZ are mixed in the ratio 4 : 4 : 1 .

Reaction	Number of events
$e^+e^- \to H_{70}\ Z$	4500
$e^+e^- \to H_{80}\ Z$	4500
$e^+e^- \to H_{90}\ Z$	4500
Total signal	13500
$e^+e^- \to q\bar{q}gg$	12000
$e^+e^- \to W^+W^-$	12000
$e^+e^- \to Z\ Z$	3000
Total background	27000

Table 4: Number of events from each subgroup used for learning.

Like in the case of kink recognition [9] the performance of the nets is slightly better if they are trained with a mixture of masses (or energies in the case of the kinks) and not for each mass separately. One single net can be trained for the whole mass-range between 70 and 90 GeV/c^2 by mixing signal events of different masses in the same proportions. Very satisfactory performance is achieved with signal events from m_H = 70, 80 and 90 GeV/c^2. In agreement with other NN applications in High Energy Physics a few 10^4 events are needed for optimal training. About 4 10^4 events are necessary here (see Table 4).

Table 5 contains the statistical significance for nets with 7 and 10 input variables trained for each mass separately (sep) or with a mixture of signal events from m_H = 70, 80 and 90 GeV/c^2 (mix).

N_{in}	Learning	H_{70}	H_{90}
7	sep	12.3	7.9
7	mix	12.6	8.0
10	sep	12.6	8.2
10	mix	14.2	8.4

Table 5: Statistical significance at m_H = 70 and 90 GeV/c^2 for nets with 7 and 10 input variables trained with signal events from the same mass (sep) or with a mixture of signal events from different masses (mix).

A comparison of the performance between the mass-independent net (mix) and the mass-dependent nets (sep) at m_H = 75, 80 and 85 GeV/c^2 shows similar results. Hence only the mass-independent nets are used in the following. This is a big advantage over the standard method where the cuts have to be adjusted to each mass for which the search is carried out.

5.3 Network analysis

There exists a variety of methods to analyse NNs during the learning in order to optimize the layout (number of hidden neurons, connections, etc.). But the best architecture for simple classification tasks can usually be found with a few trials. The challenge in physics analysis is to understand the functioning of the already trained net.

5.3.1 Partial derivatives

The partial derivatives of the NN output with respect to the input variables are a powerful tool to analyse a NN [10]. This method has to be adapted to the special requirements of a particle search. The basic ideas are described in this section.

The state (activation) of a neuron in a *standard* feed-forward net is a continuous and differentiable function (the so-called *activation function*) of the state of the neurons in the preceding layer (see e.g. [8]). This means that also the state of the output neurons S_k^{out} is a continuous and differentiable function of the state of the input neurons (i.e. of the input variables) S_i^{in}. Hence the partial derivatives of the output neurons with respect to the input neurons ($\partial S_k^{out}/\partial S_i^{in}$) exist. They measure the sensitivity of the output k for changes in the input i.

The calculation of the derivatives is straight forward. The frequent case of feed-forward nets with one hidden layer and connections only between the input neurons and the hidden neurons and between the hidden neurons and the output neurons is described in the following. It can easily be extended to more complicated nets.

For convenience the input neurons are labelled with i, the hidden neurons with j, and the output neurons with k. S denotes the state of a neuron and B its bias input. W_{ij} is the weight of the connection between input neuron i and hidden neuron j, W_{jk} the weight between hidden neuron j and output neuron k. The activation function is called f. It describes the state of a neuron for a given input I: S = f(I) .

The states S_j of hidden neuron j and S_k of output neuron k are obtained by summing over their inputs

$$S_j = f(\sum_i S_i W_{ij} + B_j)\,, \quad S_k = f(\sum_j S_j W_{jk} + B_k)\ .$$

Substituting the left equation into the right one and differentiating with respect to S_i, or using directly the *chain rule* $\frac{\partial S_k}{\partial S_i} = \sum_j \frac{\partial S_k}{\partial S_j}\frac{\partial S_j}{\partial S_i}$ yields

$$\frac{\partial S_k}{\partial S_i} = S'_k \sum_j S'_j W_{ij} W_{jk}\ . \tag{1}$$

The sum runs over all hidden neurons j. S′ is the total derivative of S. If the *sigmoid* function is chosen as activation function, i.e. $S(I) = (1 + e^{-I})^{-1}$, then the derivatives S′ have the simple form $S' = S(1 - S)$. Fig. 1 shows the sigmoid function f(x) (a) and its first, second, and third-order derivatives versus x (b) and versus f(x) (c).

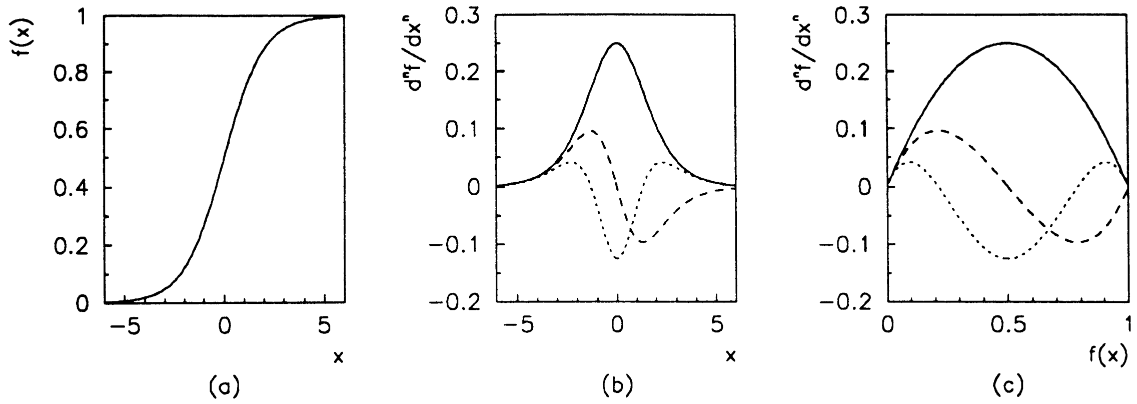

Figure 1: The sigmoid activation-function $f(x) = (1 + e^{-x})^{-1}$ (a) and its first (full line), second (dashed line), and third-order derivative (dotted line) versus x (b) and versus f(x) (c).

Eq. (1) has an interesting form since the change of output k caused by a small change in input i is factorized into two terms. The first one, S'_k, depends only on the output and the second one, the sum over the hidden neurons j, only on the hidden layer. This weighted sum over the derivatives of the states of the hidden neurons is called *hidden sum* in the following.

Eq. (1) can easily be extended to higher orders. For the sake of simplicity only the second order is given here. It reads

$$\frac{\partial^2 S_k}{\partial S_i \partial S_{i'}} = S''_k \sum_j S'_j W_{ij} W_{jk} \sum_j S'_j W_{i'j} W_{jk} + S'_k \sum_j S''_j W_{ij} W_{i'j} W_{jk} \ . \qquad (2)$$

S″ denotes the second-order derivative of S. The diagonal terms of the second and of the third order are

$$\frac{\partial^2 S_k}{\partial S_i^2} = S''_k (\sum_j S'_j W_{ij} W_{jk})^2 + S'_k \sum_j S''_j W_{ij}^2 W_{jk} \qquad (3)$$

and

$$\frac{\partial^3 S_k}{\partial S_i^3} = S'''_k (\sum_j S'_j W_{ij} W_{jk})^3 + 3\ S''_k \sum_j S'_j W_{ij} W_{jk} \sum_j S''_j W_{ij}^2 W_{jk} + S'_k \sum_j S'''_j W_{ij}^3 W_{jk} \ . \quad (4)$$

The similarities between the diagonal terms of order n can easily be seen from the above equations. The first and the last terms have the form

$$S_k^{(n)} (\sum_j S'_j W_{ij} W_{jk})^n \quad \text{and} \quad S'_k \sum_j S_j^{(n)} W_{ij}^n W_{jk} \ ,$$

respectively. $S^{(n)}$ is the derivative of S of order n. Terms with sums over powers of W_{ij} are suppressed in our application since most of the weights W_{ij} are smaller than 1. On the contrary the weights W_{jk} are generally bigger than 1. Hence the partial derivatives are dominated by the first term.

These features are demonstrated for the NN with 10 variables (N_{10}) at $m_H = 90$ GeV/c^2 (see section 6). Fig. 2 shows the sums of the first (a), of the second (b), and of the third-order partial derivatives (c) of each event versus the output. These sums run over all input variables ($D_n = \sum_i \partial^n S_k / \partial S_i^n$).

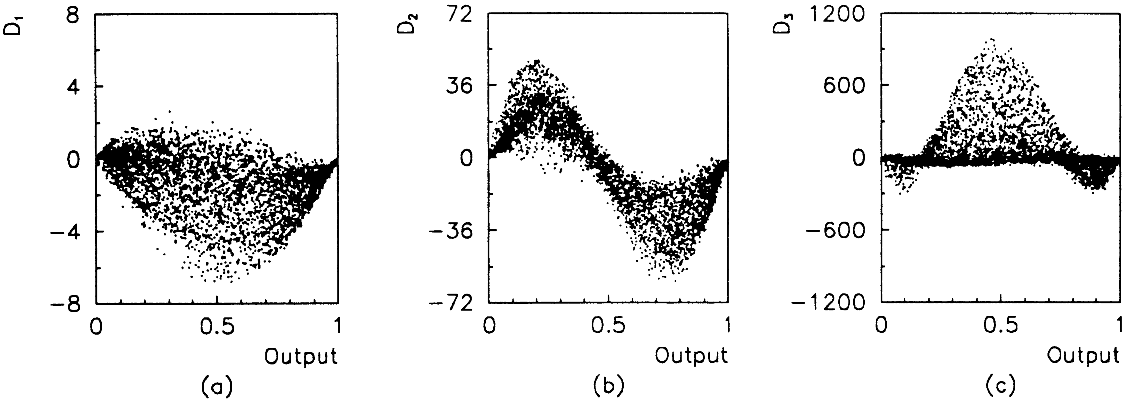

Figure 2: Sums of first (a), second (b), and third-order partial derivatives (c) versus the output for N_{10} at $m_H = 90$ GeV/c^2.

The shape of the plots follows nicely the shape of the derivatives of the activation function (fig. 1c), except from fluctuations for small mean-values and from the sign of the first and of the third-order derivatives.

5.3.2 Zero point

The difference in the sign of odd-order derivatives can easily be understood. Whether an increase of the inputs increases or decreases on average the output depends on the *zero point* (or *bias output*) of the net. This is the output of a net if all inputs vanish. Therefore it depends only on the bias inputs of the hidden and of the output neurons and on the weights between the hidden and the output neurons. For the net shown above the zero point is 0.98 . Hence an increase of the inputs tends to decrease the output. Nets with a small zero point do not change the sign of the odd-order derivatives. The zero point is usually near the edges because NNs are more stable near saturation than in the central output-region.

5.3.3 Hidden sums

In order to understand the good agreement between the shape of the partial derivatives in fig. 2 and the output derivatives in fig. 1c, the second term in the expression of the partial derivatives in eq. (1), the *hidden sum*

$$\Sigma_{ik}^{hid} = \sum_j S'_j W_{ij} W_{jk} \; , \qquad (5)$$

has to be studied in more detail. This sum plays an important role in understanding systematic changes in the inputs, like shifting or smearing. Using eqs. (1) and (5) the change of the output S$_k$ induced by a small change ΔS_i of the input S$_i$ is given by

$$\Delta_i S_k = \frac{\partial S_k}{\partial S_i} \Delta S_i = S'_k \Sigma^{hid}_{ik} \Delta S_i \ . \tag{6}$$

If correlations and non-linear effects are ignored then the effect of systematically changing the inputs onto the output can be estimated by combining the contributions of each input variable. For shifts the changes have to be added linearly and for smearing quadratically. Of special interest are *worst-case shifts* where all changes $\Delta_i S_k$ have equal sign.
Defining the *absolute hidden-sum*

$$\Sigma^{abs}_k = \sum_i |\Delta S_i \Sigma^{hid}_{ik}| = \sum_i |\Delta S_i \ \sum_j S'_j W_{ij} W_{jk}| \tag{7}$$

and the *quadratic hidden-sum*

$$\Sigma^{sq}_k = \sqrt{\sum_i (\Delta S_i \Sigma^{hid}_{ik})^2} = \sqrt{\sum_i (\Delta S_i \sum_j S'_j W_{ij} W_{jk})^2} \tag{8}$$

allows to express the effect of worst-case shifts in the form

$$\Delta S^{shift}_k = \pm \sum_i |\Delta_i S_k| = \pm S'_k \Sigma^{abs}_k \tag{9}$$

and smearing in the form

$$\Delta S^{smear}_k = \sqrt{\sum_i (\Delta_i S_k)^2} = S'_k \Sigma^{sq}_k \ . \tag{10}$$

The basic features of a NN can be studied by uniform smearing and shifting (a test with more realistic input changes based on an error estimation of each variable shows qualitatively no difference). For convenience $|\Delta S_i| = 1$ is chosen in the following.

The partial derivatives and hidden sums are studied now as function of the output S$_k$. The whole output-range [0,1] is divided into 100 intervals (bins) and mean values are calculated by averaging over the contributions of each event with an output in the same bin. The mean values of the hidden sums of each variable, Σ^{hid}_{ik}, have a rather strong dependence on S$_k$. However the mean values of the absolute, Σ^{abs}_k, and the quadratic sums, Σ^{sq}_k, depend only weakly on S$_k$ (at least for the nets used in this study). This is illustrated in fig. 3a for Σ^{abs}_k and in fig. 3b for Σ^{sq}_k using the net N$_{10}$ at m$_H$ = 90 GeV/c^2.

This means that the shape of the dependence of the partial derivatives from the output is determined by the shape of S$'_k$, the derivative of the output. The mean value of the hidden sum can in good approximation be considered as constant scaling factor.

In order to show the spread of these hidden sums in relation to the value of the derivatives two scatter plots are included in fig. 3 : the sum of the absolute value of the partial derivatives of each event (c) and the partial derivatives of each event summed quadratically over the inputs (d) versus the output. The distributions of the hidden sums are almost flat apart from edge effects for small outputs which are strongly suppressed by S$'_k$. Hence the scatter plots have the shape of S$'_k$.

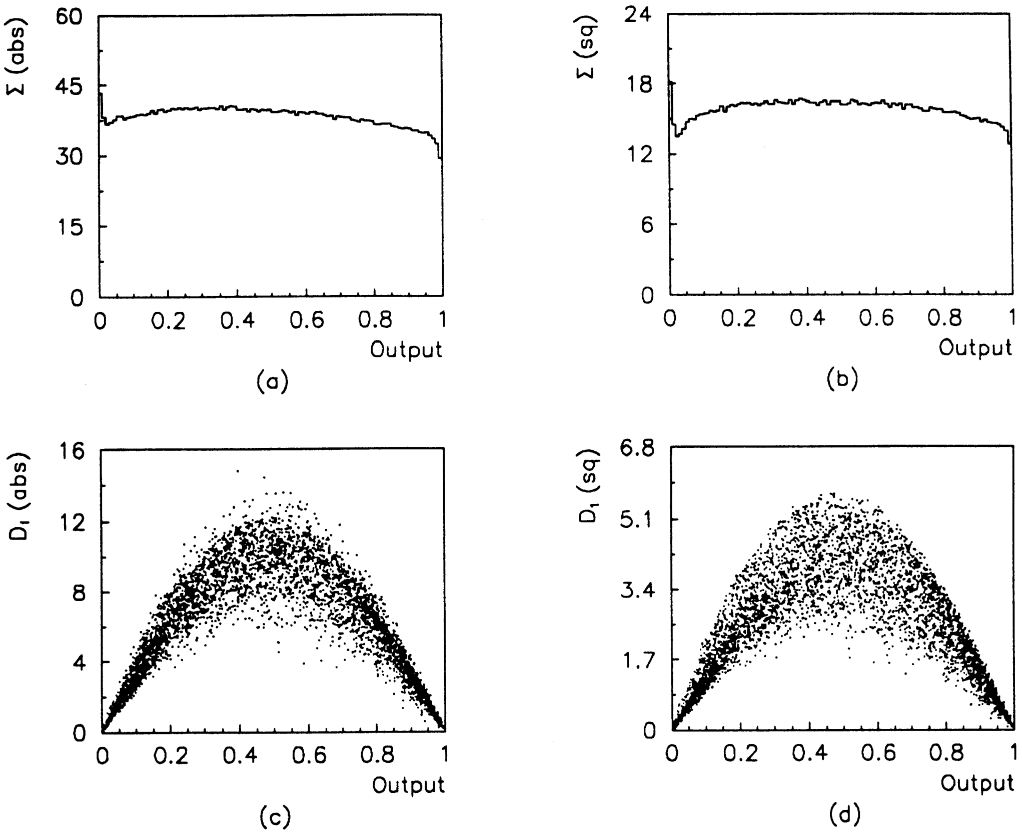

Figure 3: Mean values of the absolute (a) and of the quadratic hidden-sums (b) and scatter plots of the corresponding first-order partial derivatives (c,d) versus the output for N_{10} at $m_H = 90$ GeV/c^2.

5.4 Variable selection

The importance of variables is often estimated by analyzing their distributions and correlations. Since NNs are non-linear and since they *learn* to correlate their inputs in such a way that their output is optimized it seems to be natural to take in the selection of the input variables the special features of NNs into account. This can be done very efficiently by stepwise elimination of the weakest variables found with a combination of the following three tests:

1. Partial derivative:

The partial derivative of the output with respect to an input is a measure of the sensitivity of the output against small variations in the input. The mean value (over a test sample) of the partial derivatives for an input serves therefore as first selection criterion.

As explained in the previous section the partial derivatives depend strongly on the output value. The mean value of a first-order partial derivative is therefore dominated by the central output-region. The power of a variable near the edges can be evaluated with the second-order derivative which is positive in the first half of the output interval and negative in the second half. Since background events cluster in the first half and signal events in the second half the difference in the mean values of the second-order derivatives between background and signal events measures also the discrimination strength of a variable. For the Higgs search it turns out that the second order is well suited at the beginning of the variable selection but less powerful than the first order towards the end of the selection process. Therefore only the first-order derivative is used here.

2. Mean value:

Some variables have small partial derivatives but important correlations with other variables. Hence a complementary test is necessary to avoid the elimination of such very useful variables. It is inspired by an observation about missing inputs made in an application of NNs for kink recognition [8]. In that case the inputs are the residuals of a track fit to measured coordinates. These inputs are only weakly correlated for small kinks and it is found that the input corresponding to a missing coordinate can be fairly well substituted by 0, i.e. by its mean value.

To evaluate the correlation power of a variable a test is made where the actual value of the variable is replaced by its mean value. The decrease of the statistical significance measures the importance of the variable.

3. Correlations between partial derivatives:

The third test aims to eliminate highly-correlated variables. Since a NN builds up (during training) its own correlations between the inputs it is necessary to study the correlations in the output and not simply in the input. This can be done via the second-order partial derivatives or more efficiently by calculating the correlations between the first-order partial derivatives.

A variable is considered bad if the absolute value of the mean of its partial derivative is small and if the mean-value test shows a small decrease in performance. Among variables with strongly-correlated partial derivatives only the best one is kept.

The tests described above measure the importance of a variable in a given NN but not necessarily for the classification task itself. A variable with rather poor test-results might perform much better in a net with less inputs. It is therefore very important to eliminate only the worst variables, to retrain with the reduced variable set and to test again. In this study about a quarter of the variables can be eliminated in one step between 45 and 10 variables. Then the elimination task becomes much harder.

6 Higgs search with NNs

In order to evaluate the potential of NNs in the Higgs search a net with 45 variables is trained using a 45−22−1 layout. The emphasis of this exercise is not to achieve the optimal result by fine-tuning of the layout and of the learning procedure, but to obtain sort of an upper limit for NNs with a reduced number of input variables.

With the methods described in the previous section the number of variables (input neurons) can be reduced to 10 without losing much in statistical significance for the 90 GeV/c^2 mass. For the lower mass the loss of statistical significance is much more pronounced. This is due to the higher correlations inside the net. In this context high correlation does not mean that the variables are highly correlated (in which case the elimination of some correlated variables would not diminish the performance significantly). It rather means that more variables are nearly half-correlated (correlation coefficient $\approx \pm$ 0.5). The performance drops significantly for both masses if the number of variables falls below six.

The results for both masses are listed in Table 6 for several nets with N_{in} input neurons and N_{hid} hidden neurons. N_{hid} of the two biggest nets is not optimized, but the layouts represent an educated guess for fairly good performance (cf. [11]). The other nets have the optimal number of hidden neurons.

N_{in}	N_{hid}	H_{70}	H_{90}
45	22	19.1	8.7
20	10	15.8	8.7
15	14	15.5	8.7
10	9	14.2	8.4
9	8	13.2	8.4
8	7	13.2	8.2
7	6	12.6	8.0
6	5	10.9	8.0

Table 6: Statistical significance as function of the number of input variables for both masses.

In the following the nets will be referred to by their number of input neurons. N_{10}, for example, specifies the net with 10 inputs.

The 10 most efficient variables are the following in decreasing order of importance (the positions 2 and 3 are interchanged for the search at 70 GeV/c^2):

1. Number of secondary tracks (N_{off}).

2. χ^2 when the event is fitted with the HZ hypothesis (χ^2_{HZ}).

3. Energy of the most energetic gamma (E^{γ}_{max}).

4. Number of charged tracks (N_{trk}).

5. Thrust (T).

6. χ^2 when the event is fitted with the WW hypothesis (χ^2_{WW}).

7. Invariant mass of all secondary tracks (m_{off}).

8. Sum of the angles between the jets if the event is forced into three jets (S_θ).

9. Momentum of the most energetic electron (P^e_{max}).

10. Number of secondaries in the *Higgs* jets when the event is fitted with the HZ hypothesis (N^H_{off}).

These variables are the inputs for N_{10}. The inputs to the smaller nets N_n are the variables 1 to n of the list.

In compromising between the number of input variables and the performance two nets from Table 6 are chosen for further study. The net with 10 inputs (N_{10}) for high performance and the net with only 7 inputs (N_7) but still good performance.

A comparison with section 4 shows that all variables used in the standard analysis are contained in the list above. There are two new variables: S_θ (8) and P^e_{max} (9). Adding them to the standard analysis does not improve the statistical significance. To allow for a direct comparison between the conventional and the NN analysis, a net is trained with the 7 variables used in the standard analysis (N_{st}). 6 hidden units are used and the same learning procedure is applied as for the other nets.

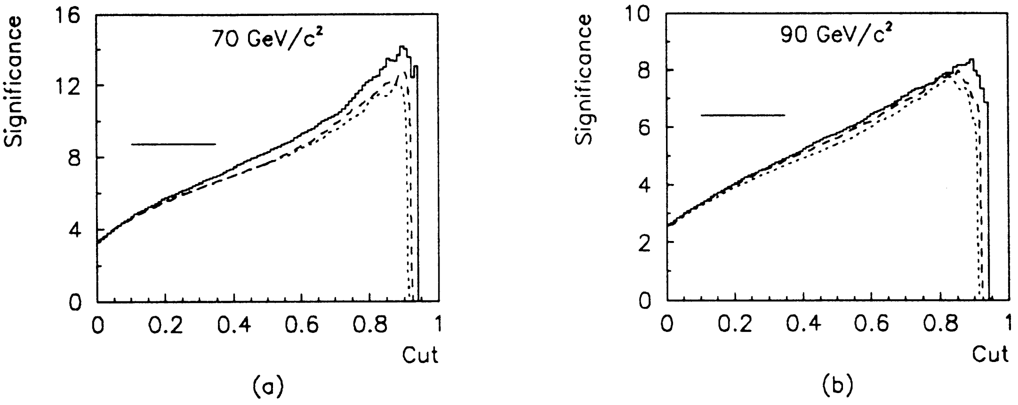

Figure 4: Statistical significance at $m_H = 70$ GeV/c^2 (a) and $m_H = 90$ GeV/c^2 (b) for N_{st} (dotted line), N_7 (dashed line) and N_{10} (full line) as function of the output cut. The result of the standard analysis (horizontal line) is included for comparison.

Fig. 4 shows the statistical significance of the three nets as function of the cut on the output for both masses. At least two background events above the cut are demanded to avoid big statistical fluctuations. The threshold at 0 is due to the preselection. The statistical significances raise almost linearly with the cut, reaching their maximum around 0.9 .

The number of accepted signal and background events of the three nets and the statistical significance are shown in Table 7 and 8 for the analysis at 70 and 90 GeV/c^2,

Reaction	Cuts	N_{st}	N_7	N_{10}
$e^+e^- \to H_{70}$ Z	60.5	29.5	23.6	40.9
$e^+e^- \to q\bar{q}gg$	22.3	2.4	1.9	3.7
$e^+e^- \to W^+W^-$	25.0	3.2	1.4	4.0
$e^+e^- \to$ Z Z	1.4	0.4	0.2	0.6
Total background	48.7	6.0	3.5	8.3
Statistical significance	8.7	12.0	12.6	14.2
Min. luminosity [pb^{-1}]	330.	174.	157.	124.

Table 7: Signal and background events left for standard and NN analyses at m_H = 70 GeV/c^2 and their statistical significance.

Reaction	Cuts	N_{st}	N_7	N_{10}
$e^+e^- \to H_{90}$ Z	38.0	30.5	26.4	22.0
$e^+e^- \to q\bar{q}gg$	11.9	5.0	3.0	1.1
$e^+e^- \to W^+W^-$	11.8	3.9	2.6	1.6
$e^+e^- \to$ Z Z	11.4	6.6	5.3	4.2
Total background	35.1	15.5	10.9	6.9
Statistical significance	6.4	7.7	8.0	8.4
Min. luminosity [pb^{-1}]	610.	422.	391.	354.

Table 8: Signal and background events left for standard and NN analyses at m_H = 90 GeV/c^2 and their statistical significance.

respectively. The cut on the NN output is chosen such that the significance is maximal. The results of the standard cuts (Table 3) are included to ease the comparison.

The NNs show a significantly higher performance than the standard method. The difference between them is considerably higher at m_H = 70 GeV/c^2. This can be explained by the higher correlation between the variables (cf. Table 6) which favours the NN technique over the standard method based on independent cuts.

The nets can be further analysed in order to better understand this behaviour. N_{10} at 90 GeV/c^2 is chosen as typical example. The output distributions of background (a) and signal events (b) are shown in fig. 5. Since the statistical significance is calculated from the number of events above the cut the output distributions have to be integrated from right to left. Fig. 6 contains the number of background (a) and signal events (c) and the square root of the number of background events (b) above the cut. The shape of the distributions in fig. 6b and 6c explains the almost linear raise of the statistical significance with the cut in fig. 4.

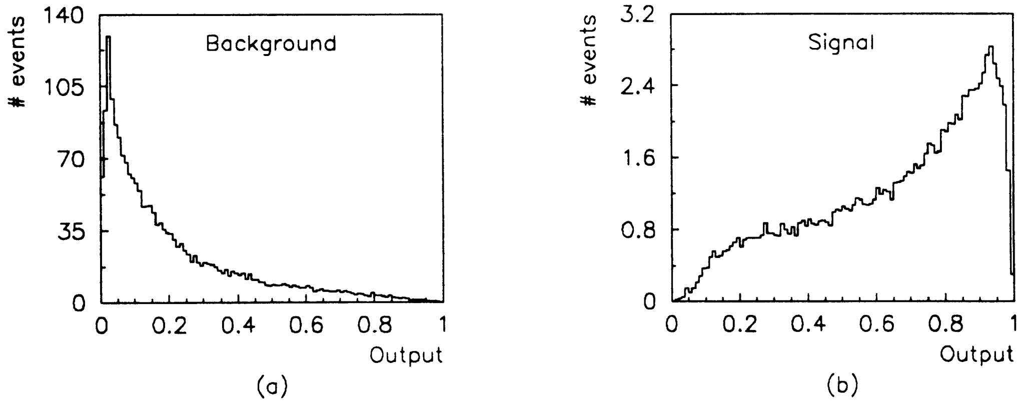

Figure 5: NN output for background (a) and signal events (b) with N_{10} at $m_H = 90$ GeV/c^2.

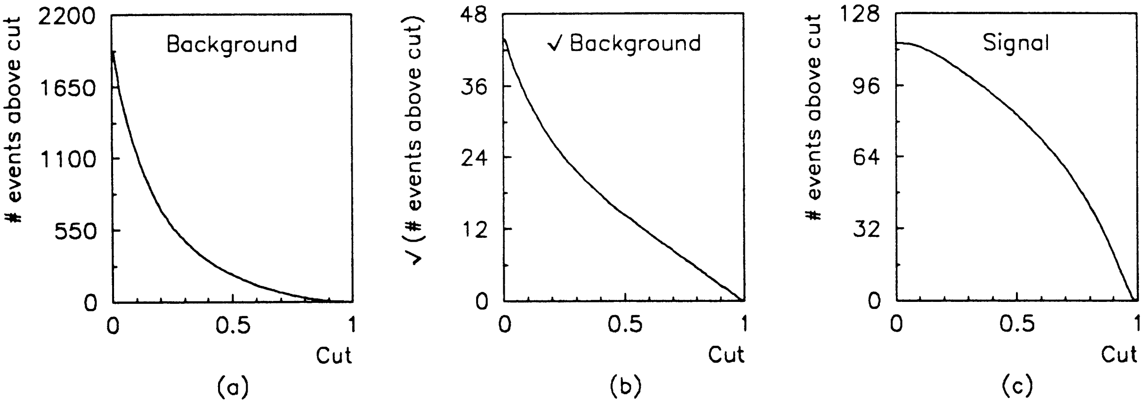

Figure 6: Number of events above cut for background (a) and signal (c) with N_{10} at $m_H = 90$ GeV/c^2. The curve in the middle (b) is the square root of the number of background events above cut.

7 Systematic effects

The results presented in the last section were obtained under the assumption that the learning data and the test data are qualitatively exactly the same. In this case the errors on the results arise only from statistical fluctuations in the test data sets due to the limited number of events. Now the question of systematic effects, i.e. effects coming from qualitative differences between the training data and the test data, is addressed. This is a crucial problem for the analysis of real data if the NNs are trained on simulated data.

7.1 Changes of the input

Systematic effects are studied by shifting and smearing the input variables and by changing parameters of the data simulation program. The nets are not retrained and the same cuts are applied as before for the unbiased samples. The statistical significances are calculated from the evaluated signal and the expected background. The expected background is the number of background events obtained with the unbiased data. The signal is evaluated by adding the number of accepted signal and background events for the biased data and by subtracting the expected background.

The variables used in the standard analysis are normalized to 1 now in order to ease the comparison with the NNs.

7.1.1 Shifting

The robustness of the methods against systematic shifts in the input is checked first. The shift of a variable is called positive if the number of accepted events increases and negative if this number decreases. The biggest effects are obtained if the shifts of all variables are either positive or negative (*worst-case shifts*). In these cases shifts of about 0.002 are tolerable for the standard and the NN methods.

7.1.2 Smearing

Now the input variables are smeared randomly by adding a normal-distributed value with mean 0 and root mean-square (rms) between 0.01 and 0.05. A change in statistical significance of about 1 is observed with rms $= 0.05$ for the standard method and with rms between 0.02 and 0.05 for the NNs. In the standard method the cut of each variable is chosen such that the distribution of the variable around the cut is fairly flat. For the NNs the cuts on the input variables are chosen implicitly by the net as function of the learning and of the cut applied to the output neuron. It is therefore not astonishing that the conventional method has an advantage in this respect. As explained in section 5.3 the total effect of shifting or smearing can approximately be decomposed into the contributions from each input (eqs. (9) and (10)).

7.1.3 Changes in the simulation

The parameters controlling the detector resolution and the energy scale are modified in order to study systematic effects. Firstly changes improving and worsening the resolution of the momentum of charged tracks and of the energy of neutral showers by roughly 25% are introduced. Then the energies and momenta are shifted systematically by +5%, +10%, -5%, and -10%.

In all cases and for both methods the changes of the statistical significances are not bigger than about 1.

7.2 Analysis of the results

So far the cut for the systematics tests was set such that the statistical significance is maximal for unbiased data. Now the differences of the statistical significances are studied as function of the output. Again N_{10} at 90 GeV/c^2 is chosen as example.

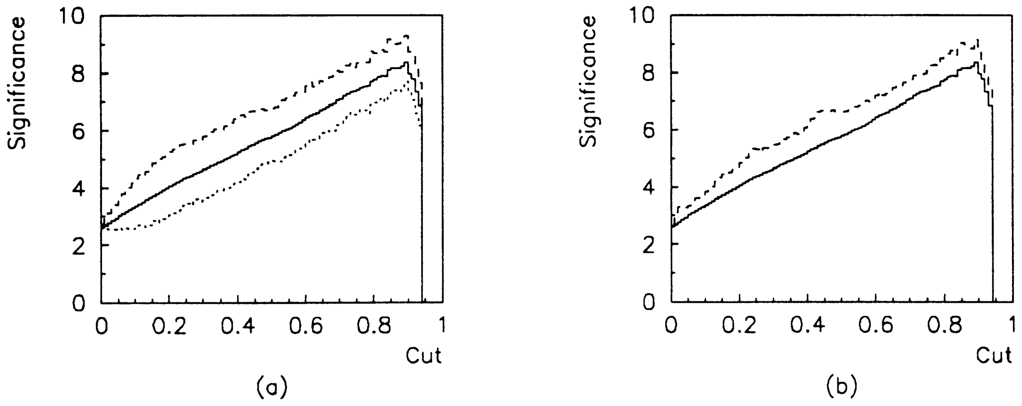

Figure 7: Changes of the statistical significance due to shifting (a) and smearing (b) of input data with N_{10} at m_H = 90 GeV/c^2. The full line represents the statistical significance for unbiased data. The results of positive and negative shifts are shown as dashed and dotted line, respectively, and the effect of smearing as dashed line.

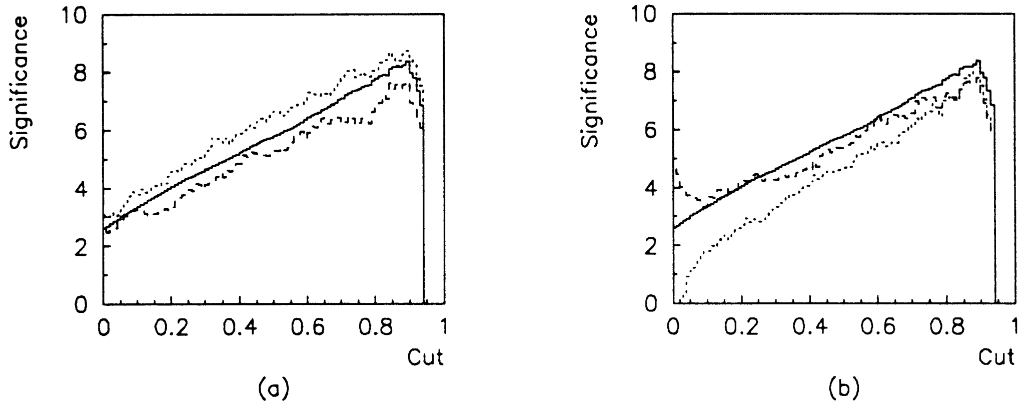

Figure 8: Changes of the statistical significance due to modifications in the Monte Carlo: detector resolution (a) and shifts in particle energy and momentum (b) with N_{10} at m_H = 90 GeV/c^2. The statistical significance for unbiased data is shown as full line, for improved resolution and positive shifts as dashed line, and for worse resolution and negative shifts as dotted line.

Fig. 7 shows the effect of worst-case shifts of ± 0.002 (a) and of smearing with rms = 0.02 (b). The influence of the modifications in the simulation program can be seen in fig. 8 for the changes in resolution of ±25% (a) and for the shifts of energies and momenta of ± 10% (b).

The differences between the unbiased and the biased samples are mainly due to changes in the background samples since the background density is much higher than the signal density over a large range of the output (see fig. 5). The big differences at small cuts in fig. 8b come from the preselection of the data. The shifts increase and decrease the number of accepted $q\bar{q}$ background events by 13% to 14%, respectively. This leads to a higher number of events with small NN output for positive shifts and to a smaller number for negative shifts. Hence the background is underestimated in the first case and overestimated in the second case. An underestimation of the background means an overestimation of the signal and vice versa. The estimated statistical significance is therefore too high for positive shifts and too low for negative shifts. The influence of the preselection becomes smaller for harder (higher) cuts.

The explanation for lower statistical significance by increased resolution in fig. 8a is analogous. Higher resolution leads to a better background rejection and therefore to an underestimation of the signal.

The regular behaviour for smearing and shifting needs further explanation. The distributions shown in fig. 5 and 6 are quite smooth apart from edge effects. Since the sigmoid function is smooth too, the output changes only slowly if the inputs are slightly changed. The effect of input smearing onto the output is shown in fig. 9 for N_{10} at 90 GeV/c^2. The differences in the output between unbiased and smeared inputs (rms = 0.02) are plotted for 3 unbiased output values: 0.1 (a), 0.5 (b), and 0.9 (c).

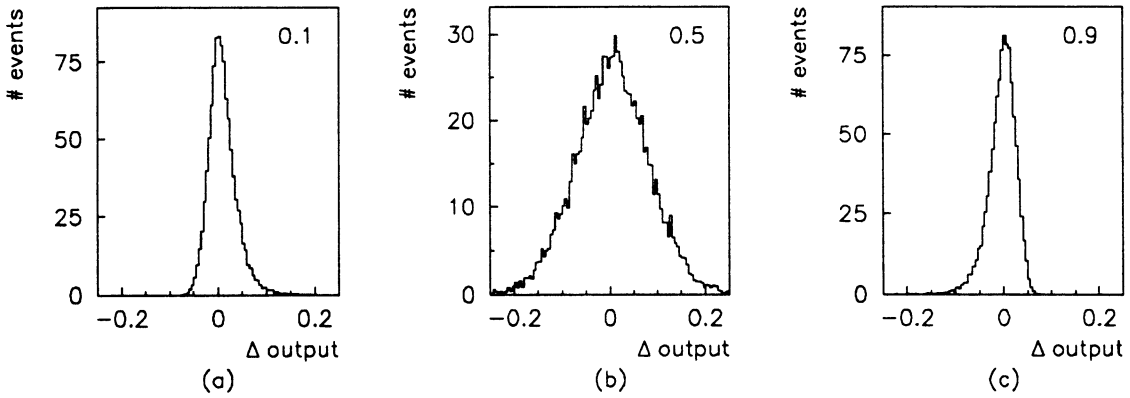

Figure 9: Effect of input smearing onto the output for unbiased output values of 0.1 (a), 0.5 (b), and 0.9 (c) with N_{10} at m_H = 90 GeV/c^2.

The differences in the widths correspond to the differences in the derivative of the activation function at these points. The shape of the distributions on the left and on the right side is asymmetric. This is due to the asymptotic behaviour of the activation function which makes it harder to push an output into the direction of saturation than into the opposite direction.

As explained in section 5.3, arguments based on the derivative of the output are only valid if the hidden sum is fairly independent from the output. Therefore a final test is made with a net which does not fulfil this condition. This is the case for the NN with 10 inputs specially trained for the 70 GeV/c^2 mass (N_{10}^{sep}). Since the hidden sums are smaller around the cut, stronger systematic changes as before are applied: shifts of 0.005 and smearing with rms = 0.05 . The results are summarized in fig. 10. Since the absolute (a) and the squared hidden-sum (b) decrease significantly with increasing output-values the differences in statistical significance between biased and unbiased data become smaller at higher cuts (c,d). But the smaller sensitivity for input changes has the disadvantage that the statistical significance is reduced too. The difference in statistical significance between this robust mass-dependent net and the corresponding mass-independent net (trained with a mixture of signal events from different masses) is rather big: 12.6 compared to 14.2 (cf. Table 5).

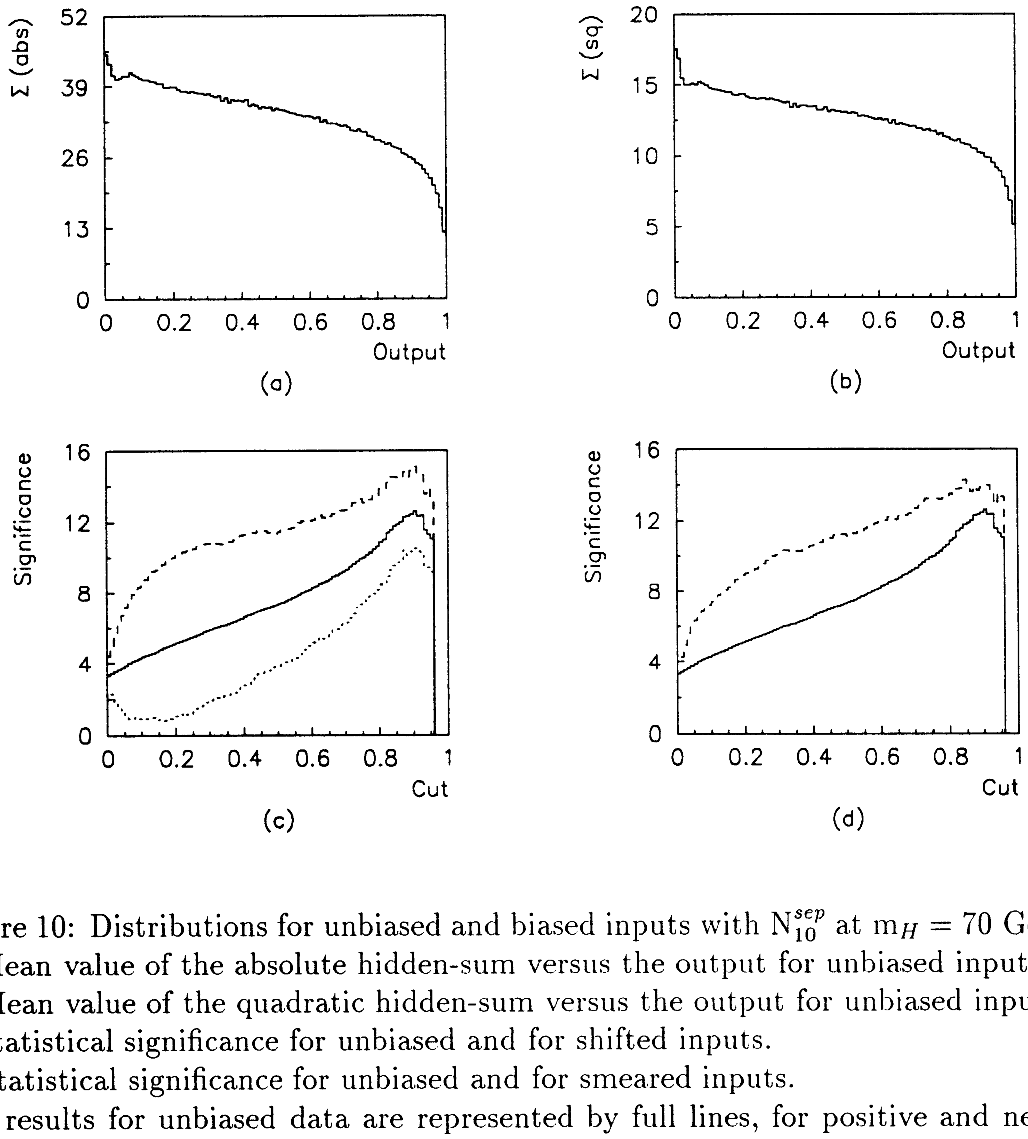

Figure 10: Distributions for unbiased and biased inputs with N_{10}^{sep} at $m_H = 70$ GeV/c^2.
a) Mean value of the absolute hidden-sum versus the output for unbiased inputs.
b) Mean value of the quadratic hidden-sum versus the output for unbiased inputs.
c) Statistical significance for unbiased and for shifted inputs.
d) Statistical significance for unbiased and for smeared inputs.
The results for unbiased data are represented by full lines, for positive and negative shifts as dashed and dotted line, respectively, and for smearing as dashed line.

7.3 Recognition of systematic effects

Although the results are fairly stable against systematic changes of the input data it is necessary to recognize significant differences between the training and the test data. The most obvious method is to check the distributions and correlations of the input variables. This can be done very efficiently with NN techniques [12].

Another possibility is to analyse the performance of the net by comparing the partial derivatives and their correlations for the learning and the test sample. In addition the significance distribution (statistical significance as function of the cut) is quite sensitive to systematic changes as shown above.

Differences in the composition of the background can further be studied with a special NN trained to discriminate between the different backgrounds. It has the same architecture and inputs as the nets described above for the classification into signal and background except from the output. Now the output layer consists of three neurons and the net is trained to answer 1 0 0 (i.e. 1 for the first output neuron and 0 for the other ones), 0 1 0 and 0 0 1 for the three background groups. Since the number of signal events is much smaller than the number of background events and since the ZZ background is rather similar to the signal it is better not to try to discriminate four classes of input events. This net detects changes in the composition of the data fairly well.

7.4 Choice of the optimal net and cut

The results of the systematics studies described above indicate that the NNs react smoothly towards small systematic changes in the input. Some nets are less sensitive to rather artificial changes, like uniform smearing and shifting, but they are in general less efficient than the more sensitive nets. This is quite understandable, since the sensitivity of the output for changes in the input is a drawback for systematic effects but necessary for the classification task. The differences in statistical significance between unbiased and biased test samples are smallest for high output-values where also the statistical significance is optimal. Therefore the cut has to be applied at the output value with the highest statistical significance.

The choice of the net has to take into account the statistical significance and the set of input variables. Because some of the variables are easier to control than other ones the final selection has to be based on a careful comparison of the Monte Carlo data used for training and the measured data. Since the operation of LEP 200 is still some years ahead this final step cannot be done yet.

8 Conclusions

Very simple NNs with about ten inputs and one hidden layer show very high performance in the search for the Higgs boson at LEP 200. The most powerful input-variables are selected by a new technique, mainly based on the partial derivatives of the state of the output neuron with respect to the state of the input neurons. This method has the advantage over others that the special features of a NN are properly taken into account in the evaluation of the utility of an input variable.

The statistical significance of the NNs is more than 60% higher than that of standard cuts for a Higgs mass of 70 GeV/c^2 and more than 30% higher for a Higgs mass of 90 GeV/c^2. This means in terms of luminosity that the NN method needs 62% (42%) less events to reach a statistical significance of 5 for a mass of 70 (90) GeV/c^2. In addition the nets are mass independent because they are trained for the whole mass-range between 70 and 90 GeV/c^2. Therefore and since the nets a quite small the learning effort for extended searches is modest.

A detailed study of the sensitivity of the nets towards systematic effects in the input data shows no significant difference between the NNs and standard cuts.

Neural nets offer also a fast and convenient way to estimate the potential of a physics analysis since many variables can be used and the cuts are *learned* by the net during training. After analysis of such a net and reduction of the variable set one-dimensional cuts can easily by determined from the distributions of the input variables of the accepted events.

References

[1] S.L. Glashow, Nucl. Phys. 22 (1961) 579.
S. Weinberg, Phys. Rev. Lett. 19 (1967) 1264.
A. Salam, Proc. Nobel Symposium, Ed. N. Svartholm (Almqvist and Wiksells, Stockholm, 1968) 367.

[2] P.W. Higgs, Phys. Lett. 12 (1964) 132, Phys. Rev. Lett. 13 (1964) 508, Phys. Rev. 145 (1966) 1156.

[3] E. Gross and P. Yepes, Int. Jou. Mod. Phys. A, Vol. 8, No. 3 (1993) 407.

[4] S. Katsanevas, *Energy and luminosity requirements for LEP 200 physics*, *The Standard Model and just beyond*, 4th Topical Seminar on the Standard Model, San Miniato, June 1992 and references therein.

[5] P. Janot, *Will a Higgs boson be found at future e^+e^- colliders ?*, *27th Rencontres de Moriond*, LAL 92-27, May 1992.

[6] T. Sjöstrand, Comp. Phys. Comm. 28 (1983) 229.
T. Sjöstrand, Comp. Phys. Comm. 39 (1986) 347.
T. Sjöstrand and M. Bengtsson, Comp. Phys. Comm. 43 (1987) 367.

[7] N.J. Kjaer and R. Moller, *Reconstruction of invariant masses in multi-jet events*, internal DELPHI note 91-17, PHYS 88, April 1991.

[8] G. Stimpfl-Abele, Comp. Phys. Comm. 67 (1991) 183.

[9] G. Stimpfl-Abele, *Neural nets for kink finding*, Proceedings of the *Second Workshop on Neural Networks: From Biology to High Energy Physics*, Elba, 1992.

[10] G. Stimpfl-Abele, *Kink recognition with neural networks*, Proceedings of the *IEEE Nuclear Science Symposium and Medical Imaging Conference*, Orlando, 1992.

[11] G. Stimpfl-Abele, *Finding the decay vertex of a charged track with neural networks*, Proceedings of the *Conference on Computing in High Energy Physics*, Annecy, 1992.

[12] L. Garrido et al., *Bayesian interpretation of the neural net output and its application to test the agreement between two empirical distributions*, Proceedings of the *Conference on Computing in High Energy Physics*, Annecy, 1992.