Pile-up studies for soft electron identification and b-tagging with DC1 data

Tomasz Bołd[1,2], *Frédéric Derue*[3], *Anna Kaczmarska*[2,3], *Ewa Stanecka*[2], *Marcin Wolter*[2,4]

1. UST-AGH, Kraków, Poland

2. Institute of Nuclear Physics PAN, Kraków, Poland

3. LPNHE - Laboratoire de Physique Nucléaire et de Hautes Énergies IN2P3 - CNRS - Universités Paris VI et Paris VII, France

4. Tufts University, Medford MA, USA

ATL-PHYS-PUB-2006-001
03 January 2006

Abstract

This note describes the analysis of the soft electrons and related b-tagging in the Higgs boson associated production (WH) events. The simulation, reconstruction and analysis are done in the framework of the ATLAS Data Challenge 1 including the simulation of pile-up. Two analysis tools, the likelihood and the neural network methods are studied.

Contents

1 Introduction **2**

2 Data samples **2**
2.1 Simulation 2
2.2 Reconstruction 2
2.3 Sources of electrons 3

3 Electron identification **4**
3.1 Overview of the soft electron identification 4
3.2 Discriminating variables 4
3.3 Results with a ratio of likelihood 9
3.3.1 Construction of the discriminating function 9
3.3.2 Performance 10
3.4 Results with a neural network 11
3.4.1 Construction of the discriminating function 11
3.4.2 Performance 14

4 Soft electron b-tagging **16**
4.1 Overview of the soft electron b-tagging 16
4.2 Performance 17

5 Conclusion **18**

6 Acknowledgements **19**

A Correlation between discriminating variables **20**

B Effect of TRT on performance **21**

C Neural network trained to equally reject pions from $H \to b\bar{b}$ $H \to u\bar{u}$ and $H \to c\bar{c}$ processes. **23**

1 Introduction

The identification of the low transverse momentum ($p_T < 20$ GeV/c) electrons inside jets was studied in [1] with the first implementation of the algorithm within the ATHENA framework. This note aims to provide a realistic evaluation of the soft electron identification and b-tagging performance. The data used include pile-up at low and high luminosities. Different analysis tools, the likelihood ratio and the neural network method, are used.

The note is organised as follows. Data samples used in this analysis are described in section 2. Performance of the electron identification procedures is given in section 3. Finally the soft electron b-tagging performance is given in section 4.

2 Data samples

2.1 Simulation

The samples contain events from Higgs-boson associated production, WH with $W \to \mu\nu$. The signal sample consists of $H \to b\bar{b}$ events and background samples of $H \to c\bar{c}, u\bar{u}$ events. The mass of the Higgs boson is set to $m_H = 120$ GeV/c^2.

Process	dataset	Nb of events		
		no pile-up	low lumi	high lumi
$H \to bb$	002055	20,000		
$H \to c\bar{c}$	002057	20,000		
$H \to u\bar{u}$	002056	50,000		25,000

Table 1: *Data samples used in this study: the process, the dataset number for* DC1*, and the available number of events for the different luminosity scenarios.*

The Monte Carlo events were generated for DC1 [2] with PYTHIA 6.203 [3] under the ATHENA framework and stored in ROOT format [4]. Those events were simulated in the ATLAS detector with GEANT 3 [5] based on ATLSIM/DICE program [6], version 3.2.1. The inputs were the event kinematics from ROOT-format files and the output was the detector hits and digits in ZEBRA-format files. Pile-up is added to study these events for the low luminosity (2×10^{33} cm^{-2} s^{-1}) and the high luminosity (10^{34} cm^{-2} s^{-1}) scenarios. The number of pile-up events per bunch crossing follows a Poisson distribution with an average of 4.6 at low and 23 at high luminosities.

2.2 Reconstruction

These DC1 data were reconstructed using ATHENA release 7.8.0. In particular following parameters were used:

- tracks in the Inner Detector were reconstructed using `xKalman`. All options were set to the default values, and the correction for internal bremsstrahlung losses was enabled;
- jets were reconstructed using the cone algorithm with the default options (cone size 0.7);
- electronics noise in the calorimeters was included.

Only tracks with transverse momentum $p_T > 2$ GeV/c and $|\eta| \leq 2.4$ are kept for the analysis. In order to reject contribution from badly reconstructed, fake or pile-up tracks the following quantities are used to select *good quality* tracks:

- at least two hits in the pixel detector, one of them in the B-layer;
- at least nine precision hits (pixels + SCT);
- impact parameter of the track in the transverse plane $|A_0| \leq 1$ mm;
- track fit quality $\chi^2_{\rm fit} \leq 3$;
- $|z_0 - z_{\rm vertex}| \sin\theta \leq 0.15$ cm. This cut is used against tracks from pile-up;
- no shared hit in the pixel detector, no more than one hit in the SCT;
- no ambiguity in the first pixel wafer.

As the TRT transition radiation information is crucial for the studies presented two further selection criteria are applied:

- at least 20 hits in the TRT detector along the track;
- at least one high energy hit (TR hit) in the TRT detector along the track.

2.3 Sources of electrons

Low momentum electrons can originate from the decays of particles from the hadronic cascade or from the interaction of particles with the detector material. The signal electrons come from direct and cascade semileptonic decays of b-hadrons, leptonic decays of J/Ψ coming from b and decays of b-hadrons to τ-leptons with subsequent decays into electrons. The background electrons arise from π^0 Dalitz decays, γ-conversions and decays of light hadrons.

	$< p_T >$ (GeV/c)			$< n_{track} >$
	signal electrons	background electrons	pions	
$H \to bb$	10.8/11.0/11.1	4.3/4.4/4.3	5.3/4.9/4.8	9.3/9.4/10.7
$H \to u\bar{u}$	5.1/5.6/5.8	4.8/4.6/5.1	6.1/6.2/6.0	8.9/9.0/10.5
$H \to c\bar{c}$	8.2/8.1/8.0	4.0/4.2/3.7	5.2/5.1/5.0	9.1/9.2/10.7

Table 2: *The mean p_T of signal and background electrons, pions and the mean track multiplicity, $< n_{track} >$, in an event in various data samples. All numbers are given for no pile-up/low/high luminosity.*

In Fig. 1 normalised distributions of the transverse momentum p_T and pseudorapidity $|\eta|$ for signal electrons from $H \to b\bar{b}$, and background electrons from background samples are presented. For comparison, distributions for pions from background samples are also shown. The background electrons have softer transverse momentum spectrum compared to the signal ones. Since γ-conversions are the main source of background electrons, they are concentrated mainly in the higher pseudorapidity region due to the increase of material in front of the electromagnetic calorimeter. In Tab. 2 the mean values of transverse momentum p_T of signal electrons, background electrons and pions in all data samples are shown for different luminosities.

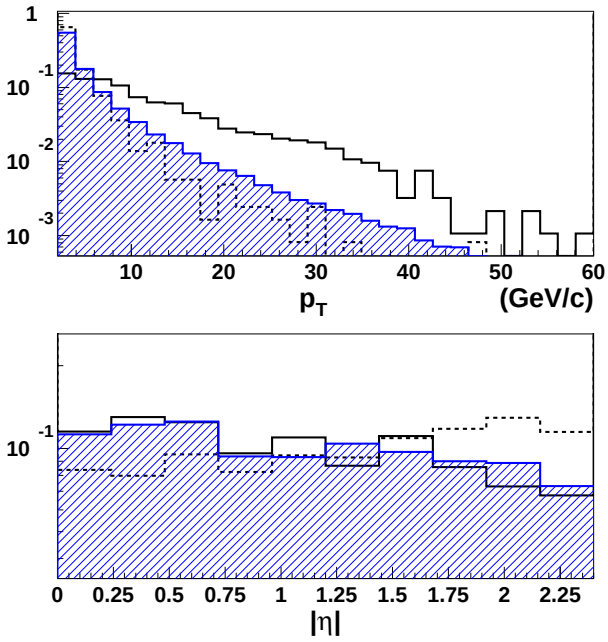

Figure 1: *Normalised distributions of p_T and $|\eta|$ for signal electrons in $H \to b\bar{b}$ (solid line), background electrons (dashed line) and pions (hatched histograms) in background samples, low luminosity data.*

As can be noticed from Fig. 1 and Tab. 2, the kinematics of electrons and pions is almost not changed in the different luminosity scenarios. The presence of pile-up events makes the pattern recognition more difficult. It increases the track multiplicity and the probability of selecting fake or lower quality track. The tracks will be also less isolated. All of these will render the electron identification more difficult in presence of pile-up.

3 Electron identification

3.1 Overview of the soft electron identification

All the *good quality* tracks are extrapolated to the subsequent calorimeter samplings. The cell with the maximum deposit is searched in the vicinity of the extrapolated position. The following η and ϕ windows are used: 3×3 in the presampler, 3×1 in the strip compartment, 3×3 in the middle sampling, 3×1 in the back sampling. In each sampling of the calorimeter the list of cells around the impact point is established in following windows: 3×5 in the presampler, 19×5 in the strip compartment, 9×13 in the middle sampling, 5×13 in the back sampling. Then a cluster is created and a set of discriminating variables is defined basing on transverse and longitudinal shapes of electromagnetic showers, track information in the Inner Detector (ID), ID-Calo matching in position and energy and transition radiation information. All variables are in the standard combined ntuple. The analysis is then performed with a set of ROOT macros/PAW kumacs.

3.2 Discriminating variables

The variables used to identify reconstructed tracks as electrons and to reject other tracks can be classified as follows:

- Variables using ID information only:

- n_{TR} – number of high energy hits in the TRT detector on the track;
- A_0 – transverse impact parameter.

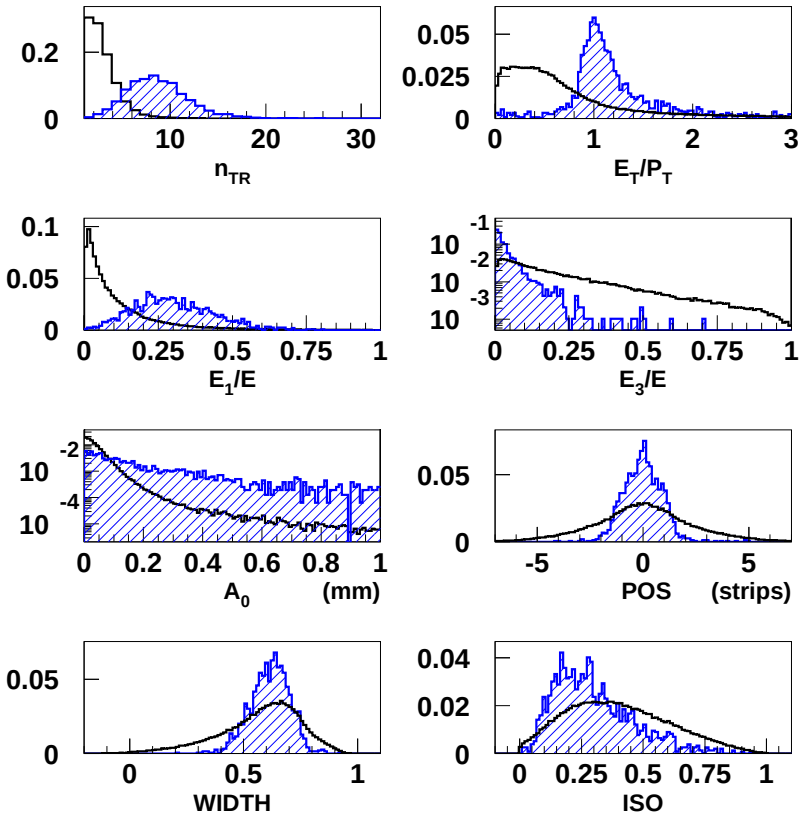

Figure 2: *Distributions of the discriminating variables obtained with low luminosity data. Hatched histograms - signal electrons from $H \to b\bar{b}$, empty histograms - pions from background samples.*

-Variables using combined information from the ID and the LArEM:

- $E_T(core)/p_T$ – the ratio of the transverse energy reconstructed in the LArEM, to the transverse momentum reconstructed in the ID;
- $POS = \sum_{i=i_m-7}^{i=i_m+7} E_i \times (i - i_m) / \sum_{i=i_m-7}^{i=i_m+7} E_i$ – difference between the track and the shower positions measured in units of distance between the strips, where i_m is the impact cell for the track reconstructed in the ID and E_i is the energy reconstructed in the i-th cell in the η direction for constant ϕ given by the track parameters.

- Variables using LArEM information only:

- E_1/E – fraction of the energy reconstructed in the first compartment of the LArEM;
- E_3/E – fraction of the energy reconstructed in the third compartment of the LArEM;
- $ISO = 1 - E_{3\times3}/E_{3\times7}$ – shower isolation in the LArEM from the ratio of energy reconstructed around the extrapolation in a 3×3 and 3×7 clusters;

- $WIDTH = \sqrt{\sum_{i=i_m-1}^{i=i_m+1} E_i \times (i - i_m)^2 \,/\, \sum_{i=i_m-1}^{i=i_m+1} E_i}$ – calculated using the highest energetic strip and its neighbour on each side.

More details on the presented variables can be found in [1]. Distributions of each variable for signal electrons from $H \to b\bar{b}$ and pions from background samples are shown in Fig. 2.

Fig. 3 and 4 show the distributions of the mean values of each discriminating variable, for signal electrons and pions from low luminosity data, as a function of $|\eta|$ or p_T. The variables show a significant dependence on the pseudorapidity and a less pronounced one on the transverse momentum. In η the changes correspond to varying granularities,

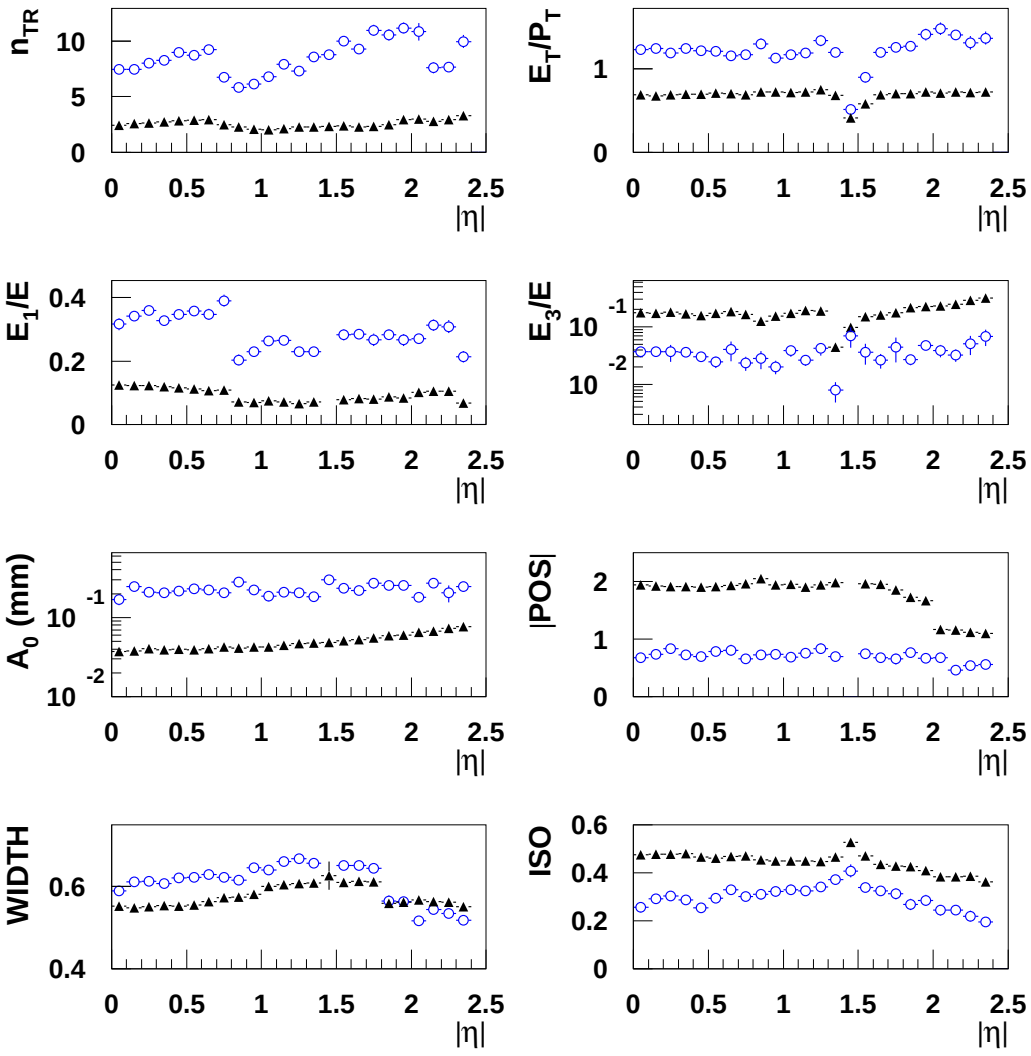

Figure 3: *Mean values of the discriminating variables for the signal electrons (circles) and pions (triangles) in function of $|\eta|$. Dependencies are shown for samples at low luminosity.*

lead thickness and material in front of the electromagnetic calorimeter. The separation between the distributions obtained for electrons and pions can vary also with η. Thus, the discriminating power of each variable varies.

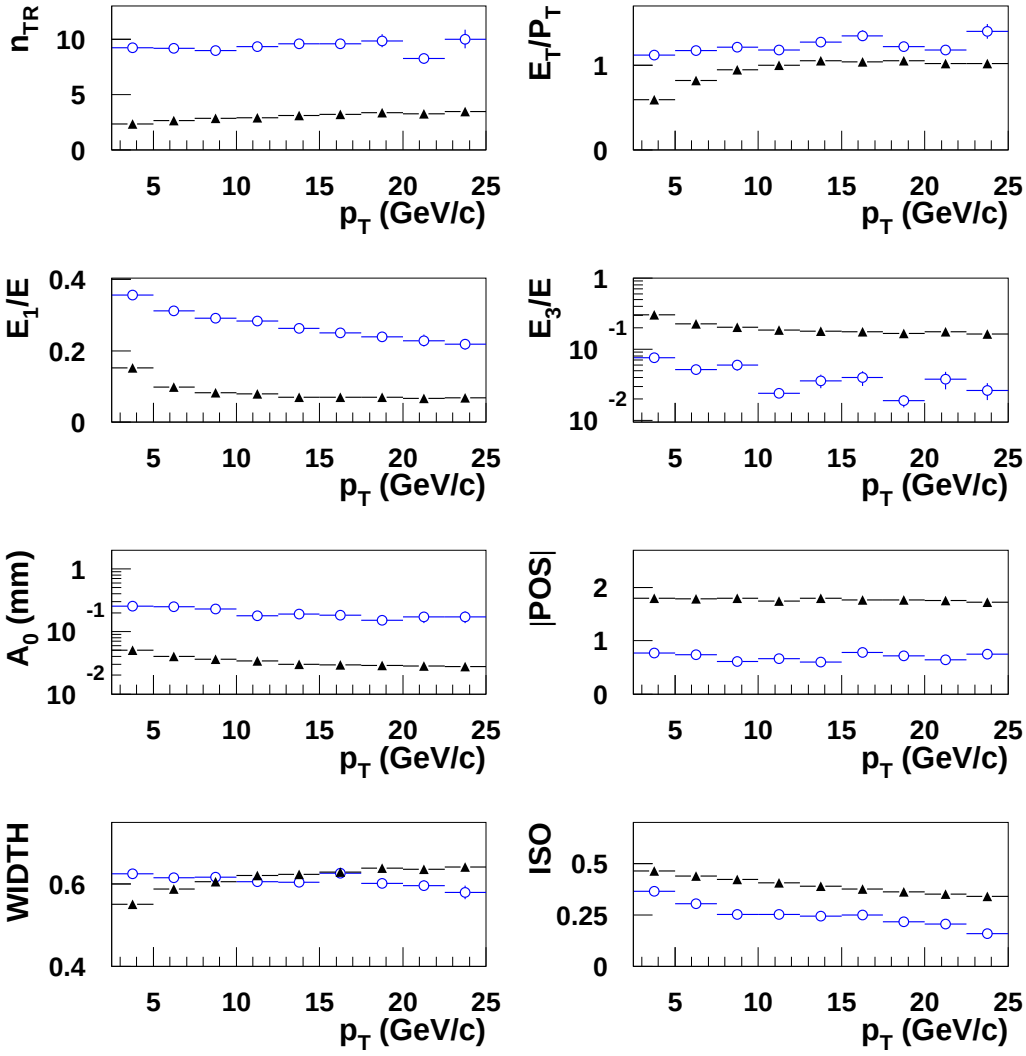

Figure 4: *Mean values of the discriminating variables for signal electrons (circles) and pions (triangles) in function of p_T. Dependencies are shown for samples at low luminosity.*

Separation variable In order to estimate the discriminating power of each variable, an estimator called raw separation is defined [11][12]:

$$< s^2_{raw}(x) >= \frac{1}{2} \int \frac{(\rho^b(x) - \rho^s(x))^2}{\rho^b(x) + \rho^s(x)} dx, \quad (1)$$

where x is the variable for which the separation is calculated. This definition assumes that $\rho^s(x)$, the probability that the track originates from a signal electron, and $\rho^b(x)$, the probability that the track originates from non-signal hadron, are normalised to the unity. Tab. 3 shows the raw separation values, ranked by decreasing order, obtained for each discriminating variable. The most discriminating is the number of TRT high threshold hits n_{TR}. It is the only variable which shows significant drop of its separation value in the presence of pile-up. It is in good agreement with statements presented in [13]. At high luminosity, other particles, including electrons from γ-conversions, deposit energy in the straws crossed by the particles of interest. This causes an increase in the number of high threshold hits for both pions and electrons. The average number of TR hits per reconstructed track increases by 1.4 for pions and by 1.2 for electrons. The pion rejection

	no pile-up	low-lumi	high-lumi
n_{TR}	0.75	0.69	0.53
E/p	0.47	0.45	0.43
E_1/E	0.49	0.49	0.44
A_0	0.32	0.32	0.31
POS	0.23	0.22	0.22
E_3/E	0.21	0.21	0.21
$WIDTH$	0.15	0.15	0.13
ISO	0.15	0.14	0.14

Table 3: *The raw separation of each discriminating variable.*

is determined by the size of the regions of overlap in Fig. 5 for the low and the high luminosity scenarios. The loss of separation corresponds to a loss of the discriminating power of this variable at high luminosity.

The variables describing the energy deposit in each sampling do not have the same separation. The E_3/E variable is less discriminating than the E_1/E. This is due to the fact that the ratio E_3/E is equal to 0 for about a third of electrons and pions. When there is energy deposit in the third sampling, its discriminating power is similar to the one of the E_1/E variable.

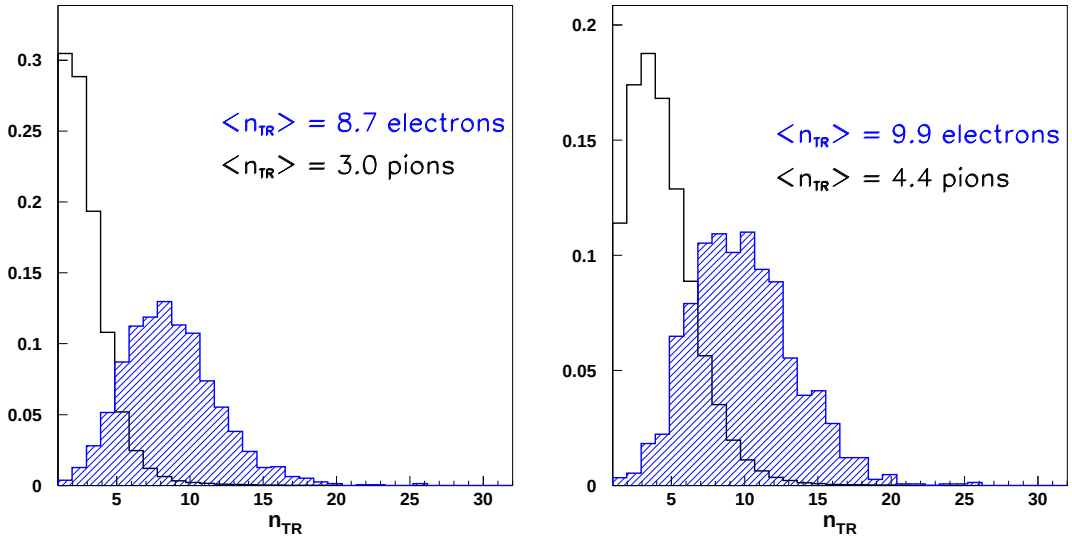

Figure 5: *The number of the TRT high threshold hits for signal electrons (hatched histograms) and for pions in background sample (empty histograms). Distributions are shown for data with the low luminosity (left) and the high luminosity (right) conditions.*

Correlation between variables The discriminating variables could be correlated. The linear correlation coefficient between variables μ and ν is given by:

$$r_{\mu\nu} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{\mu i}-\bar{x}_\mu)(x_{\nu i}-\bar{x}_\nu)}{\sigma(x_\mu)\sigma(x_\nu)}, \quad (2)$$

the indices μ and ν run over all discriminating variables, $\bar{x}$ is the mean value of the distribution of the variable x and $\sigma(x)$ its rms. All details concerning correlation coefficients can be found in Tab. 10 and 11 in appendix A.

The variables n_{TR} and A_0 obtained from the ID only are, as expected, not correlated with those obtained from the LArEM. For other variables, the correlation coefficients are rather low, never exceeding $\sim 30\%$. These low values do not exclude the more complicated, non-linear dependencies between variables, different for signal and background, which can be exploited by non-linear techniques. It can be also noticed that variables dealing with energy reconstructed in the LArEM show larger correlation for signal events than for background ones.

It should be stressed that two highly correlated variables, when combined, can be more discriminating than two other variables which are not correlated, if the correlation coefficient for signal and background is different enough.

3.3 Results with a ratio of likelihood

3.3.1 Construction of the discriminating function

The distributions of the discriminating variables are treated as probability density functions (PDFs). As already seen in section 3.2, these PDFs show a significant dependence on the pseudorapidity. In previous study [1] a dedicated data sample of $H \rightarrow b\bar{b}$ filtered for electrons from semileptonic decays has been produced. The statistics was high enough to obtain PDFs in five pseudorapidity bins (but still not enough to optimise PDFs also in p_T). No pile-up has been added on this filtered sample. It is thus not possible to optimise PDFs with the pseudorapidity, for all luminosity scenarios, as they would have to be determined on typically few tens of events.

A simple method is based on the use of relative likelihood and constructed from the distribution of the discriminating variables for the two classes of events (s for signal and b for background) which should be disentangled. Ideally, for N discriminating variables, the ratio of likelihood can be written as:

$$X_{RL} = \frac{\mathcal{L}^s(x_1, ...x_N)}{\mathcal{L}^s(x_1, ...x_N) + \mathcal{L}^b(x_1, ...x_N)}, \tag{3}$$

where $\mathcal{L}^s = g^s(x_1, ...x_N)$ is the N dimensional density distribution (or likelihood) for signal hypothesis, and $\mathcal{L}^b = g^b(x_1, ...x_N)$ for background hypothesis. Obtaining the distributions from a histogram in N dimension being a difficult task, the following approximation is usually made:

$$X_{RL} = \frac{\prod_i \rho_i^s(x_i)}{\prod_i \rho_i^s(x_i) + \prod_i \rho_i^b(x_i)}, \tag{4}$$

where $\rho_i^s(x_i)$ (resp. $\rho_i^b(x_i)$) is the one dimensional density distribution of the variable x_i for events of signal hypothesis (resp. background). Several points should be underlined:

- X_{RL} tends to 1 for signal events and tends to 0 for background events;
- if a discriminating variable x_i is not useful ($\rho_i^s(x_i) \sim \rho_i^b(x_i)$) over all the range considered for x_i, it does not dilute the information from the other discriminating variables;
- if there is no correlation between the different discriminating variables, the combined variable X_{RL} is optimal;
- in case of correlations between the discriminating variables, some information is lost. Nevertheless, the use of X_{RL} as a new discriminating variable insures that no

bias is introduced in the analysis. However, care should be taken not to incorporate too many correlated variables to avoid a dilution of the discriminating power.

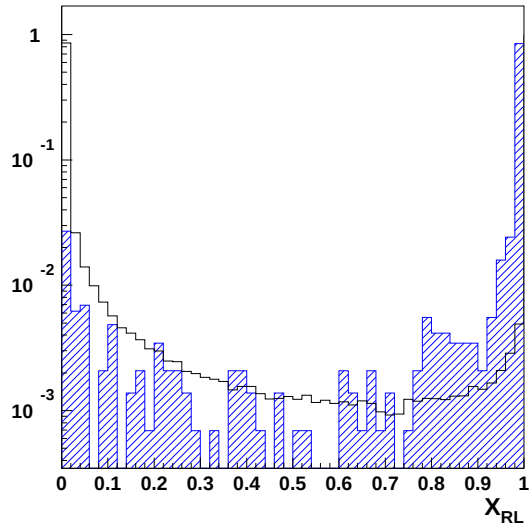

Figure 6: *Normalised distributions of the discriminating function X_{RL} for signal electron tracks in $H \to b\bar{b}$ (hashed) and pions in $H \to u\bar{u}$ (empty). Low luminosity data.*

To avoid any bias of performance the samples to estimate this discriminating function has to be different from the ones which provided the PDFs. For the PDFs preparation we used half of available $H \to b\bar{b}$, $H \to u\bar{u}$ and $H \to c\bar{c}$ samples.

Distributions of X_{RL} obtained for signal electron and background pion tracks are shown in Fig. 6. The ratio of likelihood does not take into account correlations between the variables. It is reflected in the the distributions obtained for electrons and pions which overlap and present large tails.

3.3.2 Performance

The identification of a candidate track as originating from a signal electron is based on the value of the discriminating function X_{RL} assigned to the given track. The identification efficiency versus the rejection power of the algorithm is obtained by varying the value of the threshold on X_{RL}.

To quantify the performance of the identification procedure, the following quantities are defined:

- electron identification efficiency: $$\varepsilon_e = \frac{N_e^t}{N_e}, \tag{5}$$

where N_e is the number of *good quality* signal electron tracks in the signal samples and N_e^t is the number of *good quality* signal electron tracks identified as a signal electrons track;

- charged pion rejection: $$R_\pi = \frac{N_\pi}{N_\pi^t}, \tag{6}$$

where N_π is the number of *good quality* pion tracks, and N_π^t is the number of *good quality* pion tracks misidentified as a signal electron track. The pion rejection factors obtained for various electron identification efficiency ε_e are presented in Tab. 4 and in Fig. 7 for

luminosity	ε_e	R_π		
		$H \to b\bar{b}$	$H \to u\bar{u}$	$H \to c\bar{c}$
	90 %	71 ± 3	129 ± 4	121 ± 6
no	80 %	245 ± 17	561 ± 35	508 ± 54
	70 %	514 ± 51	1439 ± 142	1040 ± 158
	90 %	39 ± 1	66 ± 1	60 ± 2
low	80 %	180 ± 11	420 ± 21	345 ± 28
	70 %	455 ± 43	1312 ± 116	862 ± 109
	90 %	20 ± 1	30 ± 1	28 ± 1
high	80 %	73 ± 3	133 ± 5	117 ± 5
	70 %	171 ± 10	405 ± 26	308 ± 22

Table 4: *Rejection of pions in various samples for varying identification efficiency ε_e and different luminosities using likelihood.*

the various types of samples. For no pile-up data and $\varepsilon_e = 80\%$ the rejection of pion tracks in $H \to b\bar{b}$ is 245 ± 17, in $H \to u\bar{u}$ is 561 ± 35 and in $H \to c\bar{c}$ is 508 ± 54. Differences between pion rejection in different jets are due to the A_0 variable which is the most discriminating in case of pions from u-jets. Compared to previous study [1] performance is higher by about 40% (this number can vary with the efficiency) as the analysis code was improved and some bugs fixed. The same figure and table show also performance obtained at low and high luminosity. For the low luminosity case, pion rejection is lower by about 30%. As shown in Tab. 3 the number of TRT high threshold hits variable is less discriminating when taking into account pile-up events. Appendix B shows performance obtained without using the n_{TR} variable, and clearly demonstrates, that the loss of efficiency is due to this particular variable. The effect is even more pronounced at high luminosity, where the rejection of pions in $b\bar{b}$ sample is 70% lower than in the case of no pile-up, or 60% compared to the low luminosity case. Once again the loss is essentially due to the n_{TR} variable.

3.4 Results with a neural network

3.4.1 Construction of the discriminating function

The neural network is a non-linear discriminating method (a detailed description of the neural network techniques can be found in [8]). The Stuttgart Neural Network Simulator [9] was used in the analysis. The architecture of the network is optimized to give the proper classification of signal and background and to avoid over-training at the same time. The neural network is built as follows (Fig. 8):

- an input layer with 8 input nodes corresponding to 8 discriminating variables;
- two internal hidden layers, each containing 16 nodes;
- an output layer containing a single neuron, since the output is a single discriminating variable.

To each neuron j in the hidden layer n inputs $x_k, k = 1, .., n$ and one output variable (the answer of the neuron) z_j are associated. For the first hidden layer the inputs are the discriminating variables, for next layers the inputs are the outputs of the preceding layer. All input variables are normalized to be within the range $[-1, 1]$.

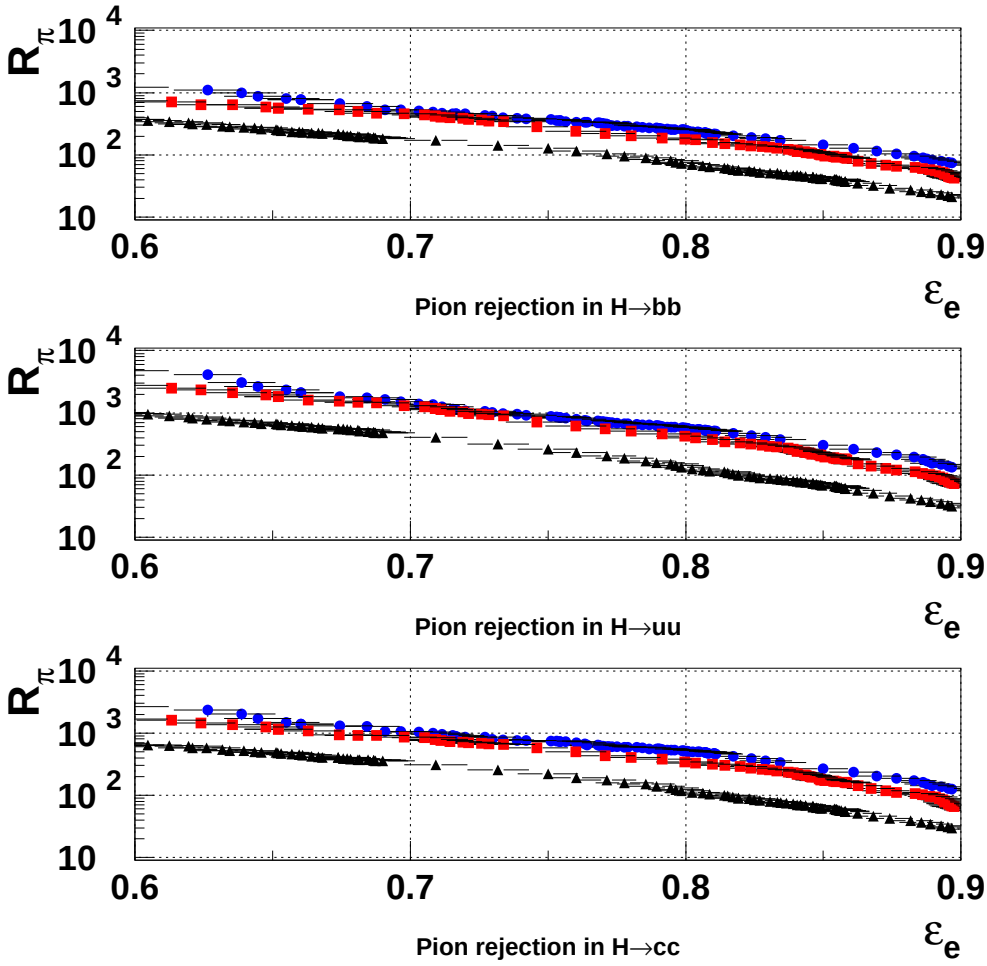

Figure 7: *Rejection of pions, R_π, in various samples as a function of the efficiency for identifying signal electrons, ε_e, for data with no pile-up (circles), low luminosity (squares) and high luminosity (triangles) using using likelihood method.*

The neuron sums up the input variables y_k, weighted by a factor w_{jk}, plus a threshold θ_j. This defines the signal Z_j:

$$Z_j = \sum_{k=1}^{N} w_{jk} y_k + \theta_j. \tag{7}$$

The output of the neuron is a function of Z: $z_j = a(Z_j)$, where a is called the activation function, and is chosen to be of the form $a(x) = 1/2(1 + tanh(x))$. The training phase of the neural network consists in determining the weighting factors w_{jk} and the thresholds θ_j. This is done by minimizing the following error function:

$$E = \frac{1}{2} \sum_{i=1}^{n} \left(X_{NN}^i - t_1^i \right)^2, \tag{8}$$

where t_1^i is the expected output (0 for background, 1 for signal), X_{NN}^i the actual value returned by the network and n is a number of events used for training. The training is

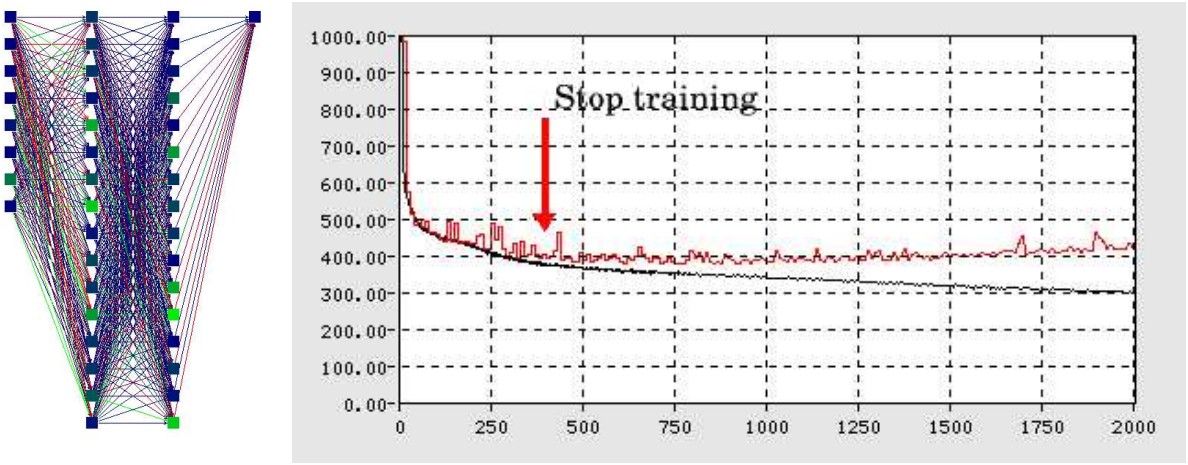

Figure 8: *Schematic view of the neural network (left) and the error function E as a function of the training cycles for the training and verification samples (right). The training was stopped after 350 training cycles to avoid network over-training.*

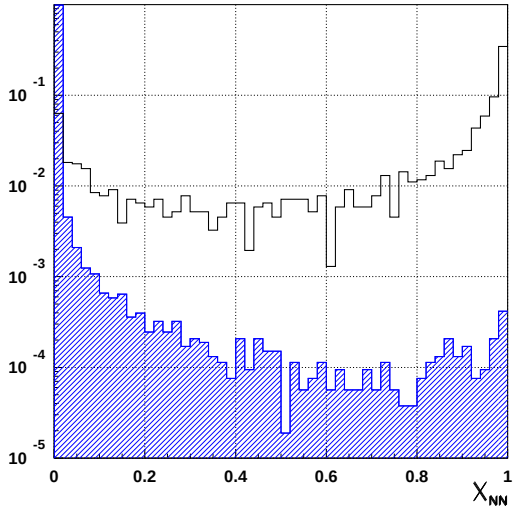

Figure 9: *Normalized distributions of the discriminating function X_{NN} for signal electron tracks in $H \to b\bar{b}$ (empty) and pions in $H \to u\bar{u}$ (hashed).*

performed using half of the available signal data (electrons from $H \to b\bar{b}$ process) and half of the $H \to u\bar{u}$ and $H \to c\bar{c}$ samples. The remaining data are used to obtain the signal detection efficiency and background rejection. It is also used as a verification sample to check, whether the values E obtained for training and verification samples are similar. This gives a useful information when to interrupt the network training (see Fig. 8). The training is stopped after about 350 training cycles to avoid over-learning. Since majority of the background part of the training sample are pions from the $H \to u\bar{u}$ process the network is tuned to reject most efficiently this type of background. The results obtained using a network optimized to reject pions from $H \to b\bar{b}$ are shown in Appendix C. Distributions of the neural network output X_{NN} obtained for signal electron and background pion tracks are shown in Fig 9.

3.4.2 Performance

The pion rejection factors obtained for various electron identification efficiency ε_e are presented in Tab. 5 and in Fig. 10 for various background types and for no-pileup, low luminosity and high luminosity data. For no pile-up data and $\varepsilon_e = 80\%$ the rejection of

luminosity	ε_e	R_π		
		$H \to bb$	$H \to u\bar{u}$	$H \to c\bar{c}$
	90 %	96 ± 4	320 ± 15	277 ± 16
no	80 %	268 ± 19	1245 ± 114	877 ± 181
	70 %	519 ± 51	2850 ± 395	2032 ± 835
	90 %	77 ± 3	206 ± 7	177 ± 9
low	80 %	205 ± 13	854 ± 61	523 ± 85
	70 %	364 ± 30	1915 ± 205	1068 ± 296
	90 %	29 ± 1	66 ± 2	54 ± 2
high	80 %	99 ± 4	261 ± 14	214 ± 20
	70 %	196 ± 11	586 ± 47	401 ± 85

Table 5: *Rejection of pions in various samples for varying identification efficiency ε_e and different luminosities using a neural network.*

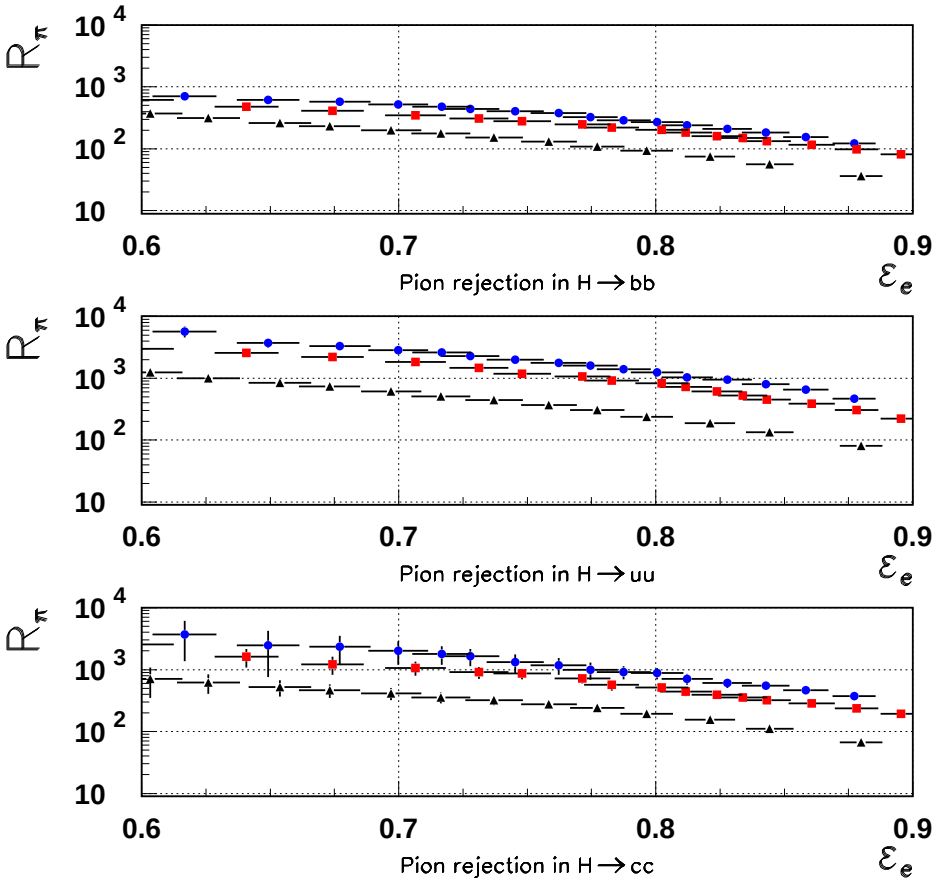

Figure 10: *Rejection of pions in various samples as a function of the efficiency for no pile-up (circles), low (squares) and high (triangles) luminosity data using a neural network.*

pion tracks in $H \to b\bar{b}$ is 268 ± 19, in $H \to u\bar{u}$ is 1245 ± 114 and in $H \to c\bar{c}$ is 877 ± 181. The use of neural network gives $\sim 120\%$ gain in the pion rejection in $H \to u\bar{u}$ sample and $\sim 70\%$ in $H \to c\bar{c}$ sample comparing to the ratio of likelihood method. In the Fig. 11 and Tab. 6 results obtained with a neural network are compared with the ones obtained using the ratio of likelihood (for $H \to u\bar{u}$ sample). It can be clearly seen that the neural network approach gives significantly better rejection for the same signal identification efficiency.

		R_π		
luminosity	ε_e	$H \to b\bar{b}$	$H \to u\bar{u}$	$H \to c\bar{c}$
no	80 %	8%	122%	72%
low	80 %	14%	103%	52%
high	80 %	36%	96%	83%

Table 6: *Gain in the pion rejection in various samples for identification efficiency* 80% *and different luminosities when using the neural network comparing to the likelihood method.*

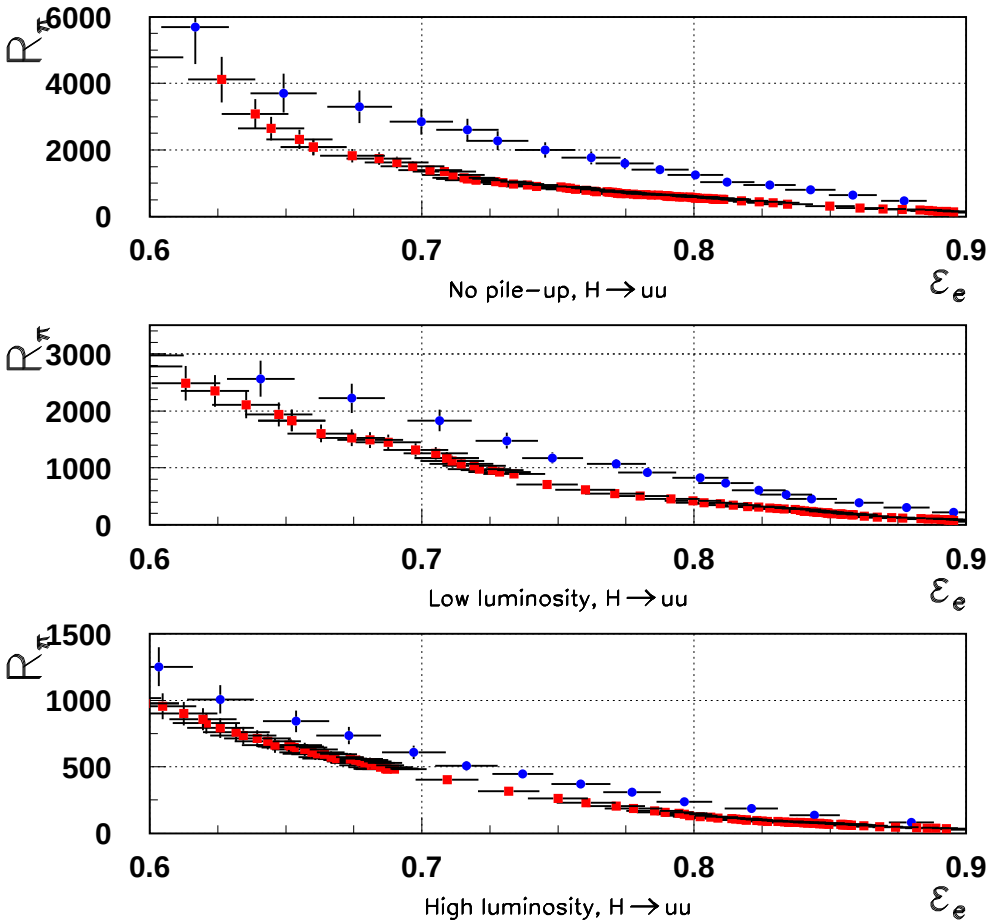

Figure 11: *Comparison of pions rejection R_π for tracks from $H \to u\bar{u}$ sample using a neural network (circles) and a ratio of likelihood (squares).*

4 Soft electron b-tagging

The soft electron b-tagging method complements the secondary vertex tagging [10] via its sensitivity to the leptonic b-decays. The branching ratio of B-meson decays to leptons is about 10.5% per lepton family.

4.1 Overview of the soft electron b-tagging

The soft electron b-tagging method is based on the electron identification procedure:

- for each track in the jet the value of the discriminating function D_{track} is calculated using one of two methods:

$$D_{track} = \begin{cases} X_{RL} & \text{likelihood method} \\ X_{NN} & \text{Neural Network method} \end{cases}$$

- for each jet the track with the highest value of D_{track} is chosen and its value of the discriminating function is taken as the value of the discriminating function for the jet, $D_{jet} = max(D_{track})$;

- for a given threshold D_{jet}^{thr} a jet with $D_{jet} \geq D_{jet}^{thr}$ is tagged as a b-jet.

Jets are reconstructed with the standard cone algorithm and labeled as originating from a b-quark, if a quark with $p_T^b > 5$ GeV/c (after FSR) is found in a cone $\Delta R = \sqrt{[\Delta\eta(e,b)]^2 + [\Delta\phi(e,b)]^2} < 0.2$ around the jet axis. The labeling for c- and u-jets is done in the same way.

The mean multiplicity of tracks inside jets with at least one reconstructed track with $p_T > 2$ GeV/c inside a cone $\Delta R < 0.4$ around the jet axis increase from the low to the high luminosity scenario: for b-jets from 3.7 to 4.3, for u-jets from 3.4 to 3.8 and for c-jets from 3.5 to 4.0. This will make b-tagging more difficult in the high luminosity samples.

The fraction of labeled jets containing electron tracks with $p_T > 2$ GeV/c is given in Tab. 7 for all types of jets and for different luminosity scenarios. The dominant source of electrons in b-jets are electrons from semileptonic decays of b-hadrons and d-hadrons. The semileptonic decays of d-hadrons occur frequently in c-jets. The u-jets are almost free from signal electrons.

Jet type	b-hadrons	d-hadrons	γ-conversions and π^0 Dalitz	Other sources
b-jets	6.6/6.7/6.5	4.2/4.3/4.2	1.5/1.6/2.0	0.4/0.4/0.4
u-jets	−/−/−	−/−/−	1.6/1.7/2.4	0.1/0.1/0.1
c-jets	−/−/−	4.8/4.7/4.6	1.4/1.4/1.4	0.1/0.1/0.1

Table 7: *Fraction of jets with electrons (in %) of a given origin Numbers are given for no pile-up/low luminosity/high luminosity events.*

In u-jets the main sources of electrons are γ-conversions and Dalitz decays. The presence of any electron in the background jets makes their rejection difficult (as the method uses electrons as tagging particles). As can be concluded from Tab. 7 the best rejection factor can be achieved for u-jets and the worst for c-jets. For events with pile-up the fraction of jets with signal electrons does not change, but the fraction of jets with background electrons increases in case of $H \to u\bar{u}$ and $b\bar{b}$ events making b-tagging less effective.

The efficiency of the b-tagging algorithm is defined as:

$$\varepsilon_b^{alg} = \frac{N_b^t}{N_b^e}, \tag{9}$$

where N_b^t is number of the tagged b-jets and N_b^e is the number of labeled b-jets with at least one electron with $p_T > 2$ GeV/c after *good quality* cuts.

The jet rejection factor is calculated as following:

$$R_{jet} = \frac{N_j}{N_j^t}, \tag{10}$$

where j is a given jet type (u,c), N_j is the number of labeled jets of type j in the $H \to j\bar{j}$ sample with at least one *good quality* track with $p_T > 2$ GeV/c inside the jet cone, and N_j^t is the number of labeled j-type jets tagged as b-jets.

The fraction of reconstructed b-jets with an electron track within a cone $\Delta R \leq 0.4$ around the jet axis is defined as the effective branching ratio BR and dependents on the performance of the detector. This value is normalised to the total number of reconstructed b-labeled jets, with at least one *good quality* tracks with $p_T > 2$ GeV/c inside the jet cone. For the sample of $H \to b\bar{b}$, this fraction is $BR \sim 13\%$. The soft electron b-tagging efficiency, ε_b^{soft}, can be calculated by multiplying the efficiency of the $b-$tagging algorithm, ε_b^{alg}, by the obtained branching ratio BR. Thus, 60% efficiency of the b-tagging algorithm with $p_T^{track} > 2$ GeV/c, corresponds to $\varepsilon_b^{soft} \sim 7.8\%$.

4.2 Performance

Fig. 12 shows the rejection factors for u- and c-jets as a function of the b-tagging algorithm efficiency for different luminosity scenarios and for both methods, the ratio of likelihood and the neural network. The achieved rejection factors for non-b jets are presented in Tab. 8. For a b-tagging efficiency of 60%, the rejection of u-jets is 151 ± 11 for the

	method	no pile-up	low lumi	high lumi
R_u	RL	151 ± 11	136 ± 9	104 ± 11
	NN	166 ± 13	144 ± 10	123 ± 15
R_c	RL	35 ± 2	36 ± 2	33 ± 3
	NN	35 ± 2	35 ± 2	33 ± 3

Table 8: *Jet rejection factors for efficiencies of b-tagging algorithm* $\varepsilon_b^{alg} = 60$ *% obtained with the ratio of likelihood (RL) and the neural network (NN) methods for different luminosity scenarios.*

likelihood and 66 ± 13 for the neural network method when no pile-up is taken into account. It decreases by $\sim 10\%$ for the low luminosity case and by $\sim 30\%$ for the high luminosity scenario. As already explained in section 3.3.2 and detailed in annexe B, this loss is due to the high threshold hits in the TRT variable which is less discriminating when pile-up is important. The rejection of c-jets shows only small degradation in the high luminosity events. The gain from using a neural network can be seen only for high efficiencies of b-tagging (see Fig. 12) when background jets are tagged mainly by pions. For low efficiencies of the algorithm where jets are tagged mainly by background electrons there is no improvement from a method giving better pion rejection and both methods, the likelihood and the neural network, give similar results.

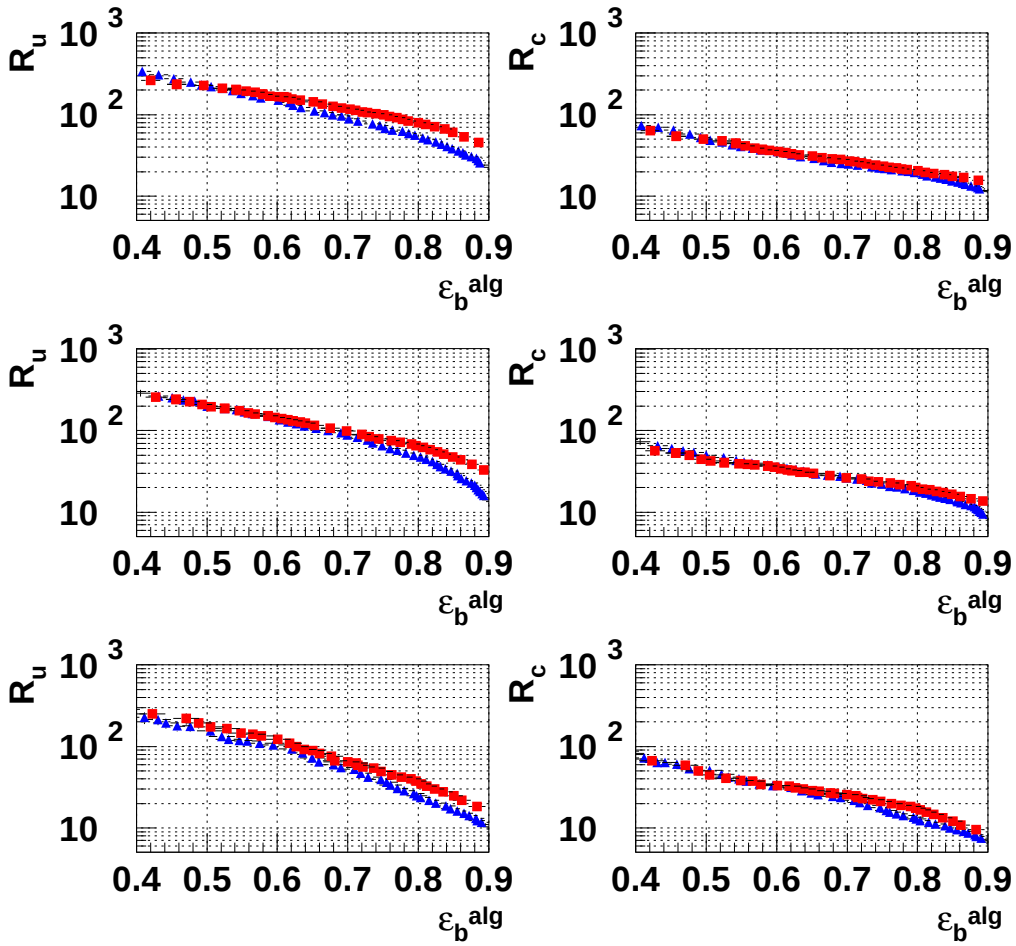

Figure 12: *Jet rejection, R_{jet}, for u- and c-jets as a function of the b-tagging algorithm efficiency, ε_b^{alg}, obtained with the likelihood (triangles) and the neural network (squares) methods. Plots for events from no-pile up samples (top panel), the low luminosity samples (middle panel) and the high luminosity samples (bottom panel).*

Tab. 9 shows the fraction of jets tagged by a given type of track for $\varepsilon_b^{alg} = 60\%$. Most of the jets are tagged by true electron tracks, independently of the jet type. It indicates, that the electron identification procedure works with a high efficiency and good purity. Most of the electron tracks which tag jets, except for u-jets, are signal electron tracks (from b- or d-hadron decays). For the b-jet sample, these tracks tag 89.9% of all tagged jets. The sample of u-jets almost does not contain electrons from b and d, but these jets are tagged, in 65.7%, by the electron tracks from γ-conversions and Dalitz decays.

5 Conclusion

In this note the soft electron identification and b-tagging results are discussed. The performance is detailed on samples WH, $H \to b\bar{b}$, $c\bar{c}$ and $u\bar{u}$ produced during the ATLAS Data Challenge 1. Compared to previous study, the emphasis of the note is put on performance obtained for the low and high luminosity scenarios. Performance obtained

Jet type	Fraction of jets tagged by a specified track [%]						
	all e	e from b	e from d	e from *	other e	π	others
b	99.4	61.5	28.4	6.6	2.9	0.3	0.3
u	69.1	0.0	0.0	65.7	3.4	16.9	14.0
c	82.4	0.0	58.2	13.5	0.7	7.7	9.9

Table 9: *Fraction of jets tagged by a specified track for* $\varepsilon_b^{alg} = 60\%$. *The e from * denotes electron tracks from conversions and Dalitz decays. Results obtained with the likelihood method on the low luminosity data.*

are similar for the low luminosity scenario and for no pile-up events. On the contrary performance decreases by 70% for the high luminosity case for rejection of pions and by 30% for rejection of u-jets. The loss is due to the number of high threshold hits in the TRT which becomes less discriminating in presence of pile-up.

To deal with non-linear dependencies between discriminating variables the neural network method is used to obtain the optimal performance. It gives $\sim 100\%$ improvement in the pion rejection. Using this method for b-tagging purpose a gain can be observed only for high b-tagging algorithm efficiencies.

6 Acknowledgements

This work was supported in part by Marie Curie Intra-European Fellowship FP6-2002/Mobility-5/MEIF-CT-2003-501408 with the LPNHE (Paris) and by Polish Government grant 620E-77SPBCERNP-03DZ 1102003-2005.

A Correlation between discriminating variables

Tables 10 and 11 show the correlation matrix between the discriminating variables used to identify the soft electrons, as detailed in section 3.2, for the signal electrons and the pions from background. The matrices are obtained on the samples at low luminosity.

	n_{TR}	A_0	E_T/p_T	POS	E_1/E	E_3/E	ISO	$WIDTH$
n_{TR}	1.000	0.005	0.094	−0.041	−0.071	0.044	−0.091	−0.078
A_0	0.005	1.000	0.124	0.046	0.049	0.038	0.136	0.002
E_T/p_T	0.094	0.124	1.000	0.183	−0.308	0.216	0.225	0.022
POS	−0.041	0.046	0.183	1.000	−0.158	0.041	0.146	−0.030
E_1/E	−0.071	0.049	−0.308	−0.158	1.000	−0.154	0.053	0.129
E_3/E	0.044	0.038	0.216	0.041	−0.154	1.000	0.123	0.015
ISO	−0.091	0.136	0.225	0.146	0.053	0.123	1.000	0.256
$WIDTH$	−0.078	0.002	0.022	−0.030	0.129	0.015	0.256	1.000

Table 10: *Correlation matrix of the discriminating variables of the signal electrons at low luminosity.*

	n_{TR}	A_0	E_T/p_T	POS	E_1/E	E_3/E	ISO	$WIDTH$
n_{TR}	1.000	−0.084	0.208	−0.020	−0.079	0.010	−0.052	0.127
A_0	−0.084	1.000	−0.021	−0.032	0.089	0.000	−0.001	−0.003
E_T/p_T	0.208	−0.021	1.000	−0.075	−0.178	−0.066	0.030	0.246
POS	−0.020	−0.032	−0.075	1.000	−0.058	0.082	0.089	−0.143
E_1/E	−0.079	0.089	−0.178	−0.058	1.000	−0.254	0.256	0.053
E_3/E	0.010	0.000	−0.066	0.082	−0.254	1.000	−0.057	−0.160
ISO	−0.052	−0.001	0.030	0.089	0.256	−0.057	1.000	0.040
$WIDTH$	0.127	−0.003	0.246	−0.143	0.053	−0.160	0.040	1.000

Table 11: *Correlation matrix of the discriminating variables of the pions from background events at low luminosity.*

B Effect of TRT on performance

This section is devoted on the individual effect of the high threshold hits variable n_{TR} on the performance of the algorithm. The likelihood method is used.

As described in Tab. 3 the n_{TR} variable is the most discriminating variable used by the algorithm. Fig. 13 shows the pion rejection as a function of the electron identification efficiency, ε_e, without the use of the n_{TR} variable. In the $H \rightarrow b\bar{b}$ sample, for low luminosity case and $\varepsilon_e = 80$ %, the pion rejection is 30, about 5 times smaller than when using it. The second observation, is that rejection is then similar for the different luminosity scenarios. Thus, the difference of performance observed in Fig. 7 and Tab. 4 for the different luminosity scenarios, in particular for the high luminosity case, is essentially due to the loss of rejection power of the n_{TR} variable.

Fig. 14 shows performance of the algorithm in terms of rejection of u-jets (left panel) and c-jets (right panel) as a function of the b-tagging algorithm efficiency ε_b^{alg}, for different luminosity scenarios.

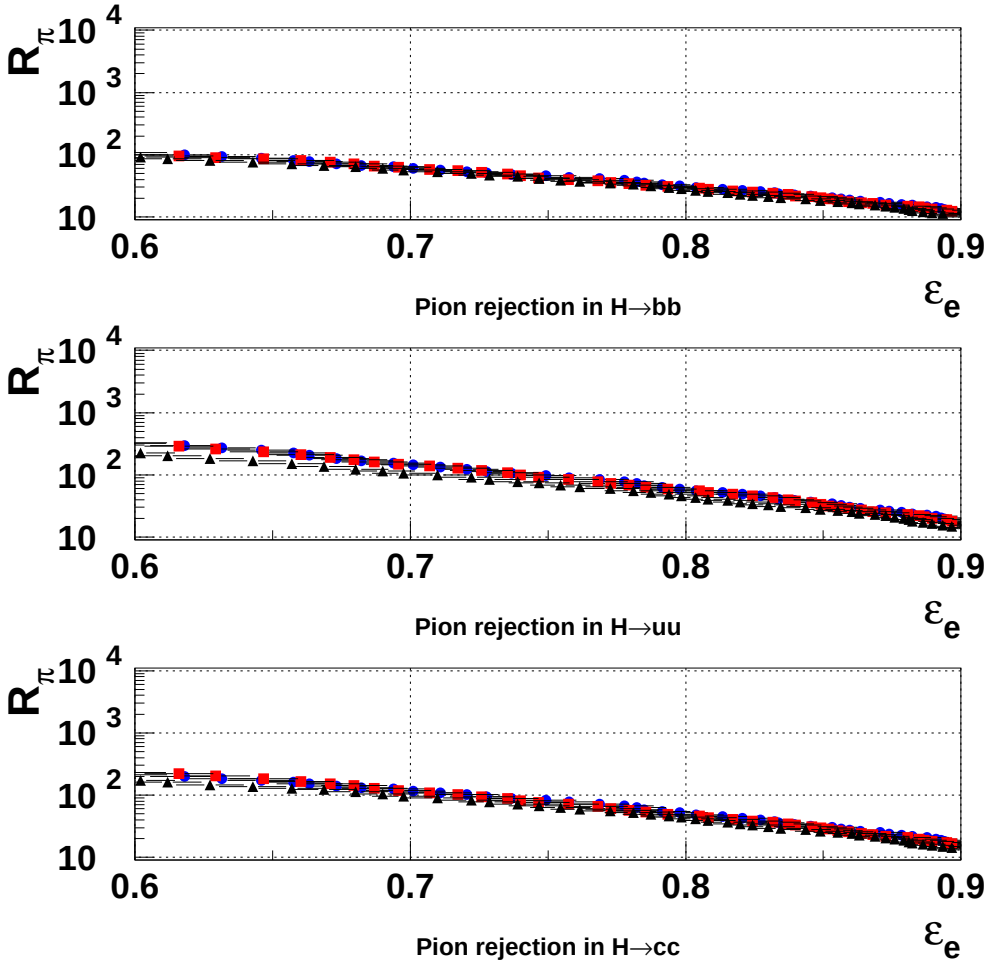

Figure 13: *Rejection of pions, R_π, in various samples as a function of the efficiency for identifying signal electrons, ε_e, for data with no pile-up (circles), low luminosity (squares) and high luminosity (triangles). The n_{TR} variable is not used.*

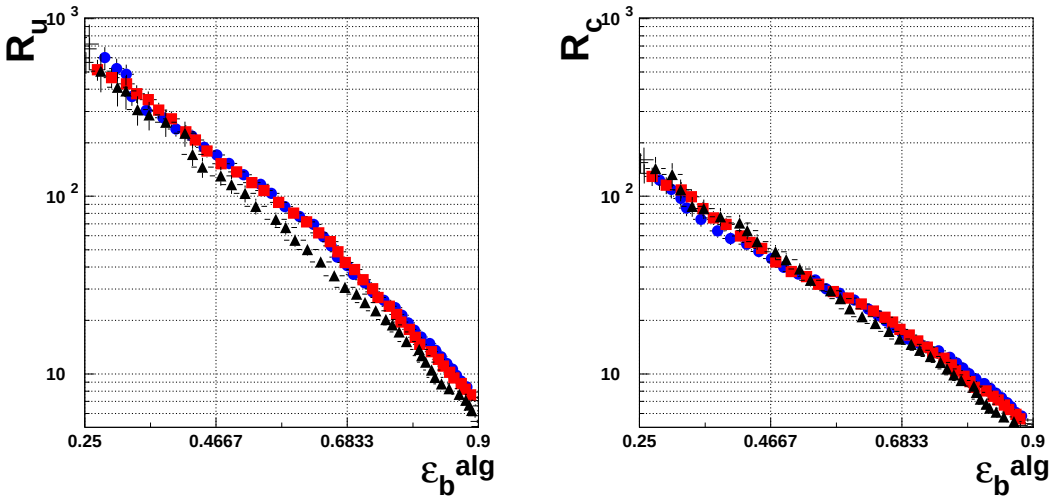

Figure 14: *Jet rejection, R_{jet}, for u-jets (left panel) and c-jets (right panel) as a function of the b-tagging algorithm efficiency ε_b^{alg}. Results presented are shown for events with no pile-up (circles), for the low (squares) and the high (triangles) luminosity scenario. The n_{TR} variable is not used.*

C Neural network trained to equally reject pions from $H \to b\bar{b}$ $H \to u\bar{u}$ and $H \to c\bar{c}$ processes.

For the neural network training the equal numbers of events from the processes $H \to b\bar{b}$, $H \to u\bar{u}$ and $H \to c\bar{c}$ were used. The network achitecture remains unchanged. The network therefore is no longer specially optimised against pions from $H \to u\bar{u}$. The resulting network suppresses better pions from $H \to b\bar{b}$ than the network described in section 3.4. Fig. 15 and Tab. 12 show the results for various luminosities. Tab. 13 compares the rejection for 80% efficiency between networks trained on different data samples.

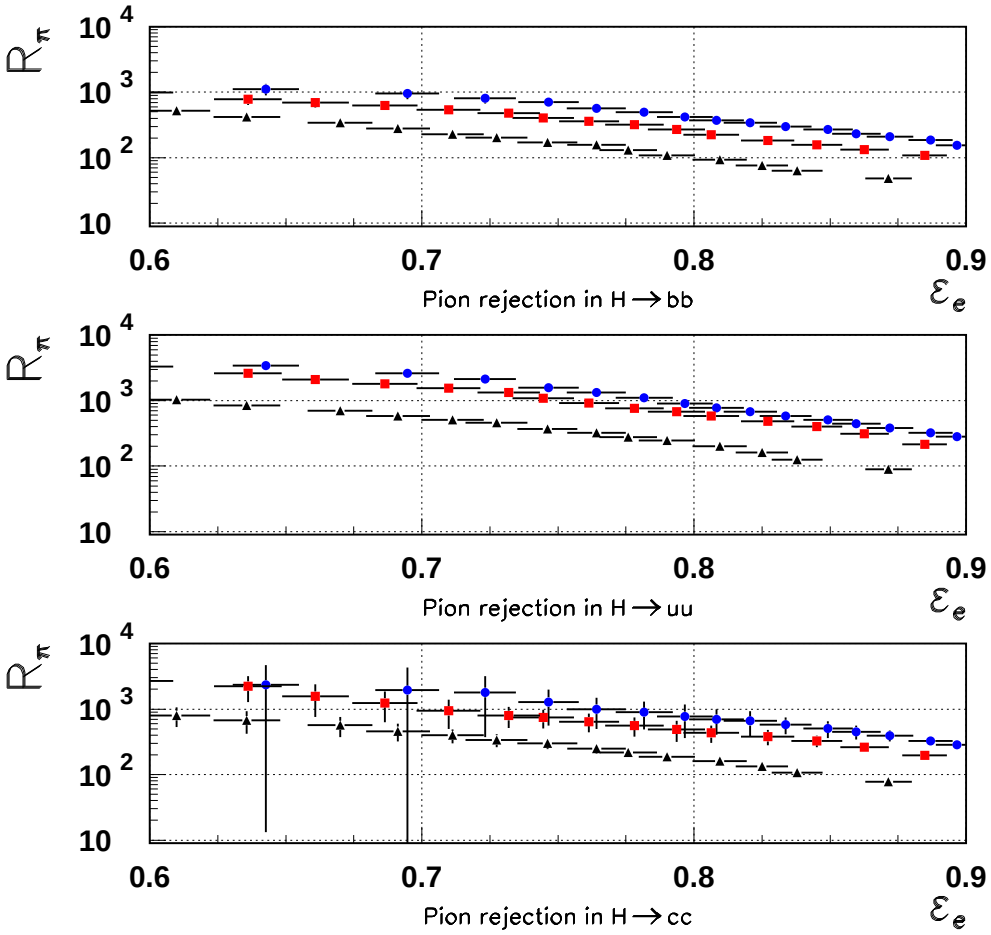

Figure 15: *Rejection of pions R_π for tracks in various samples as a function of the efficiency, ε_e, for data with no pile-up (circles), low luminosity (squares) and high luminosity (triangles) using a neural network. The network is trained on equal samples of $H \to b\bar{b}$, $H \to u\bar{u}$ and $H \to c\bar{c}$.*

luminosity	ε_e	R_π		
		$H \to bb$	$H \to u\bar{u}$	$H \to c\bar{c}$
	90 %	148 ± 11	268 ± 11	263 ± 26
no	80 %	389 ± 46	851 ± 64	757 ± 315
	70 %	895 ± 163	2429 ± 311	1863 ± 2340
	90 %	88 ± 5	165 ± 5	156 ± 13
low	80 %	254 ± 25	653 ± 41	472 ± 156
	70 %	558 ± 82	1683 ± 169	1068 ± 490
	90 %	32 ± 1	57 ± 1	50 ± 3
high	80 %	100 ± 5	214 ± 10	174 ± 19
	70 %	246 ± 22	554 ± 42	447 ± 124

Table 12: *Rejection of pions in various samples for varying identification efficiency ε_e and different luminosities using a neural network trained on equal samples of $H \to b\bar{b}$, $H \to u\bar{u}$ and $H \to c\bar{c}$.*

luminosity	training sample	ε_e	R_π		
			$H \to bb$	$H \to u\bar{u}$	$H \to c\bar{c}$
	standard	80 %	268 ± 19	1245 ± 114	877 ± 181
no	$H \to bb$, $H \to u\bar{u}$, $H \to c\bar{c}$	80 %	389 ± 46	851 ± 64	757 ± 315
	standard	80 %	205 ± 13	854 ± 61	523 ± 85
low	$H \to bb$, $H \to u\bar{u}$, $H \to c\bar{c}$	80 %	254 ± 25	653 ± 41	472 ± 156
	standard	80 %	99 ± 4	261 ± 14	214 ± 20
high	$H \to bb$, $H \to u\bar{u}$, $H \to c\bar{c}$	80 %	100 ± 5	214 ± 10	174 ± 19

Table 13: *Rejection of pions in various samples for different luminosities using a neural network trained on a standard sample (mostly $H \to u\bar{u}$ and some $H \to c\bar{c}$) and on equal samples of $H \to b\bar{b}$, $H \to u\bar{u}$ and $H \to c\bar{c}$.*

References

[1] Derue F., Kaczmarska A., *Soft electron identification and b-tagging with DC1 data*, 2004, ATLAS-PHYS-2004-026

[2] ATLAS Collaboration, *ATLAS Data Challenge DC1*, ATL-SOFT-2003-012

[3] Sjöstrand T. et al., Comp. Phys. Comm. **82** (1994) 74; Sjöstrand T., *Pythia 6.206*, LU TP 01-02 [hep-ph/0108264] 2002

[4] R. Brun et al., http://root.cern.ch

[5] Application Software Group, CERN program library long writeup W5013 http://wwwasd.web.cern.ch/wwwasd/geant/index.html

[6] http://atlas.web.cern.ch/Atlas/GROUPS/SOFTWARE/DOCUMENTS/simulation.html#atlsim

[7] Corréard S. et al., *b-tagging with DC1 data*, 2004, ATL-PHYS-2004-006

[8] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press 1999

[9] A. Zell et al., http://www-ra.informatik.uni-tuebingen.de/SNNS/

[10] Wolter M., Kaczmarska A., *Combining b-tagging methods using a neural network approach*, 2000, ATLAS-INDET-2000-023.

[11] S. Versillé and F. Le Diberder, *Estimating the tagging performances: the separation*, 1998, BaBar Tagging Note 11

[12] Versillé S., *La violation de CP dans Babar: étiquetage des mésons B et étude du canal* $B \to 3\pi$, 1999, Thèse de doctorat, Université Paris XI.

[13] ATLAS collaboration, *TDR ATLAS inner detector vol 1*, 1997, CERN/LHCC 97-16, ATLAS-TDR-4