
Machine learning techniques for background discrimination at the ATLAS experiment

CERN summer student report

Author
Morten Kuhlwein

Supervisors
Michela Biglietti
Luigi Longo



23rd of August, 2019

Abstract

The use of artificial neural network in discrimination between signal and background is investigated for two different processes in the ATLAS detector. For the WH signal against the $WZ/W\gamma^*$ background the neural network shows improvement over the boosted decision trees previously used in Aad et al. [1]. Also for the pair produced stop quark process with two leptons in the final state is achieved good discrimination between the signal and the background of fake/non-prompt leptons.

Introduction

The purpose of this project is to investigate the use of machine learning, in particular artificial neural networks, to discriminate between signal and background for different processes at the ATLAS detector at CERN.

Artificial neural networks are a category of machine learning models based on collections of connected nodes or "neurons". We will consider fully connected feed forward networks, which are networks organized in layers. Each neuron takes a linear combination of the outputs on the previous layer as inputs, then gives some output in accordance to the activation function of that neuron. Such a network is visualized on figure 1. The first layer is the "input"

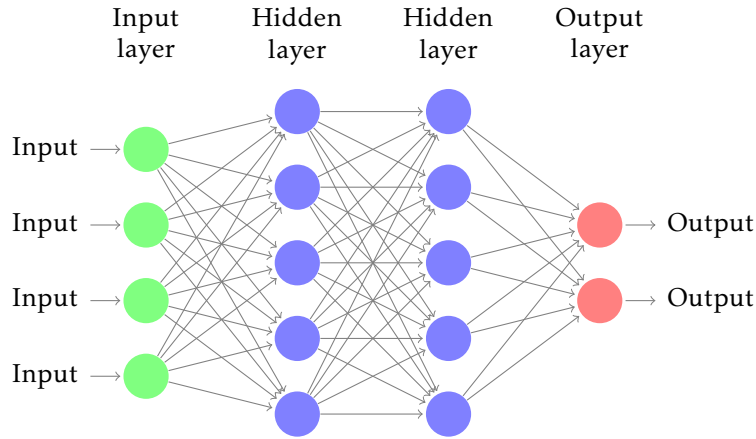


Figure 1: A fully connected feed forward network. Each neuron is connected to all neurons on the previous layer,

layer, which consists of some number of features for each data point/event. The last layer is the "output" layer, which outputs the result. The layers in between are called hidden layers. In the case of logistic regression, typically the softmax function is used as activation function for the output layer. A popular activation function for the hidden layers is the Rectified Linear Unit (ReLU), although there are many alternatives.

For the network to give good results, the weights connecting each neuron to the outputs on the previous layer must have some appropriate value. The process of optimizing these weights is called training. In order to train the network, one needs to define how "good" the output is for a given sample, this

is done by defining some loss function that can be optimized. Since this project concerns only binary classification problems, we will use binary cross entropy as our loss function, as it takes into account not only if an event is correctly classified but also how certain the network is.

To optimize the loss function during training we will use stochastic gradient descent with momentum. This introduces a couple of hyperparameters, parameters that are not determined through training, but must be chosen manually. The first is the learning rate, the size of the step taken in the direction of the gradient for each iteration. Without momentum each step is independent and determined only from the gradient. When using momentum each step is a linear combination of the previous step and the new step calculated from the gradient.

Choosing good hyperparameters is done through grid search – basically just brute-forcing a manually defined interval of hyperparameters. Since the network might overfit the data, the performance should not be evaluated using the same data that was used for training. For the purpose of estimating the out-of-sample error, the complete data set is split into a training set and a testing/validation set.

For binary classification in particular, one can evaluate the performance of the network using the receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate against the true negative rate by varying the discrimination threshold. In physics terms this corresponds to plotting the signal efficiency against the background rejection. The area under the curve (AUC) can then be interpreted as the probability of the classifier ranking a randomly chosen signal event higher (more signal like) than a randomly chosen background event.

In order to avoid overfitting, two different regularization techniques are used. The first is early stopping, which simply consists of only saving the model with the best validation loss. The second is dropout, a technique where individual neurons are temporarily removed or "dropped out" each iteration during training. This also introduces a new hyperparameter – the dropout rate [2, 4].

The models are implemented using ROOT and TMVA with Keras. Tensorflow is used as backend.

Part I

The coupling of the Higgs boson to the weak bosons can be investigated through Higgs boson production in association with a W or Z boson. Particularly the WH mode with a subsequent $H \rightarrow WW^*$ decay couples only to W bosons. This part of the project concerns the W associated Higgs boson production with three charged leptons (electrons or muons) in the final state:

$$WH \rightarrow W(\ell\nu)WW^* \rightarrow \ell\nu\ell\nu\ell\nu.$$

The leading-order Feynman diagram for the WH production process with three charged leptons in the final state is shown on figure 2.

The most prominent background processes with three prompt leptons are the $WZ/W\gamma^*$ production and the ZZ^* production with an undetected lepton. Since the leptons are both prompt and isolated, these backgrounds cannot be

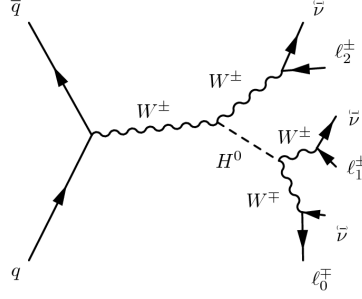


Figure 2: Tree-level Feynman diagram for WH associated production with three charged leptons in the final state.

reduced using tight lepton identification. However, these backgrounds are characterised by having either one or two pairs of Same Flavor Opposite Sign (SFOS) leptons. Other important background processes are $Z\gamma$, $Z + \text{jets}$ in the case of misidentified or non-prompt leptons and also other processes with three prompt leptons in the final state, such as WWW and $t\bar{t}V$.

Since the signal-to-background ratio differs between 0SFOS events and events with at least one SFOS pair, these should be handled separately. Since the dominant background consists of $Z \rightarrow \ell\ell$ decays, events with either one or two SFOS pairs will be denoted as the Z -dominated category while events with zero SFOS pairs is known as the Z -depleted category. In the Z -depleted signal region the background can be mostly reduced only through lepton identification and isolation, since it comes mostly from top quark processes. This is the case for about 1/4 of the signal. The Z -dominated signal region contains about 3/4 of all signal events, and here the background is harder to separate from the signal [1, 6].

In particular, this part of the project aims to use machine learning, specifically artificial neural networks, to distinguish the signal from the $WZ/W\gamma^*$ background in the Z -dominated signal region. The same Monte-Carlo samples are used as in Aad et al. [1].

For the WH channel, exactly three isolated leptons with $p_T > 15\text{GeV}$ are required with a total charge of ± 1 . The leptons are labelled as follows: The lepton with unique charge is labelled ℓ_0 , the lepton closest to ℓ_0 in angular distance ΔR is labelled ℓ_1 , and the remaining lepton is labelled ℓ_2 . For signal events, ℓ_0 and ℓ_1 will originate from the $H \rightarrow WW^*$ with probabilities of 99% and 85% respectively. In order to only select final states with neutrinos in it, E_T^{miss} is required to be at least 30 GeV. The invariant masses $m_{\ell\ell}$ of all SFOS pairs must satisfy the Z -veto selection, $|m_{\ell\ell} - m_Z| > 25\text{GeV}$, this should suppress $WZ/W\gamma^*$ events.

The network is trained using the WH process as the signal and the $WZ/W\gamma^*$ process as the background. Only events with no SFOS are used for the signal, so as to make sure the sample for training is independent. This might introduce biases and should be investigated further. For the background, we already have an independent sample.

After performing the previously mentioned cuts, we are left with $\sim 2.1 \times 10^6$ background events and $\sim 2.6 \times 10^5$ signal events. From these events, 20% is reserved for testing and validation while the remaining 80% is used for

training the neural network.

As features for our model we can use the most basic features available such as the transverse momentum $p_T^{\ell_i}$, the azimuthal angle φ_{ℓ_i} and the pseudorapidity η_{ℓ_i} for each of the three leptons respectively. Furthermore we have the missing transverse energy E_T^{miss} , the azimuthal angle of the missing transverse momentum φ_{miss} , the magnitude of the vectorial sum of the lepton transverse momenta $|\sum p_T^{\ell_i}|$ and the number of jets N_{jets} . From these basic features it is

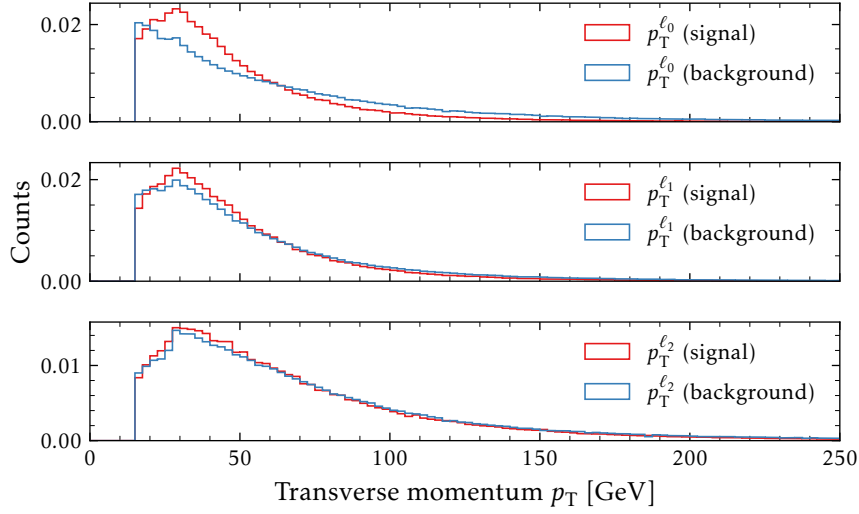


Figure 3: The normalized distribution of the transverse momentum $p_T^{\ell_i}$ of each lepton for both the signal and the background.

possible to construct more advanced features of interest, such as:

- Angular separation $\Delta R_{\ell_i \ell_j} = \sqrt{\Delta \varphi_{\ell_i \ell_j}^2 + \Delta \eta_{\ell_i \ell_j}^2}$ between any two leptons ℓ_i and ℓ_j . In particular the two leptons from the Higgs boson are expected to have a low angular separation because of the spin of the Higgs boson.
- Invariant mass $M_{\ell_i \ell_j}$ between any two leptons ℓ_i and ℓ_j .
- Transverse mass of the W boson $m_T^W = \sqrt{2p_T^{\ell_i} E_T^{\text{miss}} (1 - \cos \Delta \varphi(\ell_i, p_T^{\text{miss}}))}$ for any lepton ℓ_i assumed to originate from the W boson.
- Transverse mass of the W boson $M_{T,W}^W$, calculated by identifying the lepton that grants transverse mass closer to the expected value. The expected value can be obtained by inspecting 1 SFOS events in the background data, where the lepton not belonging to the SFOS pair will originate from the W boson.
- Transverse mass of the W boson $M_{T,Z}^W$, calculated by identifying the lepton which does not belong to the lepton pair with invariant mass closer to the Z boson mass.

Plots showcasing the distribution of transverse momentum p_T and angular separation ΔR can be seen on figure 3 and 4 respectively.

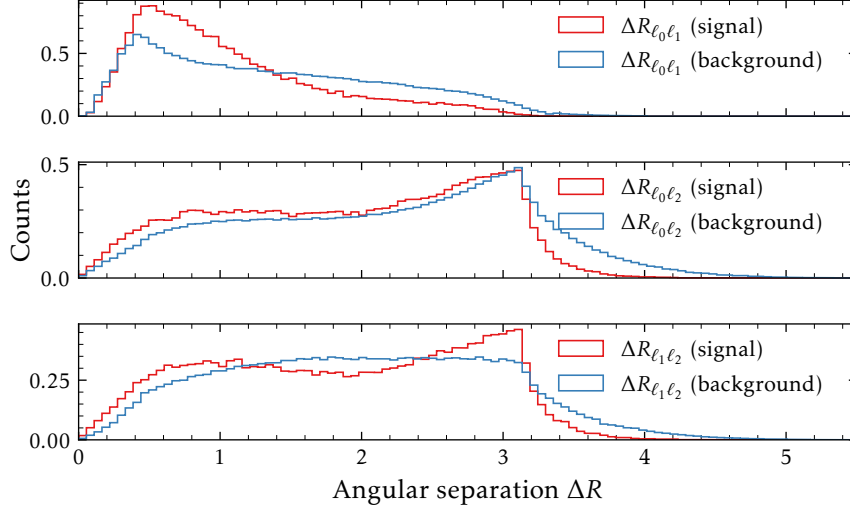


Figure 4: Normalized distributions of the angular separation $\Delta R_{\ell_i \ell_j}$ for each possible pair of leptons ℓ_0, ℓ_1 , and ℓ_2 for both the signal and the background.

As for the structure of the neural network, ReLU is used as activation function for the hidden layers while the output layer uses Softmax as activation function. The loss function is binary cross entropy, which is suited for binary classification problems. The various hyperparameters are optimized using grid search. These parameters include the number of layers, the number of neurons in each layer, learning rate, dropout rate and momentum. In addition to this, we can also experiment with the selection of features used to train the network.

In the end, the best configuration is found to be four layers of 128 neurons each, a learning rate of $\text{lr} = 0.01$, dropout rate of 15% and a momentum value of 70%. All features are used. It was also attempted to use different subsets of the features since the network might be able to more efficiently learn the advanced features by itself, but it was found that using all features gave better results. The output distribution of the trained neural network can be seen on figure 5. The network is largely able to separate the signal from the background. The distributions are skewed towards the background with a "flat" signal distribution. This is caused by the abundance of background events used for training, the background to signal ratio is almost 10:1. There is no significant difference between the output distribution of the test set and the training set, indicating that there is no overfitting.

In order to evaluate the performance of the model we can construct the ROC curve for the network. The ROC curve is shown on figure 6 along with the ROC curve from Aad et al. [1]. The AUC for the final model is 0.929, an improvement from the value of 0.856 that is obtained using the BDT model from Aad et al. [1].

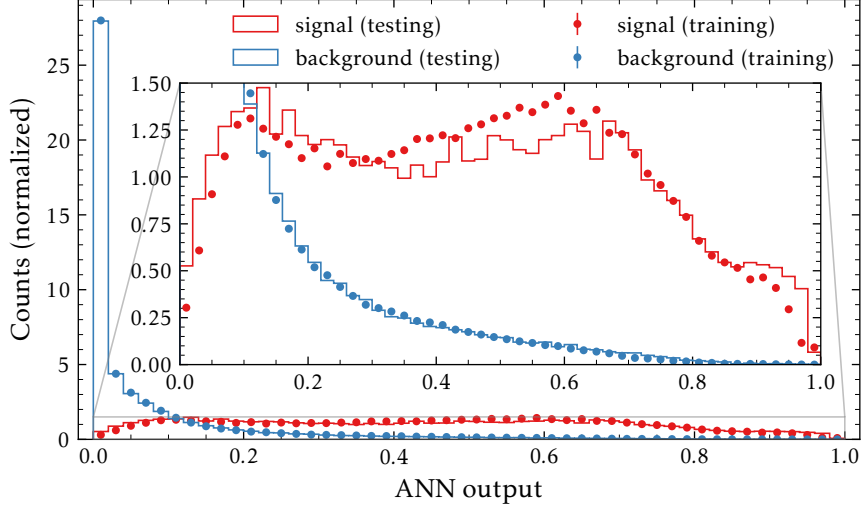


Figure 5: The output distribution of the neural network, the network is trained to output 0 for background events and 1 for signal events. The network is largely able to separate the signal from the background.

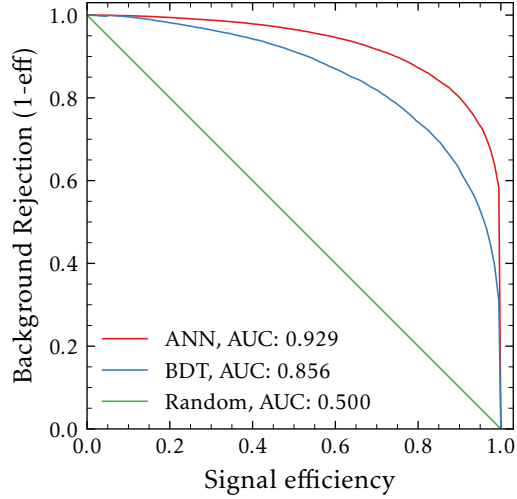


Figure 6: ROC curves for the artificial neural network (ANN) trained in this project and the boosted decision trees (BDT) from Aad et al. [1], distinguishing the WH signal from the $WZ/W\gamma^*$ background. The ANN shows improvement over the BDT.

Part II

Supersymmetry is one of the most interesting extensions of the Standard Model since provides possible solutions to unanswered questions in the Standard Model. For instance it gives a solution to the hierarchy problem and also viable dark matter candidates. The Minimal Supersymmetric Standard Model predicts a new bosonic partner for each fermionic Standard Model particle and vice versa. These supersymmetric particles would have identical quantum numbers compared to their Standard Model partners, with the exception of spin – differing by half a unit.

One can define R -parity as $R = (-1)^{3(B-L)+2S}$ with B being baryon number and L being lepton number. All Standard Model particles will then have R -parity of +1 while supersymmetric particles will have R -parity of -1. If R -parity is conserved, that would imply the existence of a stable lightest supersymmetric particle, a possible candidate for dark matter. In particular the lightest supersymmetric particle should be electrically neutral. A favored candidate is the lightest neutralino, with neutralinos $\tilde{\chi}_{i=1,2,3,4}^0$ being mixed states for the electroweak gauginos and higgsinos [3, 5].

This analysis considers the light supersymmetric partner to the top quark, called the stop squark or top squark, $\tilde{t}$, and its decay to this lightest neutralino. The available decay mode(s) depends on the splitting in mass between the stop and the neutralino, $\Delta m = m_{\tilde{t}} - m_{\tilde{\chi}_1^0}$:

- If $\Delta m > m_t$ then the stop can decay via $\tilde{t} \rightarrow t + \tilde{\chi}_1^0$.
- If $m_b + m_W < \Delta m < m_t$ then the top is too massive to be produced in the previous decay. Instead, the stop can decay via $\tilde{t} \rightarrow b + W + \tilde{\chi}_1^0$ through an off-shell top.
- If $\Delta m < m_b + m_W$ then the stop can decay via $\tilde{t} \rightarrow b + f + \bar{f} + \tilde{\chi}_1^0$ with f and $\bar{f}$ now originating from an off-shell W . This decay is also known as a four-body $\tilde{t}$ decay.

Considering a splitting mass of $\Delta m = 15\text{GeV}$, the stop can decay via $\tilde{t} \rightarrow b + f + \bar{f} + \tilde{\chi}_1^0$, as shown on figure 7.

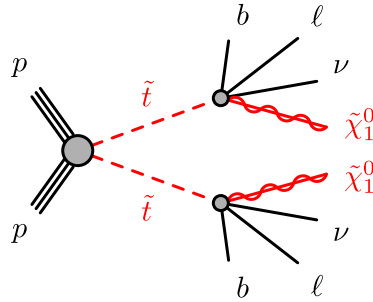


Figure 7: Stop quark pair production. Each stop decays via $\tilde{t} \rightarrow b + f + \bar{f} + \tilde{\chi}_1^0$, also known as a four-body $\tilde{t}$ decay.

This part of the project aims to use neural networks to discriminate the signal of this reaction from a background of fake / non-prompt leptons caused

by jets misidentified as leptons, and semileptonic hadron decays. Monte-Carlo samples are used for both the signal and the background. For the signal is used stop quark pair production with subsequent four-body decays. The splitting mass is $\Delta m = 15 \text{ GeV}$, with $m_{\tilde{t}}$ ranging from 260 GeV to 520 GeV . For the background is only used events enriched in fakes. To generate the background is used events from Wt , $t\bar{t}$, $t\bar{t}Z$, VV and $Z + \text{jets}$.

The leptons are both required to be well identified and isolated. The transverse momentum of the leptons should be at least 4 GeV but no more than 25 GeV for ℓ_1 or 20 GeV for ℓ_2 . The invariant mass of the leptons should be at least 10 GeV and the energy of the primary jet should be at least 150 GeV . To only select final states with neutrinos and neutralinos we require E_T^{miss} to be at least 250 GeV .

After performing the aforementioned cuts we are left with 4.0×10^3 signal events and 20.4×10^3 background events. Once again, 20% is reserved for testing and validation while the remaining 80% is used for training.

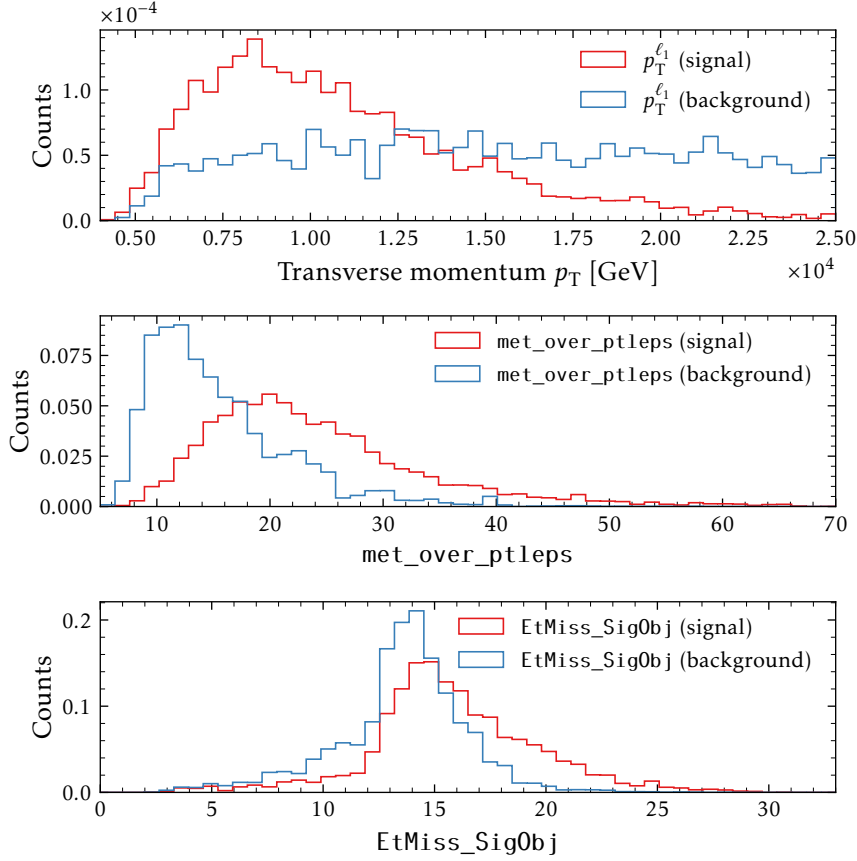


Figure 8: Normalized distributions for some of the features used.

The features used for training are:

- The transverse momentum for the two leptons, $p_T^{\ell_i}$.

- The pseudorapidity of the two leptons, η_{ℓ_i} .
- The azimuthal angle of the two leptons, φ_{ℓ_i} .
- The angular distance between the two leptons, ΔR .
- topoetcone20Lep – Sum of transverse energy of the calorimeter cells within a cone centered around the direction of the leptons, calculated using topological clusters.
- ptvarcone20Lep and ptvarcone30_TightTTVA_pt1000Lep – Variations on topoetcone20Lep using variable cone size depending on the transverse momentum of the lepton in question.
- met_over_meff – The missing transverse energy divided by $\sum p_T^{\ell_i} + E_T^{\text{miss}} + \sum p_{\text{jet}}$.
- met_over_ptleps – Missing transverse energy divided by $\sum p_T^{\ell_i}$.
- DPhiLepJet – Azimuthal angle between primary lepton and closest jet.
- EtMiss_Sig0jb – Missing transverse energy in units of calorimeters resolution.
- pb11mod – Vectorial sum of the missing transverse energy and the transverse momenta for the two leptons.
- DPhiMetPb11 – Azimuthal angle between missing transverse energy and Pb11.

Comparisons of distributions for select features are shown on figure 8.

Once again ReLU is used as activation function for the hidden layers while softmax is used for the output, the hyperparameters are optimized using grid search as well. This results in a model with 4 layers of 128, 64, 64, and 32 neurons respectively. Additionally is found a learning rate of $\text{lr} = 10^{-4}$, a dropout rate of 15% and a momentum value of 95%. The output distribution can be seen on figure 9. Once again the network is largely able to separate the signal from the background. And much like in the previous part, the distributions are skewed towards the background because of the imbalance in the training set. The signal distribution does not even reach all the way up to 1, and there are also signs of slight overfitting, most likely caused by the small size of the dataset. This is not necessarily a problem, as the chosen model is still the one with lowest estimated out-of-sample error.

Using the output distribution of the test set, it is possible to construct the ROC curve shown on figure 10. The AUC for the model is 0.885.

Summary

Artificial neural networks have been applied to two different pairs of signal/background processes. For the WH signal vs the $WZ/W\gamma^*$ background, the neural network was able to archive a ROC AUC of 0.929 improving on the value of 0.856 obtained by Aad et al. [1] using boosted decision trees.

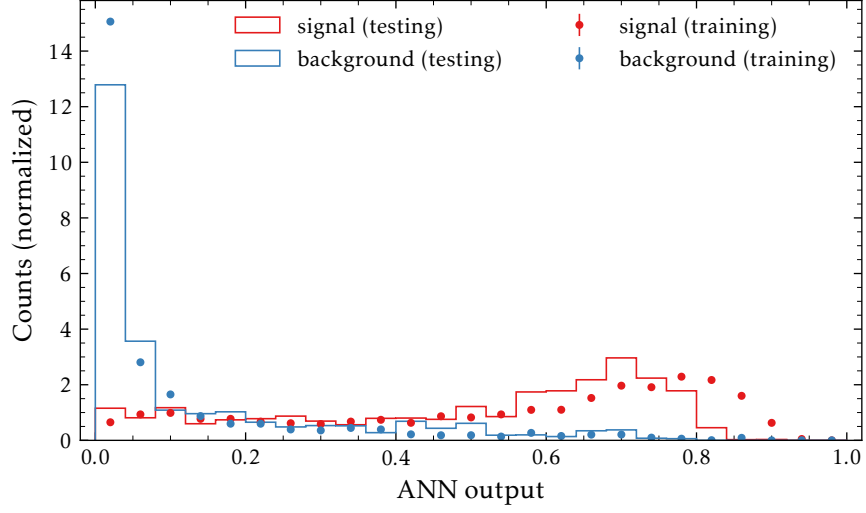


Figure 9: The output distribution of the neural network, events closer to 0 are classified as background while events closer to 1 are classified as signal. The distribution for the training set and for the testing set differ slightly, indicating slight overfitting.

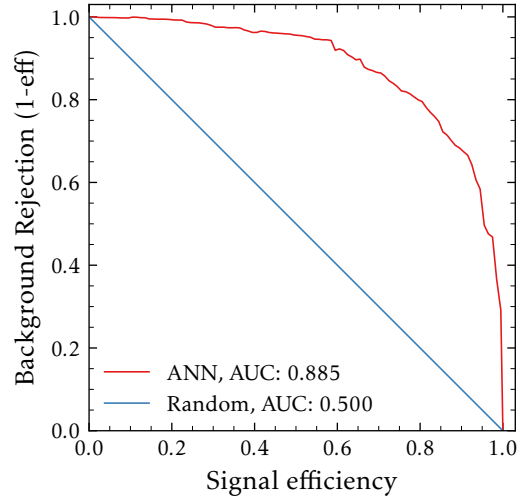


Figure 10: The ROC curve for the ANN when distinguishing the signal from the background of fake/non-prompt leptons.

Also for the pair produced stop quark process with two leptons in the final state is the neural network able to discriminate the signal from the background of fake/non-prompt leptons with an AUC of 0.885.

References

- [1] Georges Aad et al. “Measurement of the production cross section for a Higgs boson in association with a vector boson in the $H \rightarrow WW^* \rightarrow \ell \nu \ell \nu$ channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector”. In: (2019). arXiv: 1903. 10052 [hep-ex].
- [2] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012. ISBN: 1600490069, 9781600490064.
- [3] D.J. Antrim et al. “Search for pair produced top squark and Dark Matter in Final state with two leptons in p p collisions at $\sqrt{s} = 13$ TeV”. Unpublished. N.D.
- [4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016.
- [5] Luigi Longo. “Searches for charginos and neutralinos in ATLAS”. In: (July 2019). URL: <https://cds.cern.ch/record/2680901>.
- [6] Daniele Puddu. “Study of $(W/Z)H$ production using $H \rightarrow WW^*$ decays with the ATLAS detector at LHC”. PhD thesis. Roma Tre University, N.D. Chap. 4.