
# Search for the optimal neural network input for CMS T2tt SUSY data

Iuliia Sheleva, National Research Nuclear University MEPhI, Russia

September 5, 2018

**Abstract**

This work presents the study of the different inputs of the neural network which are applied to the separation of stop signal and $t\bar{t}$ background. First a short introduction to the SUSY and Machine learning in HEP is given. Then the variables that are used as input of the neural network are described. The correlation coefficients between the input variables are presented. Finally, Asimov significance for the different set of variables and the different number of layers of the neural network is shown.

# Contents

# 1 Introduction

The Standard Model (SM) has been extremely successful in the description of nature for the last decades. Despite its great success, the SM does not answer all questions, e.g. the masses of the leptons and quarks and their hierarchies are not explained, the question why there are three generations of them is not answered, not all forces of nature are included in the SM. There are also experimental results, for which the SM does not provide answers, e.g. neutrino oscillation experiments give clear evidence for non-zero neutrino masses; observations in astrophysics lead to the postulation of dark matter and dark energy. There are many experimental and theoretical arguments, which motivate the expectation that the SM is not the final answer, but rather an approximation (or effective theory) for the underlying, fundamental theory.

One very promising and arguably the best studied candidate for an extension of the Standard Model is Supersymmetry (SUSY). SUSY proposes a relationship between two basic classes of elementary particles: bosons (with integer-valued spin) and fermions (with half-integer spin). Each particle from one group would have an associated particle in the other, which is known as its "supersymmetric" partner, the spin of which differs by a half-integer. Standard particles and their "supersymmetric" partner are shown on Fig. 1. Since no SUSY partners of the fundamental SM particles have been observed yet, SUSY, if it exists at all, must be broken.

The model studied in this paper predicts the two lightest SUSY particles as the partner of top quark, $\tilde{t}$ ($m_{\tilde{t}}$ =600GeV), and the lightest supersymmetric particle LSP ($m_{LSP}$ =400GeV).

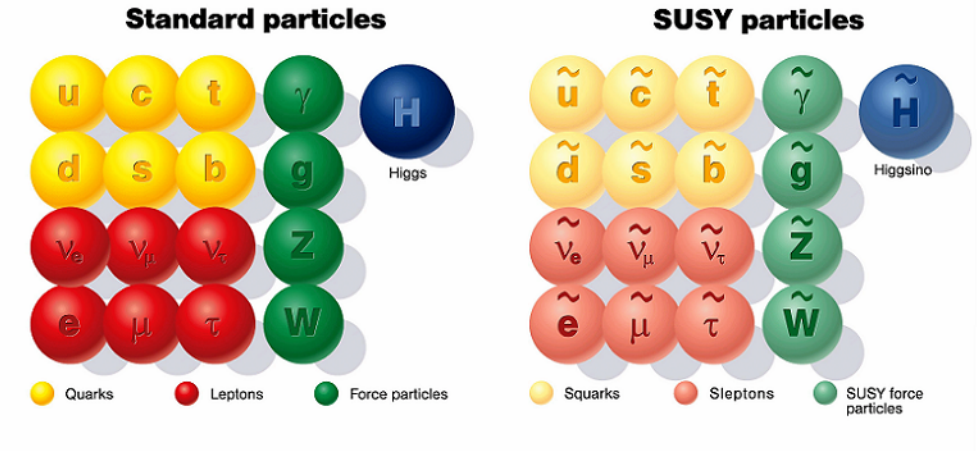


Figure 1: Table of the SM and SUSY particles

# 2 Machine learning in HEP

General goal of using Machine learning in HEP is separation between signal and background. Binary classification are used to separate signal and background. Every event is characterized by some event variables $\vec{x}$ and true class $y$. We can predict class $\hat{y}$ (signal or background) based on input variables $\vec{x}$.
We use supervised learning and use a neural network as Machine learning algorithm.
A neural network (NN) consist of neurons which distributed into layers. Each neuron in a certain layer is connected with all neurons in the previous layer. Neuron takes a input $x$ and takes the weighted sum $z = \sum_i x_i w_i + b$, where the bias b is an extra parameter. Then it pass through activation function. The structure of a NN is shown in Fig. 2.
As activation function we use ReLU function which is shown in Fig. 3. Difference between NN output $\hat{y}$ and true class $y$ measures through loss function. As a loss function we use cross entropy:

$$C = -\frac{1}{n}\sum_x [y \ln \hat{y} + (1-y)\ln(1-\hat{y})], \tag{1}$$

where $n$ is the size of the test set.
The training of a NN consist of minimizing a loss function.

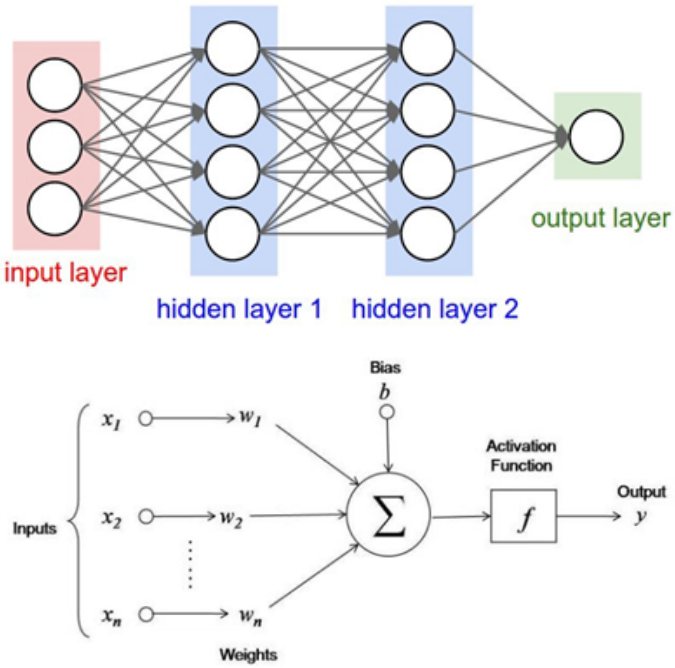


Figure 2: The structure of a neural network

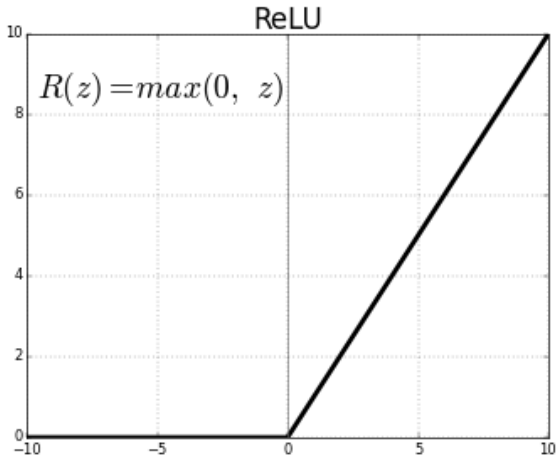


Figure 3: ReLU function

# 3 Analysed data

## 3.1 Signal

We consider signal is due to stop pair production, which decays into top and LSP. The top quark decays into a b quark and W boson. W boson can either decays into a lepton + neutrino or quark-antiquark pair. The corresponding diagram is shown on Fig. 4.

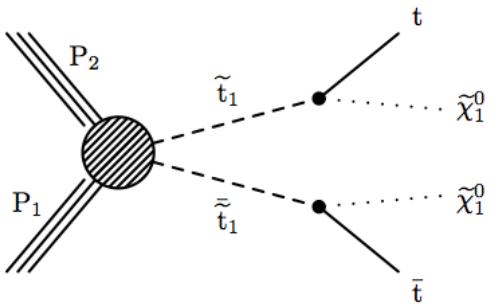


Figure 4: Signal decay (stop)

## 3.2 Background

As background, we consider the dominant process of top-antitop ($t\bar{t}$) production. The top quark decays into a b quark and W boson as in the signal case. The corresponding diagram is shown on Fig. 5

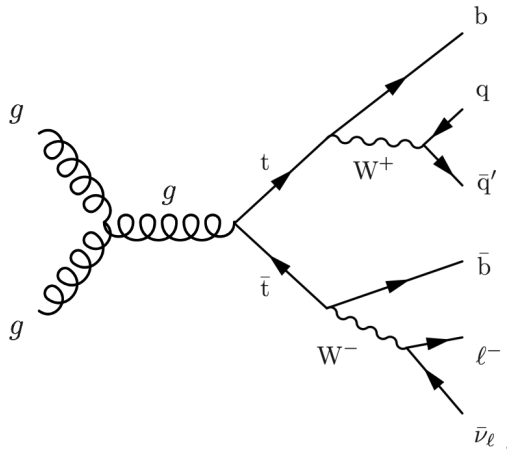


Figure 5: Background decay ($t\bar{t}$)

## 3.3 Data

For our study we use data, which was generated via Monte Carlo simulations. We consider a top squark mass of 600 GeV and a LSP mass of 400 GeV. PYTHIA8 [1] is used for the event simulation and DELPHES3 [2] is used to model of the CMS detector. We require at least one lepton with transverse momentum $p_T$ >30GeV. Each event must contain at least four jets with $p_T$ >40GeV, at least one of the jets must be tagged as originating from a bottom quark. More detailed description of used data is given in [3].

# 4 Choice of variables

The choice of the input variables is important for training of the neural network.
The variables can be divided in two subsets:

- Low level variables: consist of basic properties of the reconstructed physics objects measured directly by the detector
- High level variables: constructed from the low level variables

List of the variables used in this analysis is presented in Table 1.
This variables are used as input of the neural network.

## 4.1 Variables definitions

### 4.1.1 Low level variables

As low level variables we have:

- the pseudorapidity $\eta_l$, transverse momentum $p_{T,l}$ and azimuthal angle $\phi_l$ of the selected lepton

Table 1: Table of the used variables

| Low level | High level |
|---|---|
| $p_{T,l}$ | $H_T$ |
| $\eta_l$ | $m_T$ |
| $\phi_l$ | $m_{T2}^W$ |
| $p_{T,jet(1,2,3,4)}$ | |
| $\eta_{jet(1,2,3,4)}$ | |
| $\phi_{jet(1,2,3,4)}$ | |
| $p_{T,bjet1}$ | |
| $n_{jet}$ | |
| $n_{bjet}$ | |
| $\not{E}_T$ | |

- the pseudorapidity $\eta_{jet(1,2,3,4)}$, transverse momentum $p_{T,jet(1,2,3,4)}$ and azimuthal angle $\phi_{jet(1,2,3,4)}$of the 4 leading jets
- the transverse momentum $p_{T,bjet1}$ of the leading b jet
- the number of jets $n_{jet}$ and b-tagged jets $n_{bjet}$
- the missing energy in the transverse plane $\not{E}_T$

### 4.1.2 High level variables

As high level variables we have:

- the scalar sum $H_T$ of the module of the transverse momentum of the selected jets
- the transverse mass $m_T = \sqrt{2p_{T,l}\ \not{E}_T(1-\cos\Delta\phi(l,\not{E}_T))}$, where $\Delta\phi(l,\not{E}_T)$ is the azimuthal angle between the lepton and the $\not{E}_T$ vector
- the $m_{T2}^W$ is the minimal mass of the mother particle compatible with the event in the topology shown in Fig. 6 (we consider that one lepton has been lost). More detail description of this variable is given in [4].

Plots for the missing transverse energy $\not{E}_T$, the transverse mass $m_T$ and the first jet invariant mass for the background and for the signal can be found in Fig. 7, 8 and 9, respectively.

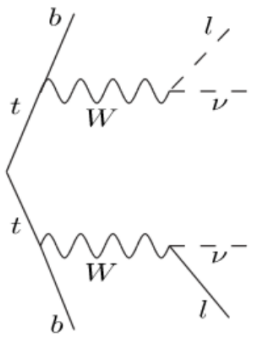


Figure 6: $t\bar{t}$ dileptonic decay.

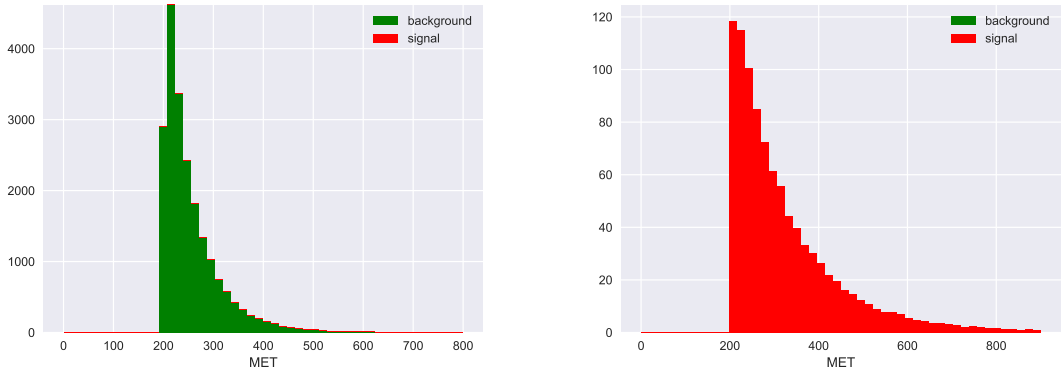


Figure 7: Distributions of the missing transverse energy $\not{E}_T$ for the background (on the left) and for the signal (on the right)

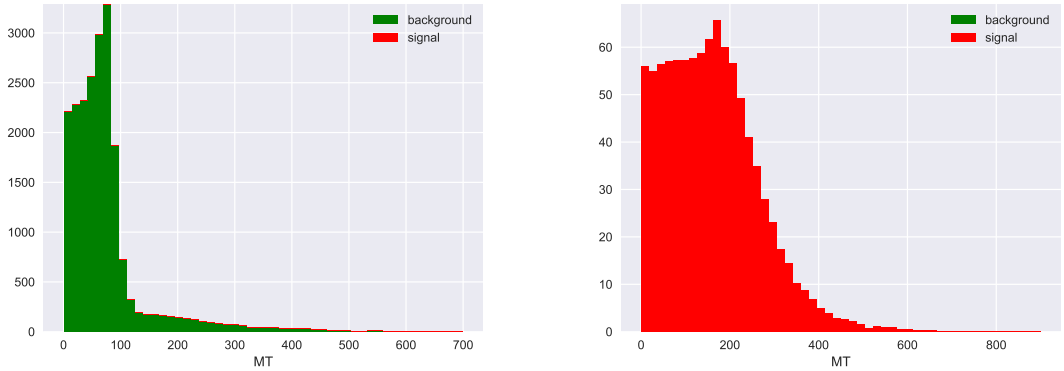


Figure 8: Distributions of the transverse mass $m_T$ for the background (on the left) and for the signal (on the right)

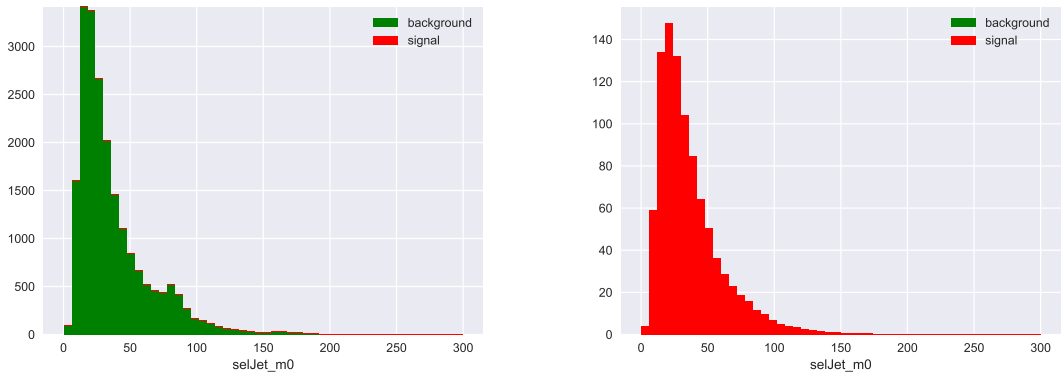


Figure 9: Distributions of the first jet invariant mass for the background (on the left) and for the signal (on the right)

# 5 Correlation matrixes

We can see a relation between variables through correlation matrix. Correlation coefficients are defined as Pearson correlation coefficients between two variables:

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2 \sum_{i=1}^{m}(y_i - \overline{y})^2}} \tag{2}$$

## 5.1 Correlation matrixes of the input variables

The correlation matrixes of the input variables for the signal and for the background are shown in Fig. 10.

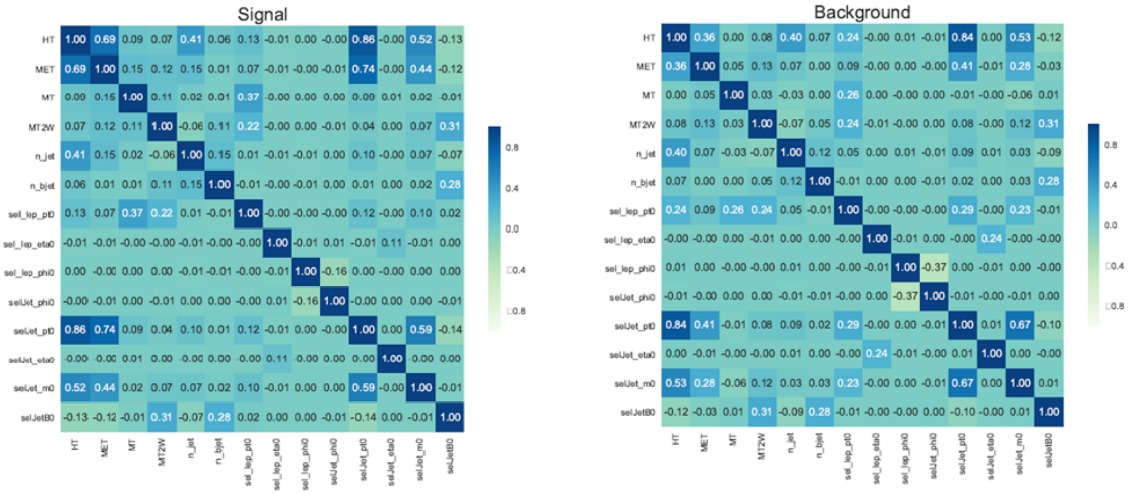


Figure 10: The correlation matrixes of input variables for the signal (on the left) and for the background (on the right).

As seen in these matrixes, there is strong correlation between $H_T$ and $\not{E}_T$; $H_T$ and $n_{jet}$; $H_T$ and $p_{T,jet1}$; $H_T$ and $m_{jet1}$; $\not{E}_T$ and $p_{T,jet1}$; $\not{E}_T$ and $m_{jet1}$.

## 5.2 Correlation matrix on output of classifier

The correlation matrix on output of classifier after training is shown in Fig. 11. $y_{pred}$ is signal or background on output of classifier, its value can be from 0 to 1.
As seen in this matrix, some variables don't correlate or very weakly correlate with $y_{pred}$ ($\phi_l$, $\phi_{jet1}$ and $m_{jet1}$). Next we can remove this variables.

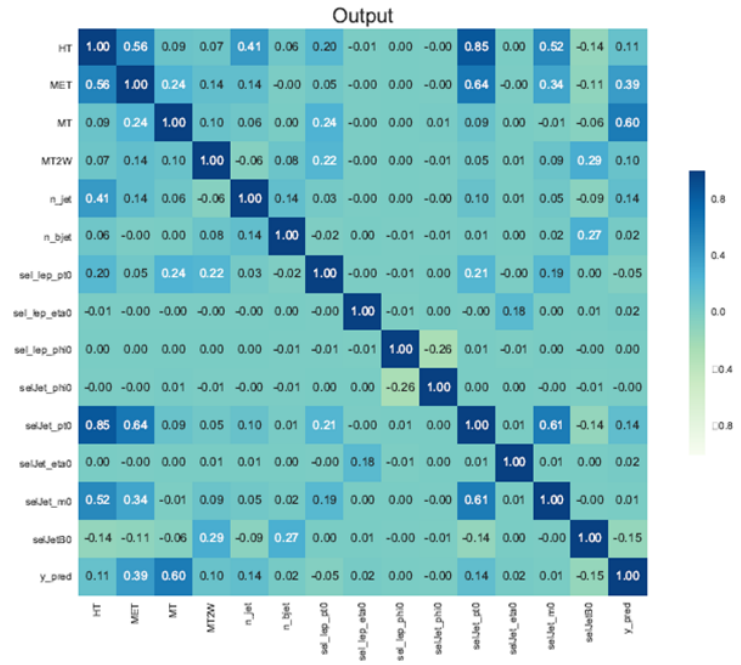


Figure 11: The correlation matrix on output of classifier after training.

### 5.3 Variables list

Full list of used variables in descending order of the correlation coefficient with $y_{pred}$ is presented in Table 2.

## 6 Asimov significance

To study the performance of the implemented algorithm we use Asimov significance. Asimov significance is an approximate form of discovery significance for a Poisson-distributed background and signal (see [5]):

$$Z_A = [2((s+b)\ln[\frac{(s+b)(b+\sigma_b^2)}{b^2+(s+b)\sigma_b^2}] - \frac{b^2}{\sigma_b^2}\ln[1+\frac{\sigma_b^2 s}{b(b+\sigma_b^2)}])]^{1/2}, \quad (3)$$

where s - signal events and b - background events with background uncertainty $\sigma_b$.
The error on the Asimov significance is calculated based on the Poisson uncertainty of the background and signal counts.
We want to find out how the choice of input parameters of NN affects the significance. We consider following cases:

- The different variables(low level, high level, low level and high level) and the different number of layers

Table 2: Full list of used variables

| variable | $C(var, y_{pred})$ | variable | $C(var, y_{pred})$ |
|---|---|---|---|
| $m_T$ | 0.596051 | $\eta_{jet4}$ | 0.015207 |
| $MET$ | 0.389345 | $\eta_l$ | 0.012802 |
| $B0_{jet}$ | -0.154295 | $\eta_{jet1}$ | 0.009998 |
| $p_{T,jet1}$ | 0.141983 | $m_{jet1}$ | 0.008761 |
| $n_{jet}$ | 0.137892 | $p_{T,jet2}$ | -0.007572 |
| $H_T$ | 0.112880 | $\phi_{MET}$ | -0.007009 |
| $m_{T2}^W$ | 0.089710 | $\phi_{jet3}$ | -0.004985 |
| $p_{T,l}$ | -0.056083 | $\phi_{jet2}$ | 0.003581 |
| $m_{jet2}$ | 0.044926 | $p_{T,jet3}$ | -0.003479 |
| $m_{jet3}$ | 0.039461 | $\phi_{jet1}$ | 0.003127 |
| $m_{jet4}$ | 0.034022 | $\eta_{jet2}$ | 0.002225 |
| $p_{T,jet4}$ | 0.026675 | $\phi_{jet1}$ | 0.000996 |
| $\eta_{jet3}$ | 0.016211 | $\phi_{jet4}$ | -0.000309 |
| $n_{bjet}$ | 0.015554 | | |

- The variables for the different number (1-4) of leading jets and the variables with the different correlation coefficient with $y_{pred}$

## 6.1 Asimov significance for the different number of layers

Asimov significance for low level, high level, low level and high level variables and for the different number of layers is shown in Table 3, 4 and 5, respectively. Asimov significance for same events and same random seeds is presented.

Table 3: Asimov significance for the different number of layers for low level variables. The best significance is shown in green.

| $N_{layer}$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $s$ | 11.7 | 8.2 | 7.5 | 9.5 | 10.6 |
| $b$ | 16.6 | 9.7 | 4.2 | 8.3 | 6.9 |
| Asimov significance | $1.94 \pm 0.36$ | $1.94 \pm 0.42$ | $2.7 \pm 0.7$ | $2.38 \pm 0.52$ | $2.87 \pm 0.63$ |

Asimov significance is the best for the large number of the NN layers if we use low level variables. Opposite, Asimov significance is the best for the small number of the NN layers if we use high level and low level+high level variables.

Table 4: Asimov significance for the different number of layers for high level variables. The best significance is shown in green.

| $N_{layer}$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $s$ | 71.5 | 64.5 | 58.8 | 64.5 | 83.2 |
| $b$ | 29.1 | 25.0 | 27.7 | 26.3 | 36.1 |
| Asimov significance | $6.16 \pm 0.77$ | $6.22 \pm 0.81$ | $5.46 \pm 0.71$ | $6.03 \pm 0.78$ | $6.10 \pm 0.72$ |

Table 5: Asimov significance for the different number of layers for low level + high level variables. The best significance is shown in green.

| $N_{layer}$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $s$ | 108.1 | 97.1 | 67.4 | 56.0 | 52.0 |
| $b$ | 43.0 | 44.4 | 29.1 | 22.2 | 18.0 |
| Asimov significance | $6.68 \pm 0.72$ | $6.05 \pm 0.66$ | $5.89 \pm 0.75$ | $5.96 \pm 0.81$ | $6.29 \pm 0.90$ |

## 6.2 Asimov significance for the different set of variables

As shown in 5.2, some variables don't correlate or weakly correlate with $y_{pred}$. We try to remove variables for which correlation coefficient with $y_{pred}$ ($C(var, y_{pred})$) less than $\epsilon$, where $\epsilon$ is 0.01, 0.02 or 0.03. This can increase Asimov significance.
Results for the two NN layers are shown in Table 6. Notice, that $N_{jets}$ in Table 6 means no cut on number of jets, but variables of first jet ($p_{T,jet1}$, $\eta_{jet1}$, $\phi_{jet1}$), first and second jets etc.

Table 6: Asimov significance for the different set of variables. The best significance is shown in green.

| $N_{jets}$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| All variables | $6.30 \pm 0.70$ | $5.69 \pm 0.58$ | $6.94 \pm 0.87$ | $6.61 \pm 0.85$ |
| Var. for which $C(var, y_{pred}) > 0.01$ | $7.20 \pm 1.00$ | $7.00 \pm 0.85$ | $7.07 \pm 1.01$ | $6.31 \pm 0.88$ |
| Var. for which $C(var, y_{pred}) > 0.02$ | $6.75 \pm 0.95$ | $7.06 \pm 0.83$ | | $7.59 \pm 1.21$ |
| Var. for which $C(var, y_{pred}) > 0.03$ | | $7.23 \pm 1.07$ | | $7.36 \pm 1.15$ |

As seen in Table 6, the best Asimov significance is $7.59 \pm 1.21$ (for 4 jets variables with correlation coefficient $C(var, y_{pred}) > 0.02$).
Also in Table 7 Asimov significance for the different set of variables and different number of NN layers is presented.

Table 7: Results for Asimov significance for different set of variables. The best significance is shown in green.

| $N_{jets}$ $N_{layer}$ | | 1 | 2 | 3 |
|---|---|---|---|---|
| 1 | All variables | $6.59 \pm 0.85$ | $6.30 \pm 0.70$ | $6.08 \pm 0.66$ |
| | Variables for which $C(var, y_{pred}) > 0.01$ | $6.98 \pm 1.14$ | $7.20 \pm 1.00$ | $6.08 \pm 0.62$ |
| | Variables for which $C(var, y_{pred}) > 0.02$ | | $6.75 \pm 0.95$ | $6.77 \pm 1.05$ |
| 2 | All variables | $6.81 \pm 1.03$ | $5.69 \pm 0.58$ | $6.42 \pm 0.88$ |
| | Variables for which $C(var, y_{pred}) > 0.01$ | $7.20 \pm 1.02$ | $7.00 \pm 0.85$ | $6.41 \pm 1.04$ |
| | Variables for which $C(var, y_{pred}) > 0.02$ | $6.57 \pm 1.13$ | $7.23 \pm 1.07$ | $6.82 \pm 1.12$ |
| 3 | All variables | $6.71 \pm 1.01$ | $6.94 \pm 0.87$ | $6.44 \pm 0.82$ |
| | Variables for which $C(var, y_{pred}) > 0.01$ | $6.48 \pm 0.96$ | $7.07 \pm 1.01$ | $6.62 \pm 0.68$ |
| | Variables for which $C(var, y_{pred}) > 0.02$ | $7.11 \pm 0.90$ | | $6.93 \pm 1.10$ |
| 4 | All variables | $6.81 \pm 0.87$ | $6.61 \pm 0.85$ | $6.28 \pm 0.81$ |
| | Variables for which $C(var, y_{pred}) > 0.01$ | $7.06 \pm 1.19$ | $6.31 \pm 0.88$ | $5.83 \pm 0.73$ |
| | Variables for which $C(var, y_{pred}) > 0.02$ | $6.57 \pm 1.13$ | $7.59 \pm 1.21$ | $7.48 \pm 1.05$ |

As seen in Table 7, the best Asimov significance is $7.59 \pm 1.21$ (for two NN layers, 4 jets variables with correlation coefficient $C(var, y_{pred}) > 0.02$).
Plots of Asimov estimate of significance and comparison between training and test sets as result of classifier output for two NN layers, 4 jets variables with correlation coefficient $C(var, y_{pred}) > 0.02$ are shown in Fig. 12.

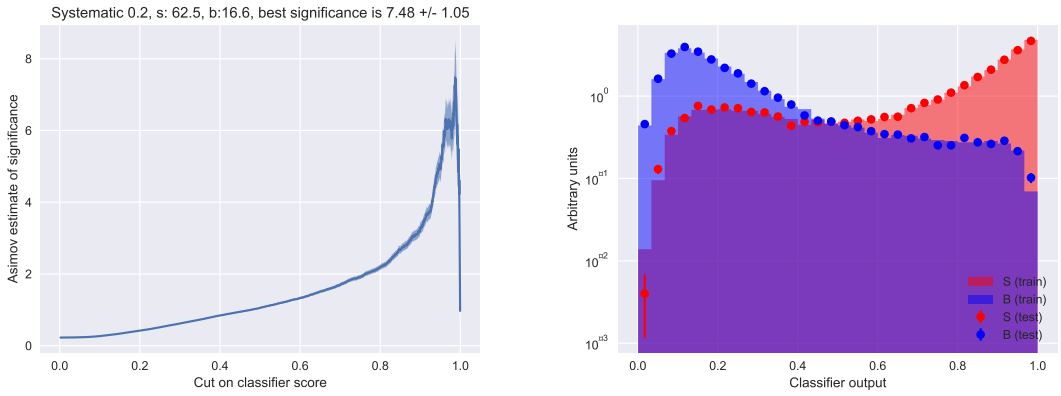


Figure 12: Plots of Asimov estimate of significance (on the left) and comparison between training and test sets (on the right) for the best Asimov significance

# 7 Conclusion

Study of the input variables of the neural network were carried out on the CMS T2tt SUSY model MC data. To study were selected some important input variables. Correlations between the selected variables were studied. Asimov significance for the different set of variables and different number of layers of the neural network were considered.
For set of high level and high level + low level variables the best result obtained with the NN with small number of layers. For set of low level variables the best result obtained with the NN with large number of layers. Also the optimal set of variables for the best Asimov significance is found (for two layers of NN, 4 jets variables with correlation coefficient $C(var, y_{pred}) > 0.02$).

# 8 Acknowledgments

Thanks to my supervisors Dirk Kruecker, Adam Elwood, Oleksii Turkot and Isabell Melzer-Pellmann for their help, support and helpful advice. Also thanks to Mykyta Shchedrolosiev for help and good working atmosphere.

# References

[1] Torbjorn Sjostrand and Stephen Mrenna and Peter Skands, "PYTHIA 6.4 physics and manual", JHEP 0605:026,2006, doi:10.1088/1126-6708/2006/05/026

[2] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaitre, A. Mertens, M. Selvaggi, "DELPHES 3, A modular framework for fast simulation of a generic collider experiment", doi:10.1007/JHEP02(2014)057

[3] Adam Elwood, Dirk Kruecker, "Direct optimisation of the discovery significance when training neural networks to search for new physics in particle colliders", DESY-18-082, arXiv:1806.00322 [hep-ex]

[4] Yang Bai, Hsin-Chia Cheng, Jason Gallicchio, Jiayin Gu, "Stop the Top Background of the Stop Search", JHEP 1207 (2012) 110, doi:10.1007/JHEP07(2012)110

[5] Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells, "Asymptotic formulae for likelihood-based tests of new physics", Eur.Phys.J.C71:1554,2011, doi:10.1140/epjc/s10052-011-1554-0