# 1 Introduction

## 1.1 Motivation

Goal of this project is the improvement of the distinction process between $t\bar{t}\gamma$ and $t\bar{t}$ events in the signal region. Depending on the success the results can be further used for other projects, which would improve the simulations and their predictions regarding the standard model (SM) and beyond standard model (BSM).

## 1.2 Approach

Attempt was perfomed through a multivariate analysis and by developing and training models. At the beginning different variables were analysed and compared and the most promising were used for training. Type of models, denoted MVA in the following, used are boosted decision tree (BDT) and mutilayer perceptron (MLP). The results of the MVAs for three different variable categories were created and in the course of this project analysed and presented.

# 2 Background information

## 2.1 ATLAS detector and data

The ATLAS detector is part of the Large Hadron Collider, a particle accelerator, which was built by the research organisation CERN in Genf, Siwtzerland. LHC has a circumference of 27km and is up to 175m deep. Besides ATLAS there are other experiments as well such as CMS, ALICE, LHCb. ATLAS can be described by three sub-detectors. All detectors surround the beam pipe, where the particles are accelerated and collided against eachother.

In figure a sketch of the detectorsystem can be seen. The inner detecor is a tracking chamber for charged partices. It is surrounded by a superconducting solenoid with a magnetic field of 2T used to divert the particle trajectory to measure the momentum. The calorimeter is made up of alternating layers. One is a dense absorber material, which is used to prompt electromagnetic and hadronic showers (ECAL, HCAL). The other layer is of active material used for energy measurement. The muon spectrometer (MS) is positioned in the outermost layer since muons may pass inner layers as well as the calorimeters. It is surrounded by a magnetic field and the diversion is measured by a multiple layer of high precision tracking chambers.

The data are filtered by a two-step triggersystem. The hardware-based system is triggered by xxx (criteria). This reduces the data amount from 40MHz to 100kHz. The software-based system is triggered by xxx (criteria), which reduced the data amount from 100kHz auf 1kHz. The data used during this project was recorded in 2016 at a centre-of mass energy of 13TeV. The signal and background processes were modelled using Monte Carlo generators and passed through a detector simulation (DOUBLE CHECK).

### 2.2 Proton-Proton-collision and $t\bar{t}\gamma$

The protons are accelerated in bunches, each containing $10^{11}$ particles, through different beam pipes in opposite directions and then collided with each other. The $t\bar{t}\gamma$ events were analysed, therefore top quark pair events with additional photons.

The top quark is the heaviest quark in the known elementary particles and is the only quark so far whose properties can be analysed, since the lifetime is very short and it forms no bound state before it decays. By that the electromagnetic coupling, therefore the number of radiated photons after interaction with other charged particles, can be measured. This is a test of the standard model, since it is a direct measurment of the charge of the top quark and shows possible anomalous structure of its electromagnetic interaction.

During the pp-process a top quark pair is produced. The top quark then decays in a W-boson and a b-quark. The W-boson disintegrates into a charged lepton and a corresponding neutrino or a pair of up-type quark and down-type antiquark. The remaining quarks decomposes into jets.

## 3 Challenge in categorisation

As seen in figure (XXX), the background makes up nearly 50% of the signal region. These are made up by particles, which were misconstructed or not detected due to the limited detector's acceptence. Other reasons are photons not being radiated from the $t\bar{t}$ process, but rather from an initial charged patron, an intermediate top quark or any charged final state particles or decay products. Further sources could be hadrons misreconstructed as electrons or hadrons and electrons misidentified as photons.

### 3.1 Current status

First a neural network was developed to seperate the hadronic-fake photons, which are hadrons misidentified as photons. The result was then used as an input for the neural network created to split the $t\bar{t}\gamma$ signal and the background further. What still needs to be further investigated is the possibility of improvement in the discrimination between $t\bar{t}\gamma$ and $t\bar{t}$ events in the signal region through MVAs.

## 4 Analysis of variables

In the following table are all relevant variables and their description listed, categorised by their type:

- Jets
- Lepton
- Photon

- Rest

Variables are categorized by their type of particles and the ending or prefix show, what information is saved. The description of variables, which don't follow this logic, are stated at the end.

Type of particles:

- Jets
- Lepton
- Photon
- Rest

Prefix and suffix:

- _eta
- _phi
- _pt
- dR
- n

Other variables:

- ht
- m3
- mT
- PhotonGood0
- MVA_ID
- MET: Missing transverse momentum

These were analysed by comparing the distribution of events with uniform scaling. A uniform scaling was needed to have a clearer view of relative differences between signal and background. If the variable shows a different behaviour for each type of signal, this could be an indication of discriminatory power.

As seen in the graphs above, figure x - x are almost identical and therefore not further analysed.

But for figure x - x the are some differences visible. These were candidated for the first set of selected variables.

The variable MVA_ID, seen in figure x, is currently used as a parameter to differentiate beween a good and a bad photon. It shows quite a good distinction and is therefore used as the second selection.

Figure x - x shows the photon variables. These show the most promising behaviour for discrimination between $t\bar{t}\gamma$ and $t\bar{t}$ events and were put into the third selection category.

# 5 MVA

Trying to split signal and background through a variable with a cut-off value faces the problem, that this criteria doesn't apply for every event. If it works for one event, it doesn't mean that it works for the next one. Therefore mutliple variables need to be used to distinguish between these two categories, therefore called multivariate analysis. As already mentioned in chapter xxx, the chosen MVA-types are boosted decision trees (BDT) and multilayer perceptron (MLP).

## 5.1 BDT - Boosted Decision Trees

The decision tree starts with the root node containing the whole sample For a binary tree, the node is then split into two branches using a variable and a corresponding cut-off value. The cut-off value should be a value, which seperates the signal from the background the best. This process will be repeated for each branch for all relevant variables, including already used variables, until an end criteria is met. The final branch is called a leaf and this will be assigned to one of the two categories. End criterias could be a minimum size of a leaf, perfect separation, insignificant improvement after split and maximal tree depth. Boosted decision trees uses additional information, so called weak classifieres. These are variables with a low discriminatory power and they will be used to improve the main decision tree to a more stable model with a lower error rate.

## 5.2 MLP - Multilayer Perceptron

The model is built out of multiple perceptrons, which are arranged in layers. The value of one perceptron is the sum of all weighted inputs. This value is then transformed through an activation function and a treshhold. The activation funtion is a linear function for the input- and outputlayer, while for the hiddenlayers it is usually the log-sigmoid transfer function, seen in figure xxx. All layers have a treshhold, except the inputlayer.

(FUNKTION).

In the MLP every perceptron of a layer is connected to every perceptron of the previous and next layer. The first and layer are called input- and outputlayer respectively, while the intermediate layers are the hiddenlayers. A MLP has a minimum of one hiddenlayer, therefore contains at least three layers.

If the signal is only transmitted in one direction, therefore the MLP doesn't have loops, it is described as feed forward. The learning process of the model is called backpropagation algorithm. It learns by minimizing the error between network output and expexted output. Initially all weights are assigned random values, then every event will go through the network and the weights are adapted. Either all events will be put through the network first and then the values are adapted or after each event, but for the latter the order of the events might be important. After one epoch the end criteria will be tested. If it fails, the whole learning process will be carried out again. If it is met, the algorithm is finished.

## 5.3 Training of MVA

The following table shows the default values of each model.

- BDT
  - x
  - y
  - z
- MLP
  - x
  - y
  - z

These parameters were adapted to improve the result:

- BDT
  - x
  - y
  - z
- MLP
  - x
  - y
  - z

The different selection categories were used to train the models. The distribution and ROC-curves are shown in the following chapters. adaptieren: In figure x - x the distribution of signal can be seen. The blue line is BDT, the red line is MLP. The dashed out line is from the test sample, the solid line is from the training sample. (Analog für ROC-cuve) The left figures show the results with these parameters xxx, on the right side the best parameters. By calculating the AUC (area under the curve) the discriminatory power can be calculated.

The first selection category were put into the model. An improvement of x% is visible.

The second selection category were put into the model. An improvement of x% is visible.

The third selection category were put into the model. An improvement of x% is visible.

In all categories the MLP show a better performance than the BDT. The distribution shows a weird shape because of reasons. Nevertheless the MLP was chosen.

## 6 Conclussion

It is very difficult to differentiate between $t\bar{t}\gamma$ and $t\bar{t}$ events because they show very similiar behaviour in the signal region. Nevertheless, a slight improvement was possible by training a MVA through relevant variables. Especially photon variables who a bigger improvement, which is expected. MLPs show a better performance than BDTs, which suggests the selection of this model.

## 7 Bibliography

## 8 List of figures

## 9 Appendix