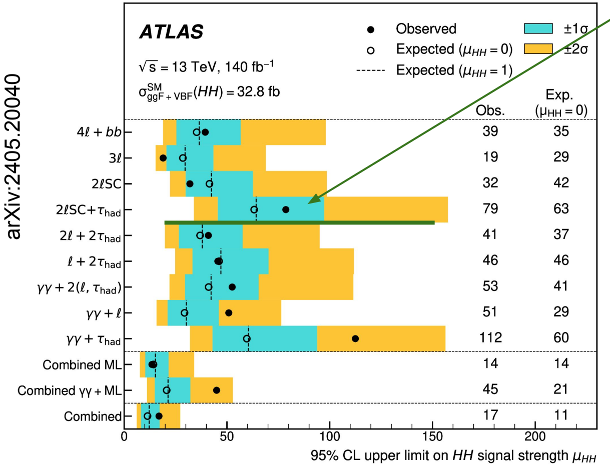
Double Higgs-boson signal strength measurement in the multi-leptons production channel in the ATLAS experiment

Motivation

Of the 9 channels, this one (2lSC+1τ) has the second last significance.
Due to two major reasons:

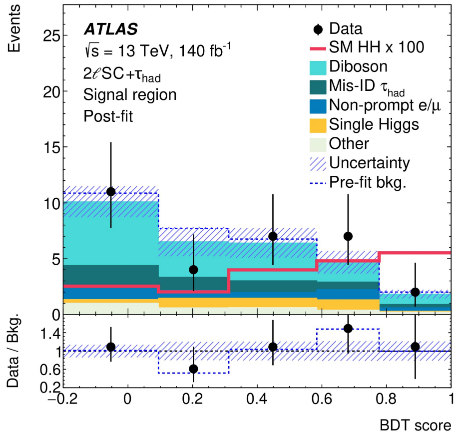
- The small statistic of the signal events - small cross-section
- Irreducible background- mostly from the WZ (VV) bosons production channel.

arXiv:2405.20040



The latest signal strengths limits in the various di-Higgs multi-leptons channels.

arXiv:2405.20040

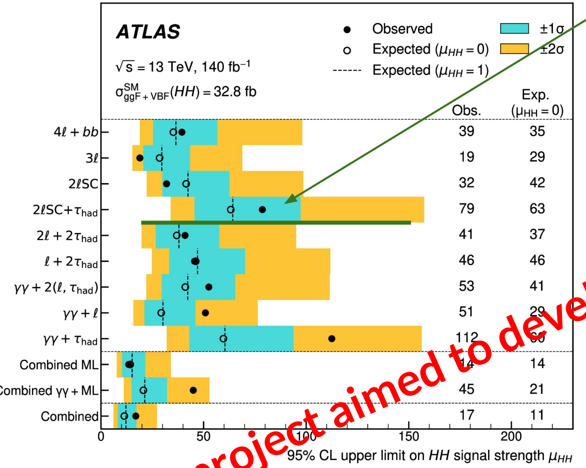


Motivation

Of the 9 channels, this one has the second last significance.
Due to two major reasons:

- The small statistic of the signal events- small cross-section
- Irreducible background- mostly from the WZ (VV) boson production channel.

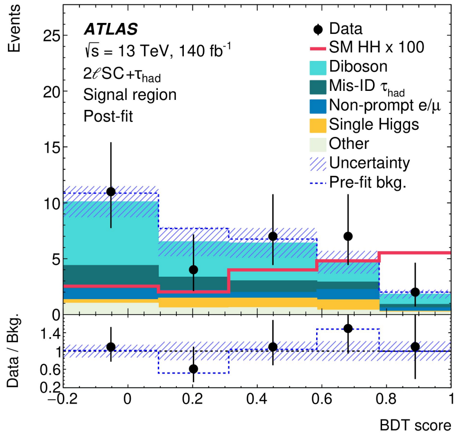
arXiv:2405.20040



The latest signal strengths limits in the various di-Higgs multi-leptons channels.

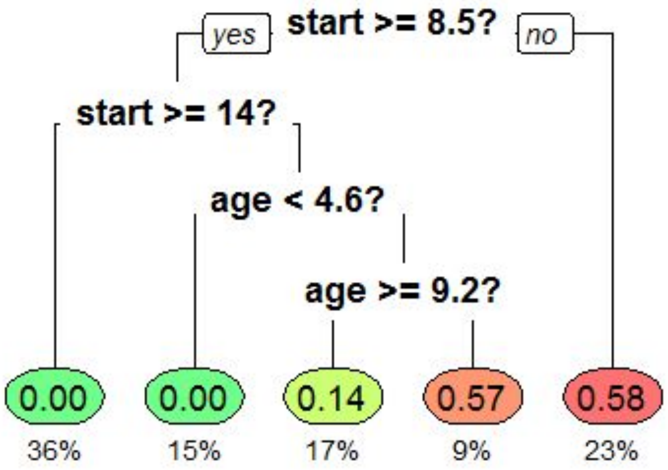
The project aimed to develop a new method for signal and background separation.

arXiv:2405.20040



Boosted Decision Trees (BDT)

A usual approach to separate background and signal data. A decision tree takes a set of input features and splits input data recursively based on those features. Boosting is a method of combining many weak learners (trees) into a strong classifier.



Methods of boosting:

- AdaBoost (Adaptive Boosting)
- BDTG (Gradient Boosted Decision Tree)
- XGBoost (Extreme Gradient Boosting)

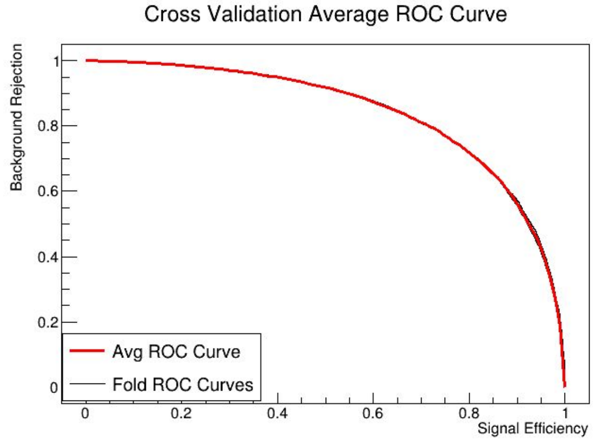
https://indico.fnal.gov/event/15356/contributions/31377/attachments/19671/24560/DecisionTrees.pdf
Figure by Stephen Milborrow - Own work, CC BY-SA 4.0,https://commons.wikimedia.org/w/index.php?curid=68192267

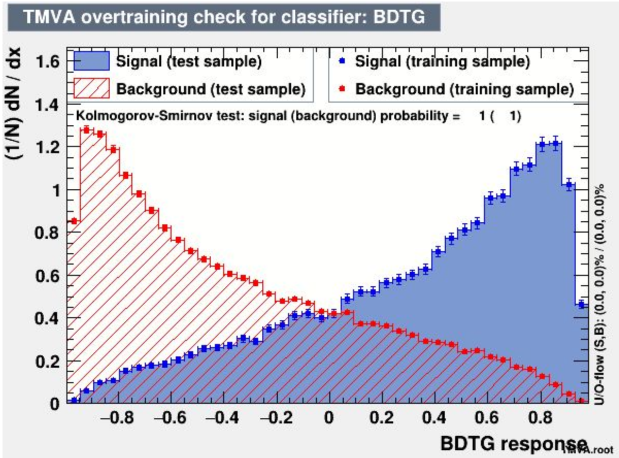
BDT analysis

The BDTG process allowed to reduce the number of useful variables to 21 out of 272. It was possible to use the shrunk down data to signal and background separation.

In the BDTG calculation process a *k-fold* cross-validation method was used to check if the results are applicable to independent datasets.

A ROC curve was generated to assert if the performance of the model is satisfactory. The plot integral = 0.841 ≈ 1.

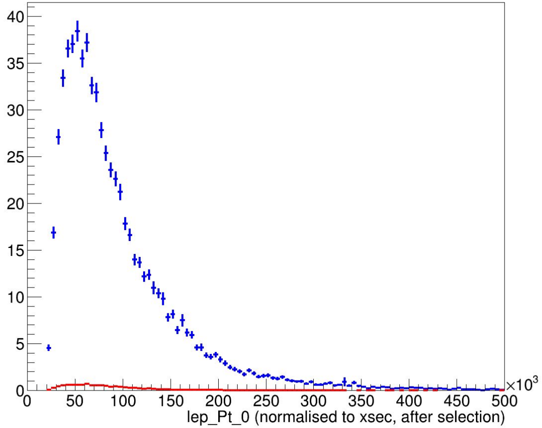


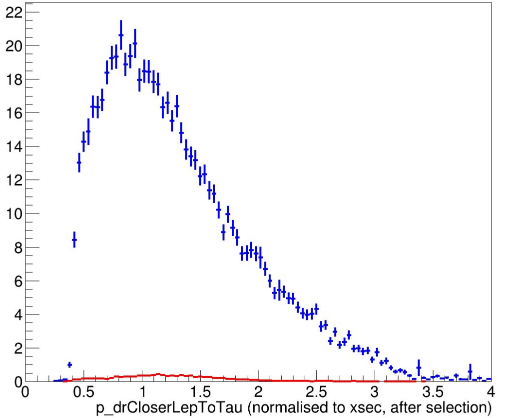


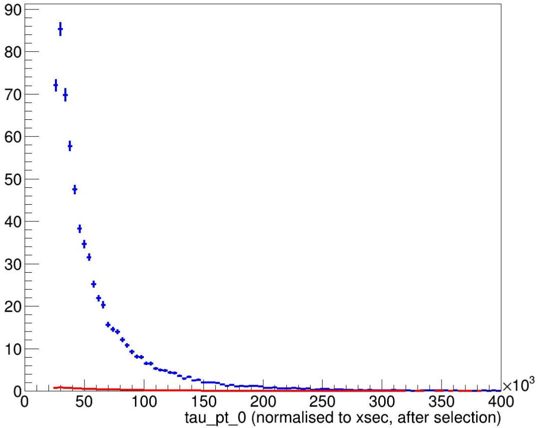
Results – BDT method

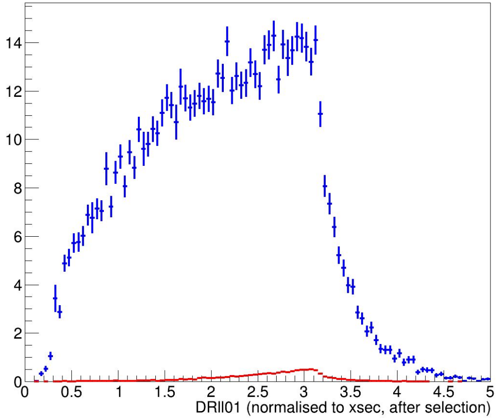
Histograms shown on this slide (of 4 chosen variables) were normalised to cross sections with proper conditions applied.

It is visible that the background distributions (blue) are greater than the signal (red) distributions.









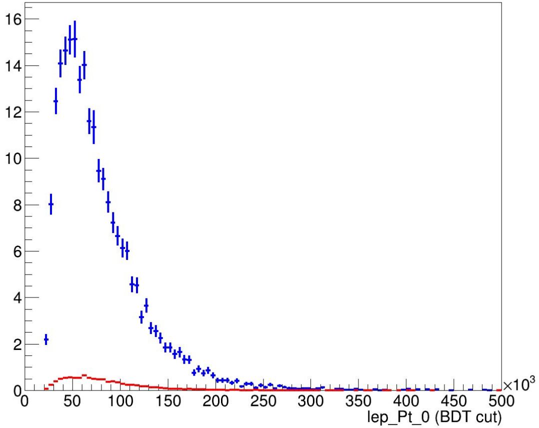
signal

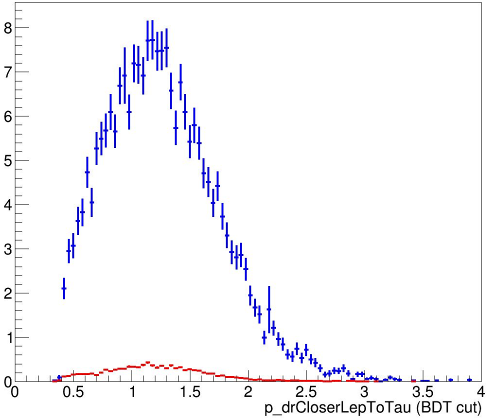
background

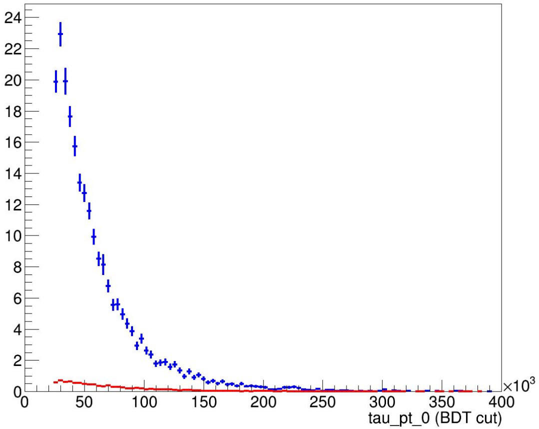
Results – BDT method

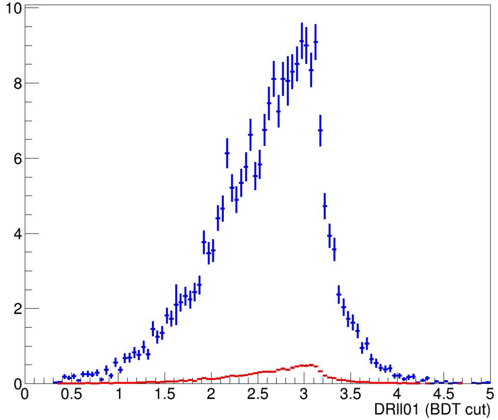
According to the calculation results, it is possible to obtain the highest background reduction efficiency when cutting at -0.2. That is, the variables for which the optimal-cut value is greater than -0.2 are cut off.

Results are shown on the histograms. 67% of background and 8% of signal was deleted.









signal

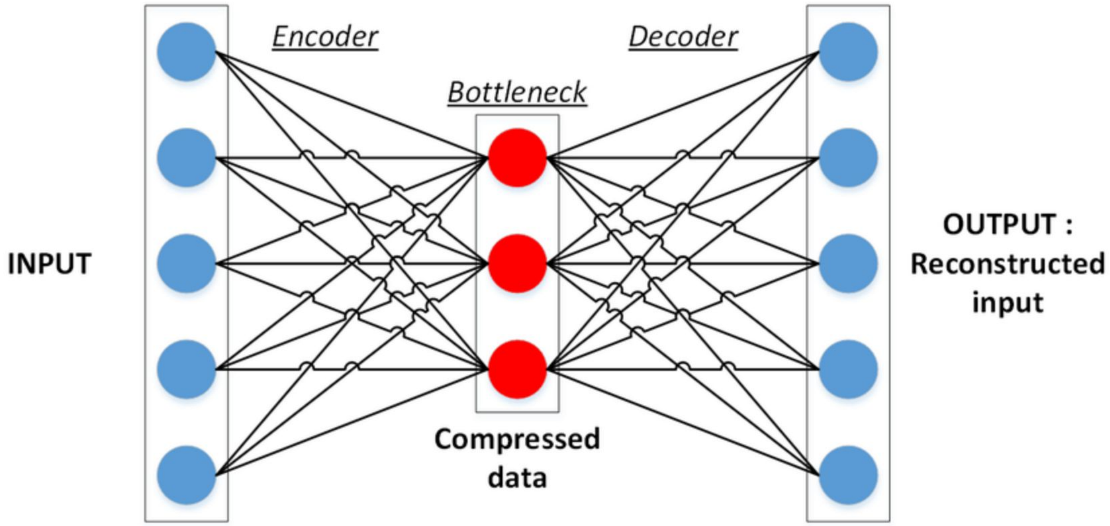
background

Outlier detection

A new approach for separating signal and background.

Autoencoder algorithm:

- Learns to minimize reconstruction error
- Reduces noise and removes anomalies
- Compresses data to capture essential features
- Useful for denoising and **anomaly detection**



Outlier analysis

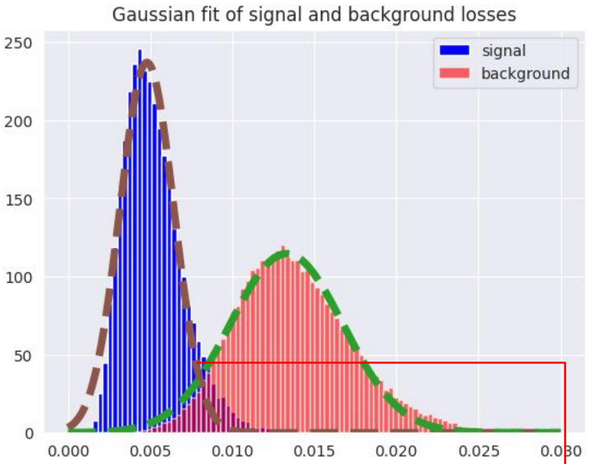
The input layer consists of 183 variables.

The encoding and decoding layers are symmetrical, each utilizing the ELU activation function. The bottleneck layer, containing 30 neurons, employs the ReLU activation function. Model performance during training is evaluated using the mean squared error loss function.

The signal sample was used for the training. Originally signal sample contains 23000 events.

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 183)	0
dense (Dense)	(None, 512)	94,208
dense_1 (Dense)	(None, 256)	131,328
dense_2 (Dense)	(None, 128)	32,896
dropout (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 30)	3,870
dense_4 (Dense)	(None, 128)	3,968
dense_5 (Dense)	(None, 256)	33,024
dense_6 (Dense)	(None, 512)	131,584
dropout_1 (Dropout)	(None, 512)	0
dense_7 (Dense)	(None, 183)	93,879

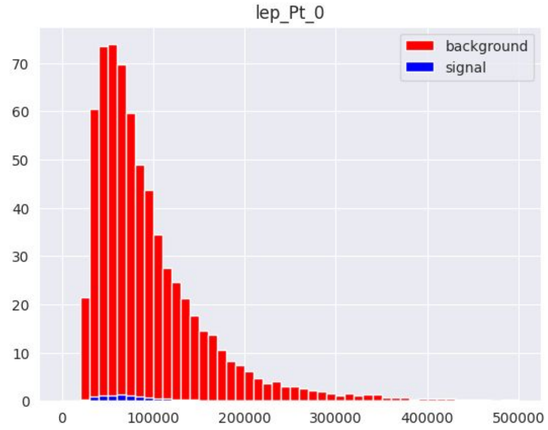
Total params: 524,757 (2.00 MB)

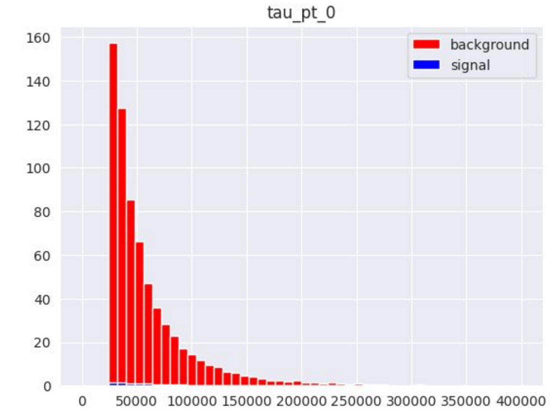


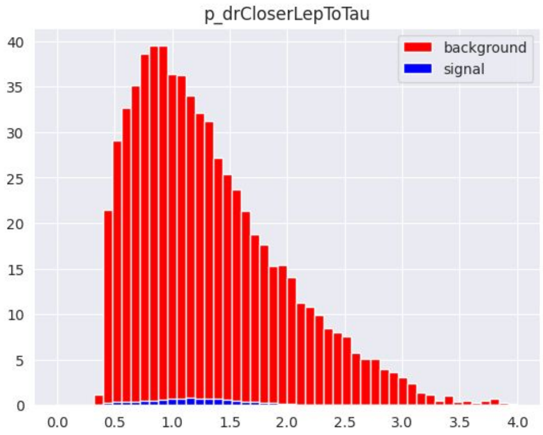
The histograms show the normalized distribution of loss values for signal (blue) and background (red) data, with a Gaussian fit applied. A cut at 0.008 is used to separate signal from background.

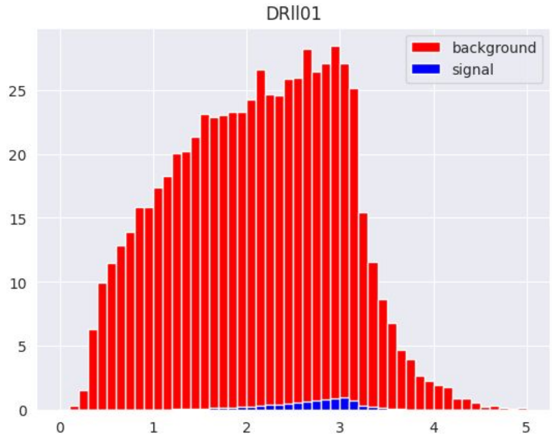
Results - outlier detection

The histograms show the original distribution of a four selected variables for both signal (blue) and background (red) data, with cross-section normalization applied



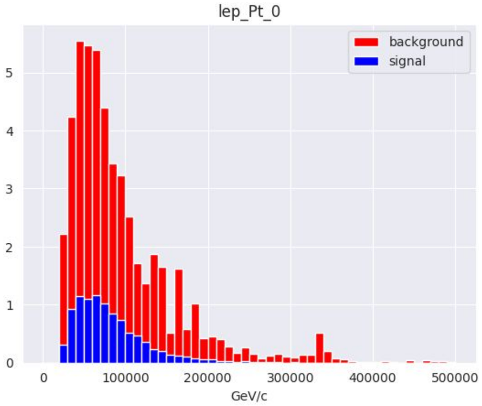


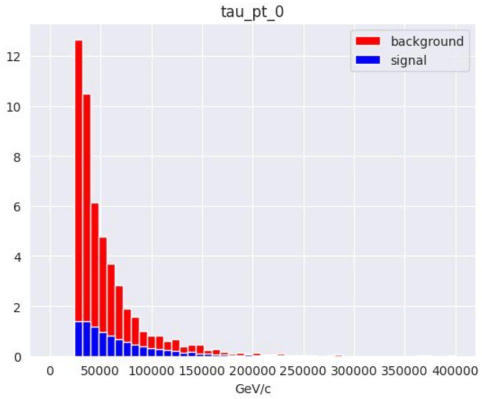


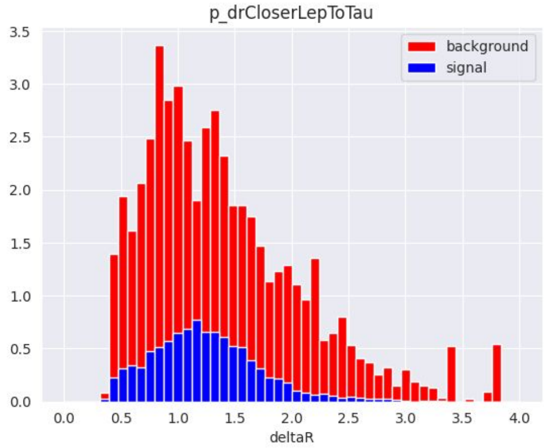


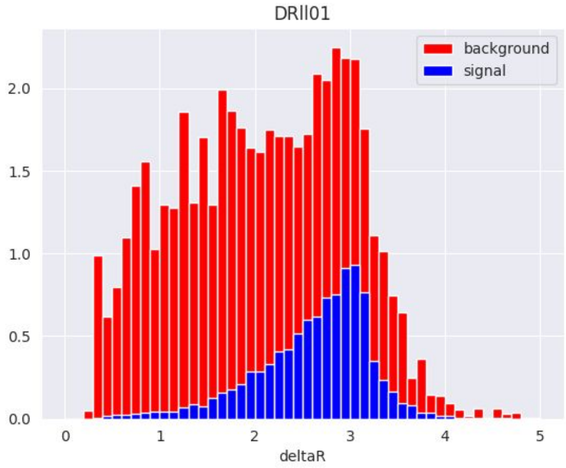
Results - outlier detection

Based on the previously determined cut, we separated the signal from the background. The histograms show the distribution of a four selected variables for both signal (blue) and background (red) data, with cross-section normalization applied.
We deleted 92% of background and 5% of signal.









Significance

Equation for significance: $\frac{s}{\sqrt{b}}$, where s is the area under the signal histogram and b is the area under the background histogram.

Outlier detection significance:

Before cut:

- Signal yield: **10.262**
- Background yield: **676.336**
- Significance: **0.394**

After cut:

- Signal yield: **9.780**
- Background yield: **50.80**
- Significance: **1.372**

BDT significance:

Before cut:

- Signal yield: **10.262**
- Background yield: **676.336**
- Significance: **0.394**

After cut:

- Signal yield: **9.478**
- Background yield: **226.379**
- Significance: **0.630**

Summary

➢ A new method for signal and background separation utilizing machine learning techniques has been developed. This method was compared to the BDT (Boosted Decision Trees) method used in previous analyses.

➢ The new method achieved better results and will be proposed for use in future analyses.

➢ Special thanks to Prof. Marcin Wolter for his invaluable support and guidance.