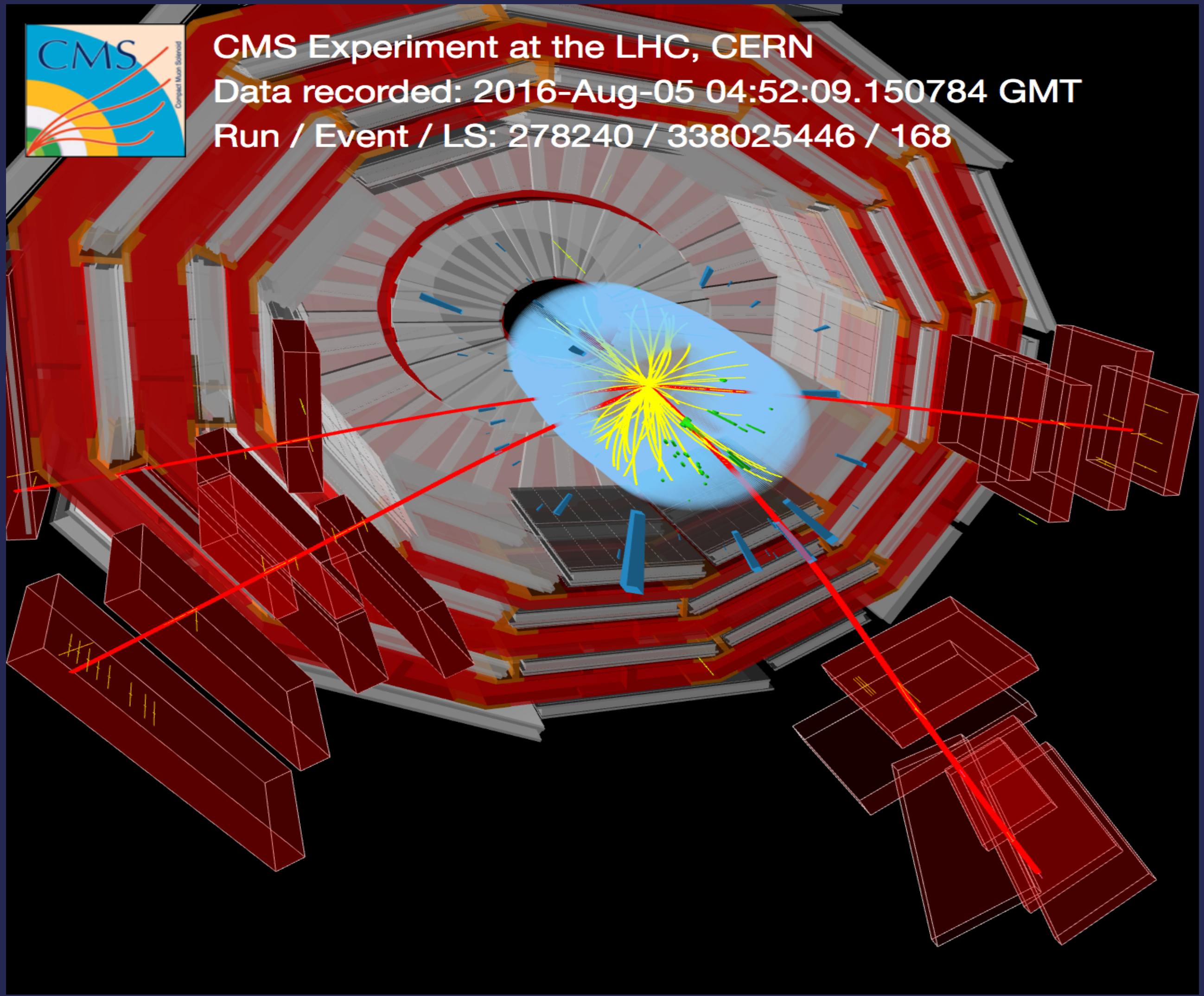


Classification of Signal and Background Using Multivariate Analysis

Presented by : Sayan Dhani |Summer Intern, SINP | M.Sc. Physics, IIT Bombay | E-mail: 22n0324@iitb.ac.in

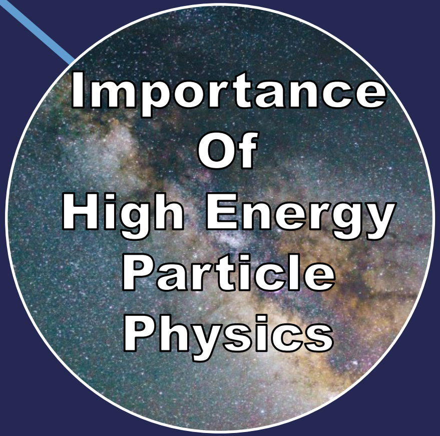
Supervisor : Prof. Subir Sarkar , SINP

Collaboration : Kuldeep Nishad(IIT Kanpur), Ratul Sarkar(IISER TVM)

Content

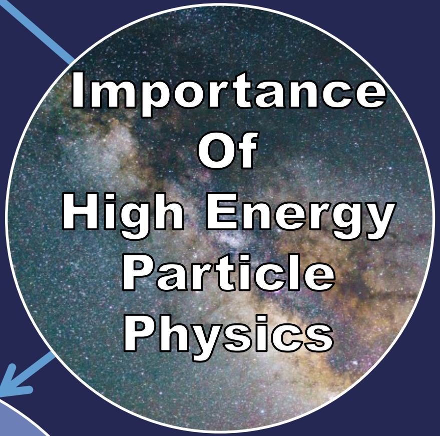
High Energy Particle Physics

Understanding the fundamental nature of the Universe



High Energy Particle Physics

Understanding the fundamental nature of the Universe

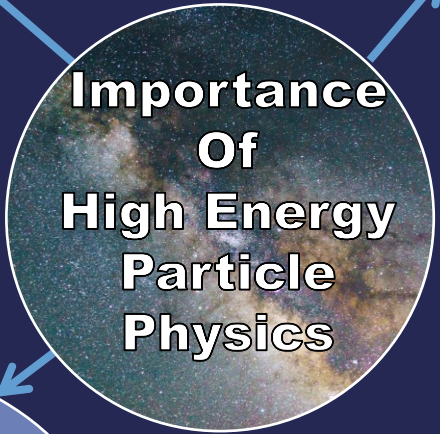


Knowledge about the fundamental forces, and particles

High Energy Particle Physics

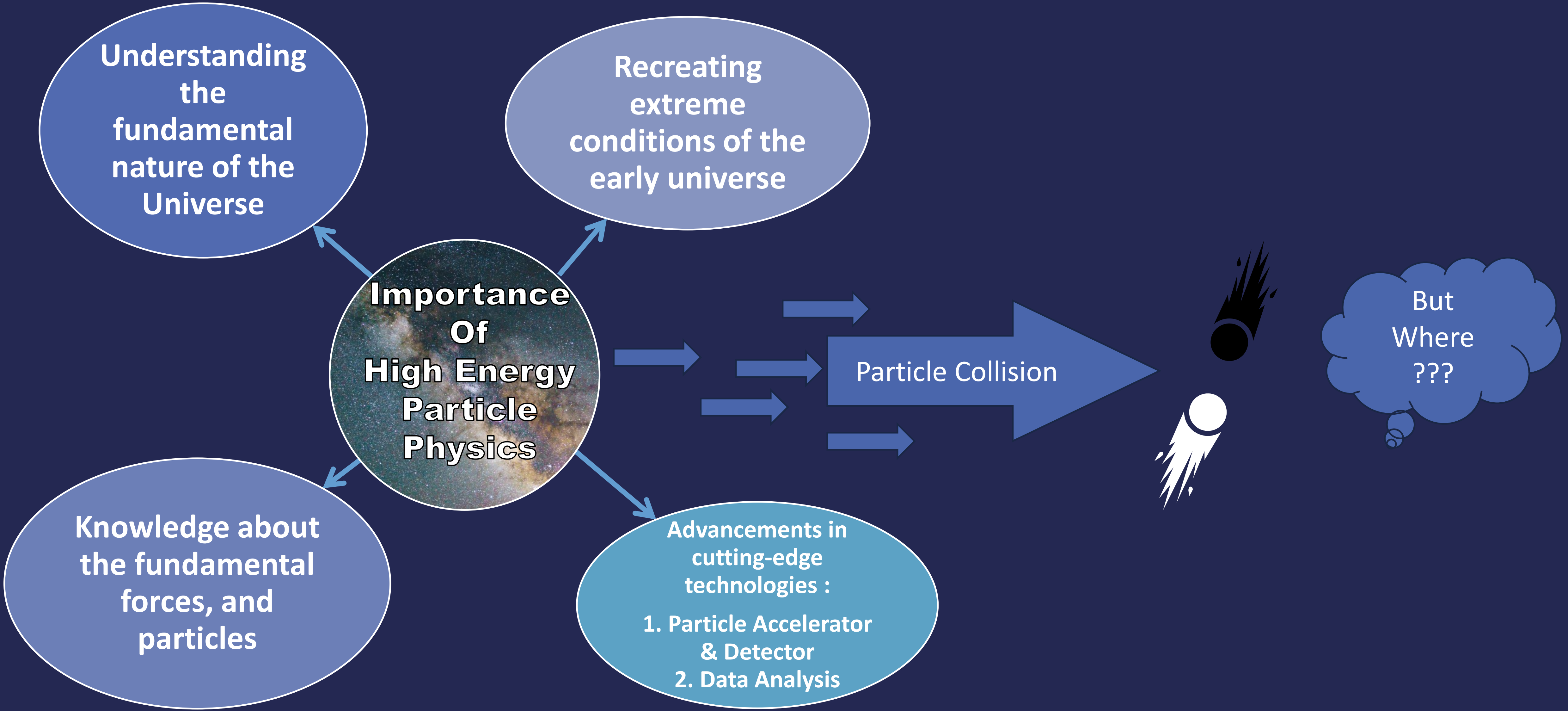
Understanding the fundamental nature of the Universe

Recreating extreme conditions of the early universe



Knowledge about the fundamental forces, and particles

High Energy Particle Physics

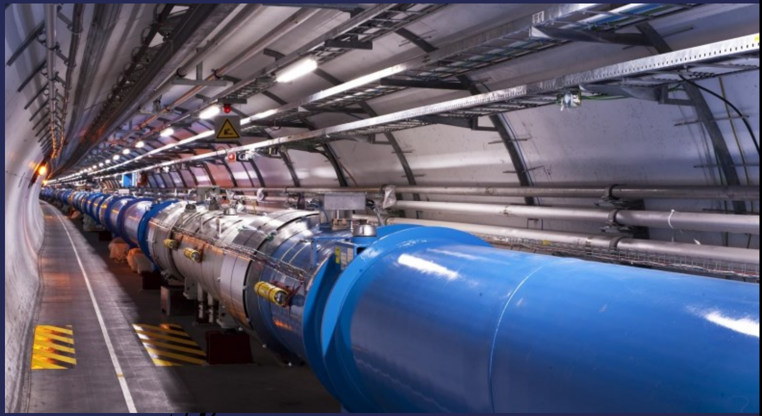


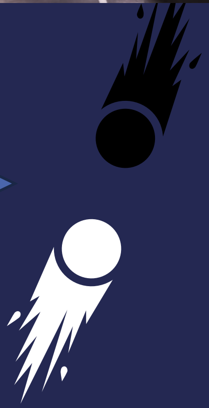
High Energy Particle Physics

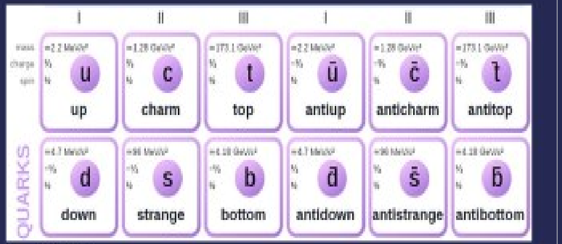
Importance Of High Energy Particle Physics

- Understanding the fundamental nature of the Universe
- Recreating extreme conditions of the early universe
- Knowledge about the fundamental forces, and particles
- Advancements in cutting-edge technologies :
  1. Particle Accelerator & Detector
  2. Data Analysis

Particle Collision

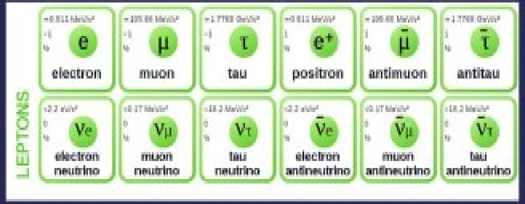


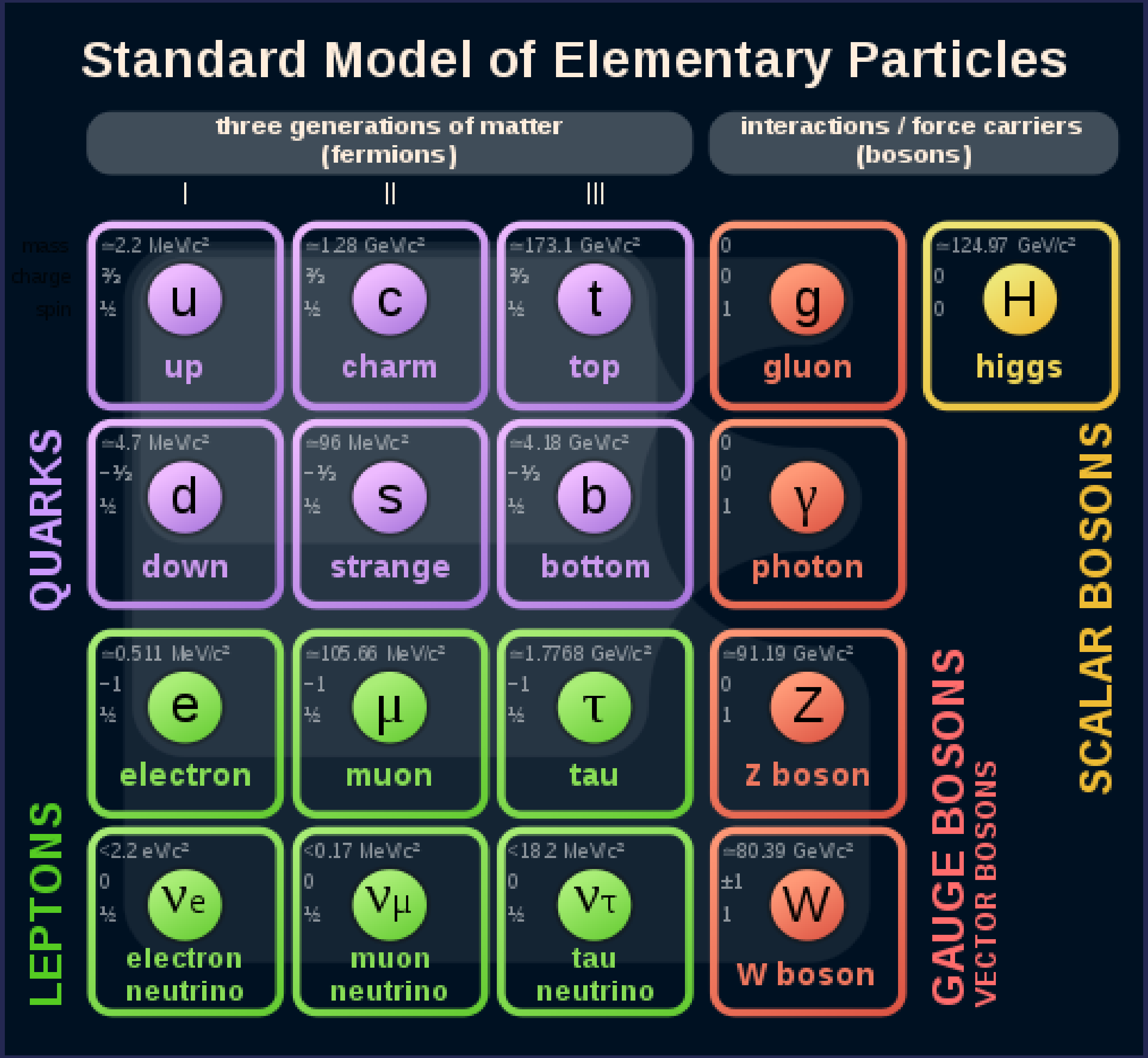




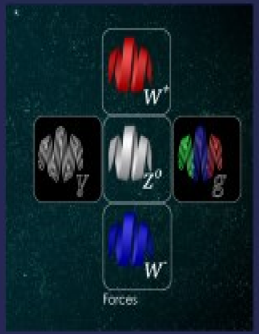
- **Quarks :**
  - The building block of Baryons(P and N)
  - 6 types with their corresponding anti-quark
  - have a fractional electric charge
  - Fermions(spin ½)

- **Leptons :**
  - Elementary particles do not participate in strong interaction
  - 3 types with their anti-lepton, neutrino, anti-neutrino
  - Have light mass
  - Fermions(spin ½)

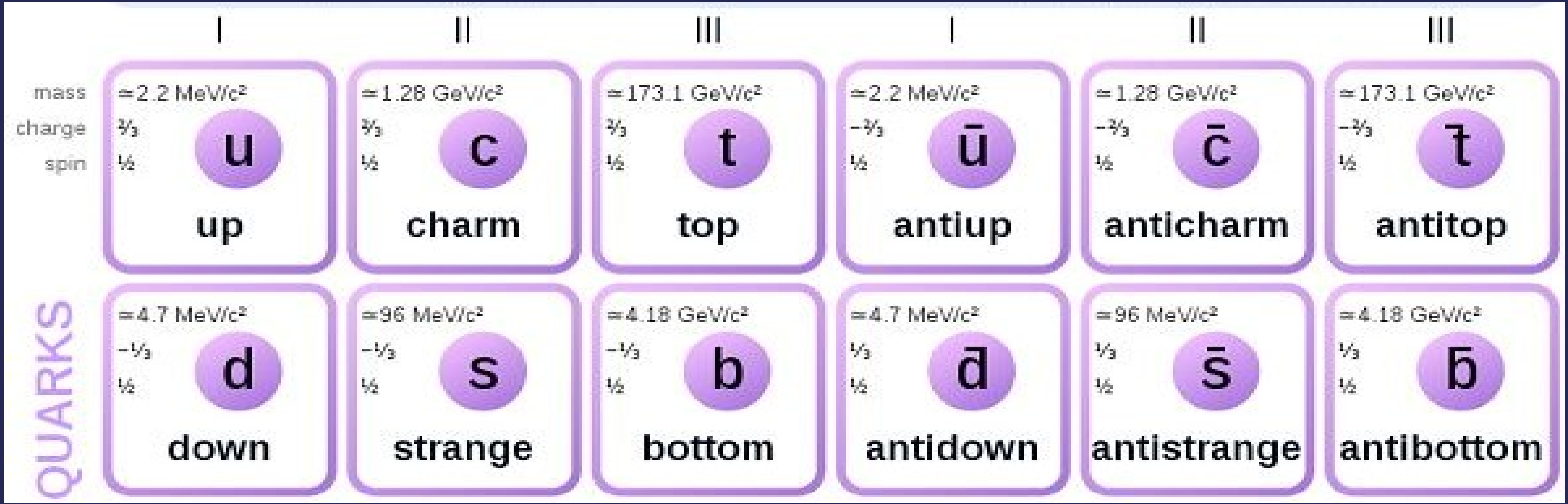




- **Higgs boson (Scalar Boson) :**
  - Spin 0 particle
  - Responsible for giving mass to all fundamental particles

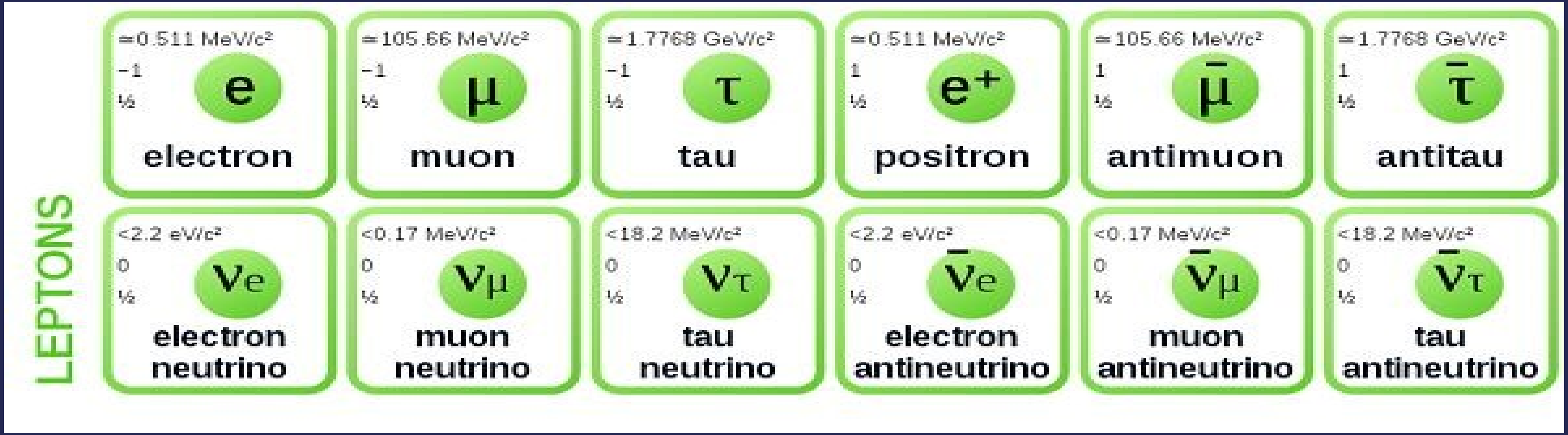


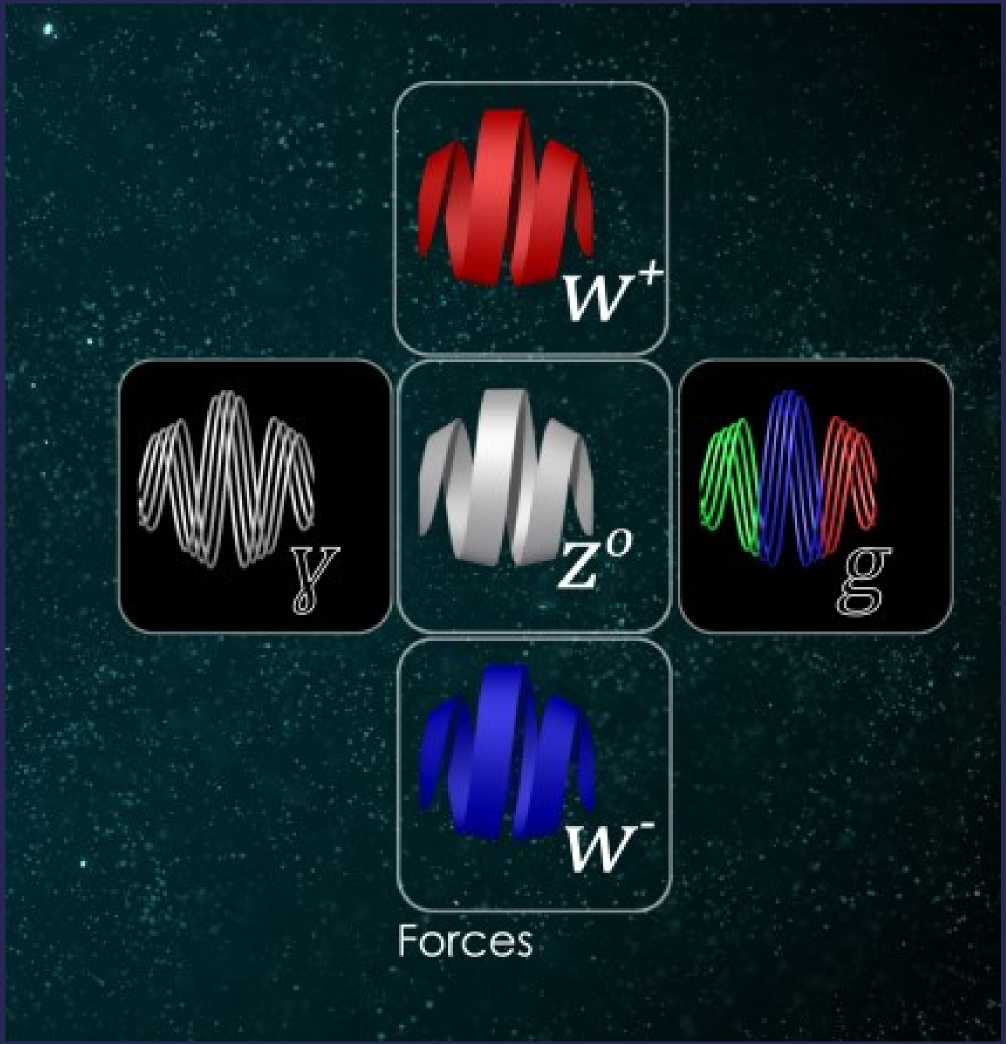
- **Gauge boson (Vector boson) :**
  - Spin 1 Particle.
- **Mediate the Fundamental forces.**
  - Photons for Electromagnetic force
  - W⁺ / Z for the weak force
  - Gluons for the strong force



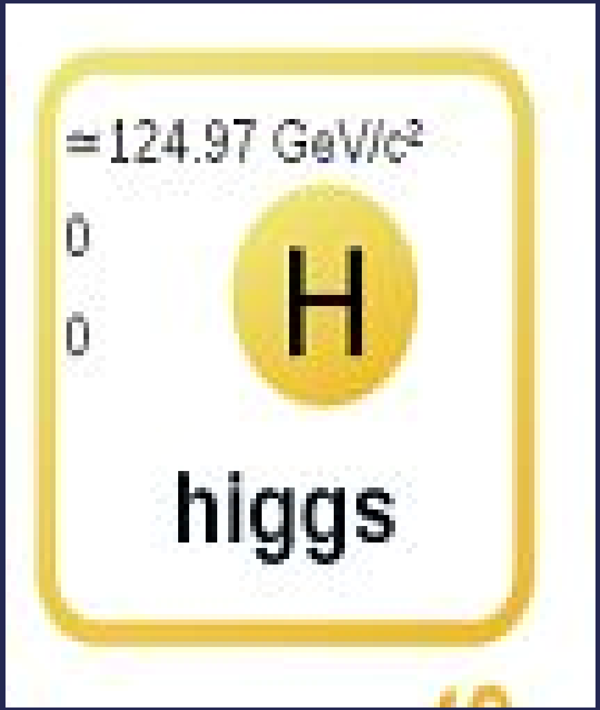
- **Quarks :**
  - **The building block of Baryons(P and N)**
  - **6 types with their corresponding anti-quark**
  - **have a fractional electric charge**
  - **Fermions(spin ½)**

- **Leptons :**
  - **Elementary particles do not participate in strong interaction**
  - **3 types with their anti-lepton, neutrino, anti-neutrino**
  - **Have light mass**
  - **Fermions(spin ½)**





- **Gauge boson (Vector boson) :**
  - **Spin 1 Particle.**
- **Mediate the Fundamental forces.**
  - **Photons for Electromagnetic force**
  - **W^{+-} / Z for the weak force**
  - **Gluons for the strong force**



- Higgs boson (Scalar Boson) :
  - Spin 0 particle
  - Responsible for giving mass to all fundamental particles

Signal and Background

Signal : specific physical processes or particles that researchers are interested in studying or discovering.

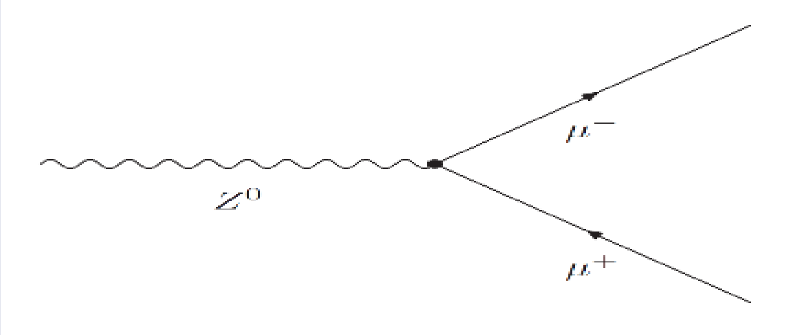
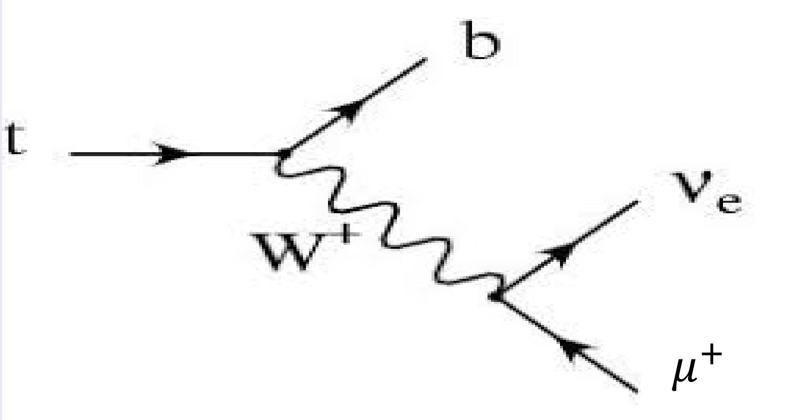
- Characterise by specific experimental signatures or properties that distinguish them from the background events
- Goal is to identify and isolate the signal events from the background events to study-measure their properties accurately and compare them with theoretical predictions or models.

Background: refers to the events or processes (like the QCD process) that mimic or resemble the signal events

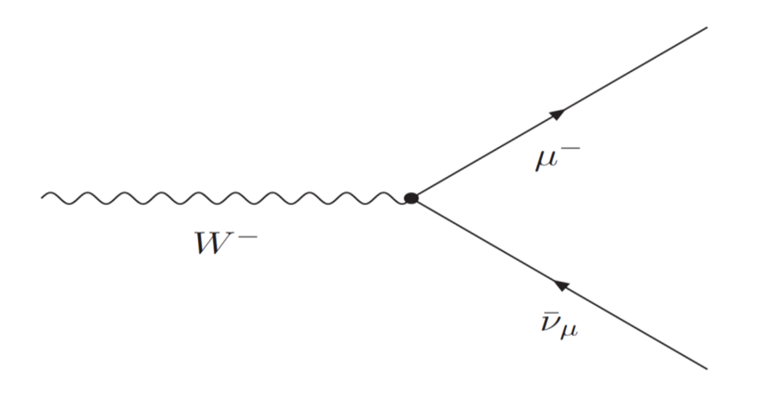
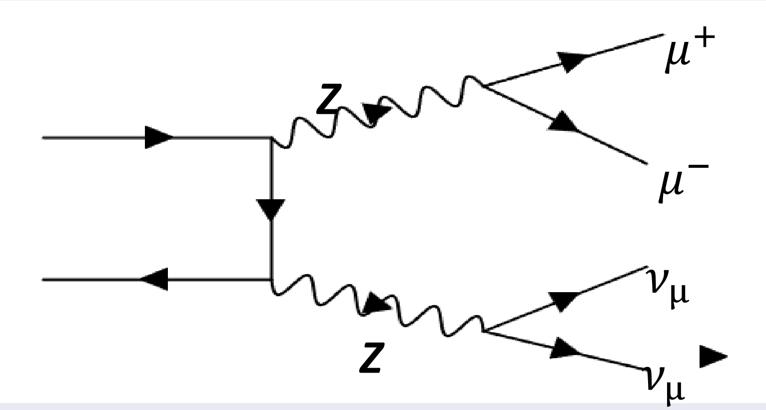
- Arise from well-understood and known physical processes that occur in particle collisions at the LHC.
- Challenge is to minimize the background events and enhance the signal-to-background ratio. Achieved through careful data analysis techniques, selection criteria, and background estimation methods.

Researchers employ various strategies, such as the use of control samples, background modelling, and statistical methods, to quantify and subtract the background contributions from the observed data.

Our Signal & Background in Project

Signal / Background	Process	Some features	Feynman's Diagram
Signal 1	$z \rightarrow \mu^{+} + \mu^{-}$	1) Occurs through the weak interaction 2) Cross-section : 2025 pb	
Background 1	tt $\rightarrow$ $WW + bb \rightarrow 2\mu + 2\nu_{\mu} + 2b$	1) Predominantly decays before it hadronized 2) Decay through the weak interaction -> WJets, bottom 3) Cross-section : 88.6 pb	

Our Signal & Background in Project

Signal / Background	Process	Some features	Feynman's Diagram
Background 2	$WW \rightarrow 2\mu + 2\nu_{\mu}$	1) produced in association with one or more jets of hadrons. 2) Decay through the weak interaction -> charged lepton and neutrino. 3) Cross-section : 61526 pb	
Background 3	$ZZ \rightarrow 2\mu + 2\nu_{\mu}$	1) Cross-section : 0.564 pb	

Data Analysis

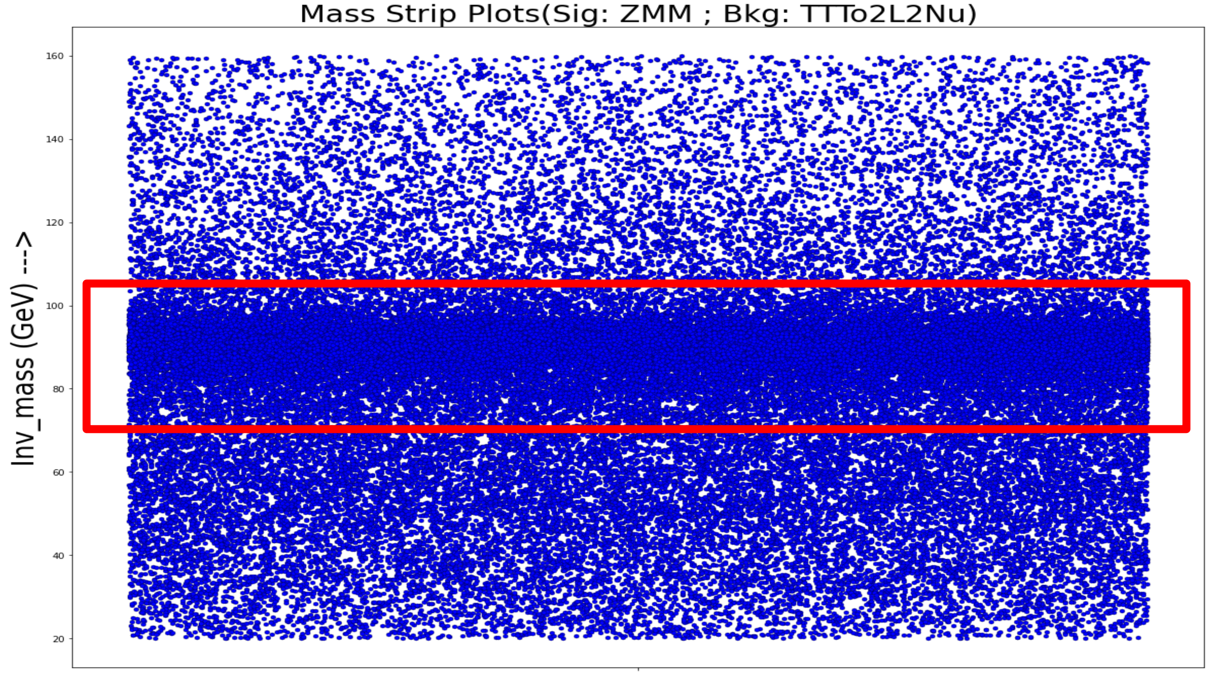
- It is a process of inspecting, cleaning, transforming, and modelling data to discover useful information, draw conclusions, and support decision-making. Involves various techniques and methodologies to extract insights from raw data, uncover patterns, and make predictions or informed recommendations.
- **Steps**: 1. Data Collection

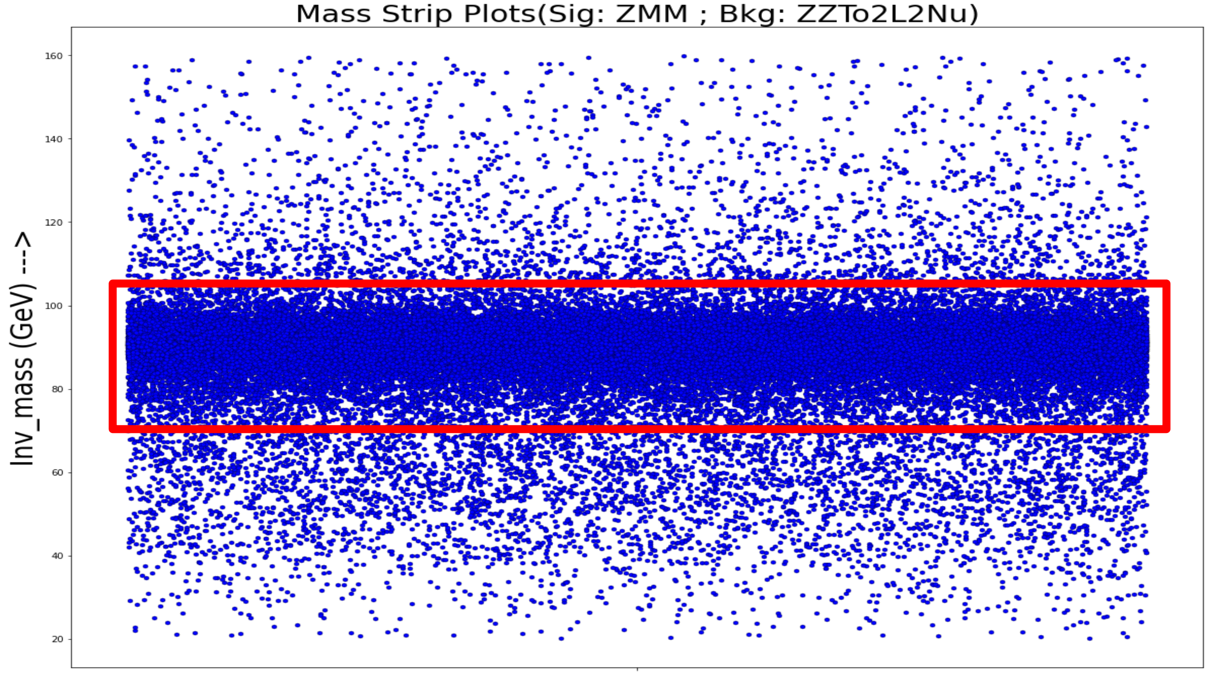
  2. Data Cleaning, Exploratory Data Analysis, Feature Engineering, Data Transformation

  3. Model building, training, testing

  4. Prediction/Inference and iteration
- Most HEP Analyses require discrimination of Signal from the background:
- Event level, Cone level, Track level, Lifetime and Flavour tagging, Parameter estimation etc.

Motivation for Multivariate Analysis

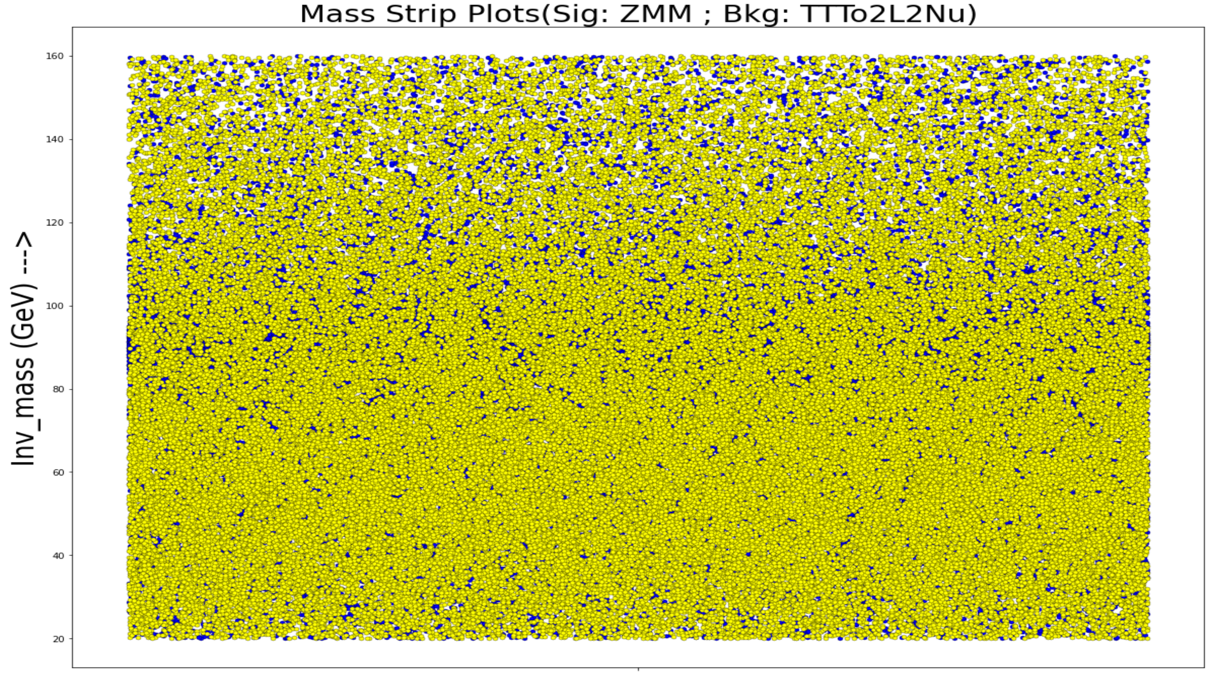


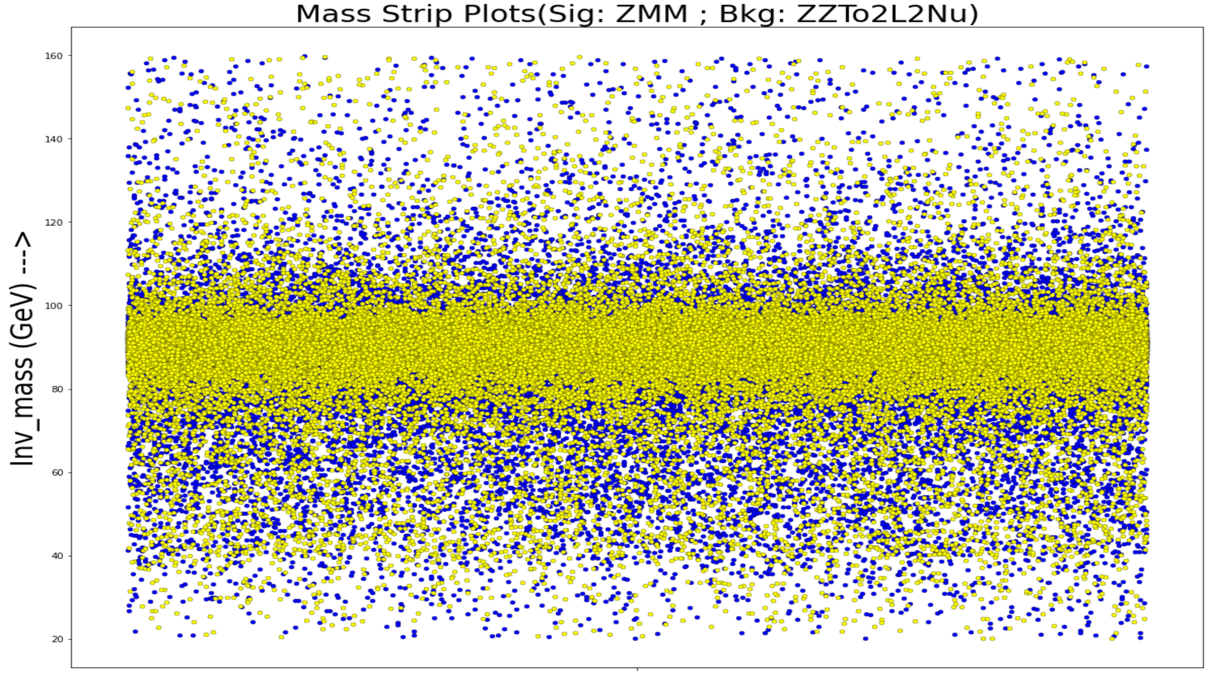


- Events’ Inv-Mass Topology, involve both signal(ZMM) and background(T-Tbar)
- Very dense in between mass range 80-100 GeV => Could be the signal region
- Traditional Cut based Analysis can be applied.

- Events’ Inv-Mass Topology, involve both signal(ZMM) and background(Z-Z)
- Very dense in between mass range 80-100 GeV => Could be signal region, but here background may also overlap
- Other masses range from very less density
- Traditional Cut based Analysis may not help much.

Motivation for Multivariate Analysis

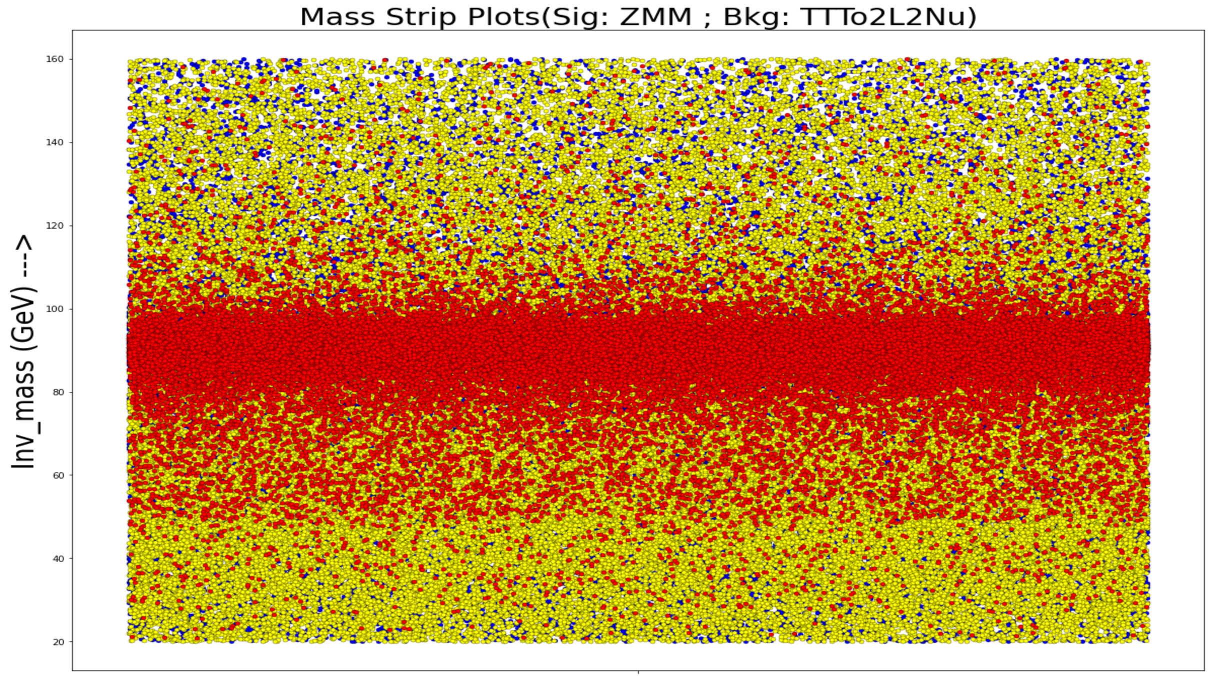


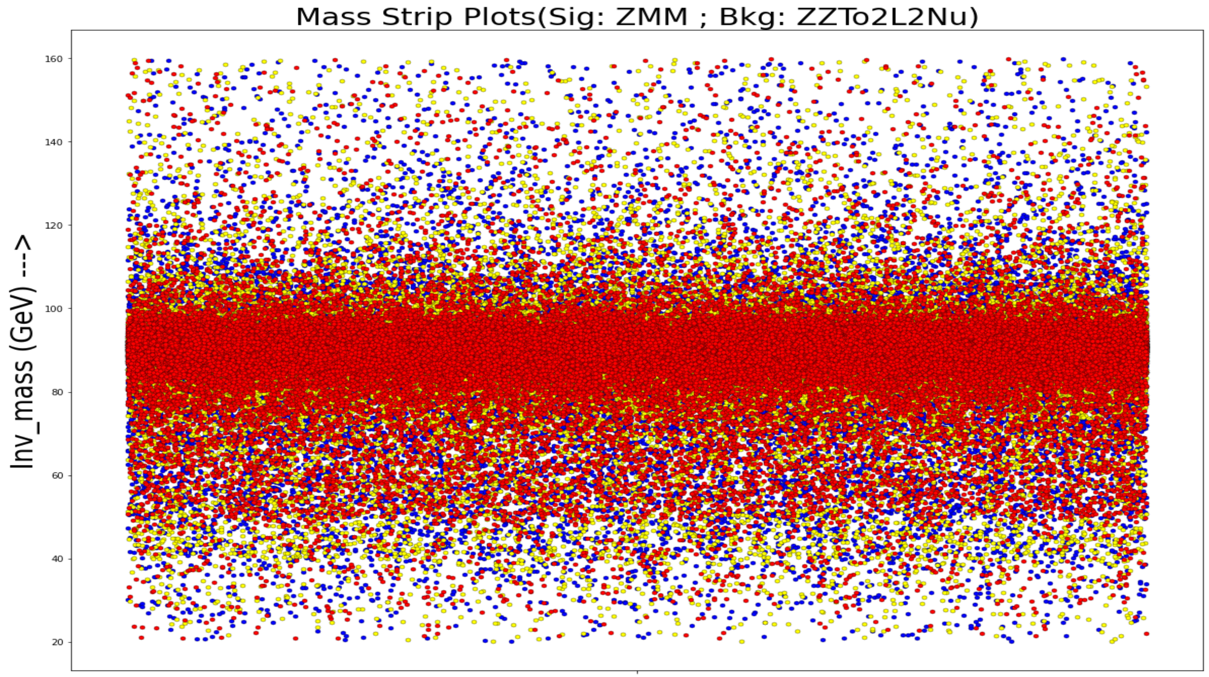


- **But the Background is everywhere,**
- **Traditional Cut Can remove many, but some still be there**
- **Better to do a more effective classification, if possible.**

- **The background is almost everywhere in the data.**
- **Heavily mixed up with signal events.**
- **Traditional cut is not effective at all, Difficult for most classifier algorithm**

Motivation for Multivariate Analysis





- Traditional cuts also remove some signal events, also
- MVA can analyze the whole topology along with other variables
- Can be very effective; give the optimal cut.

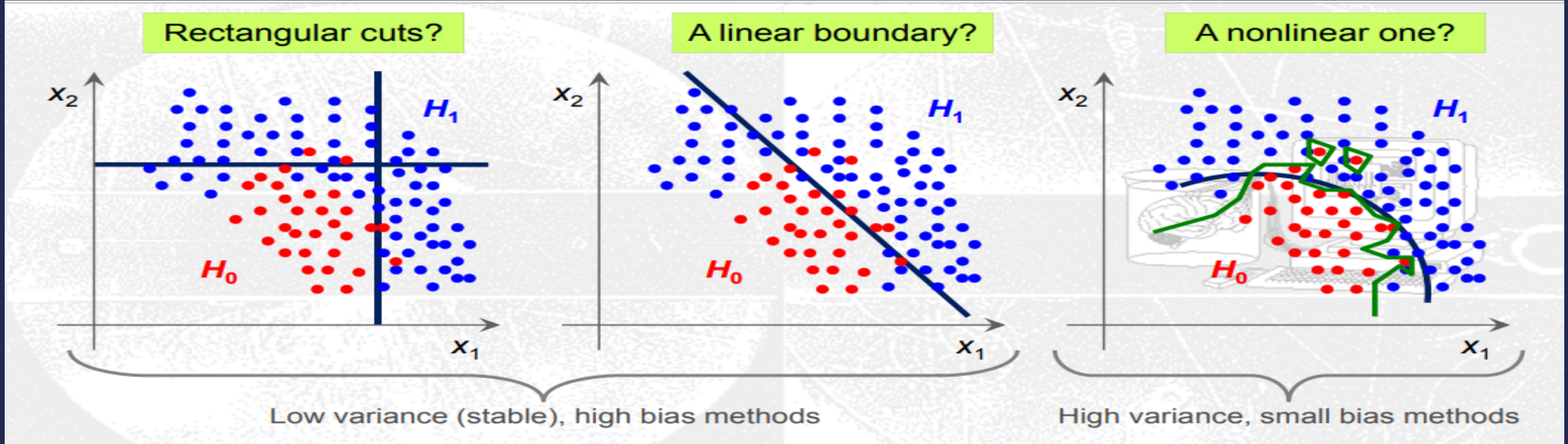
- The Signal & background almost overlapped.
- Very difficult to classify
- But MVA study the other variable simultaneously
- MVA still can classify, may not be so effective but better than other

Motivation for Multivariate Analysis

- Traditional cut-based analysis relies on selecting events based on specific criteria or thresholds applied to individual variables or combinations of variables.
- While cut-based analysis is straightforward and widely used, it may overlook complex relationships and correlations among variables.
- MVA techniques consider the full multidimensional information contained in the data, allowing for a more comprehensive analysis.
- MVA can capture subtle patterns and correlations that may be missed by individual cuts, leading to improved sensitivity and classification performance.
- By using advanced algorithms and considering multiple variables simultaneously, MVA can effectively separate signals from background events.

Multivariate(MVA) Analysis

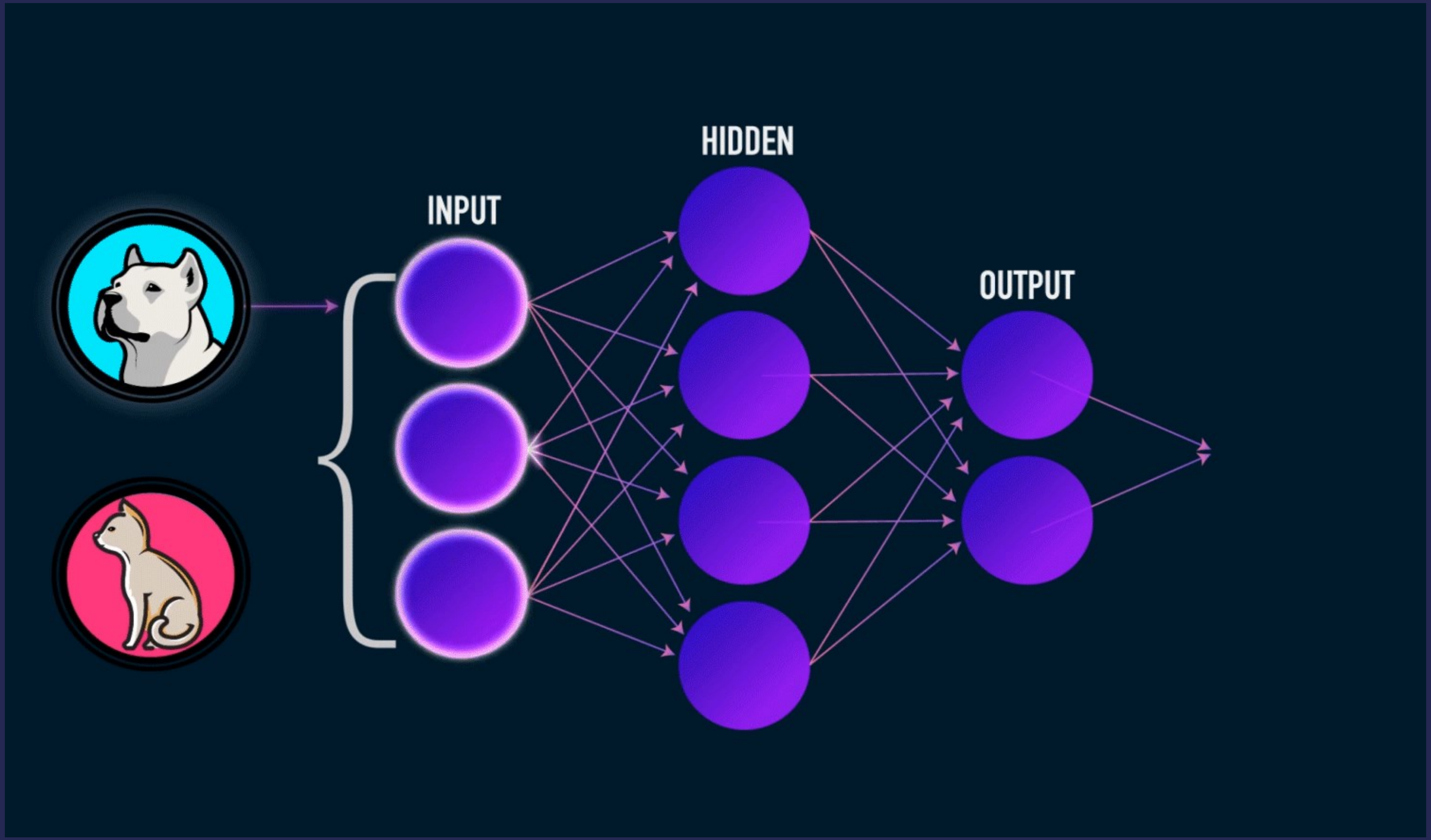
- MVA involves simultaneous analysis of multiple variables or features and sometimes makes a new variable to make predictions or classifications
- Unlike traditional analysis methods that rely on single-variable cuts or selections, MVA considers the correlations and interactions between variables.
- Objectives :
  - Data Reduction
  - Data Classification
  - Data inter-dependency
  - or, Correlation
- Aim to extract the maximum amount of information from the data, improving sensitivity and reducing background contamination.
- MVA methods often involve the use of advanced algorithms, such as Artificial Neural Networks (ANN), Boosted Decision Trees (BDT), Support Vector Machines (SVM), likelihood, DNN etc.



TMVA Package from ROOT

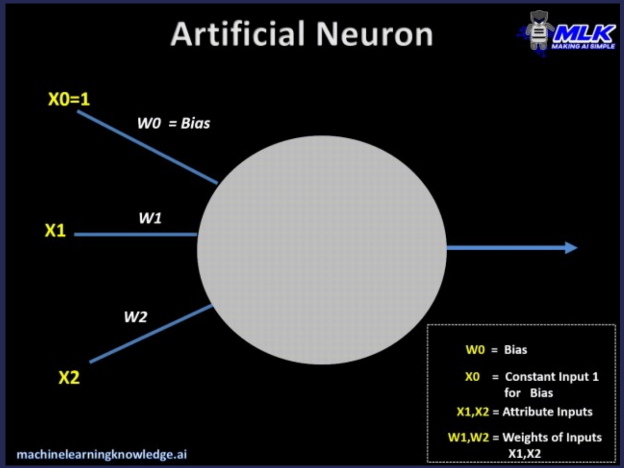
- The TMVA (Toolkit for Multivariate Analysis) package is a powerful tool within the ROOT framework for performing multivariate analysis.
- TMVA provides a comprehensive set of algorithms and tools for data preprocessing, feature selection, training, and evaluation of machine learning models.
- It offers various machine learning techniques, including decision trees, neural networks, support vector machines, and more, enabling researchers to choose the most suitable method for their analysis.
- TMVA provides efficient implementations of these algorithms, optimized for high-performance computing, allowing for quick and accurate analysis of large datasets.
- The package also offers features such as cross-validation, hyperparameter optimization, and visualization tools to facilitate the analysis and interpretation of results.

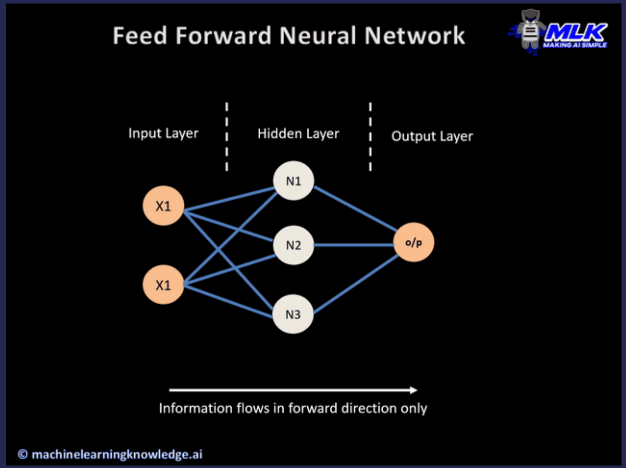
Workflow of Artificial Neural Network

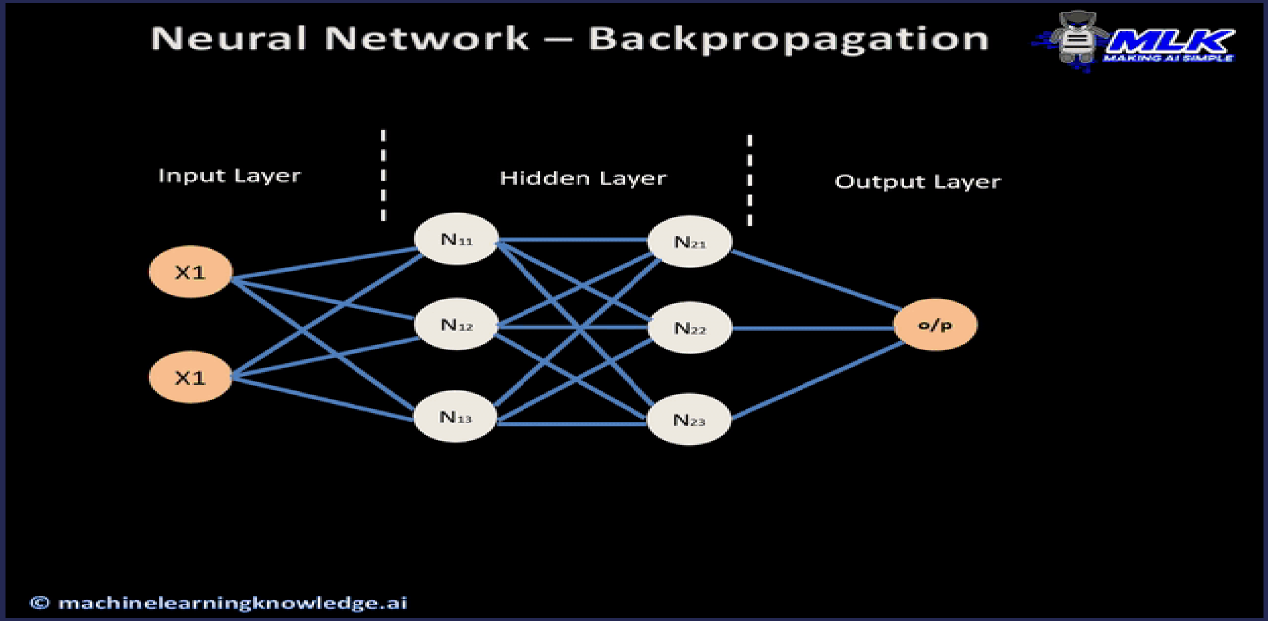


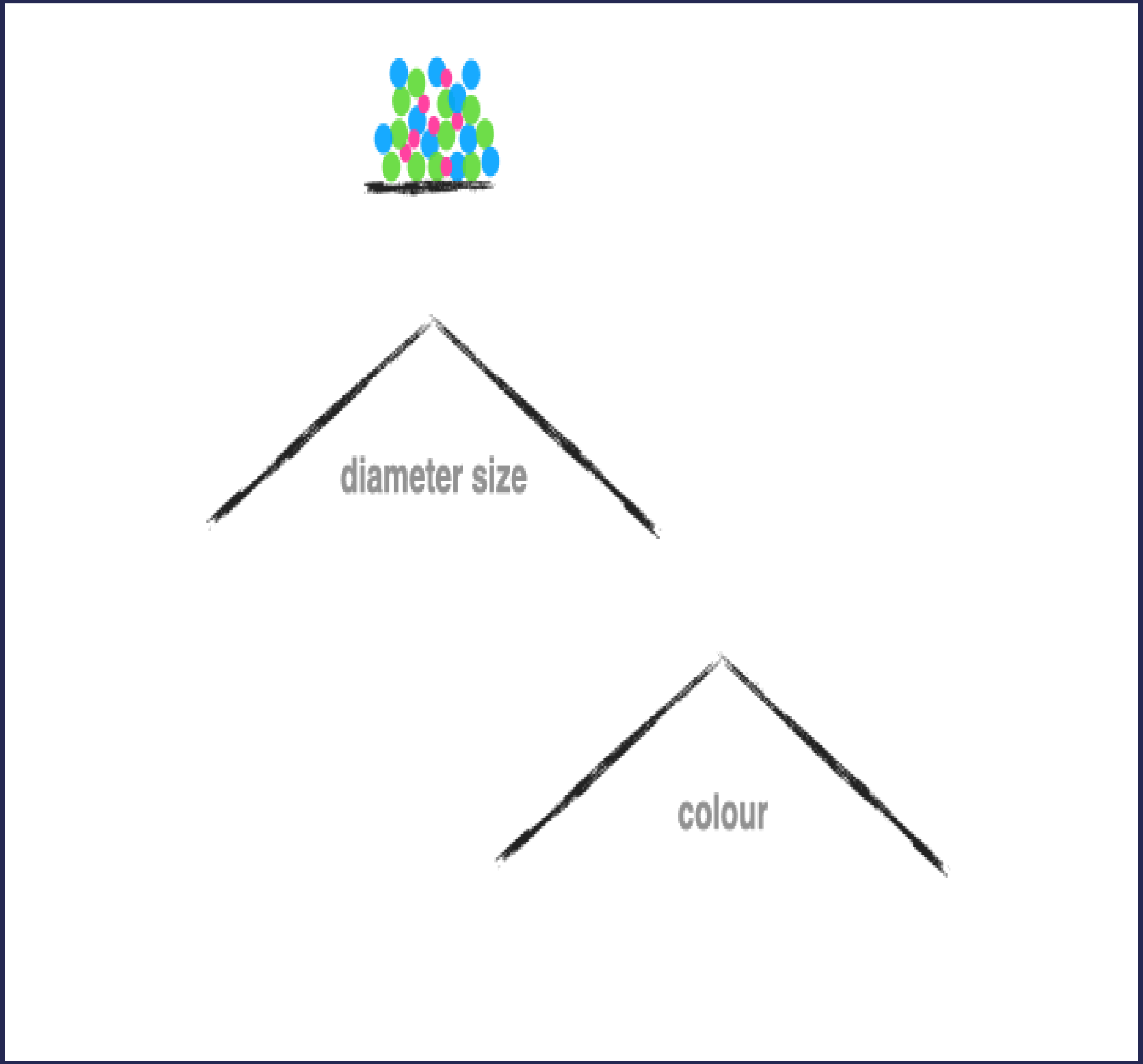
Artificial Neural Network (ANN)

- Artificial Neural Network (ANN) is a machine learning algorithm inspired by the structure and functioning of biological neural networks.
- ANN consists of interconnected nodes (neurons) organized in layers: input layer, hidden layers, and output layer.
- It uses a process called training to learn from labelled data and optimize the connections (weights) between neurons.
- ANN is widely used in High Energy Physics (HEP) to classify signal and background events based on input features.
- ANN can capture complex relationships and non-linearities in the data, making it suitable for discriminating between different event types.





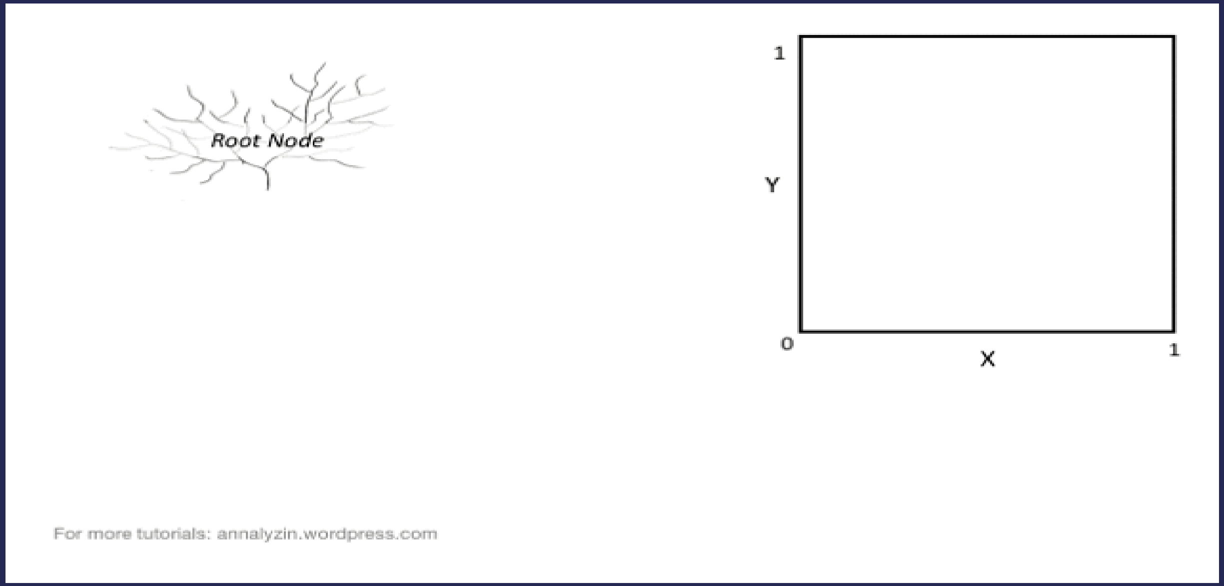




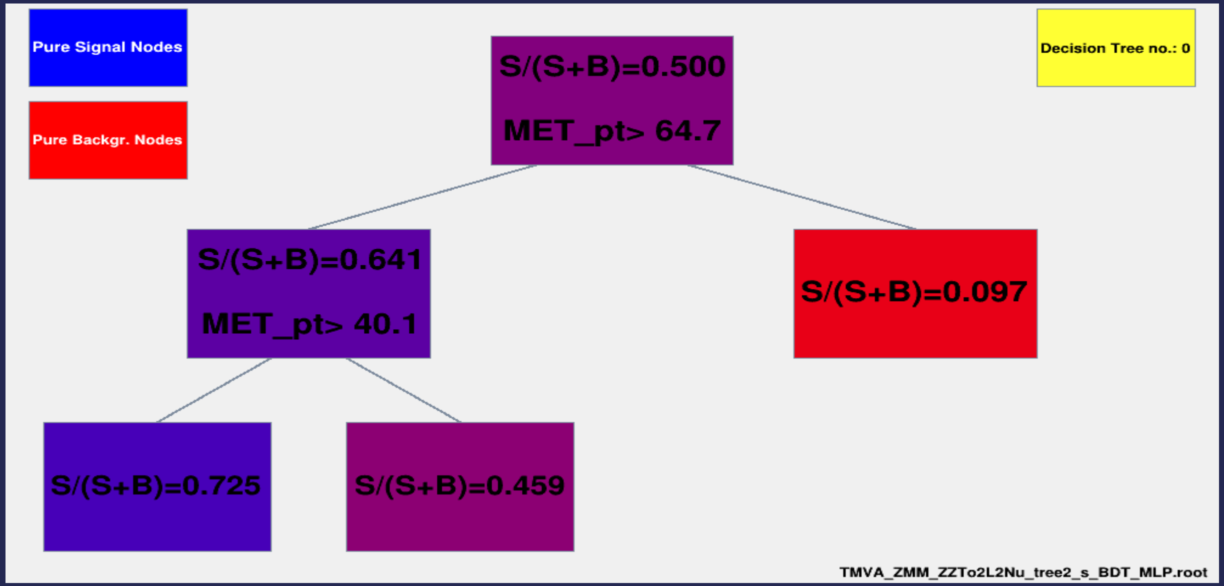
Workflow of Decision Tree

Boosted Decision Tree (BDT)

- Boosted Decision Tree (BDT) is a machine learning algorithm that combines multiple decision trees to make predictions.
- BDT sequentially builds a series of decision trees, where each subsequent tree focuses on correcting the errors made by the previous trees.
- BDT is well-suited for classification tasks and can effectively handle complex datasets with a large number of input features.
- In High Energy Physics (HEP), BDT is commonly used to classify signal and background events based on their characteristic features.



Schematic of 1 Decision in this project ->



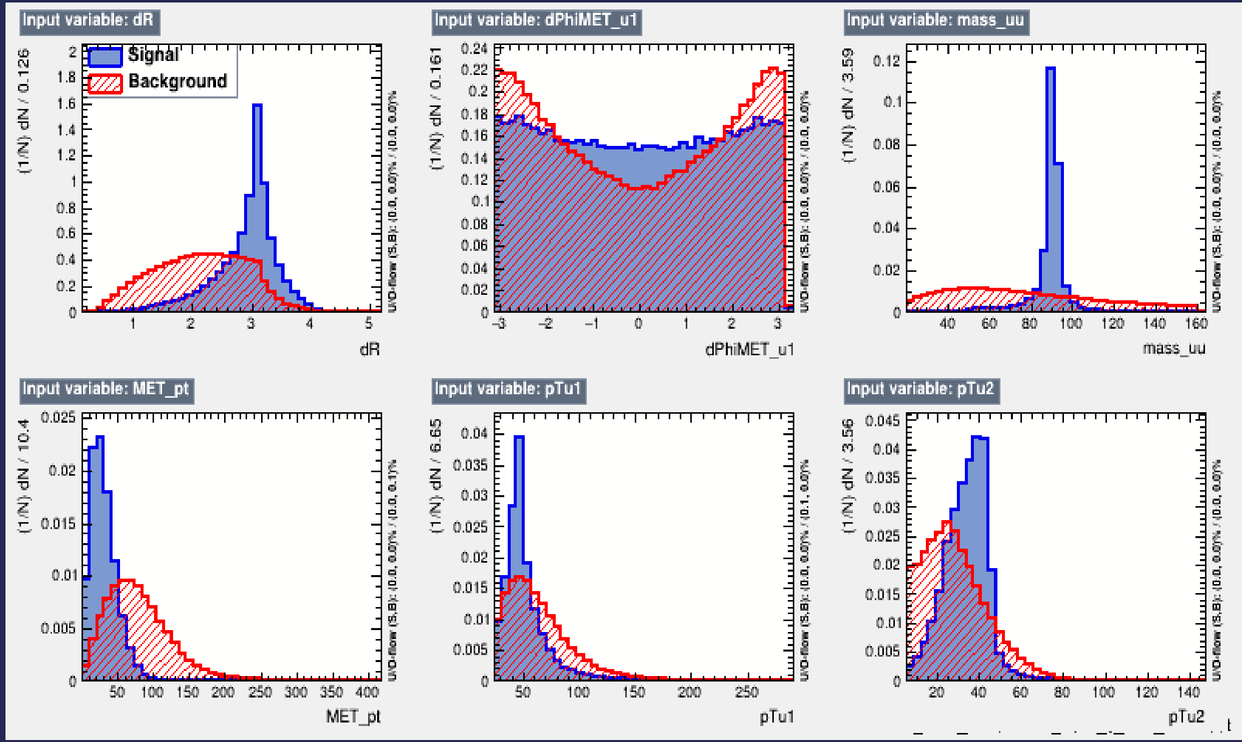
Results and Analysis

Signal : $\boldsymbol{z \rightarrow \mu^{+} + \mu^{-}}$

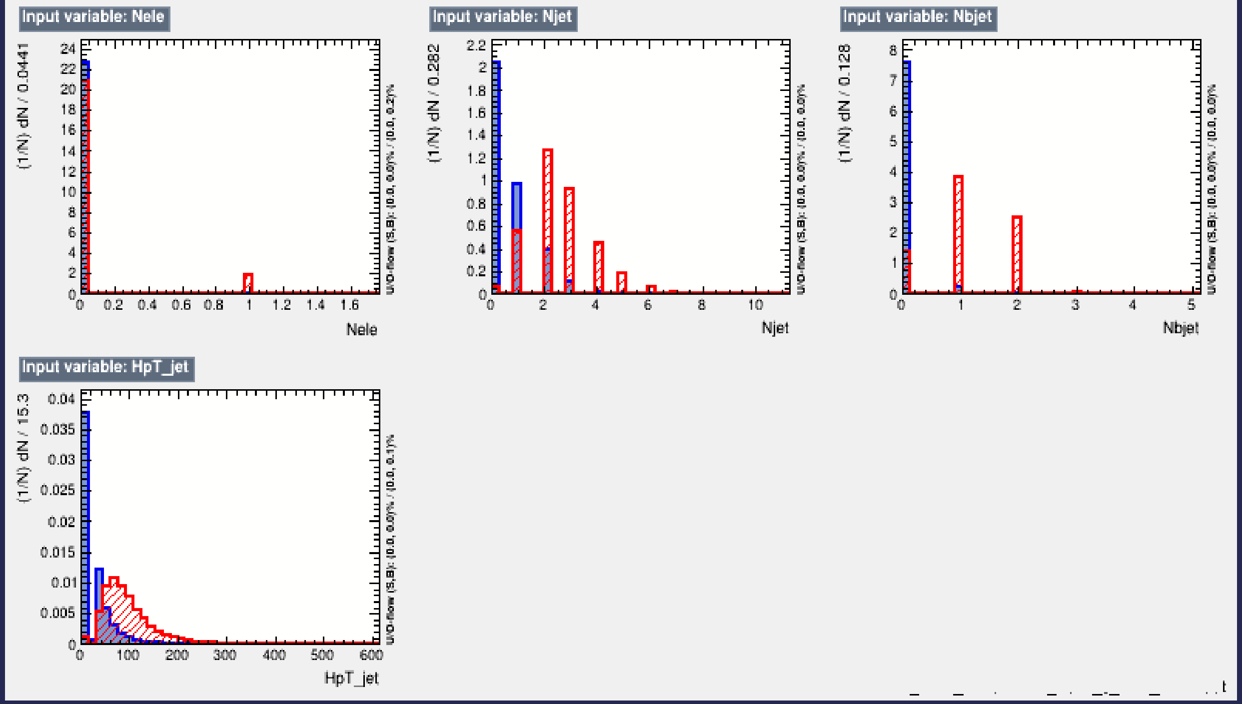
Background : $\boldsymbol{tt \rightarrow WW + bb \rightarrow 2\mu + 2\nu_{\mu} + 2b}$

Distribution of Input Variables, rank

Signal : ZMM ; Background : TTTo2Mu2Nu



Rank	Variable	Separation
1	Nbjet	6.384e-01
2	mass_uu	5.641e-01
3	Njet	5.215e-01
4	HpT_jet	5.079e-01
5	MET_pt	4.099e-01
6	dR	2.788e-01
7	pTu2	1.535e-01
8	pTu1	1.013e-01
9	Nele	3.612e-02
10	dPhiMET_u1	7.715e-03

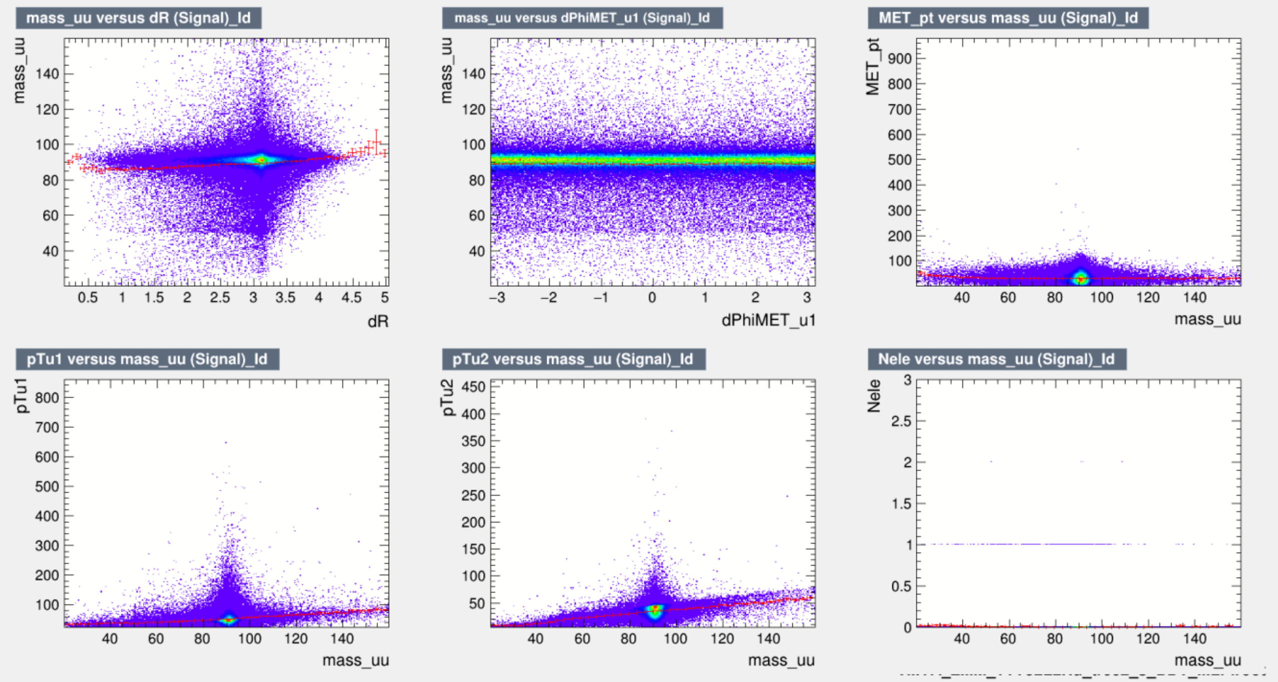


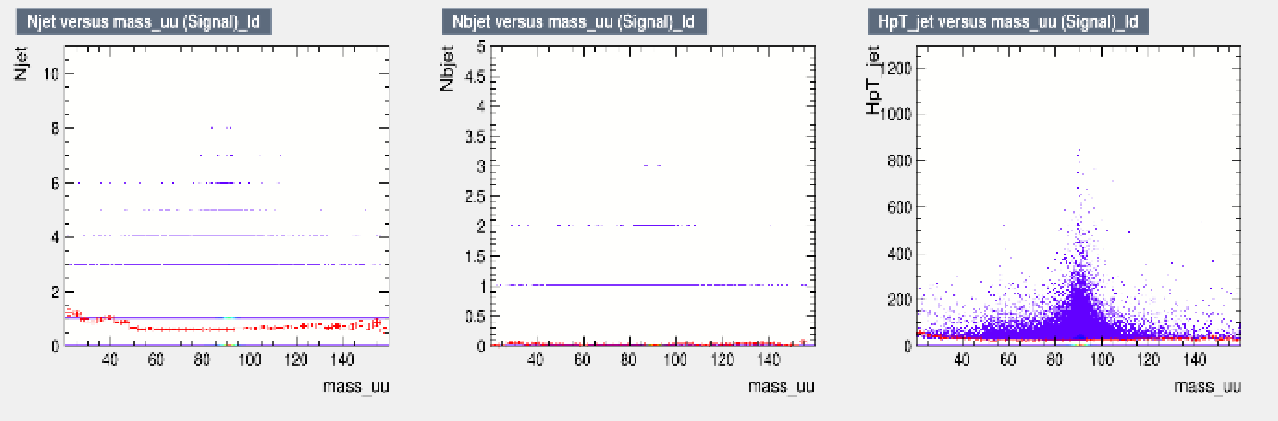
Training the algorithm

Correlation of mass with other Variables

Signal : ZMM ; Background : TTTo2Mu2Nu

Mass Correlation with other variables

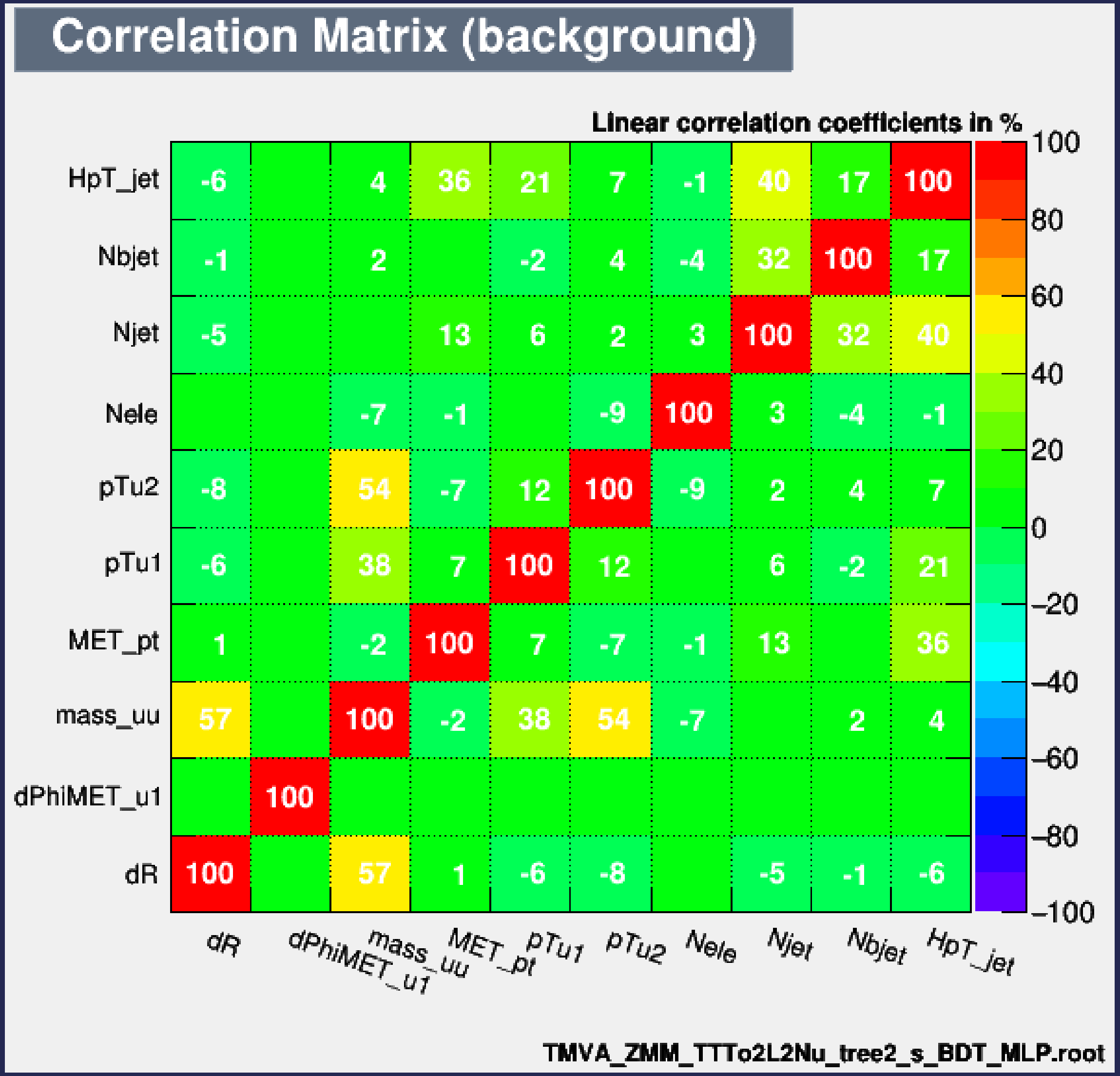


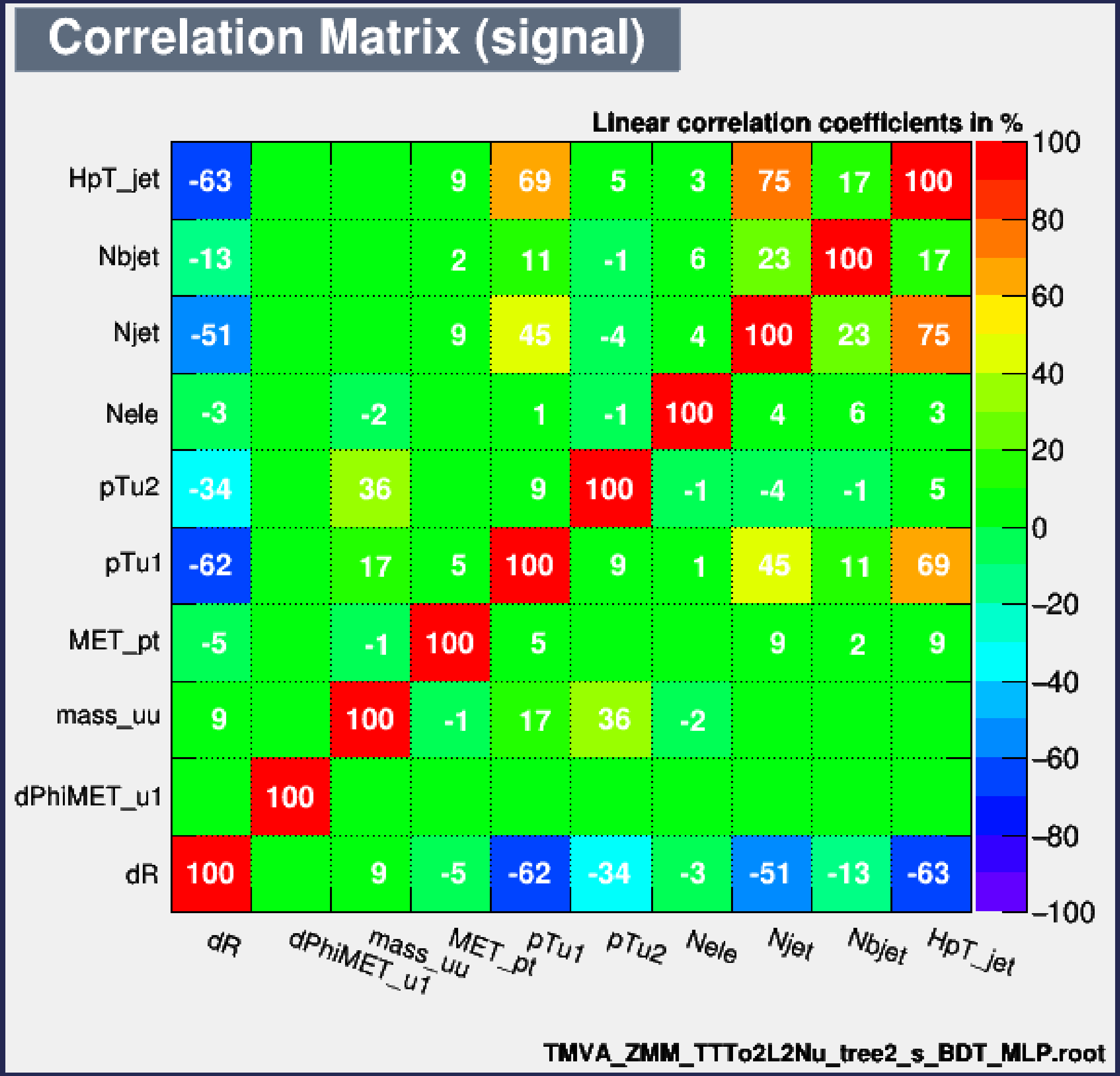


- Like every other given variable, inv_mass is actually correlated with all other variables.
- Some have –ve correlation or no correlation (like the number of electrons, jets, and b-Jets)
- Some have +ve correlations like dR, pT of Good Muon, pT of medium Muon etc.
- This correlation helps for classification data, and learning the small details of dependencies b/w the variables used

Correlation of Variables

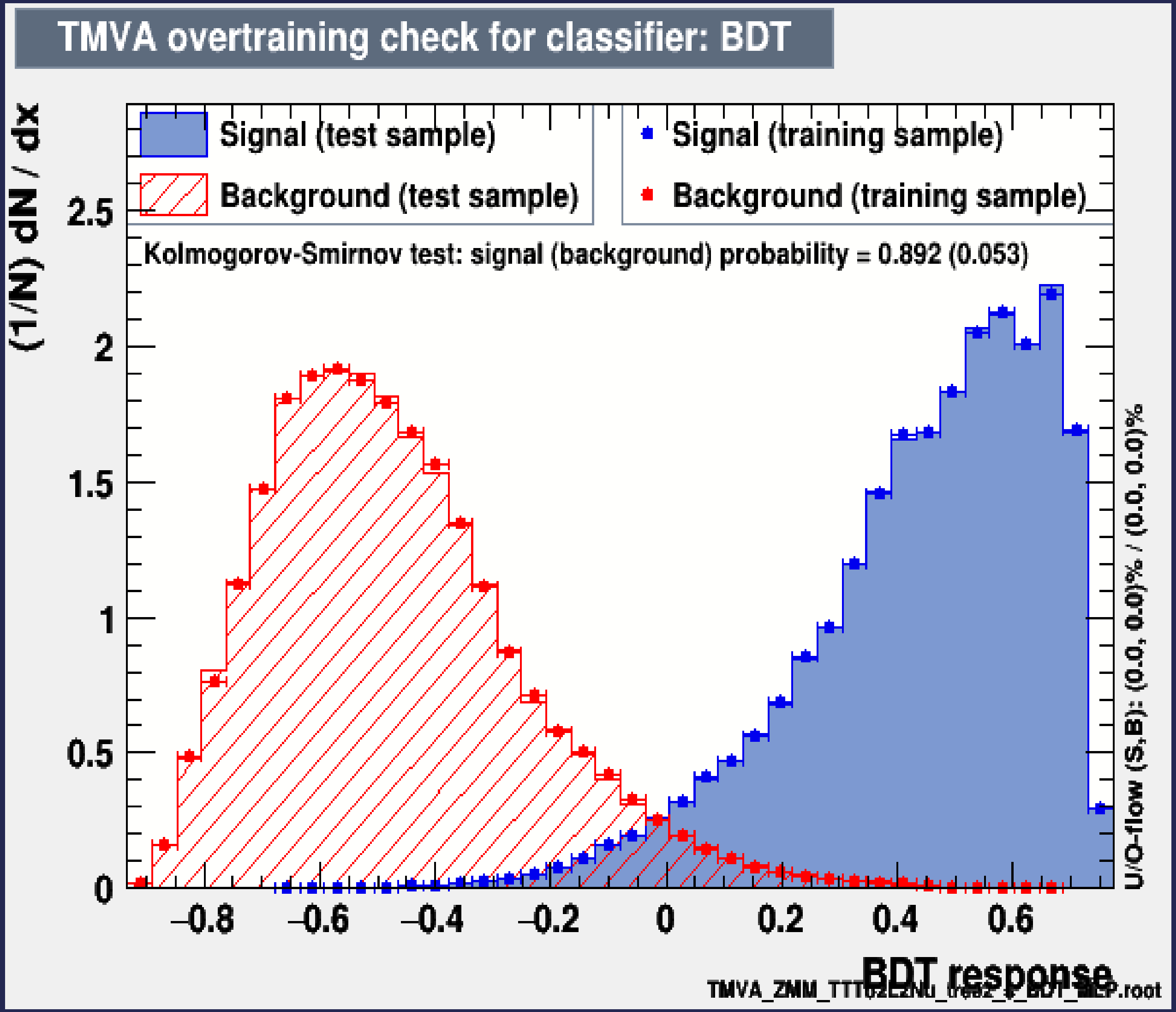
Signal : ZMM ; Background : TTTo2Mu2Nu

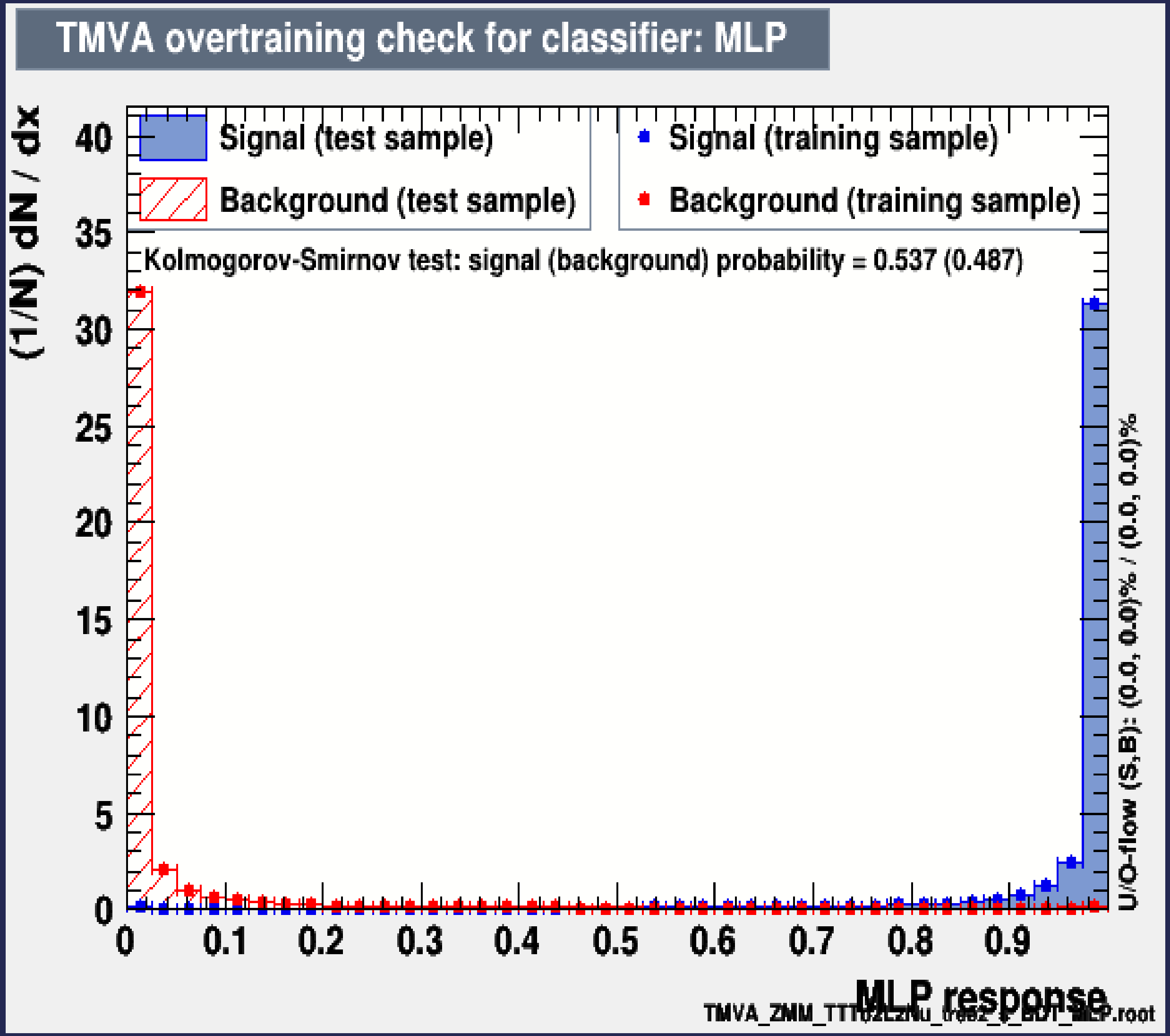




Overtraining Check & Classification

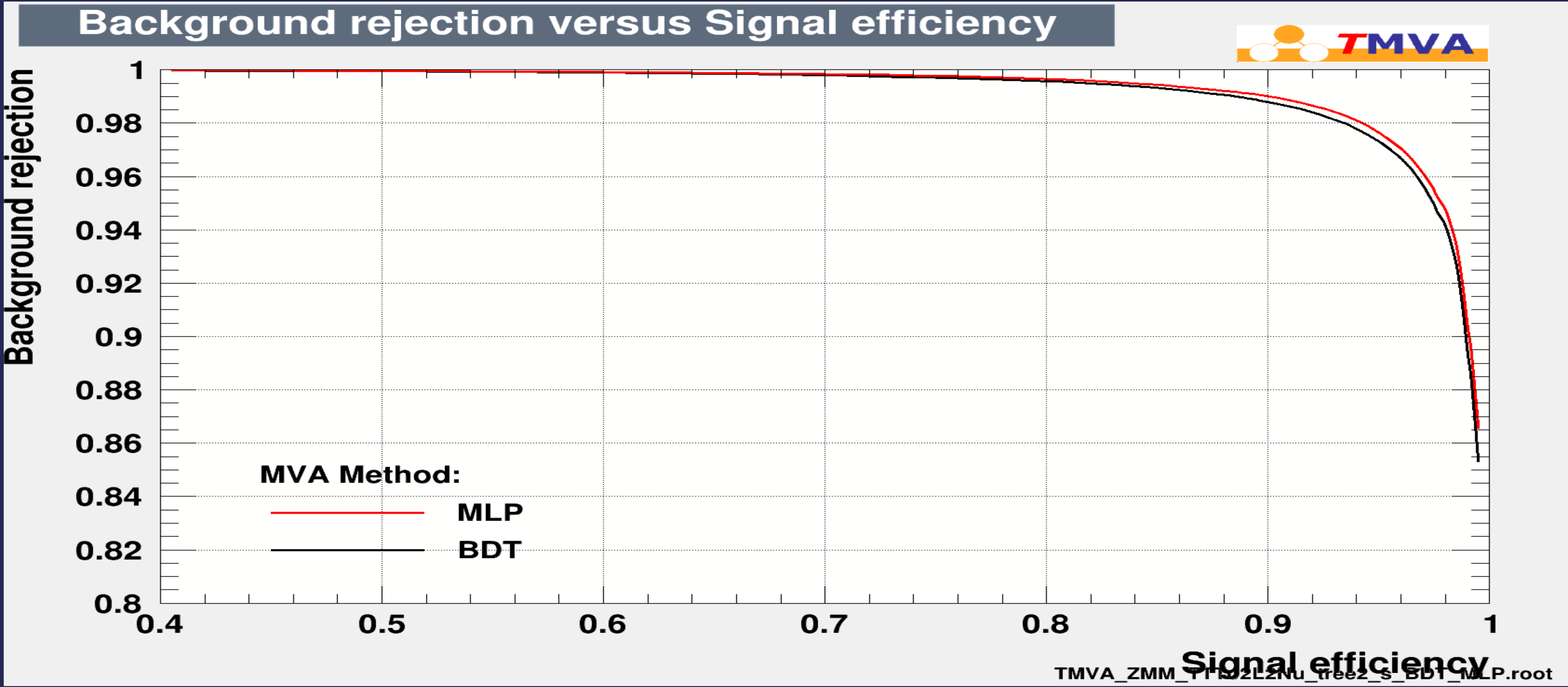
Signal : ZMM ; Background : TTTo2Mu2Nu





Network Performance : ROC

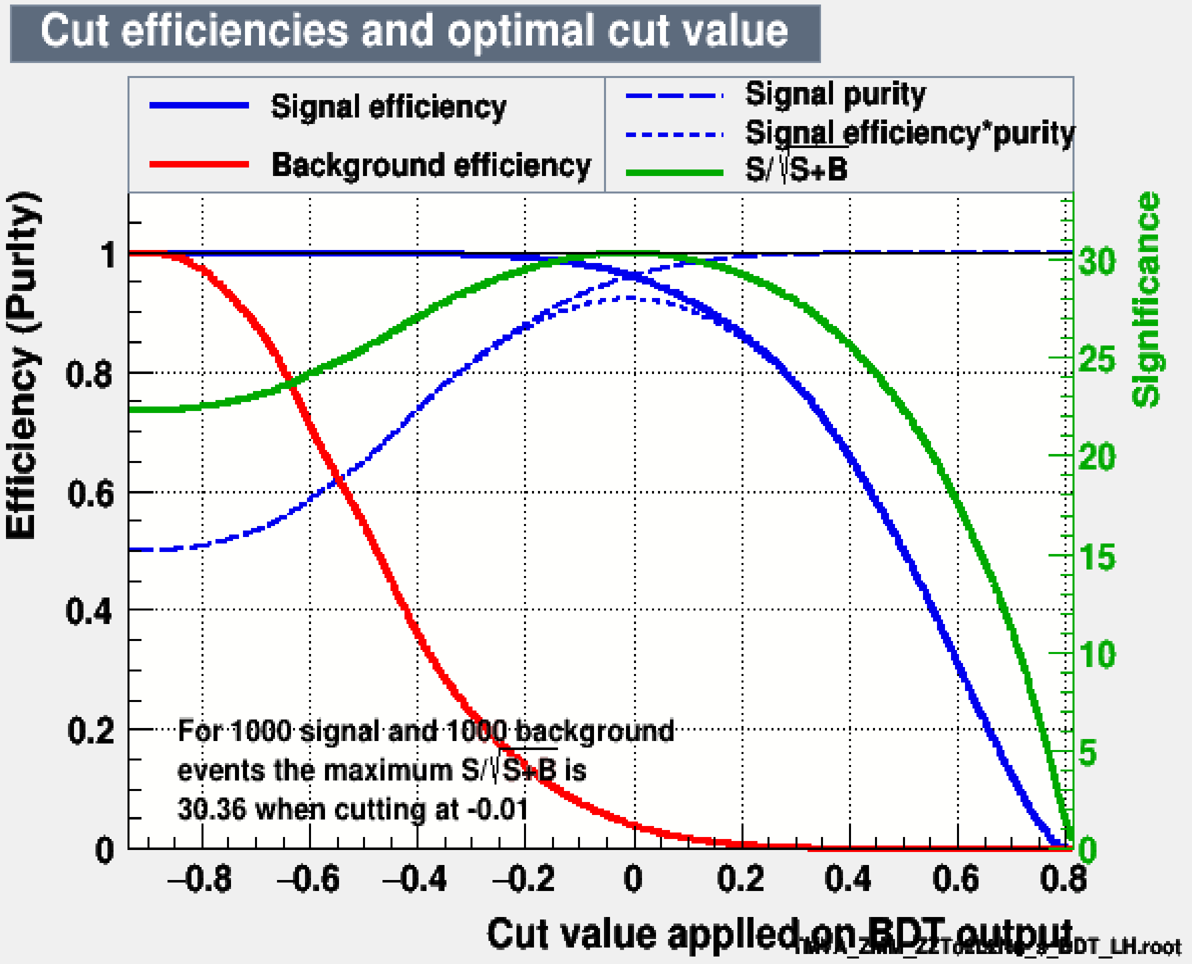
Signal : ZMM ; Background : TTTo2Mu2Nu



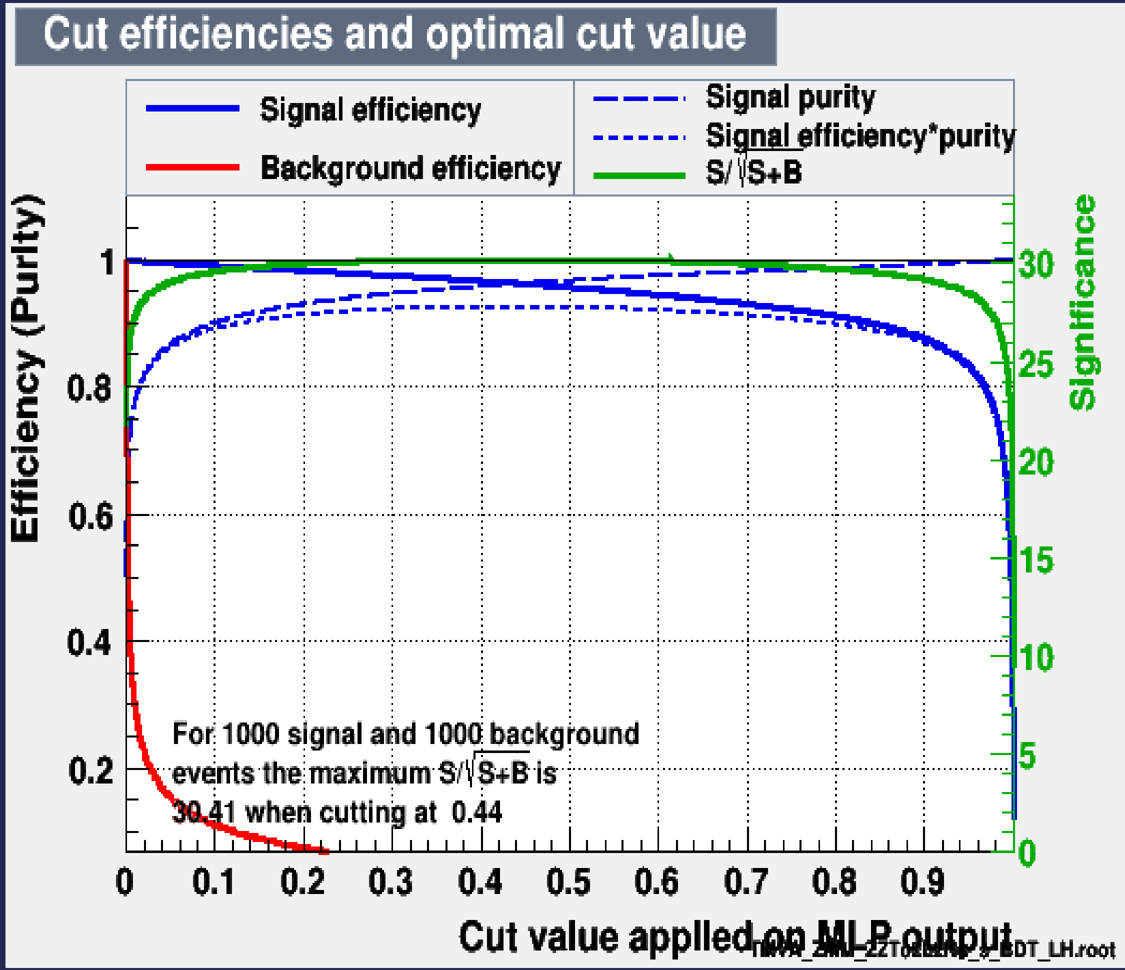
ROC Curve : receiver operating characteristic curve

Network Performance: Significance etc.

Signal : ZMM ; Background : TTTo2Mu2Nu



BDT Discriminator

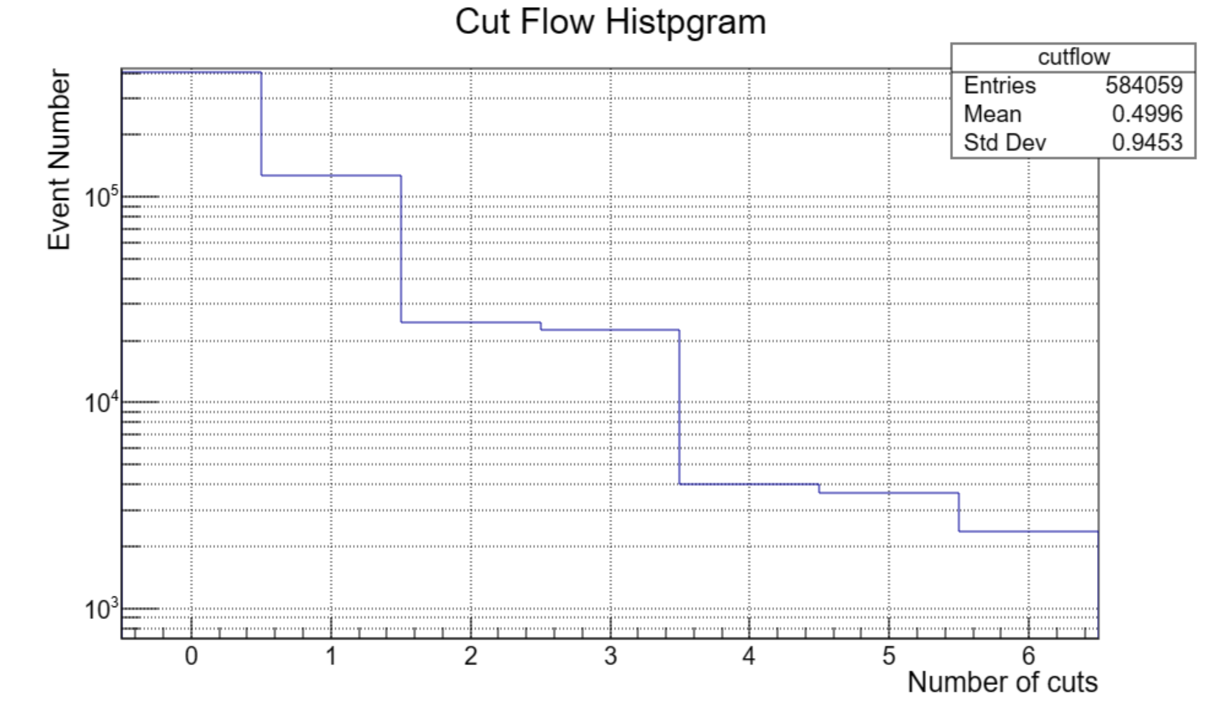


MLP Discriminator

Results from Cut based Analysis

Signal : ZMM ; Background : TTTo2Mu2Nu

TTTo2Mu2Nu Cut-flow



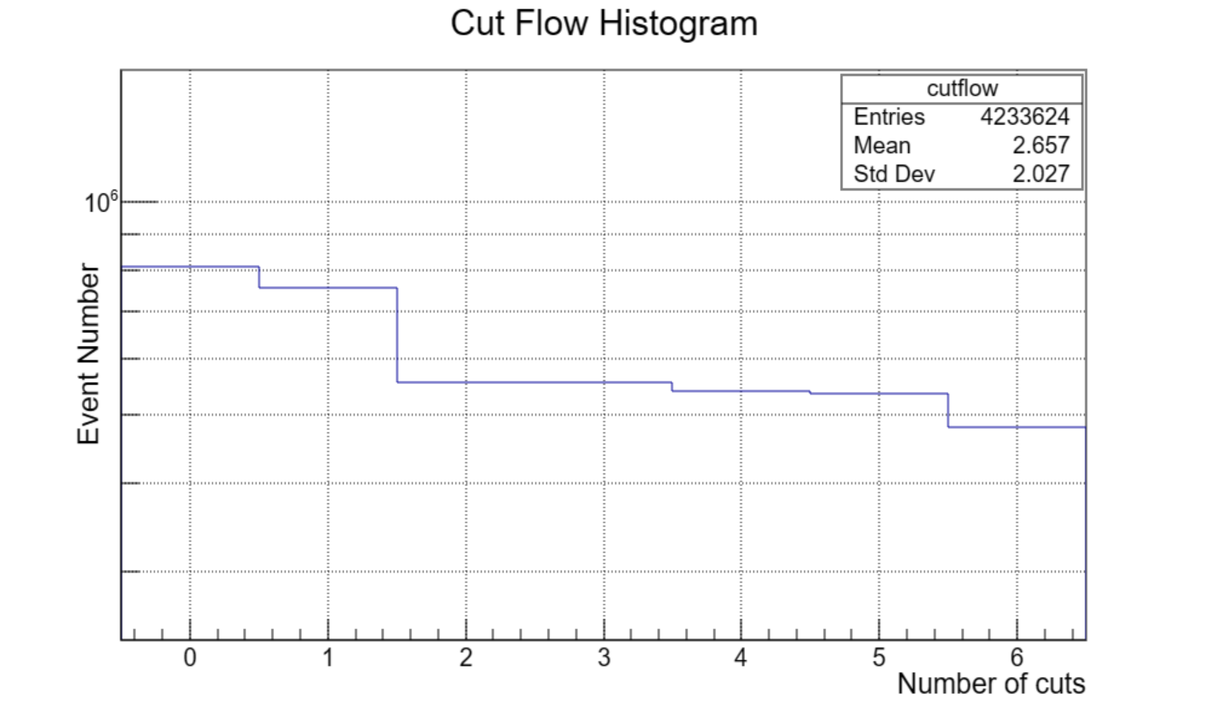
Number of bins: 7

	Bin Content	Bin Error	Bin centers	Cut_Effi(%)
Bin 1 :	400130.0	632.558	0.0	100.0
Bin 2 :	126988.0	356.354	1.0	31.737
Bin 3 :	24389.0	156.17	2.0	6.095
Bin 4 :	22549.0	150.163	3.0	5.635
Bin 5 :	3999.0	63.238	4.0	0.999
Bin 6 :	3627.0	60.225	5.0	0.906
Bin 7 :	2377.0	48.754	6.0	0.594

ZMM Cut-flow

Number of bins: 7

	Bin Content	Bin Error	Bin centers	Cut_Effi(%)
Bin 1 :	811358.0	900.754	0.0	100.0
Bin 2 :	756849.0	869.971	1.0	93.282
Bin 3 :	556232.0	745.81	2.0	68.556
Bin 4 :	554494.0	744.644	3.0	68.341
Bin 5 :	539138.0	734.26	4.0	66.449
Bin 6 :	535731.0	731.936	5.0	66.029
Bin 7 :	479822.0	692.692	6.0	59.138



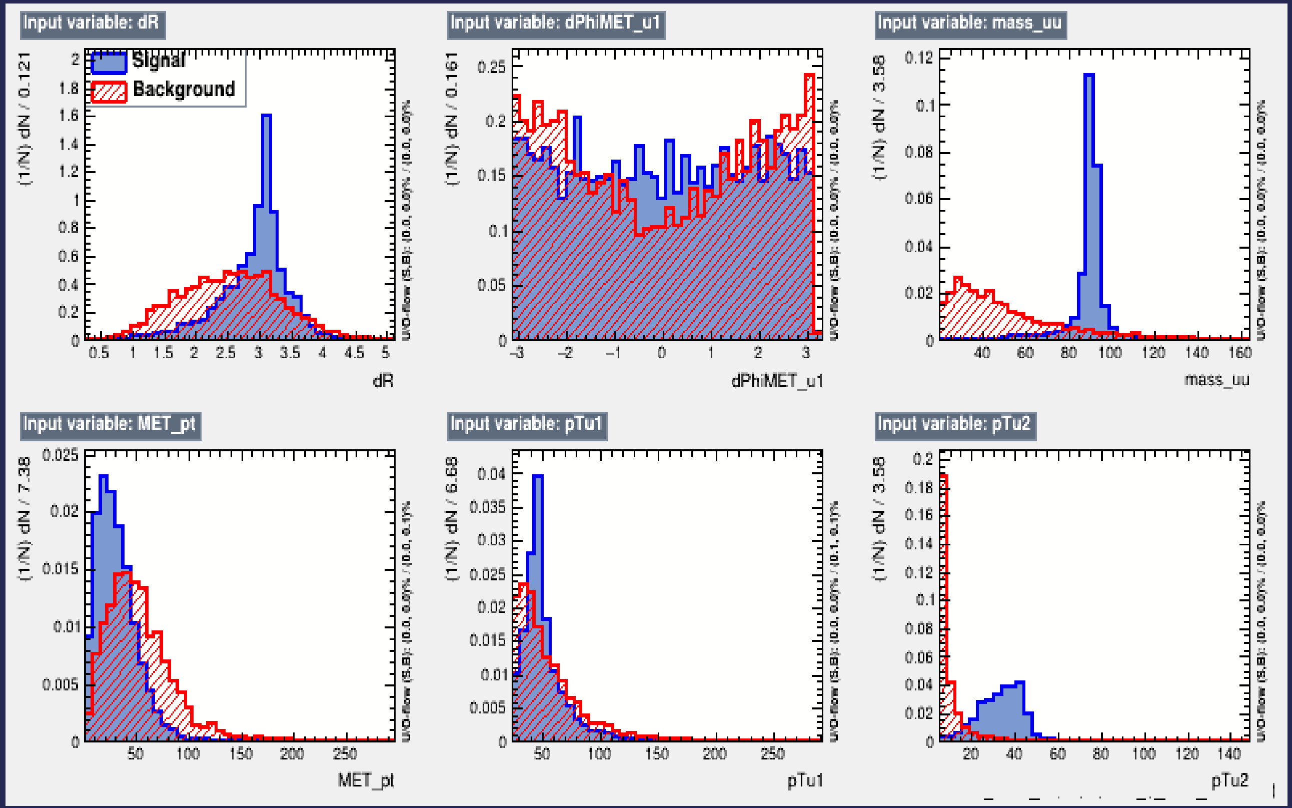
Results and Analysis

Signal : $z \rightarrow \mu^{+} + \mu^{-}$

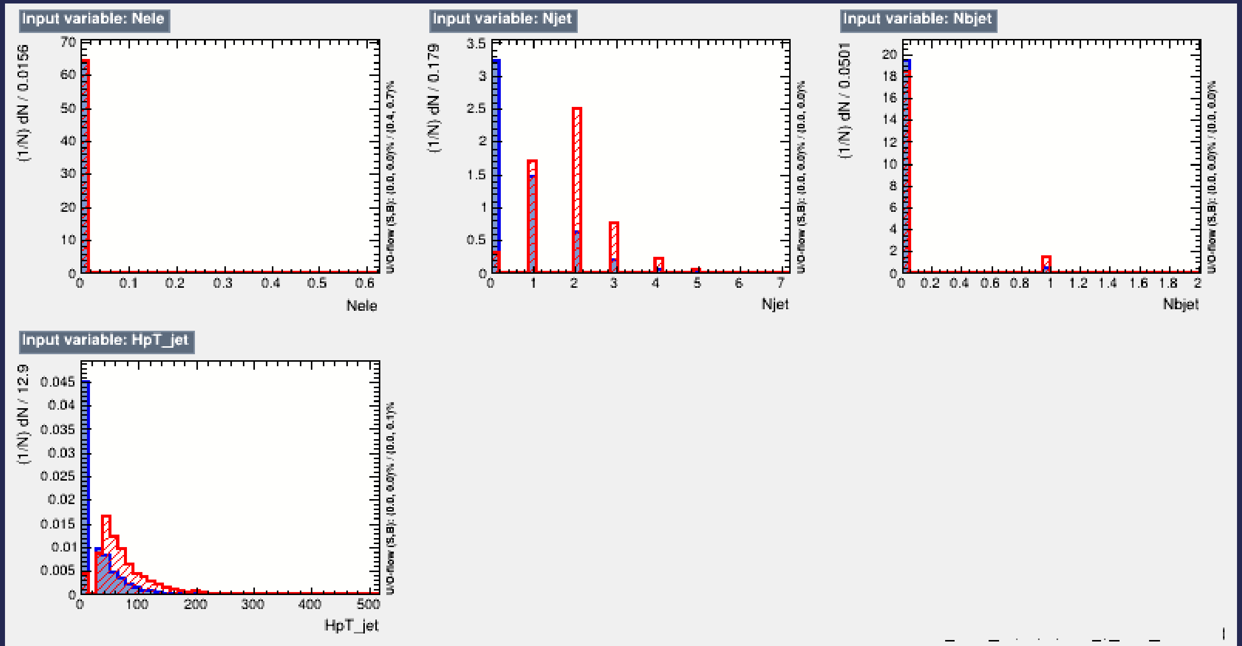
Background : $WW \rightarrow 2\mu + 2\nu_{\mu}$

Distribution of Input Variables, rank

Signal : ZMM ; Background : WJetsTo2Mu2Nu



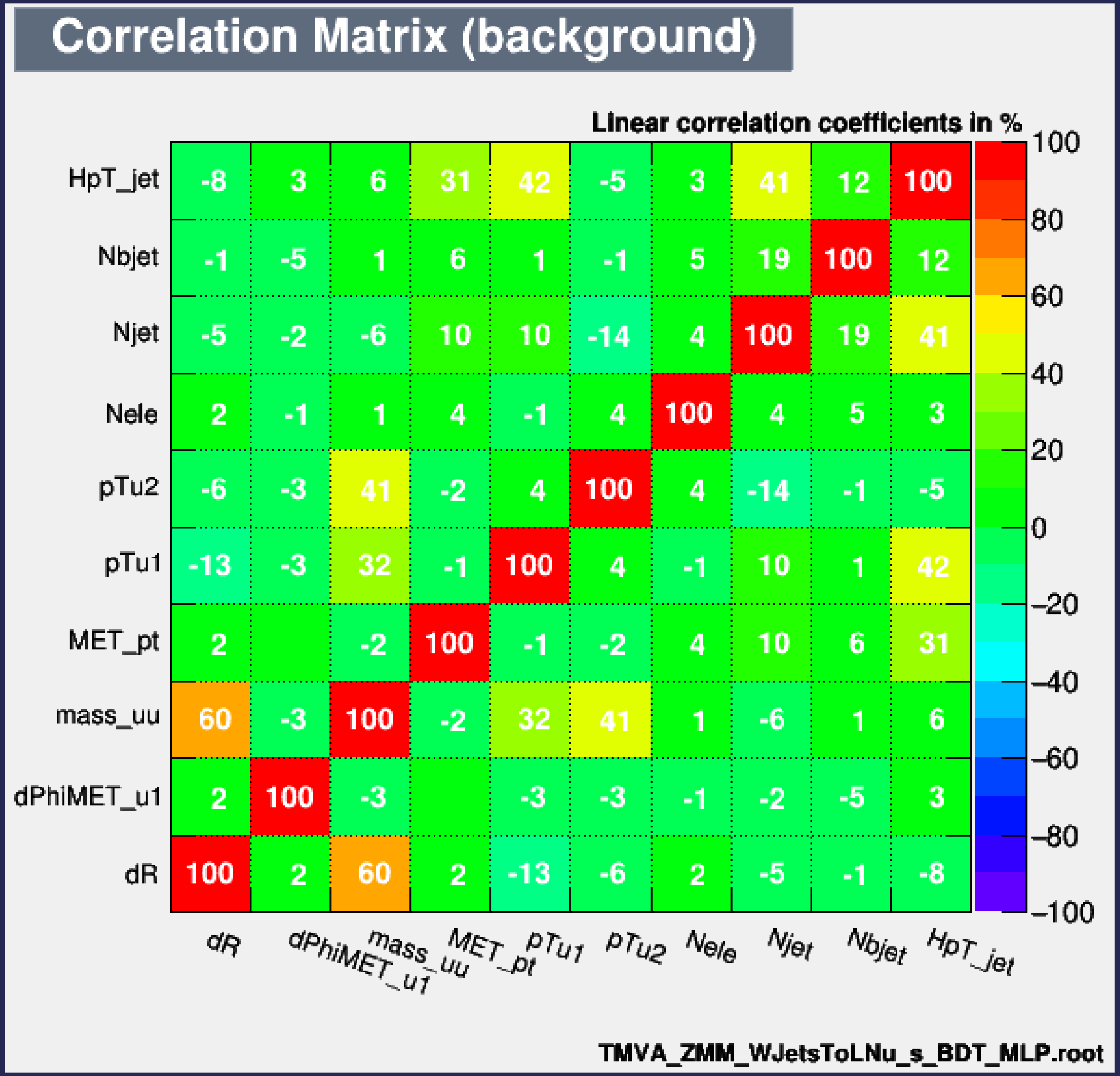
Rank	Variable	Separation
1	pTu2	7.557e-01
2	mass_uu	7.000e-01
3	Njet	3.588e-01
4	HpT_jet	3.457e-01
5	MET_pt	1.684e-01
6	dR	1.603e-01
7	pTu1	6.939e-02
8	dPhiMET_u1	1.485e-02
9	Nbjet	1.119e-02
10	Nele	0.000e+00

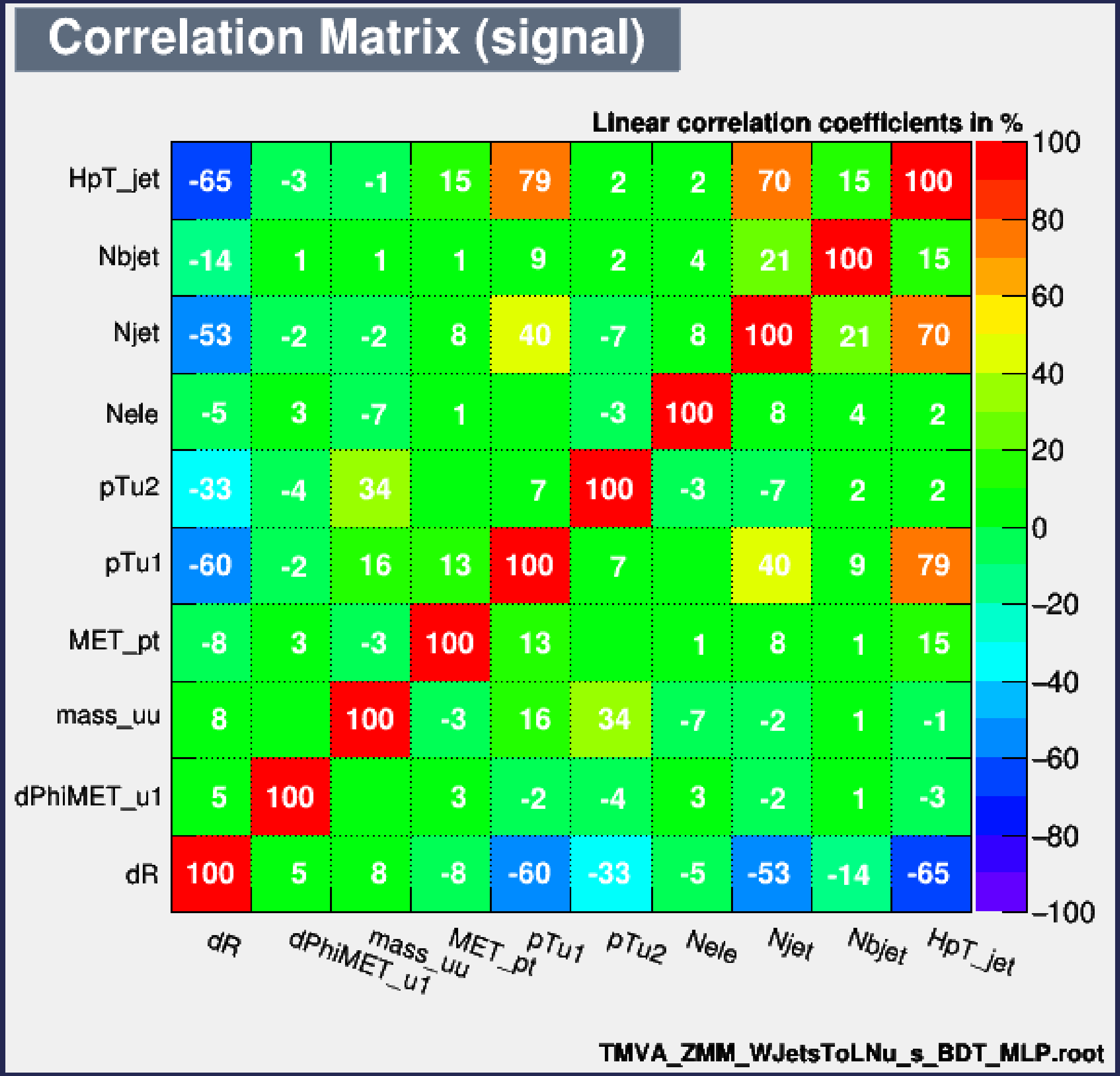


Training the algorithm

Correlation of Variables

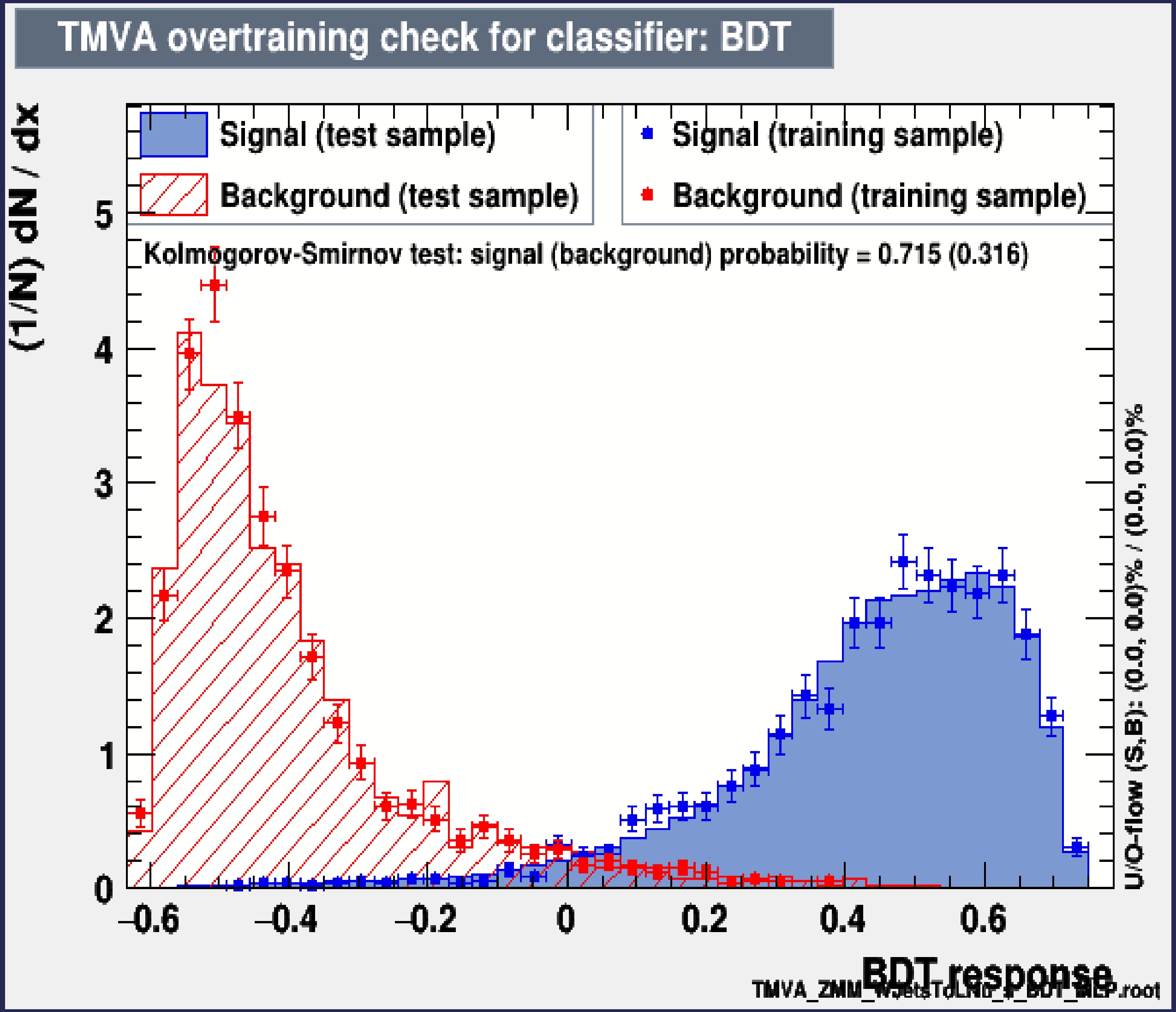
Signal : ZMM ; Background : WJetsTo2Mu2Nu

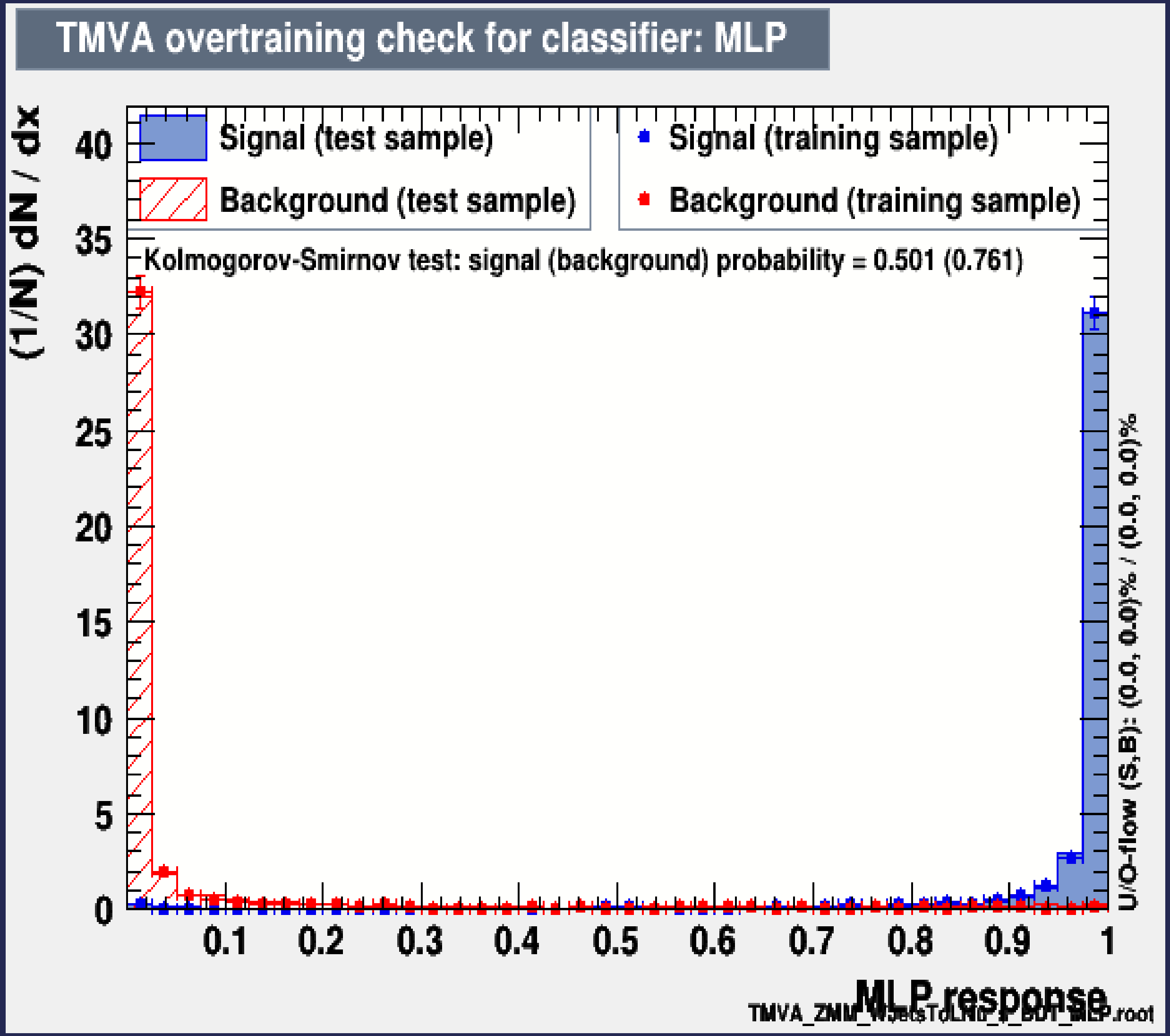




Overtraining Check & Classification

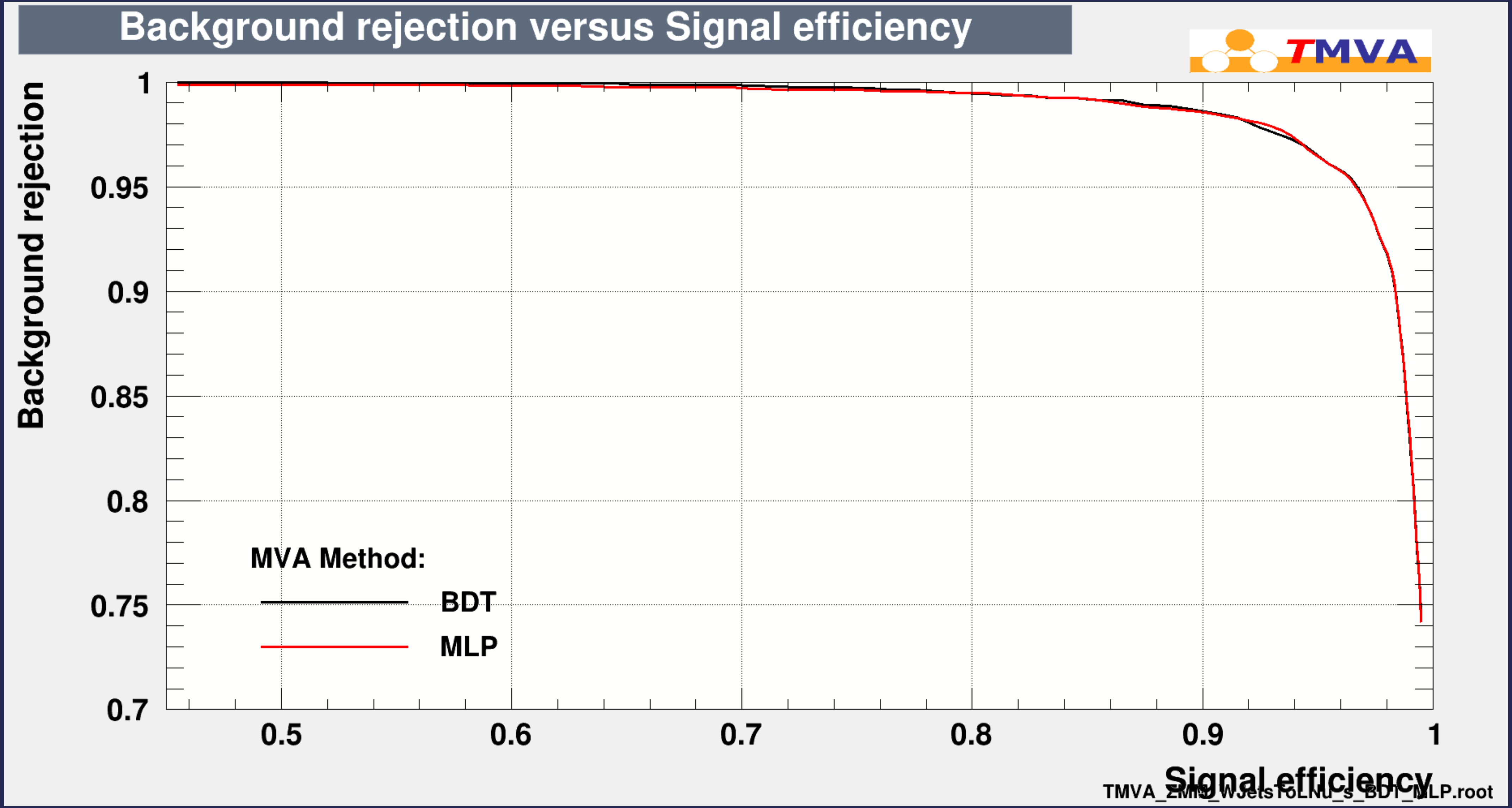
Signal : ZMM ; Background : WJetsTo2Mu2Nu





Network Performance : ROC

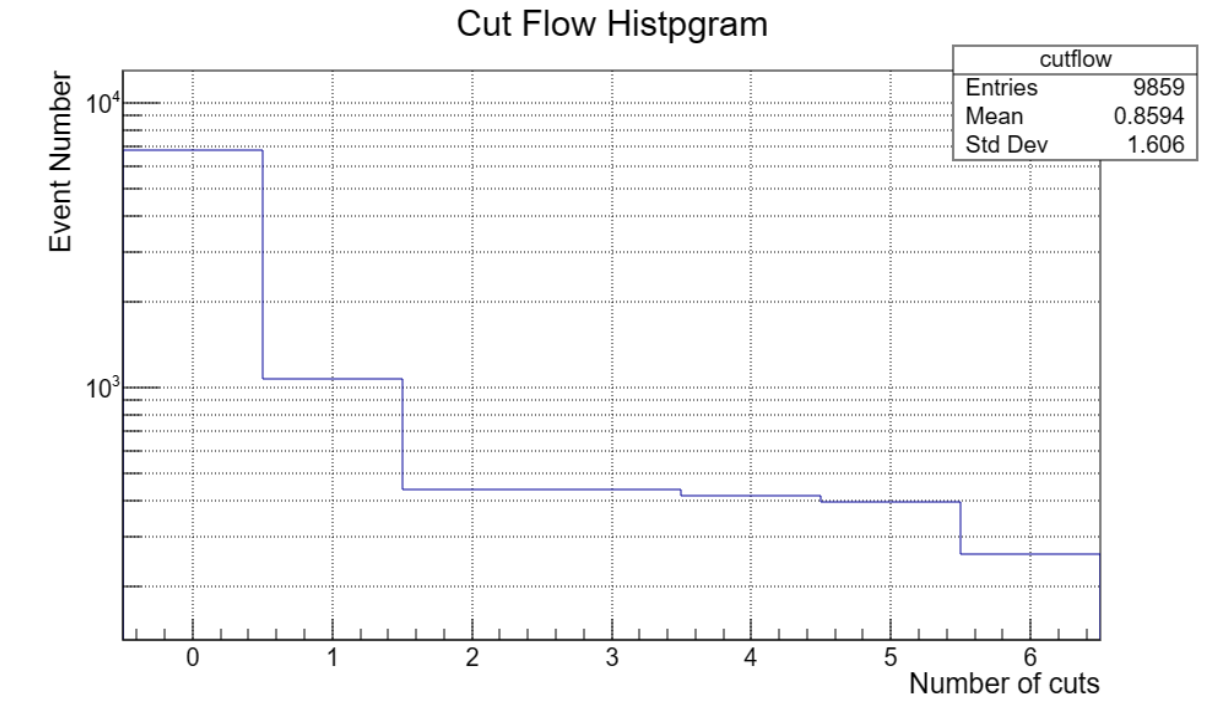
Signal : ZMM ; Background : WJetsTo2Mu2Nu



Results from Cut based Analysis

Signal : ZMM ; Background : WJetsTo2Mu2Nu

WJetsTo2Mu2Nu Cut-flow



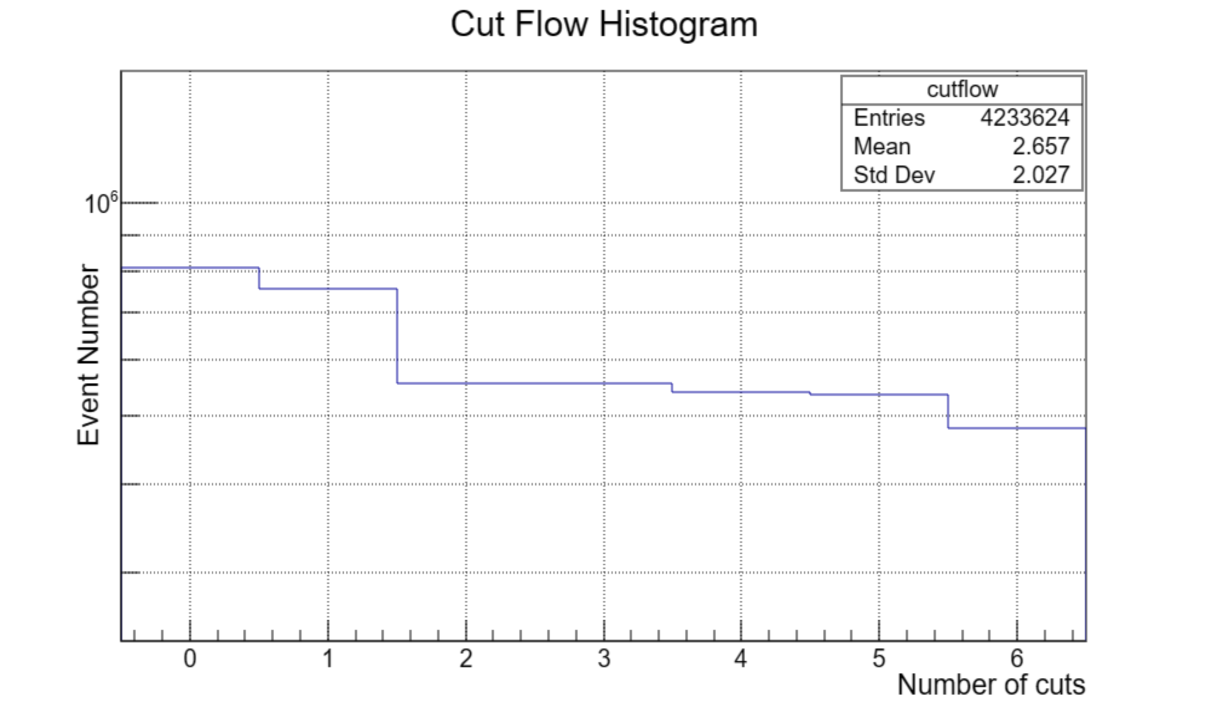
Number of bins: 7

	Bin Content	Bin Error	Bin cente	Cut_Effi(%)
Bin 1 :	6840.0	82.704	0.0	100.0
Bin 2 :	1068.0	32.68	1.0	15.614
Bin 3 :	439.0	20.952	2.0	6.418
Bin 4 :	439.0	20.952	3.0	6.418
Bin 5 :	415.0	20.372	4.0	6.067
Bin 6 :	398.0	19.95	5.0	5.819
Bin 7 :	260.0	16.125	6.0	3.801

ZMM Cut-flow

Number of bins: 7

	Bin Content	Bin Error	Bin centers	Cut_Effi(%)
Bin 1 :	811358.0	900.754	0.0	100.0
Bin 2 :	756849.0	869.971	1.0	93.282
Bin 3 :	556232.0	745.81	2.0	68.556
Bin 4 :	554494.0	744.644	3.0	68.341
Bin 5 :	539138.0	734.26	4.0	66.449
Bin 6 :	535731.0	731.936	5.0	66.029
Bin 7 :	479822.0	692.692	6.0	59.138



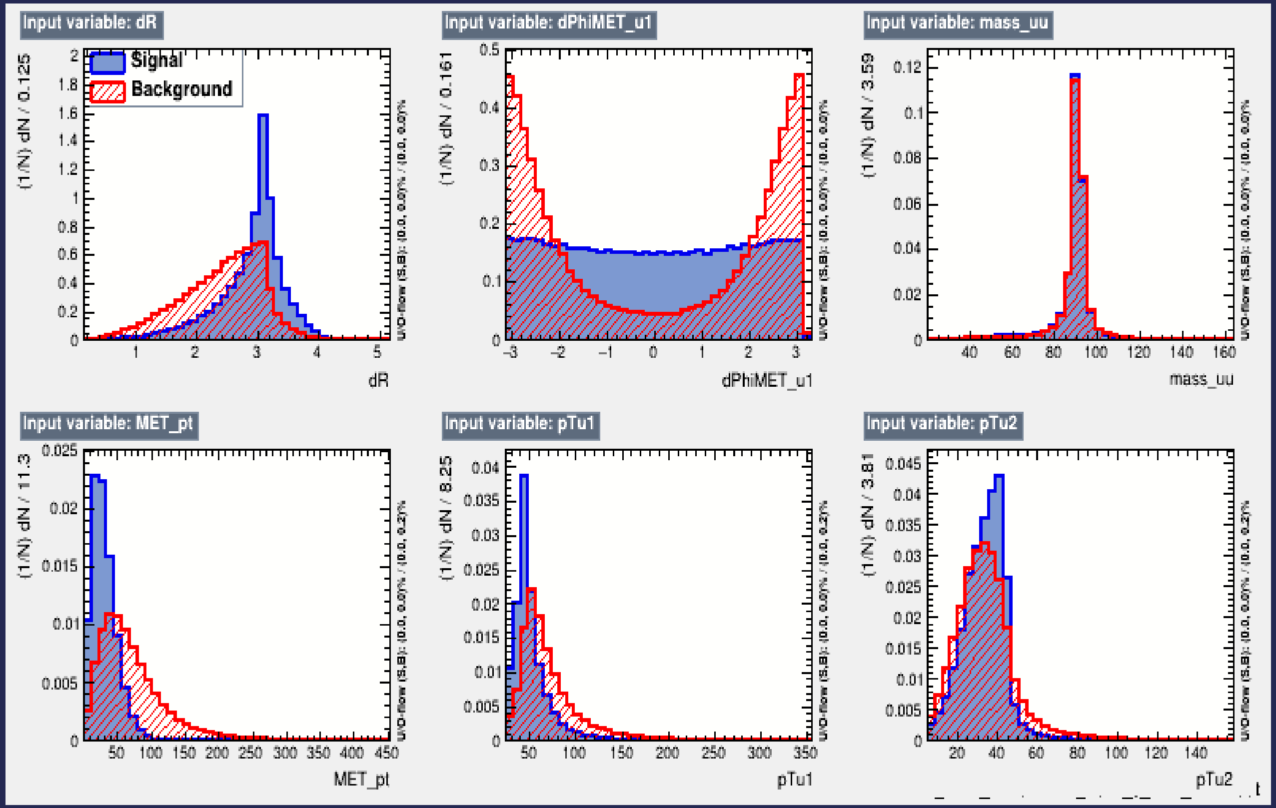
Results and Analysis

Signal : $z \rightarrow \mu^{+} + \mu^{-}$

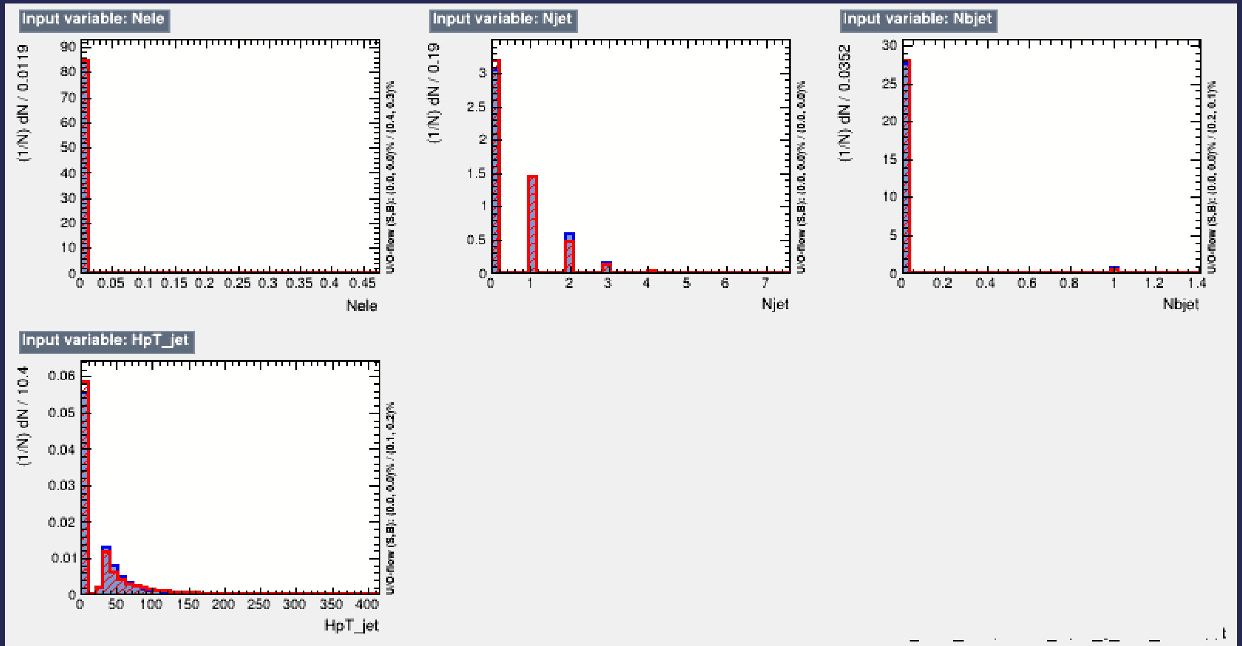
Background : $\mathrm{ZZ} \rightarrow 2\mu + 2\nu_{\mu}$

Distribution of Input Variables, rank

Signal : ZMM ; Background : ZZTo2Mu2Nu



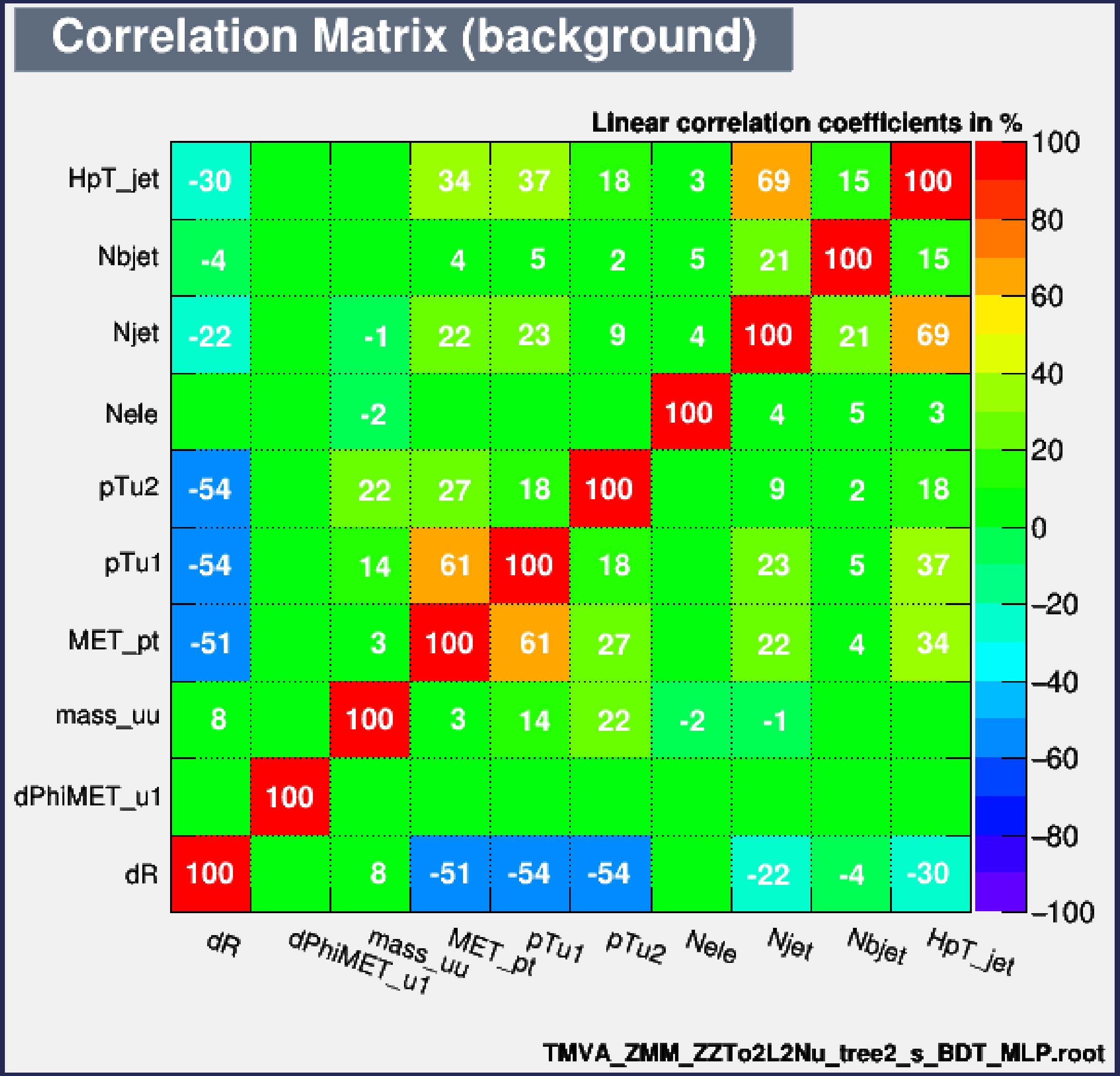
Rank	Variable	Separation
1	MET_pt	2.940e-01
2	dR	1.583e-01
3	pTu1	1.547e-01
4	dPhiMET_u1	1.276e-01
5	pTu2	3.961e-02
6	HpT_jet	4.955e-03
7	mass_uu	3.611e-03
8	Njet	2.016e-03
9	Nbjet	1.056e-03
10	Nele	0.000e+00

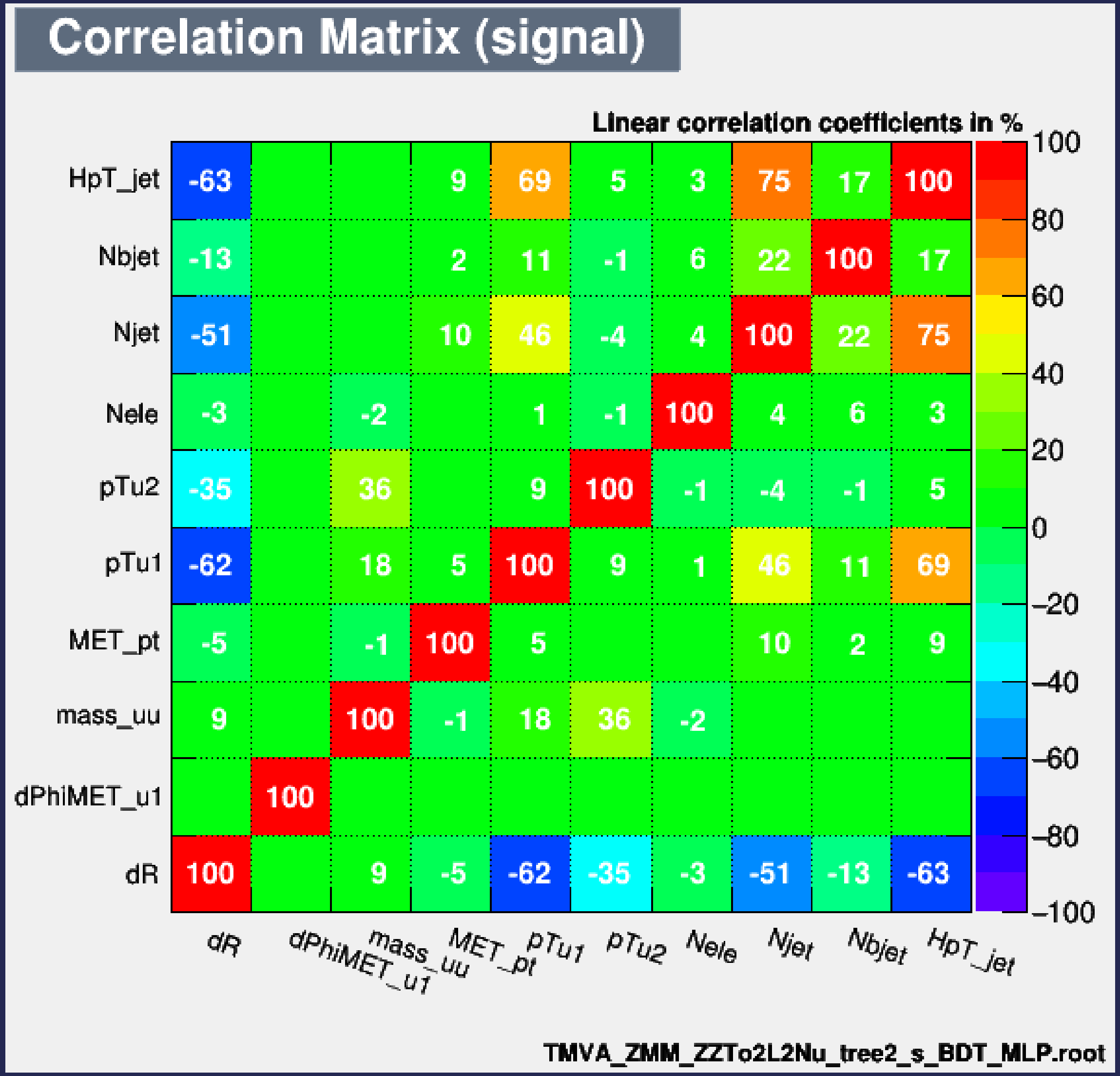


Training the algorithm

Correlation of Variables

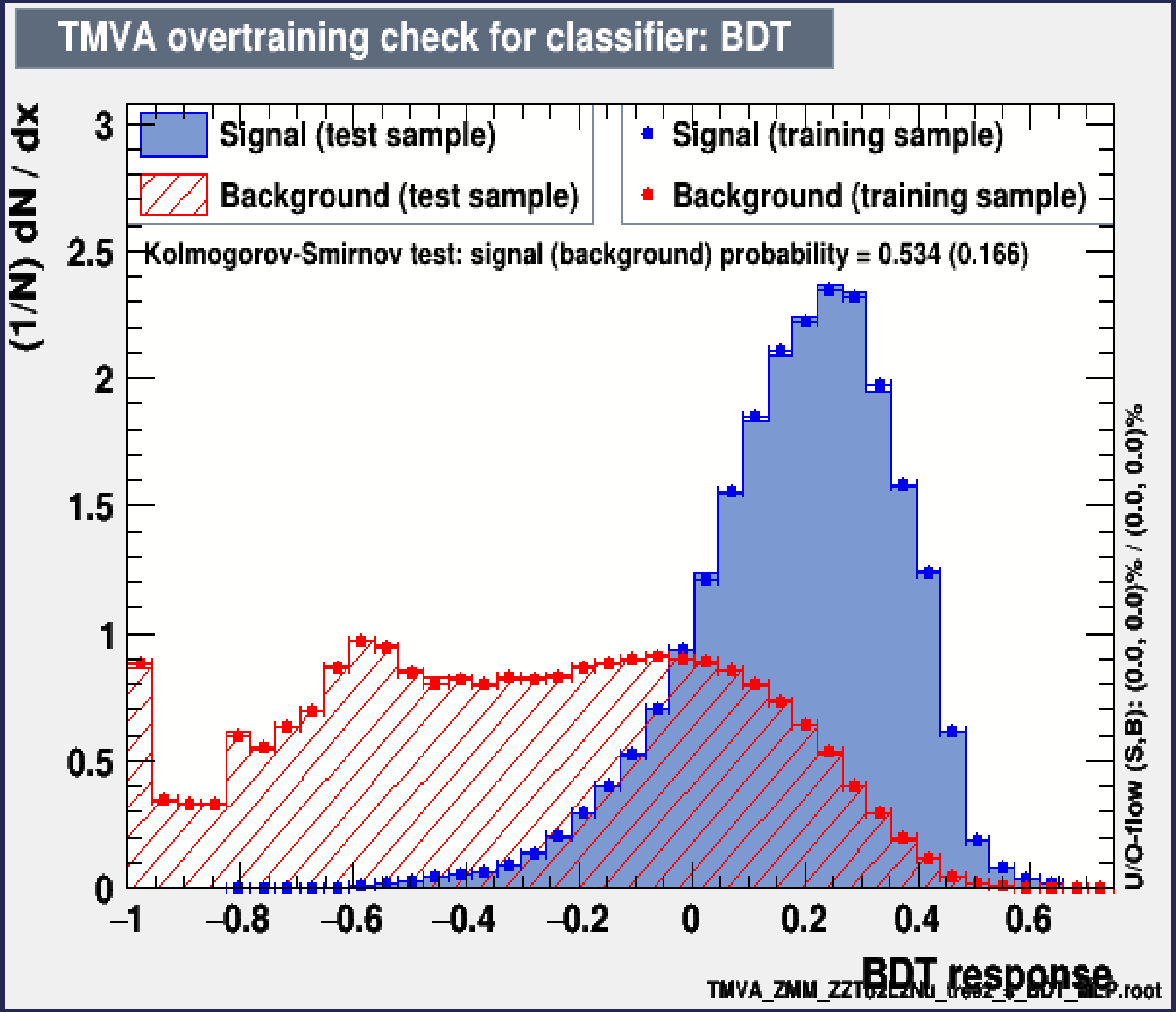
Signal : ZMM ; Background : ZZTo2Mu2Nu

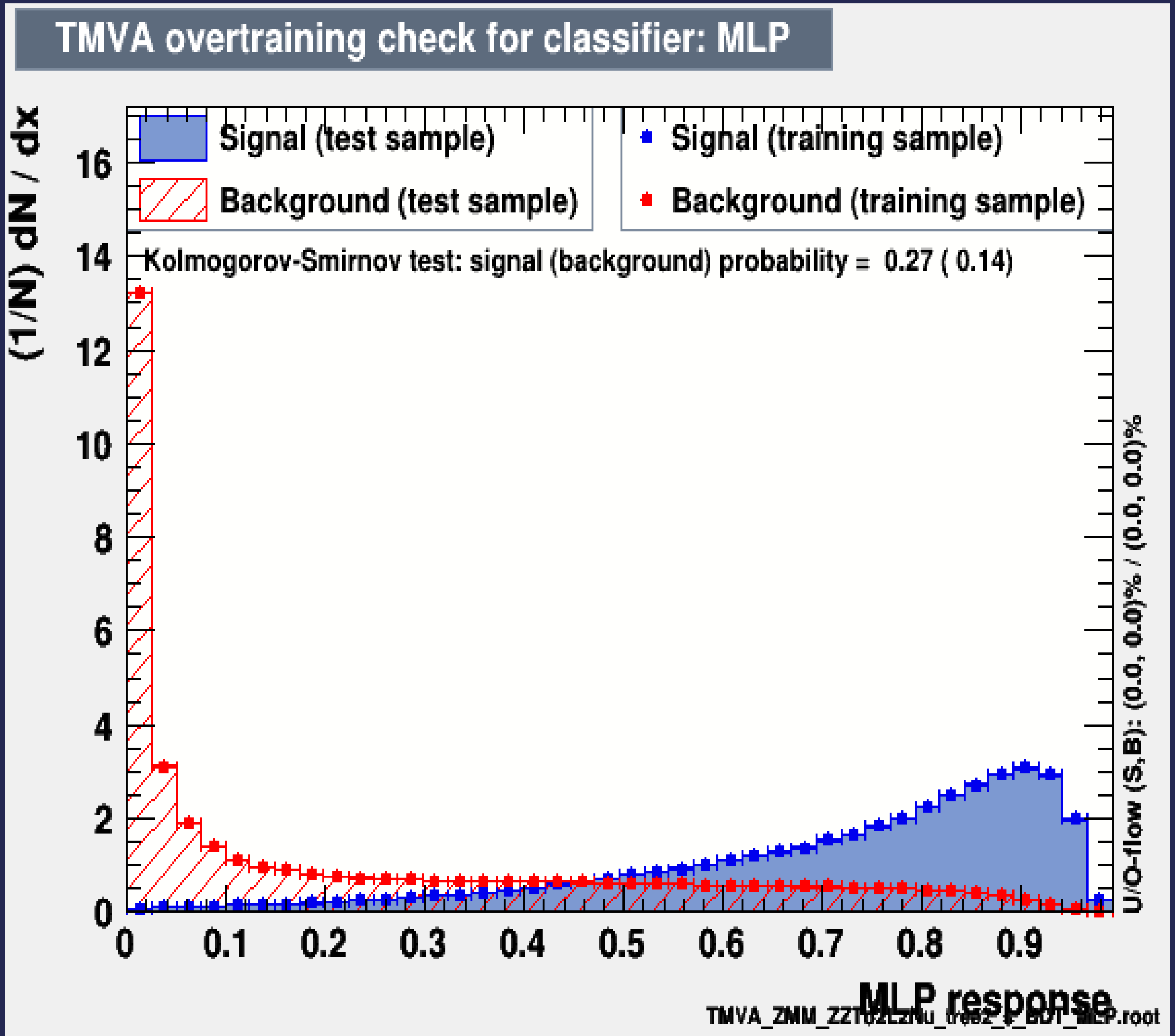




Overtraining Check & Classification

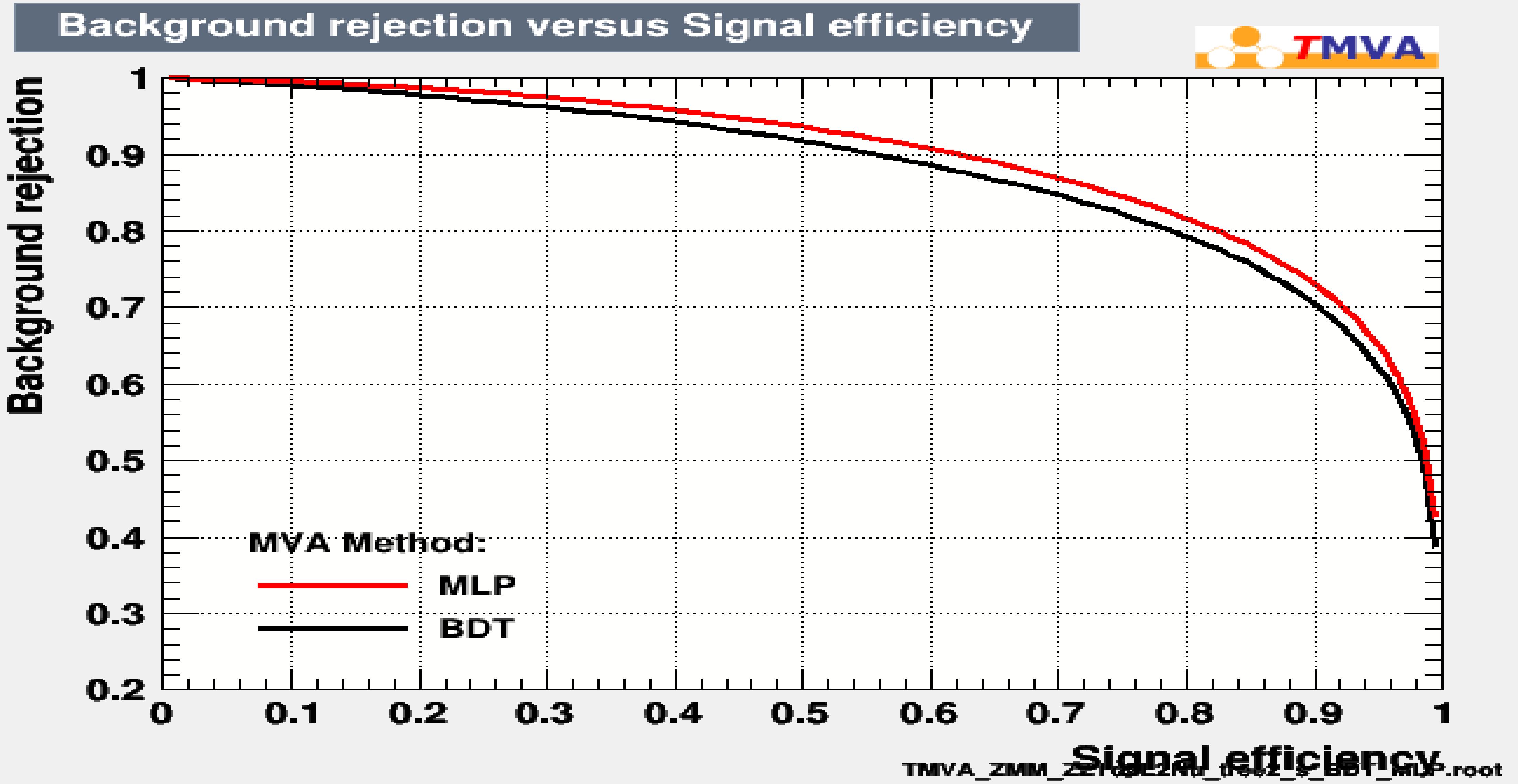
Signal : ZMM ; Background : ZZTo2Mu2Nu





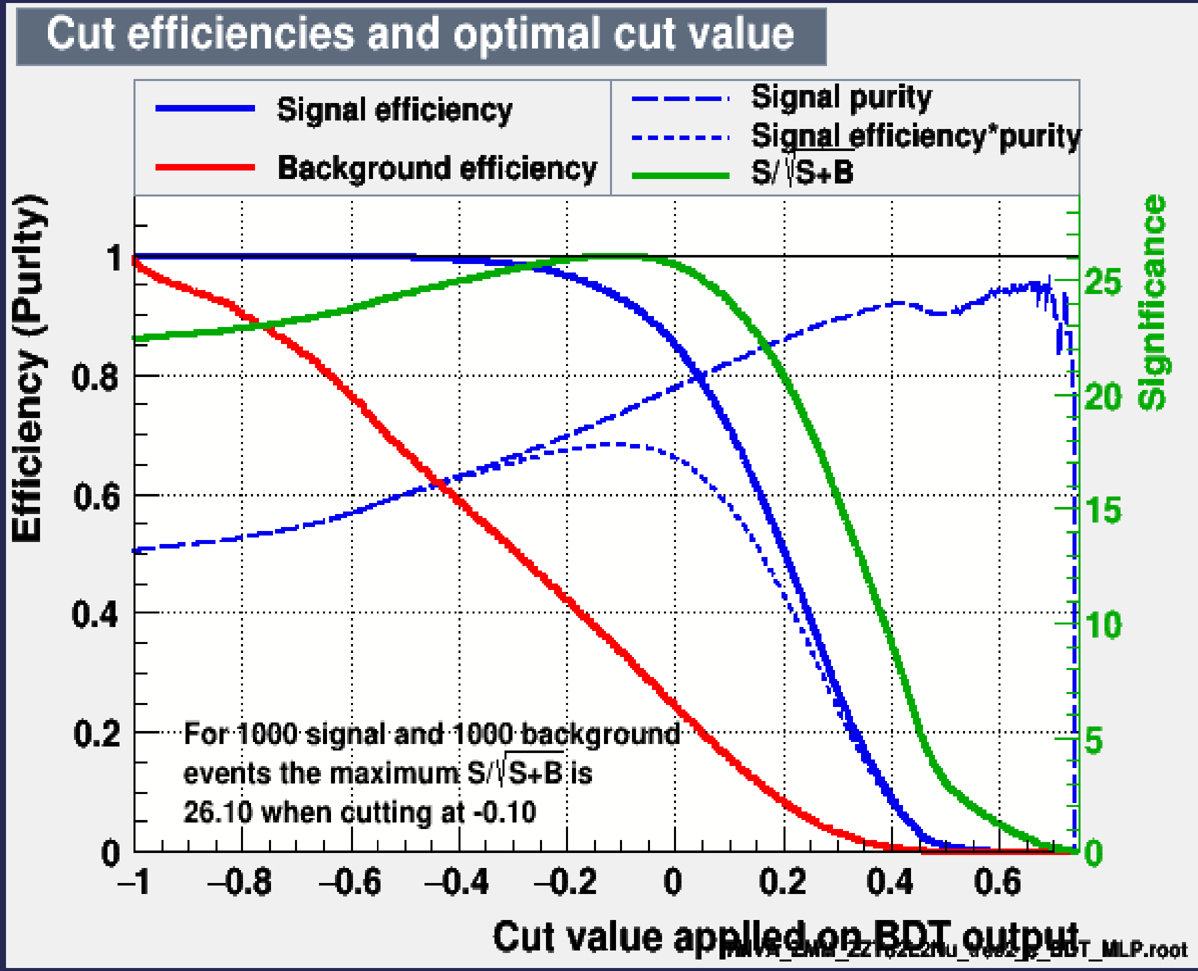
Network Performance : ROC

Signal : ZMM ; Background : ZZTo2Mu2Nu

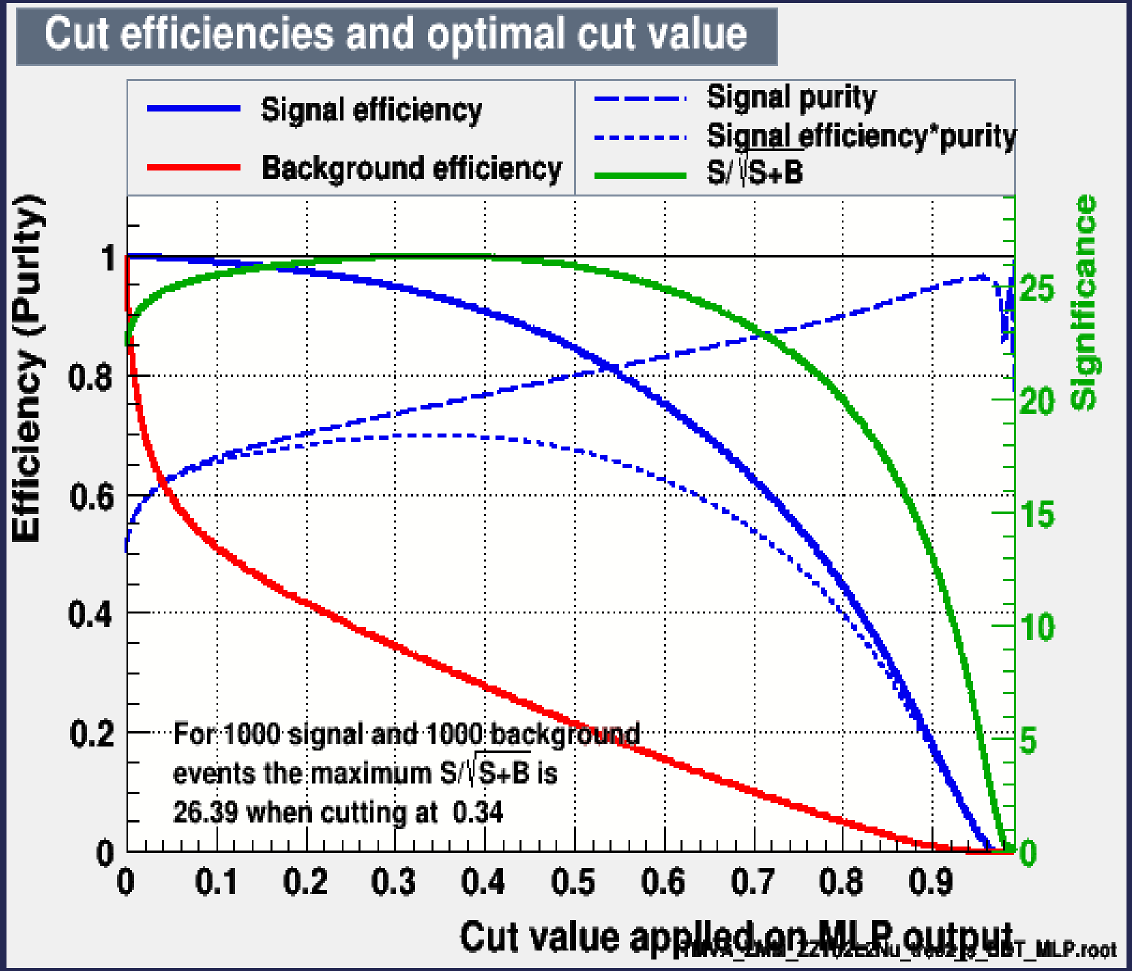


Network Performance: Significance etc.

Signal : ZMM ; Background : ZZTo2Mu2Nu



BDT Discriminator

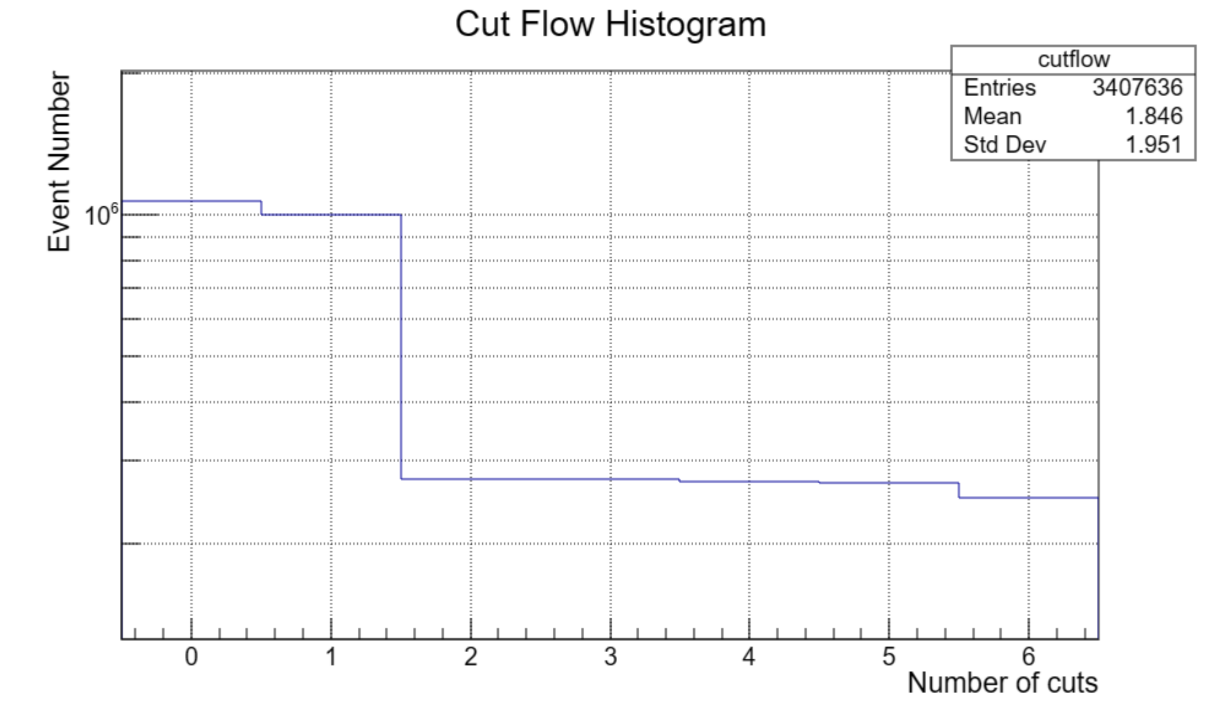


MLP Discriminator

Results from Cut based Analysis

Signal : ZMM ; Background : ZZTo2Mu2Nu

ZZTo2Mu2Nu Cut-flow



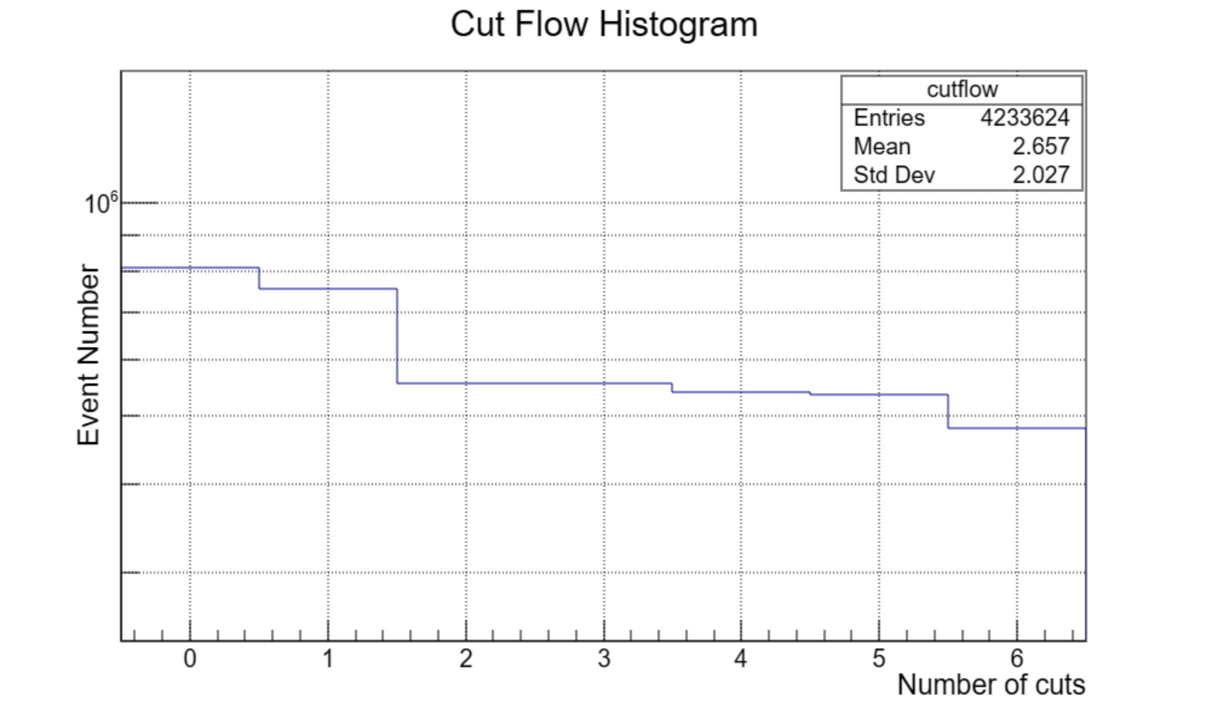
Number of bins: 7

	Bin Content	Bin Error	Bin center	Cut_Effi(%)
Bin 1 :	1071295.0	1035.034	0.0	100.0
Bin 2 :	1000268.0	1000.134	1.0	93.37
Bin 3 :	274016.0	523.465	2.0	25.578
Bin 4 :	273424.0	522.9	3.0	25.523
Bin 5 :	269830.0	519.452	4.0	25.187
Bin 6 :	269084.0	518.733	5.0	25.118
Bin 7 :	249719.0	499.719	6.0	23.31

ZMM Cut-flow

Number of bins: 7

	Bin Content	Bin Error	Bin centers	Cut_Effi(%)
Bin 1 :	811358.0	900.754	0.0	100.0
Bin 2 :	756849.0	869.971	1.0	93.282
Bin 3 :	556232.0	745.81	2.0	68.556
Bin 4 :	554494.0	744.644	3.0	68.341
Bin 5 :	539138.0	734.26	4.0	66.449
Bin 6 :	535731.0	731.936	5.0	66.029
Bin 7 :	479822.0	692.692	6.0	59.138



Conclusion

➢ Summary of objectives and Findings

- Objective was to classify the Signal and Background
- Application of MVA made the job easy
- MVA always performs better than Cut Based; even the topology is very similar for Signal and Background
- Also studied the Tag & Probe Method to find the Muon Detection Efficiency.

➢ Key takeaways from the research

- During the research I learned about Experimental Particle Physics & how research has been done.
- I learned about LHC, CMS Detector & Trigger Mechanism, Future Circular Collider and many more….
- I also learned about the application of various programming lang like Python, C++, Bash, Julia etc.
- Although all work is done in Python & PyROOT framework(Python interface of ROOT framework).
- Learned about Machine Learning Algorithms.

➢ Future directions for further investigation

- Here, ANN performance is better or equal to BDT
- But, generally, BDT performs better; need to optimize the algorithm

References

1. The CMS experiment at the CERN LHC - https://doi.org/10.1088/1748-0221/3/08/S08004
2. TMVA User Guide - https://doi.org/10.48550/arXiv.physics/0703039
3. TMVA code from ROOT Tutorial

Acknowledgement

I would like to express my heartfelt gratitude to my guide **Prof. Subir Sarkar**, and lab mates : **Kuldeep Nishad, Ratul Sarkar, Dr Pritam Palit , Shweta Baradia, Suman Dasgupta**; for their invaluable support and assistance throughout my work on the experimental high-energy project. Their guidance, collaboration, and contributions have been instrumental in shaping my research experience. I am also thankful to the SINP and SIRD for this summer internship program. It is through their collective efforts that I have been able to achieve meaningful results and grow as a researcher.

-- Sayan Dhani

Comments, Questions?

Thank you !!