Optimizing the signal over background ratio in the search for Vector Like-Leptons

Mentee: Andrea Ola Mejicanos (Berea College)

Mentor: Charis Kleio Koraka (UW-Madison)

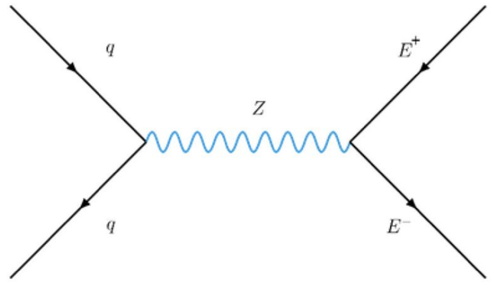
Introduction

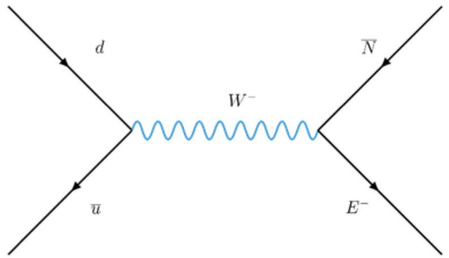
What is a Vector-Like Lepton?

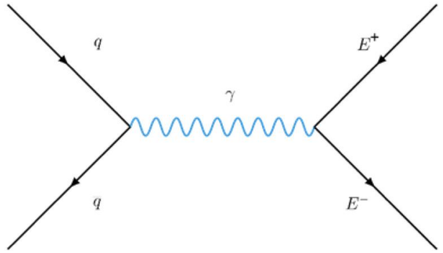
The SM has room for fermions that are non-chiral i.e fermions with right-handed and left handed components that are identical (or nearly identical).

VLLs main characteristics

- VLLs are a hypothesized particle.
- VLLs have no distinct left-handed and right handed components.
- Unlike SM leptons, VLLs electroweak interaction is indistinct for left-handed and right handed components.
- Their mass is not directly related to the Higgs mechanism.







Theoretical Motivation

Why should we search for Vector-Like Leptons?

VLLs are a extension of the SM that contemplates a new generation of Leptons, and could explain certain discrepancies that have been observed in the SM including:

- Electron and Muon anomalous magnetic, the discrepancy between the experimental and theoretical value is:

$$\Delta a_\mu = a_\mu^{\text{exp}} - a_\mu^{\text{SM}} = 288(63)(49) \times 10^{-11}.$$

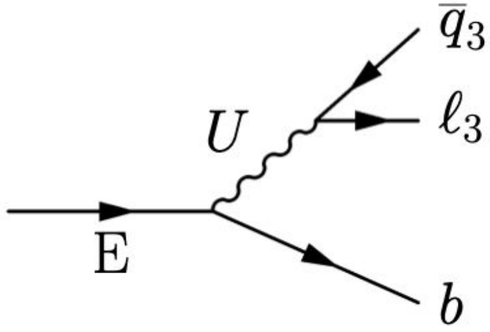
  VLLs as an extension of the SM introduces new contributions to the muon's magnetic Moment. Similarly for the electron's magnetic moment.

- Lepton flavor non-universality, VLLs introduce new interactions and couplings beyond the SM. Lepton flavor non-universality says that weak interactions involving different flavors of leptons may not be ruled by the same set of coupling constants. In the SM these couplings are the same for all leptons, i.e. the strength of the weak interaction is the same for all leptons. VLLs introduces a new interaction mediator Leptoquark U as the source of Lepton Flavor Universality Violations.

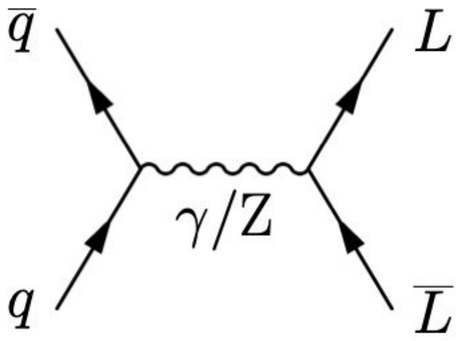
LFV Processes $\mu \to e\gamma$ $\tau \to e\gamma$

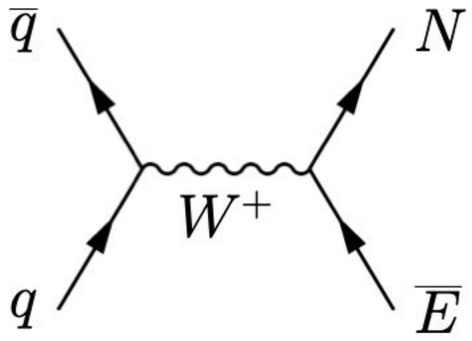
VLL production and decays

- VLLs are pairs produced through Electroweak interactions.
- We consider leptonic final states with 1st and 2nd Generation leptons .
- Decay is mediated by a vector leptoquark (**U**).



Neutral (**N**) and charged (**E**) VLL production.

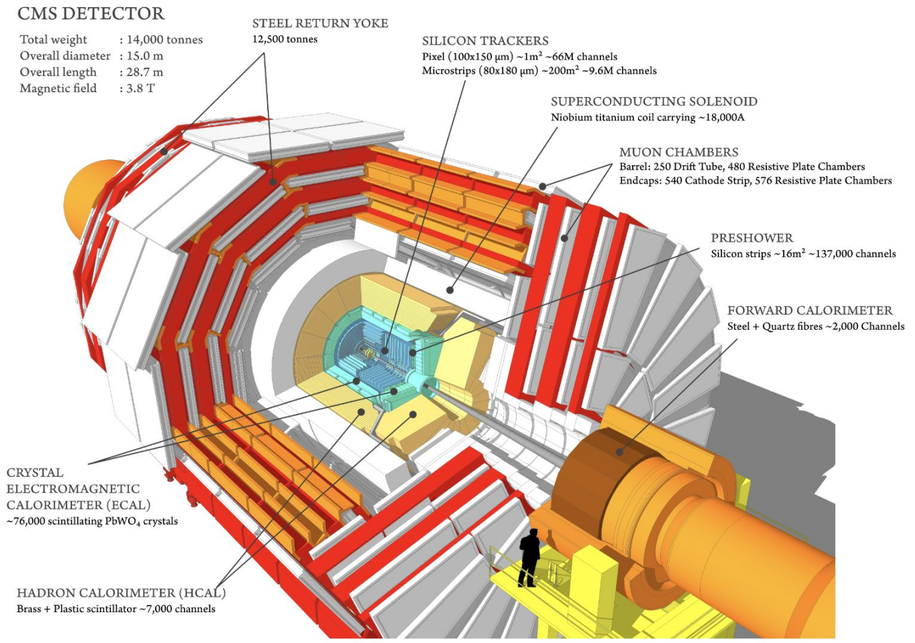




The LHC and the CMS detector

The CMS detector plays a crucial role in identifying and reconstructing the decay products of our particles (VLLs). The tracker, ECAL, HCAL, and muon system measure the properties of charged leptons and other particles resulting from VLLs decays.

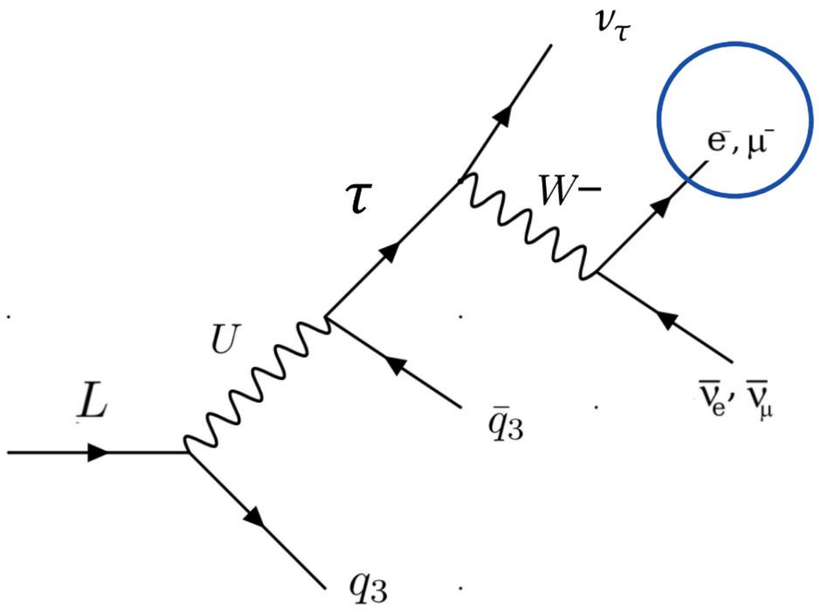
In this search, we consider Vector-Like Leptons at a Mass of 600 GeV. The search uses LHC simulated data.

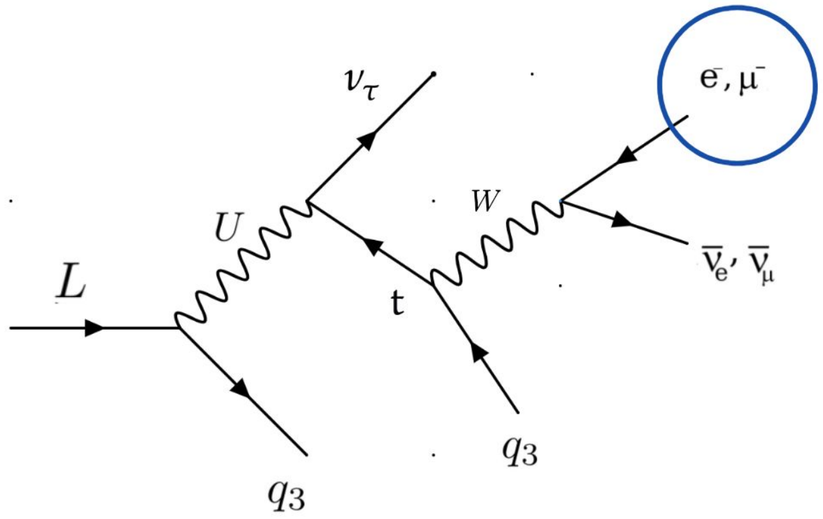


Final states VLLs pairs

Full chain decay of the event signature

In our study we consider final states with two leptons (muons or electrons) coming from top quarks and Tau decays.





SM processes with the same final state

Background process are those with similar final states (Jets/bJets and 2 Leptons), that can mimic the VLL signal. In our analysis, we specifically consider the ttbar background, as it constitutes the most dominant background in our two leptons final state signal region.

Main backgrounds:

- **tt-bar**
- DY+jets
- Di-boson production (ZZ,WW,ZW)
- Tri-boson production (www, zzz, wwz)
- tt(V/H)+jets (ttZ, ttZ, ttH)
- tt+VV ttHH, ttZZ
- 4-top (tttt)

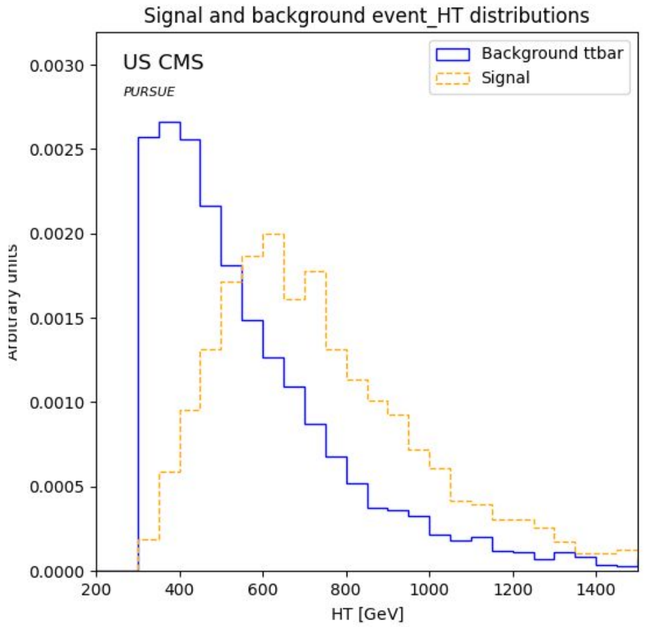
Procedures

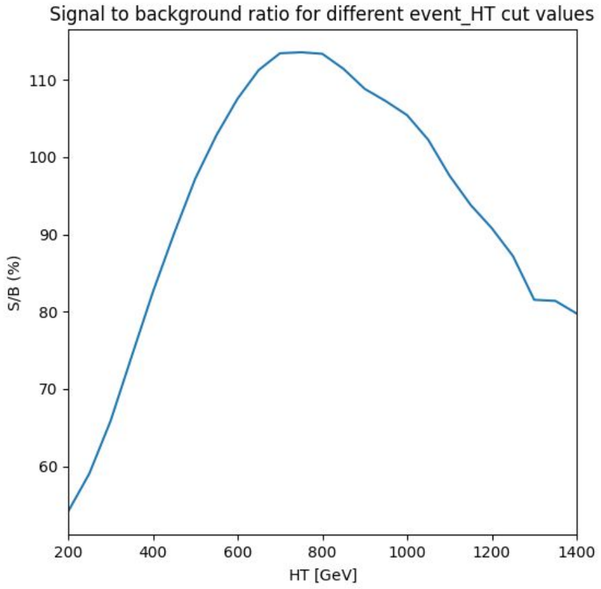
- **Goal:** We want to optimize our signal to background ratio.
  In order to achieve our goal we train a neural network so that we can optimize signal/background classification.

  Steps pre/post ML:
  - Prepare data: this includes writing python scripts for obtaining and formatting data for analysis.
  - “Clean” our data by applying basic cuts: we apply cuts that guarantee a high S/B ratio.
  - Identify and calculate kinematic variables: variables are calculated in root data analysis framework.
  - Choose best variables for ML input: Plot superimposed normalized distributions for the calculated variables and analyze variables correlations for variable selection optimization.
  - Train model.
  - Evaluate model: apply tests to evaluate classification.

Making Cuts

Signal vs Background HT





$$\frac{S}{B} = \frac{N\text{signal}}{N\text{background}}$$

Nsignal ⇨Number of signal events passing a cut value

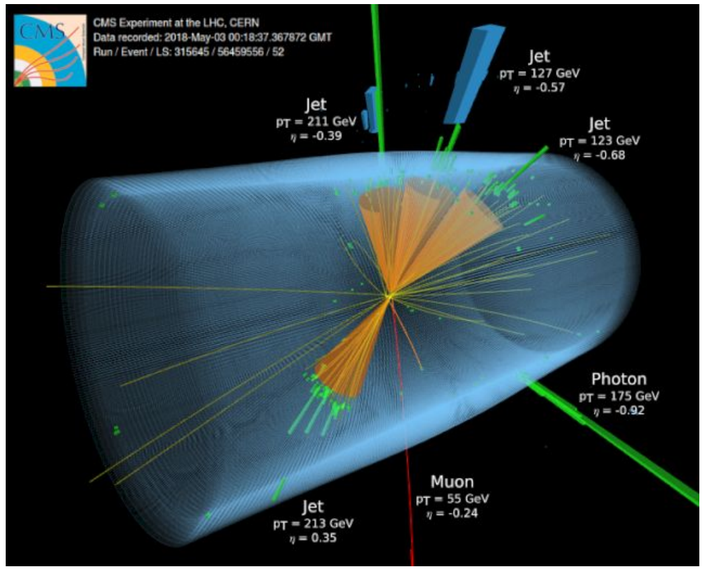
Nbackground ⇨Number of background events passing a cut value

Selection Criteria

In our data we encounter thousands of events. The number of signal events compared to the background from SM process is very low. Therefore we apply cuts to our data.

Selection Criteria

Objects	Selection
MET	> 40 GeV
Lepton Leading,e/μ p_T	> 30 GeV
Number Leptons	= 2
Number of Jets	>3
Leading Jet p_T	> 100 GeV



These basic cuts are applied to our simulated data NanoAODv9 samples in the analysis code.

- VLL: EE, EN, NN
- TTbarPowheg_Dilepton

Calculated Variables

Variable Name	Variables Description
1.MET	Missing transverse energy
2.HT	pT sum of all jets in the event
3.bHT	pT sum of all b-tagged jets
4.LeadingLeptonpT	Leading Lepton pT in the event
5.LeadingBjetpT	Leading b-tagged jet pT in the event
6.SubLeadingBJetpT	Sub-leading lepton pT in the event
7.ClossestBJetsLepMass	Invariant mass between the two closest b-tagged and the leading lepton
8.DeltaRLeptonBJet	DeltaR between leading lepton and leading b-tagged jet
9.MassDilepton	Invariant mass between leading lepton and subleading lepton
10.MinDeltaRLeptonBJet	Min DeltaR between b-tagged jet and leading lepton
11.SubleadingLeptonpT	Sub-leading lepton pT
12.ClosestBJetsMass	Invariant mass between closets b-tagged jets
13.LT	pT sum of all the leptons in the event
14.MaxDeltaRBJetLep	Max DeltaR between b-tagged jet and leading lepton
15.MinDeltaRBJetLep	Min DeltaR between b-tagged jet and leading lepton
16.DeltaRLeptonPair	DeltaR between leading and subleading lepton
17.DeltaRLeptonJet	DeltaR between lepton and leading jet
18.DeltaRBJetsPair	DeltaR between leading and subleading b-tagged jets pairs
19.LeptonPairpTratio	Leading to subleading lepton pT ratio

- Variables were calculated based on the topology of the final state signature of our signal event.
- A better understanding on 4-vectors and root analysis calculations was gained by calculating variables “manually” first and then using the methods included in root .

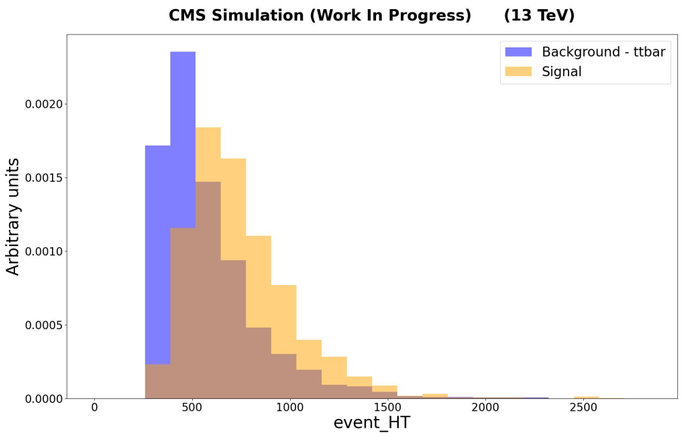
$$M^2 = 2p_{T1}p_{T2}(\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2))$$

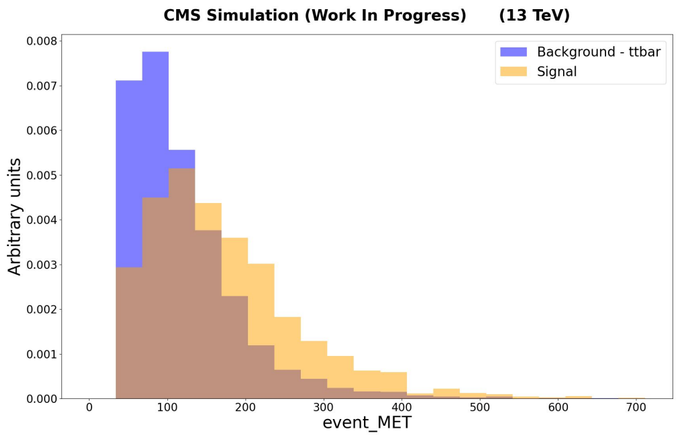
$$d = \sqrt{(\eta_2 - \eta_1)^2 - (\phi_2 - \phi_1)^2}$$

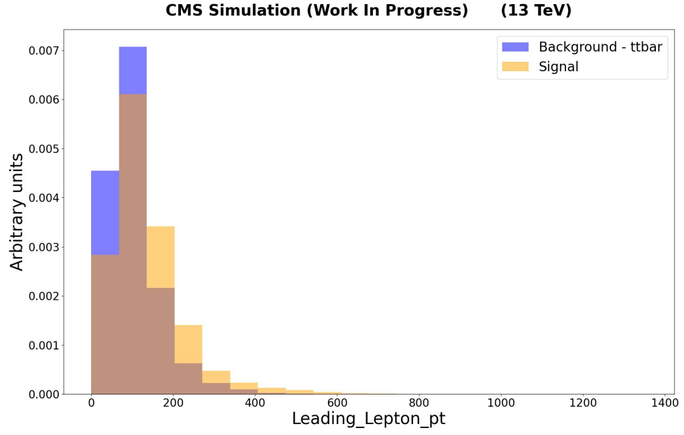
Comparing Signal vs Background Variables

In this project we train a neural network that takes distributions of variables as input and classifies events as either signal or background in a supervised training procedure. Therefore, we choose variables that yield a better separation between signal and background.

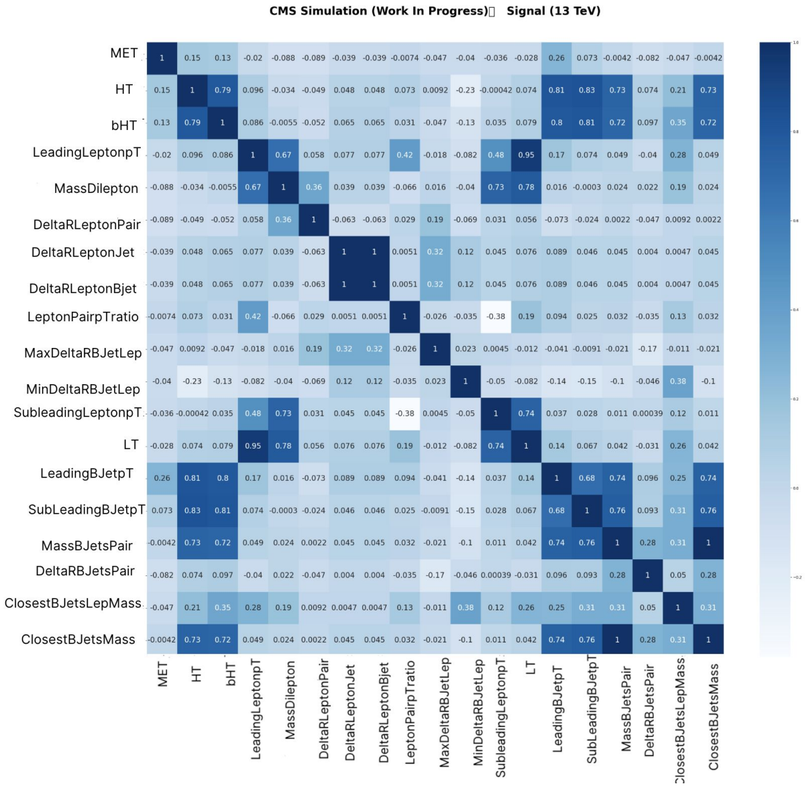
Input Variables



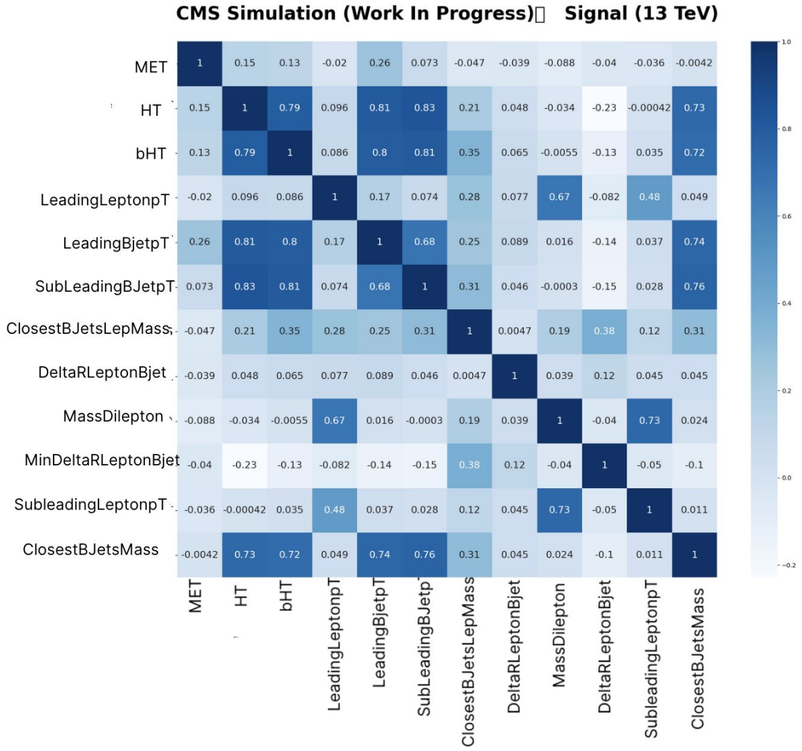




ML Parameters Correlations



- In the analysis, we started with **19** variables.



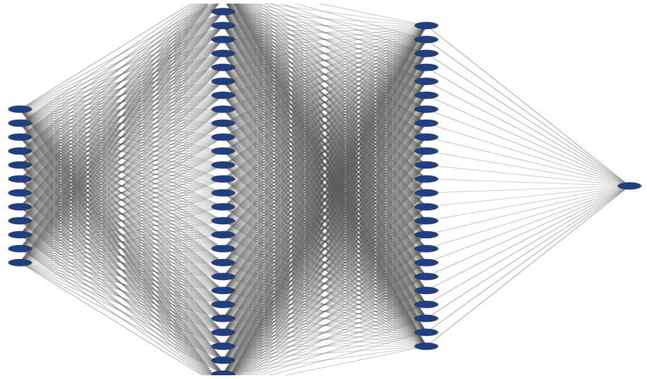
- We selected the best **12** variables by removing highly correlated and similar distributions between signal and background.

ML Hyperparameters

Number of input variables	12
Hidden Layers	2
Nodes per Hidden Layer	[249, 24]
Batch Size	1000
Number of Epochs	200
Trainable Parameters	8990
Training set size	80%
Evaluation size	20%

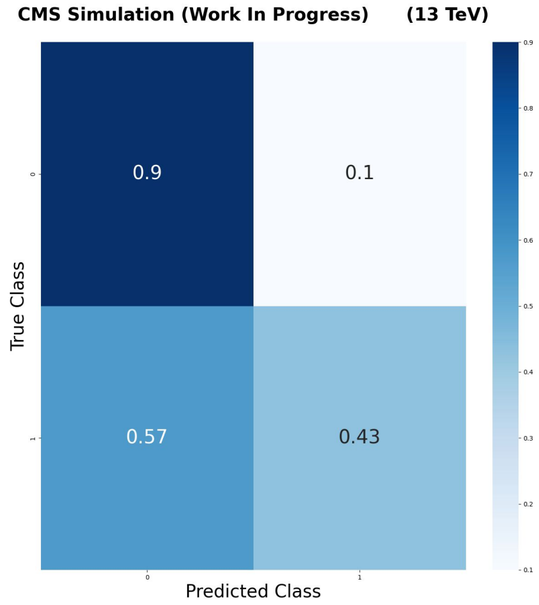
First Hidden Layer:
Number of trainable parameters = input_dim * num_nodes + num_nodes
Second Hidden Layer:
Number of trainable parameters = input_nodes_previous_layer * num_nodes + num_nodes
Output Layer:
Number of trainable parameters = input_nodes_previous_layer * num_nodes + num_nodes
Total:
Total = input_layer + first_hidden_layer + second_hidden_layer + output_layer

Trainable parameters
12 x 249 +249 = 2988
249 x 24 + 24 = 5976
24 x 1 + 1 = 25
Total = 8990

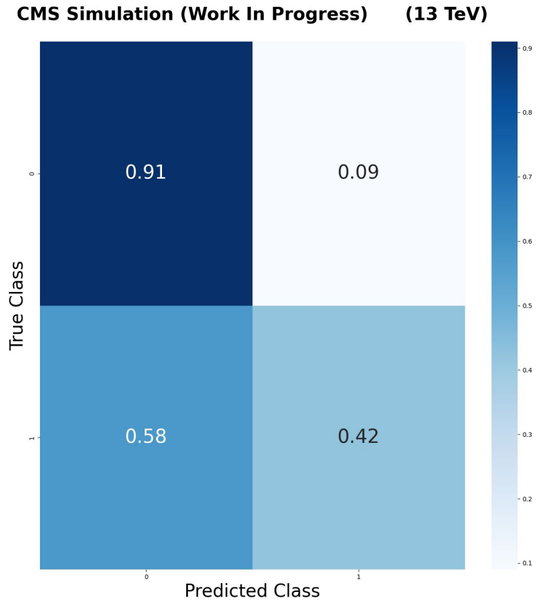


Results

Initial Input Variables: 19



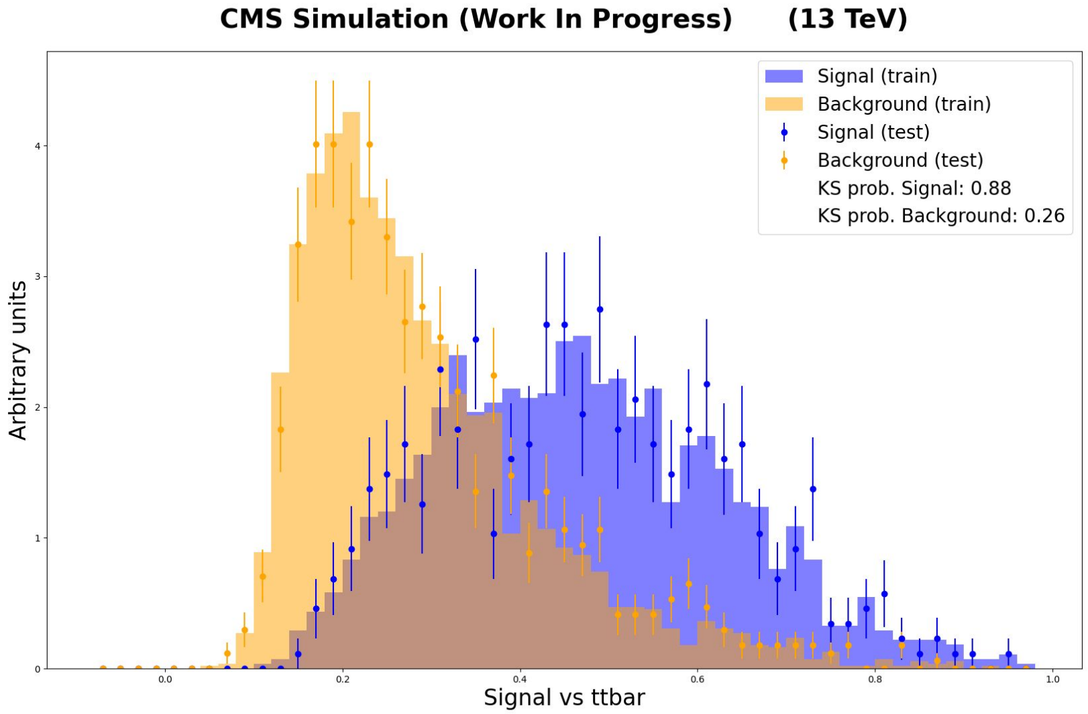
Input Variables: 12



- Despite removing highly correlated variables and variables with very similar signal and background distributions, there was no observable improvement in the neural network's performance.
- However, it was observed that keeping variables with very similar signal vs background we can slightly improve the performance of the model.

Results

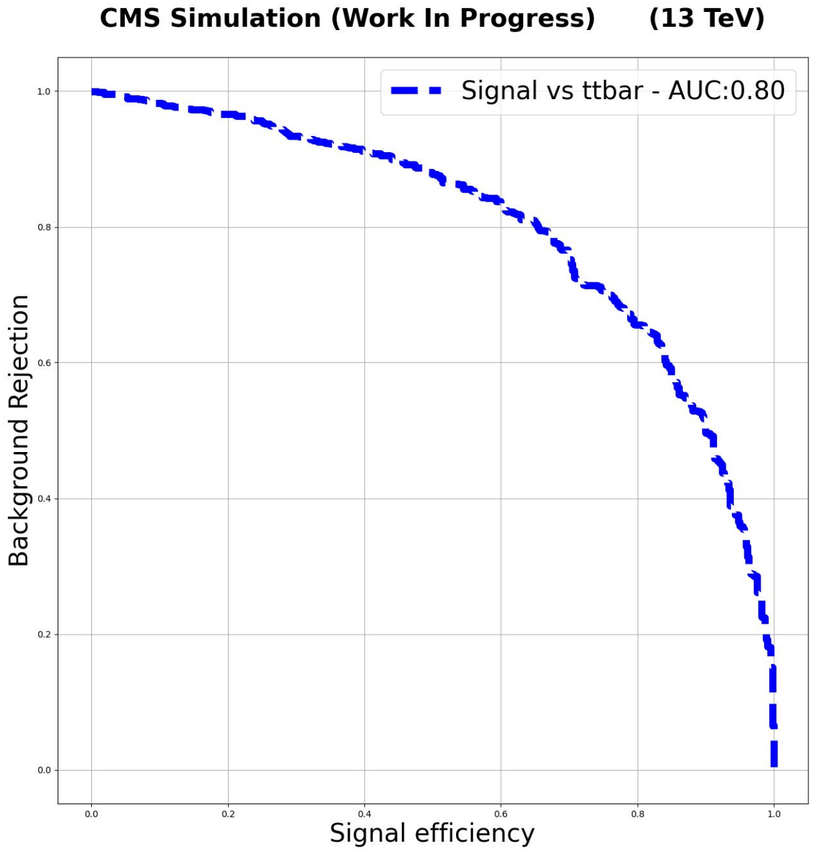
Final number of variables: 12



KS prob. Signal: 0.88
Suggests a better generalization for the signal distribution.

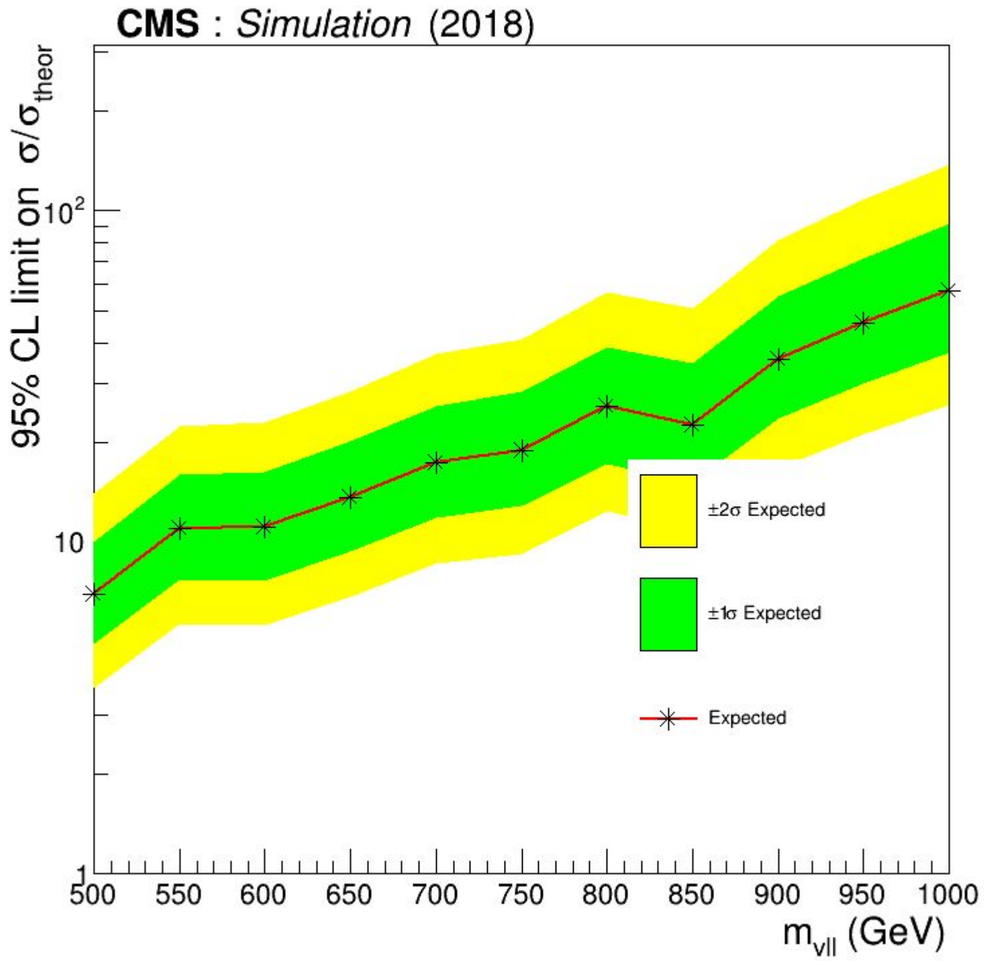
KS prob. Background: 0.26
Suggests the model may not generalize well unseen data.

ROC curve



The ROC curve for our ML model shows significant trade-offs when choosing a threshold value. In this case we have a low background rejection.

Limit Plot



- Produced data cards for each mass point and channels.
- Used combine to produce the limit plot
- The goal is to see the expected signal strength for each mass point.

This Limit plot shows the expected limit values for different mass points. We see upper bounds on the signal strength that the experiment would be able to exclude or include with a certain level of confidence (68% and 95%), assuming the background-only hypothesis.

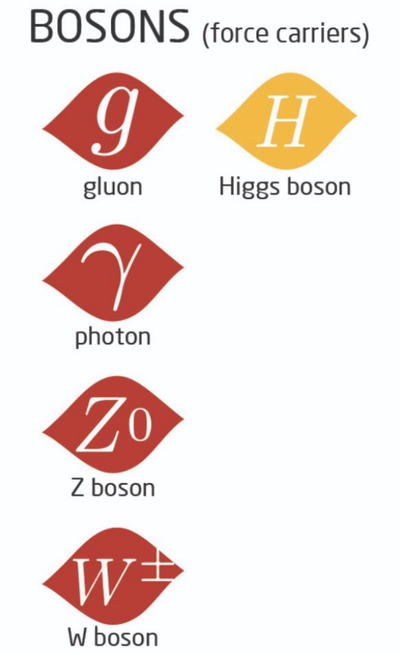
Summary

The project aimed to enhance the signal-to-background ratio in the search for vector-like leptons. This was performed by training a neural network using calculated kinematic variables. The variables demonstrating distinct distributions between signal and background were selected. The model achieved an AUC score of 0.8, indicating relatively good classification ability. Additionally, the Kolmogorov-Smirnov (KS) test demonstrated high discrimination with a KS probability of 0.88 for signal and 0.26 for background suggesting that the model does a better job in identifying signal events, as revealed by the confusion matrix. These results suggest that the model effectively separates signal and background, making it a promising tool for identifying vector-like leptons in the dataset.

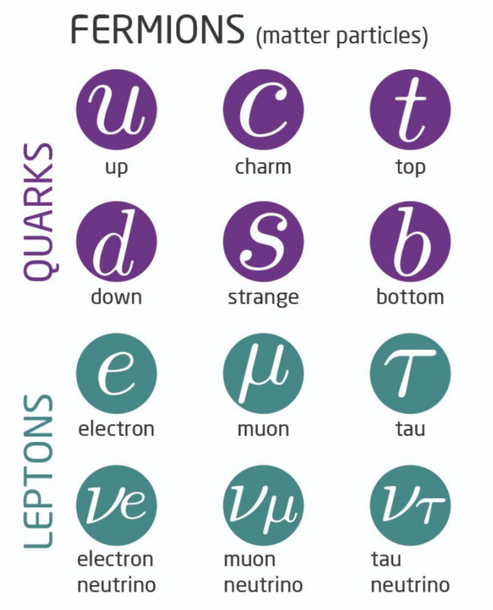
Backup

Elementary Particles

The SM accounts for two types of fundamental particles: *Bosons and Fermions.* While Gauge bosons are "carriers" of fundamental forces, the Higgs boson generates mass. Fermions are matter particles and interact by exchanging gauge bosons. Elementary particles in the SM posses anti-particles with opposite charge and parity.



Bosons:
- Integer spin quantum number.
- Gauge Bosons are carriers of fundamental forces.



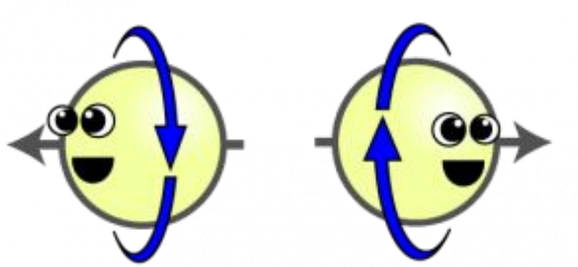
Fermions:
- Half-integer spin Quantum number.
-Fermions are categorized into quarks and leptons

Fermions

Fermions are described as Dirac spinors, which represents their quantum state. A Dirac spinor consists of four components representing two different **chiralities**:

Right-handed or right chiral particle

Spin is aligned in the same direction to the particle's momentum



Left-handed or left chiral particle

Spin is aligned opposite to the particle's momentum

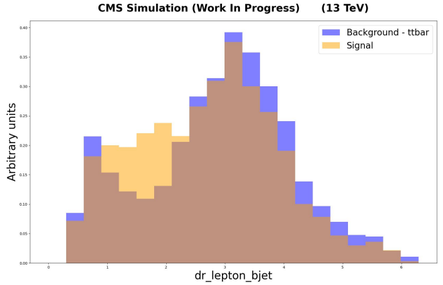
Data

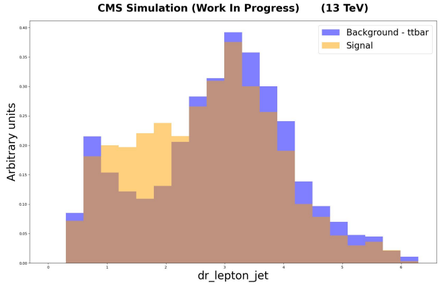
Object	Collection	Selection
Jets	Jet_*	p_T>30 GeV & |η|<2.4 & mediumJetIDbit & $\Delta R_{e,\mu}$ > 0.4
Electrons	Electron_*	p_T>10 GeV |η|< 2.4 & CutBasedIdTight & eleEtaGapVeto
Muons	Muon_*	p_T>10 GeV |η|< 2.4 & tightId & pfRelIso04_all < 0.25
MET	MET_*	> 40 GeV

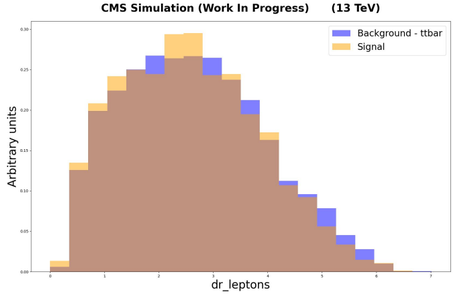
Event Selection criteria :

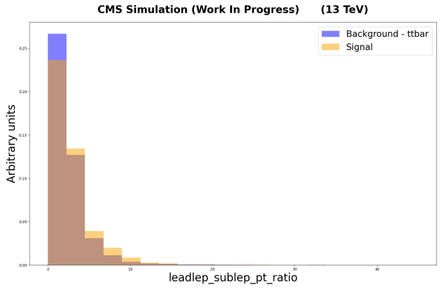
- =2 #lepton, $p_T^{lead,e/\mu}$ > 30 GeV
- m_{ll}>20 GeV & Z peak veto m_{ll} < 76 GeV & m_{ll} > 106 GeV if OS
- #jets>3 & $p_T^{jet,lead}$ > 100 GeV & $p_T^{jet,sub\text{-}lead}$ > 50 GeV
- H_T >300 GeV

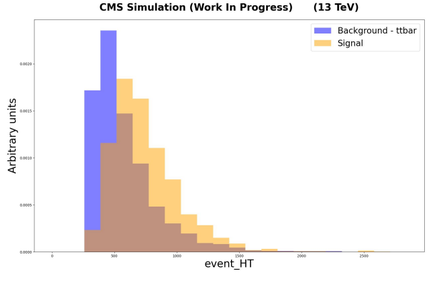
Using **RunIISummer20UL18NanoAODv2** samples

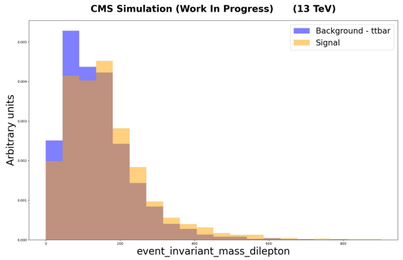

CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
dr_lepton_bjet

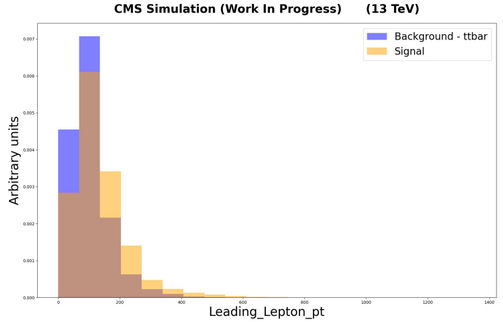

CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
dr_lepton_jet

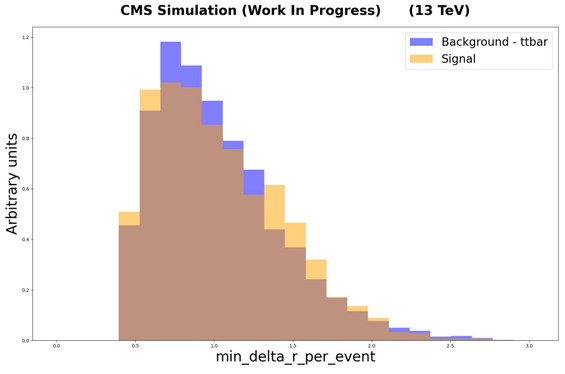

CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
dr_leptons

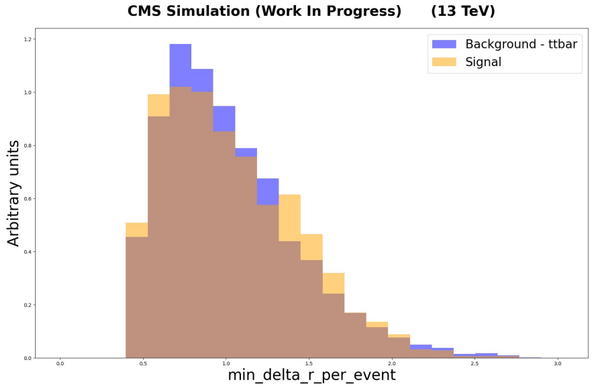

CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
leadlep_sublep_pt_ratio

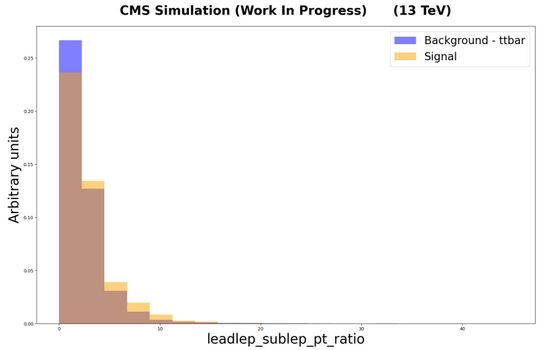

CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
event_HT

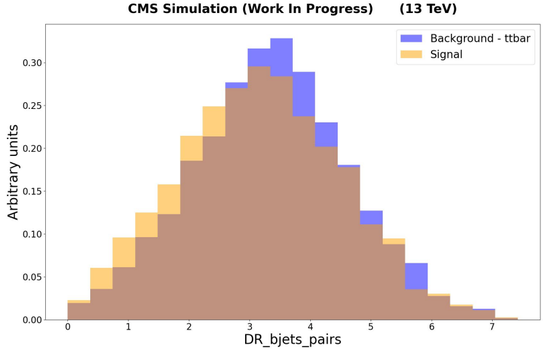

CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
event_invariant_mass_dilepton

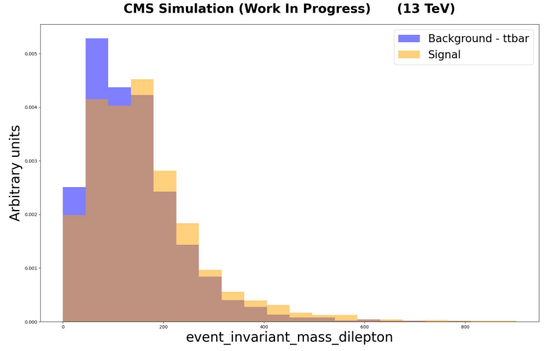

CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
Leading_Lepton_pt


CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
min_delta_r_per_event


CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
min_delta_r_per_event


CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
leadlep_sublep_pt_ratio


CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
DR_bjets_pairs


CMS Simulation (Work In Progress) (13 TeV)
Background - ttbar
Signal
Arbitrary units
event_invariant_mass_dilepton

Background Processes Description

DY+jets (Drell-Yan production): This background arises from the production of a lepton-antilepton pair (such as electron-positron or muon-antimuon) via the exchange of a virtual photon or Z boson. While this process can lead to two leptons in the final state, it usually has fewer jets than the signal.

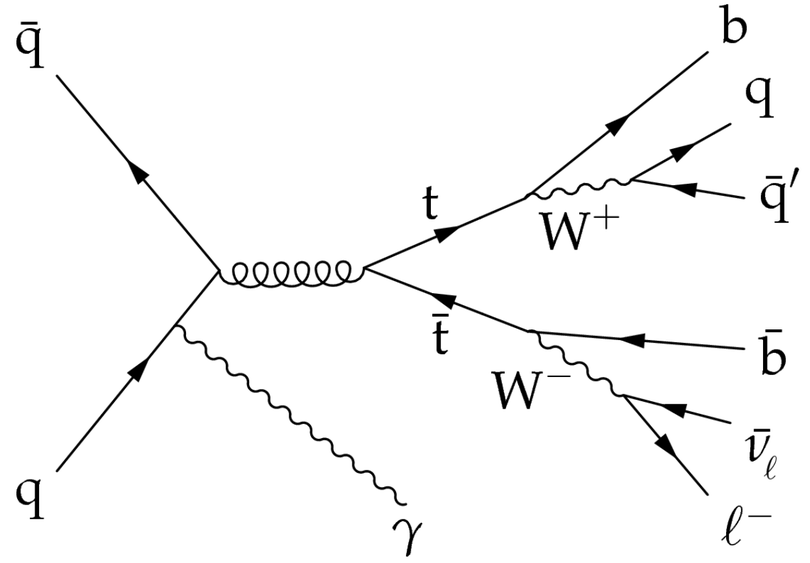
Di-boson production (ZZ, WW, ZW): Di-boson production involves the production of pairs of weak bosons (W, Z, or gamma) which can decay into leptons and jets. These processes can mimic the signal, but they often have lower cross-sections and different kinematic properties.

Tri-boson production (WWW, ZZZ, WWZ): Similarly to di-boson production, tri-boson production involves the production of three weak bosons. These processes are rarer and can lead to complex final states with multiple leptons and jets, but their low cross-sections make them less dominant.

tt(V/H)+jets (ttZ, ttW, ttH): These are processes where top quark pairs are produced in association with other vector bosons (Z, W) or the Higgs boson (H). They can produce final states with multiple leptons and jets similar to your signal, but again, their cross-sections are generally lower.

tt+VV (ttHH, ttZZ): These processes involve top quark pairs produced in association with pairs of vector bosons. They lead to final states with multiple leptons and jets, but their cross-sections are often smaller than that of tt-bar.

4-top (tttt): This process involves the production of four top quarks and can result in final states with multiple leptons and jets. However, due to its higher order in the strong coupling constant, its cross-section is much smaller than that of tt-bar.



Optimizer	Adagrad
Number of input variables	12
Hidden Layers	2
Nodes per Hidden Layer	[249, 24]
Node Activation	LeakyReLu
Output Node Activation	Sigmoid
Regularization L1	1e-5
Regularization L2	7*1e-4
Dropout Percentage	0.09
Loss Function	Binary cross-entropy
Batch Size	1000
Number of Epochs	200
Early Stopping	12
Training/Evaluation python packages	Keras and Tensorflow
Trainable Parameters	8990