Study of low-momentum taus at CMS

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Parijat Banerjee

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

March, 2024

Supervisor: Dr. Sourabh Dube

Certificate

This is to certify that this dissertation entitled "**Study of low-momentum taus at CMS**" towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Parijat Banerjee at Indian Institute of Science Education and Research under the supervision of Dr. Sourabh Dube, Associate Professor, Department of Physics, during the academic year **2023-2024**.

Sourabh Dube

Dr. Sourabh Dube

Committee:

Dr. Sourabh Dube

Dr. Saranya Ghosh

This thesis is dedicated to my grandparents.

Declaration

I hereby declare that the matter embodied in the report entitled Study of low-momentum taus at CMS are the results of the work carried out by me at the Department of Physics, Indian Institute of Science Education and Research (IISER) Pune, under the supervision of Dr. Sourabh Dube and the same has not been submitted elsewhere for any other degree. Wherever others contribute, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

Banerjee

Parijat Banerjee

Acknowledgments

First and foremost, I would like to express immense gratitude to my supervisor, Dr. Sourabh Dube, for nurturing my scientific journey through many discussions, debates, and philosophical wanderings. Your ever-supportive attitude has helped boost my self-confidence and given me the motivation to pursue a PhD. I would also like to sincerely thank Dr. Saranya Ghosh for accepting to be my TAC expert member and for your valuable comments and insights on my project.

I am also indebted to Arnab, Prachurjya, Yash, and Riya for their valuable suggestions and assistance at various stages of this study. Your collective expertise has expedited my research, time after time, and contributed significantly to its success. Thanks to my colleagues Soumya and Chitrakshee for always keeping my spirits up in the lab and contributing to a healthy research environment. Our discussions over chai will be missed the most by me.

Special thanks are due to my friends Jason, Ashli, and Reetish for their encouragement and moral support throughout the ups and downs of this journey, many of which we have faced together.

I am deeply grateful to my family - my father, Partha Pratim Banerjee; my mother, Sharmila Banerjee; and my brother Pushan Banerjee. Your unconditional love and unwavering support have helped me throughout my academic pursuit.

Finally, I would like to thank the Department of Science and Technology for granting me the INSPIRE-SHE scholarship over a period of five years. Lastly, I wish to thank IISER Pune for providing an environment conducive to research that has helped me establish a solid foundation.

Abstract

The tau lepton, which is the third-generation lepton, is the only lepton that decays hadronically as well as leptonically. Tau lepton reconstruction is thus challenging at a hadron collider such as the LHC. Standard tau reconstruction and identification at CMS typically requires tau leptons with transverse momentum (p_{T}) above 20 GeV.

However, several BSM models could benefit from the usage of lower p_{T} tau leptons. For example, in the VLL singlet model, we are unable to significantly exclude vector-like taus of any mass. The selection of lower p_{T} tau leptons would improve the acceptance of searches for the vector-like tau and thus could improve the sensitivity of the search. This is true of many other BSM searches as well. In this thesis, I will describe my study of the reconstruction and identification of taus at low momentum and improve the efficiency of these algorithms using machine learning models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Physics is the study of natural phenomena. Over the last two centuries, humans have attempted to answer some basic questions, such as what we are made of. The curiosity to discover and understand the fundamental building blocks of the universe brings us to our modern-day study of particle physics, which attempts to describe these building blocks and their interactions.

The standard model (SM) of particle physics is the most widely accepted theory describing the fundamental particles of the universe and their interactions with each other. However, it fails to explain many physical phenomena which are known to be true, such as neutrino masses, the existence of dark matter etc. The Large Hadron Collider at CERN collides bunches of protons together to study the properties of all these particles and their interactions with each other. The CMS experiment is a large experiment where a dedicated detector called the CMS detector is strategically placed at the collision point of these protons, and the particles coming out of the collisions are studied to find hints of new physics. This requires dedicated algorithms to detect and identify signatures of all the known particles.

One such particle of interest is the tau lepton. It is the heaviest lepton and the only one capable of decaying hadronically. The existing CMS identification algorithm for taus recommends using only taus above 20 GeV transverse momentum(p_{T})[1]. This significantly limits our sensitivity to any new physics hiding at low momentum! Many models of beyond standard model physics, such as the Vector-Like-Lepton singlet model, could benefit from the usage of low p_{T} taus[2].

In this thesis, I will describe the tau reconstruction and identification algorithms used at CMS and their performance for low momentum taus. Other physics objects at CMS, such as electrons and muons, can be identified with p_{T} as low as 5(3) GeV![3, 4] Why can't taus be identified at lower momentum? Is there something inherently more challenging about low p_{T} taus, or can the identification be improved by developing a better algorithm? Which algorithm should be improved; reconstruction or identification? Here, I will try to answer some of these questions.

I will describe a classification scheme to identify reconstructed tau objects by matching them to truth-level objects. I will also demonstrate that a naive overlap removal algorithm that excludes reconstructed tau objects that match an identified electron or muon candidate does not perform better than the standard tau identification algorithm(DeepTau[5]) extended to low momentum. I will also show that it is not possible to obtain better identification at low momentum just by using a different set of cuts.

I have optimized a multi-layer neural network by carefully training it specifically for taus at low momentum. Using a carefully selected set of 28 variables, I demonstrate the performance of my neural network at low momentum. I also compared my algorithm with DeepTau applied to low momentum taus, and it is clear that my algorithm provides larger signal acceptance for the same value of background rejection.

The thesis is divided into multiple chapters described as follows: Chapter 2 describes the standard model, its limitations, and the place of taus within the standard model. Chapter 3 discusses the CMS detector and the reconstruction and identification algorithms currently in place at the CMS experiment. Chapter 4 contains the technological challenges that were overcome to enable this thesis. Chapter 5 mentions a naive overlap removal algorithm, and the performance of DeepTau extended to low momentum taus. Chapters 6-7 contains a detailed explanation of training neural networks and applying it to quantify acceptance and signal-to-background ratio for various samples. Finally, chapter 8 summarizes the results, conclusions, and future work.

Chapter 2

The standard model of particle physics

The standard model (SM) is a theory describing fundamental particles in nature and how they interact with each other. There are three forces through which they can interact, namely, the strong nuclear interaction, the weak nuclear interaction, and the electromagnetic interaction. It is an extremely successful theory that incorporates multiple particles and ties them together into a neat little framework. As of today, there are a total of 61 particles and anti-particles in the SM. The SM is divided into two main families: fermions and bosons. Fermions are classified into two types: quarks and leptons. They are also the fundamental particles (with a corresponding antiparticle) that are the building blocks of matter. Bosons, on the other hand, are the mediators of the interactions between all these particles. The modern-day visualization of the SM, where all particles are grouped together into their own families according to their properties, is shown in figure 2.1.

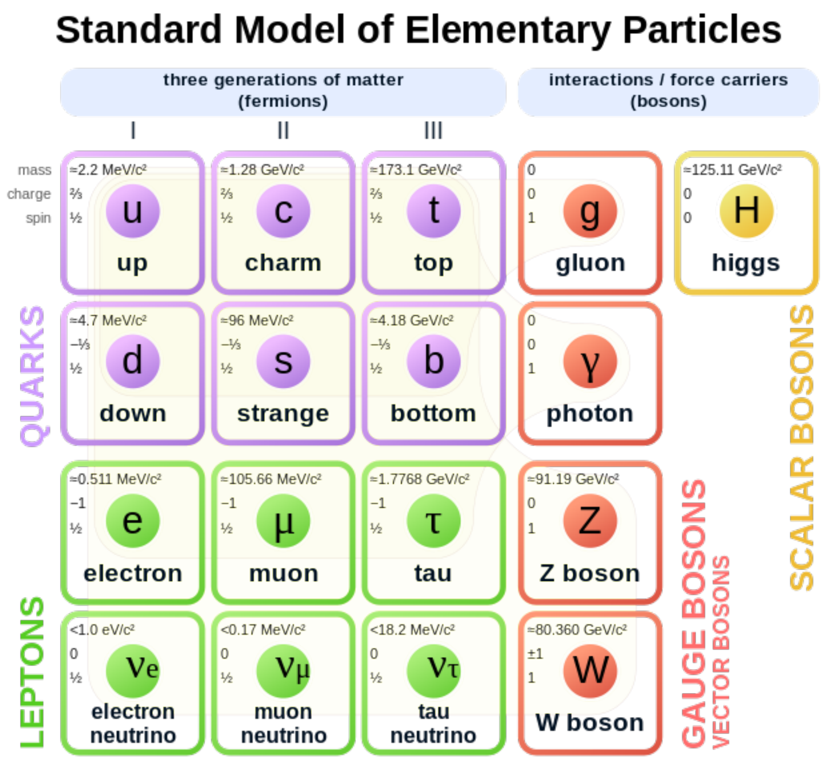


Figure 2.1: The standard model of Particle Physics[*Image Courtesy: Wikipedia*]

2.1 Fundamental interactions

The fundamental interactions are characterized based on the type of particle that experiences the interaction, the strength and range of the interaction, and the nature of the particles that mediate these interactions. The SM contains particles that are the force carriers of the strong nuclear force, the weak nuclear force, and the electromagnetic force. The strong and weak forces are short-range forces mostly confined to the nucleus, while the electromagnetic interaction is an infinite-range interaction. The carrier of the strong force is the gluon, the weak force is the W and Z bosons, while the electromagnetic force is carried by the photon. The final piece of the SM, the Higgs boson was discovered in the year 2012. Higgs is the particle that is responsible for giving mass to all other fundamental particles. The SM is an extremely successful theory that scales almost nine orders of magnitude. However, it has certain limitations.

2.2 Limitations of the SM

Despite the success of the SM in describing three out of the four fundamental interactions of nature in a single framework, there are many contradictions with the experimental observations. A few of the limitations of the SM are described below.

1. **Gravity:** The SM only describes three out of the four fundamental forces of nature and does not describe gravity. This is partly due to the reason that the SM is a quantum field theory, while a renormalizable quantum description of gravity is not known to date[6].

2. **Neutrino mass and oscillations:** The SM assumes neutrinos to be massless. However, experiments have shown that neutrinos oscillate[7] between their flavor and mass eigenstates, which is only possible if they have non-zero mass.

3. **Hierarchy problem:** The masses of various particles and their couplings with the gauge fields are free parameters in the SM framework and can only be experimentally determined. However, we see a hierarchy of masses among both the fermions and leptons. The difference spans almost six orders of magnitude, from the lightest particle with a known mass, the electron, to the top quark, the heaviest particle in the SM.

4. **Matter-Antimatter asymmetry:** Although the SM now allows for matter-antimatter asymmetry in the universe through CP violation, the amount of CP violation required to explain the observed difference from astrophysical observations is too large.

5. **Dark Matter:** There is substantial proof from cosmological observations that there is dark matter in the universe. For example, the rotational curves of the galaxies where the spiral arms are seen to be moving at a faster velocity than what the calculation predicts, based on the total visible mass of the galaxy. Observation of gravitational lensing of the galaxies is also strong evidence for the existence of dark matter.

2.3 Beyond SM Physics

Many proposed theories of beyond-the-SM (BSM) physics extensions try to address various shortcomings of the SM to provide better knowledge of the laws governing the universe. The Grand Unified Theories, or GUTs propose the unification of the electroweak and the strong force at very high energies. Supersymmetry is a neat idea that proposes new ”superpartner particles” of each fermion and boson in the SM. Supersymmetry could explain the hierarchy problem, as well as provide dark matter candidates, such as LSPs[8]. Many other BSM theories also exist, such as vector-like-leptons (VLLs). The motivation for this thesis is partly derived from recent VLL results from the CMS experiment at the Large Hadron Collider[2].

2.4 VLL and low p_{T} taus

The vector-like-lepton model [9, 10] introduces hypothetical particles that are fermions and transform under non-chiral representations, which means that their left and right-handed components transform identically under gauge symmetries of the SM [11]. These arise in many BSM scenarios, including supersymmetry models [12, 13], extra dimension models[14] as well as GUTs [15].

Like in the SM, VLLs carry an electromagnetic charge and can couple to the SM leptons. Extensions of the SM with one or more families of vector-like fermions may also provide a dark matter candiate[16]. The third generation vector-like singlet model proposes a new particle $\tau'^{\pm}$ which couples to the SM $\tau^{\pm}$, the third generation lepton. The doublet model proposes two new particles, the $\tau'^{\pm}$ and $\nu_{\tau'}$. Thus, the singlet model allows only charged VLLs, while the doublet model proposes charged as well as neutral VLLs. This thesis draws some motivation from the most recent results on the exclusion of the VLL τ' singlet model from CMS [2]. Figure 2.2 shows the Feynman diagram for the production of τ', and figure 2.3 shows the most stringent limits on such a τ' mass.

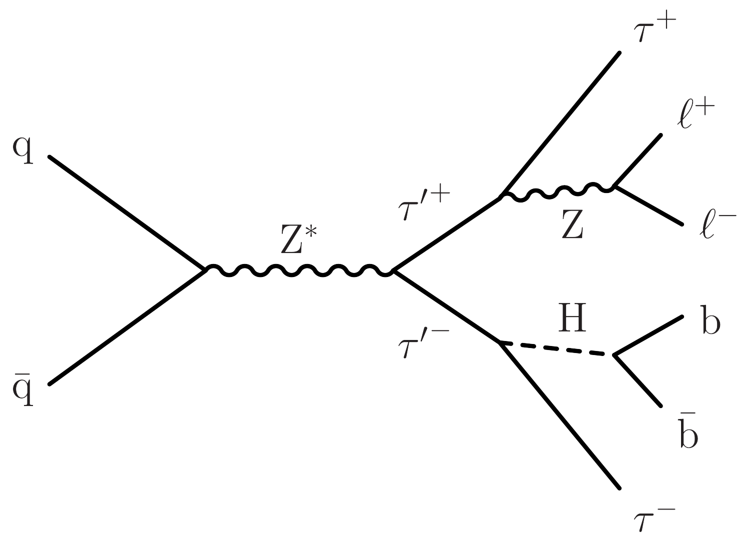


Figure 2.2: Example process illustrating the production and decay of singlet vector-like τ pair [2]

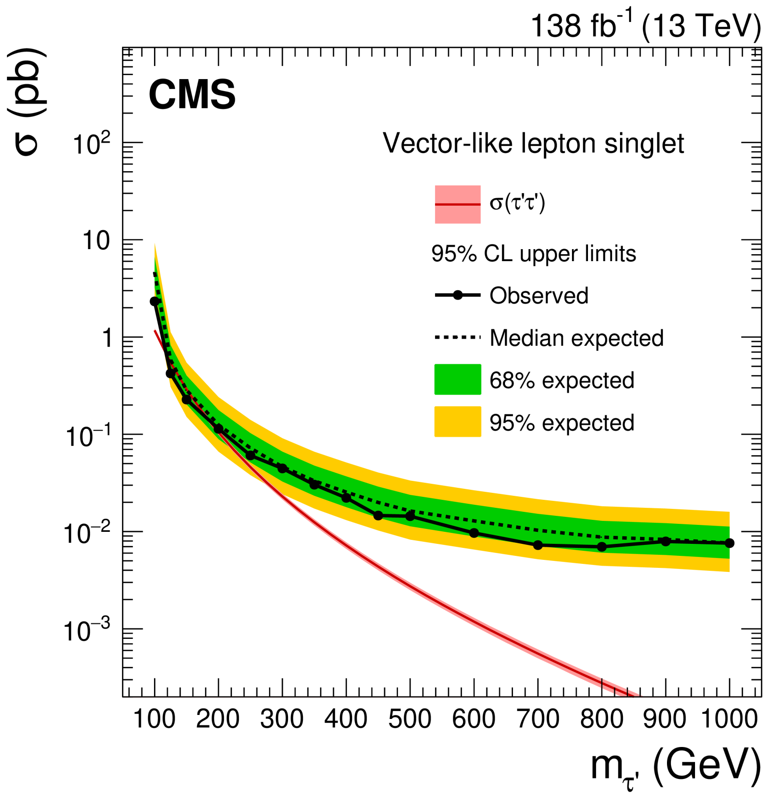


Figure 2.3: Limit plot depicting the excluded mass range for the singlet vector-like τ particle [2]

The VLL τ' decays to SM τ leptons and produces a multilepton final state. However, as we can see from figure 2.3, the τ' is only barely excluded at considerably low masses of τ' around $100 - 150$ GeV. If a 100 GeV τ' decayed to an on-shell Z and a τ, the momentum of the SM τ produced will be very low due to the process kinematics. Now since σ_{lim} is inversely proportional to $A\epsilon$ (acceptance times efficiency), a lower $A\epsilon$ will increase the σ_{lim}. This is partly the reason why the observed data line in figure 2.3 increases for low masses of τ'. The inefficiency in detecting and identifying low momentum taus contributes to the lack of a good limit.

Improved reconstruction and identification of taus at low momentum could help improve the VLL τ' limit. In this thesis, extensive studies have been performed on taus below 20 GeV, specifically on improving identification efficiencies between $10 - 20$ GeV.

Chapter 3

The CMS detector and objects

The CMS experiment is one of the two large general-purpose experiments stationed at the Large Hadron Collider(LHC) at CERN in Geneva, Switzerland. The LHC is a 27 km ring that accelerates bunches of protons at 99.999999% the speed of light and collides them at the center of mass energy 13 TeV. The CMS detector is a general-purpose detector. It detects the various particles produced in the collisions to discover hints of new physics at high energies.

3.1 The CMS Detector

The CMS detector consists of a superconducting solenoid of 6 m internal diameter, which generates a magnetic field of 3.8 T. It consists of various sub-detectors, each of which has a dedicated role. The innermost part is a silicon tracker, which records the tracks of all charged particles produced in the collisions. The next layer is the electromagnetic calorimeter (ECAL) made of lead tungstate($PbWO_4$) crystals. This causes the electrons and photons to shower and deposit all of their energy in the ECAL crystals. The next layer is the hadron calorimeter (HCAL), made of a brass absorber and plastic scintillator tiles. This detects the energy of the hadrons produced in the collisions, such as neutrons, pions, jets, etc. The outermost layers consist of gas-ionization chambers and drift tubes embedded in the steel return yoke, which allows us to track muons.

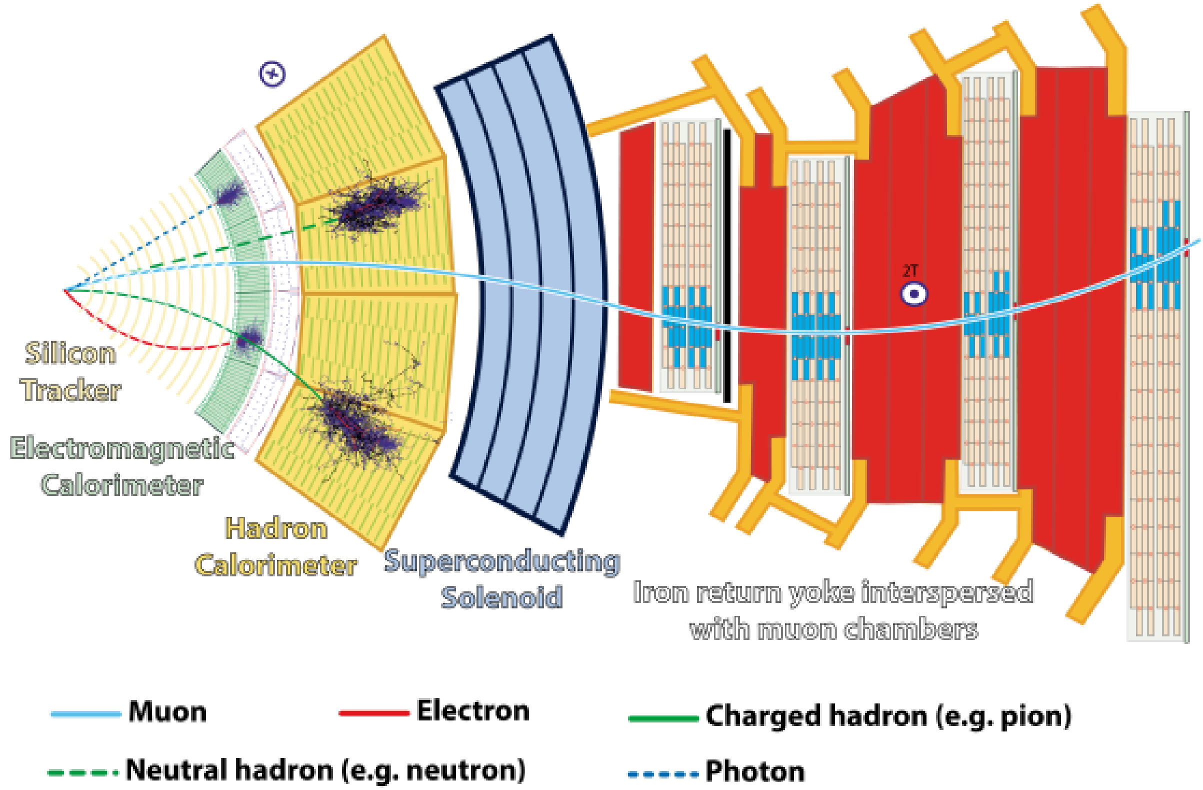


Figure 3.1: Slice showing CMS sub-detectors and how different particles interact with them[17]

3.2 Particle flow Algorithm

The particle flow (PF)[18] is a reconstruction algorithm used to classify all objects into five categories: electrons, photons, muons, neutral hadrons, and charged hadrons. Muons are reconstructed by matching tracks in the outer muon chambers with the tracks in the inner silicon tracker. The momentum is calculated by fitting the global track. Electrons are reconstructed by matching ECAL energy deposits to tracks from the silicon tracker. ECAL energy deposits that do not find a matching track are classified as photons. Energy deposits in the HCAL matching a track are classified as a charged hadron, while that which does not find a matching track is classified as a neutral hadron.

Jet clustering algorithms such as anti-k_T[19] are then run over the PF objects into PF Jets. This is important as the hadronic decays of tau produce jets.

3.3 Taus

Tau is the heaviest lepton, with a mass of 1.776 GeV, and is almost 3500 times heavier than the electron. It decays almost immediately after being produced in the collisions, with a short lifetime of $2.9 \times 10^{-13}s$. Hence, we detect the decay products of the tau and not the tau itself. The decay products of taus produce electronic signatures, which we detect in the detectors, from which the PF algorithm builds specific objects called PF objects, and the anti-k_T algorithm makes PF jets. Taus are finally reconstructed from these PF jets. The tau can decay in many ways. It is the only lepton massive enough to decay hadronically. In fact, it decays to hadrons 65% of the time. Table 3.1 summarizes the various decay modes of taus and their branching fractions. In the table, h denotes π or K.

Decay Mode	Resonance	Mass (MeV/c^2)	Branching Fraction (%)
Leptonic Modes:			
$\tau^- \to \mu^- \, \overline{\nu}_\mu \, \nu_\tau$			17.39%
$\tau^- \to e^- \, \overline{\nu}_e \, \nu_\tau$			17.82%
Hadronic Modes:			
$\tau^- \to h^- \nu_\tau$			11.51%
$\tau^- \to h^- \pi^0 \nu_\tau$	ρ^-	770	25.93%
$\tau^- \to h^- \pi^0 \pi^0 \nu_\tau$	a_0	1200	9.48%
$\tau^- \to h^- h^+ h^- \nu_\tau$	a_0	1200	9.80%
$\tau^- \to h^- h^+ h^- \pi^0 \nu_\tau$			4.76%

Table 3.1: Branching fractions for various tau decay channels[20]

3.3.1 Tau Reconstruction

Hadronically decaying taus are reconstructed using the Hadrons Plus Strips (HPS)[21] algorithm. This algorithm is designed to optimize the performance of hadronic tau reconstruction by individually considering the various decay modes. The algorithm starts by taking a PF jet seed as a hadronic tau candidate and looking for intermediate resonances that are contained in most hadronic tau decays. The charges hadrons originating from the decay are detected by the HCAL, and the photons produced in the decay of neutral pions ($\pi^0 \to \gamma\gamma$) are detected by the ECAL. To account for the high probability of the photons converting into

e^+e^- pairs, the ECAL crystals are clustered into strips, where energy in a strip is combined. The strip four-momentum is then combined with the hadrons to reconstruct the masses of the intermediate resonances.

Strip Reconstruction

The HPS algorithm pays special attention to photon conversions in the tracker, which leads to a spread of the electron/positron due to the magnetic field. This requires the momentum of each "strip" in the ECAL to be reconstructed to account for the converted objects.

The first step is centering to center a strip on the most energetic EM particle inside the PF jet. The strip size is fixed to $\eta \times \phi = 0.05 \times 0.20$. The algorithm then looks for other EM particles within the strip. If found, the four-momentum of the strip is recalculated. This process is repeated until no other particles can be combined within the strip. This algorithm has a static strip size and hence does not take into account the dependence of the size of the spread on the momentum of the particle. This algorithm is improved upon by making the size of the strip dependent on momentum. Figure 3.2 shows the ECAL strips in η-ϕ plane.

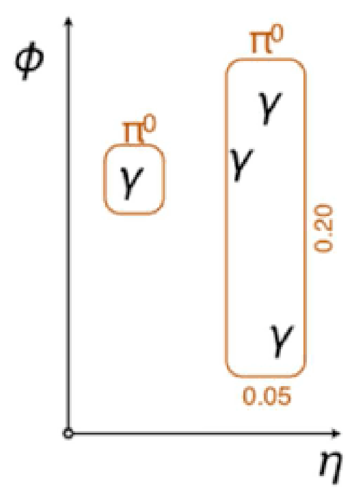


Figure 3.2: Strips in eta-phi plane [22]

Dynamic Strip Reconstruction

In dynamic strip reconstruction, the size of the strip is not fixed. This helps in certain cases where a photon coming from $\pi^0 \rightarrow \gamma\gamma$ converts in the tracker, and the electron/positron bends a large distance and falls outside the fixed strip window. Charged pions can also interact with the tracker material and produce several low-momentum electrons and photons, which can bend a lot and fall outside the strip. Dynamic strip reconstruction calculates an optimal strip size based on the momentum of the particle to better account for these effects.

3.4 Tau Identification

Since the first step of reconstruction is meant to be an inclusive selection, it is also essential to identify the correct physics objects after reconstruction. CMS uses a convolutional neural network (CNN) called DeepTau[5], which uses over 200 object-level and event-level variables. It classifies a reconstructed tau candidate object as a tau, a QCD jet, an electron, or a muon. The output is various scores for each node, which is then used to define discrimination scores for taus against jets, electrons, and muons. For example, if the raw output of the multiclassifier is y_i where $i =$ (tau, jet, electron, muon), the discrimination scores are defined as:

$$D_\alpha = \frac{y_\tau}{y_\tau + y_\alpha} \tag{3.1}$$

where $\alpha =$ (jet, electron, muon). At NanoAOD, only the Discrimination scores $vsE, vsMu$, and $vsJet$ are available. One can invert the equation to find the raw scores as:

$$y_\tau = \frac{1}{\left(\Sigma_\alpha \left(\frac{1}{D_\alpha}\right)\right) - 2} \tag{3.2}$$

$$y_\alpha = \left(\frac{1}{D_\alpha} - 1\right) y_\tau \tag{3.3}$$

Note that $\Sigma y_i = 1$, where $i \in$ (tau, jet, electron, muon).

Chapter 4

Workflow

4.1 CMS Data Tiers

Since this study is specifically aimed at taus with p_{T} between 10 and 20 GeV, the first step was to obtain these taus in the first place. In CMS, there are three main data tiers[23] to optimize storage, as well as physics information. These are AOD, MiniAOD, and NanoAOD formats. The AOD and MiniAOD formats require CMS Software[24] usage. The NanoAOD format does not require any additional software apart from ROOT[25] to analyze physics data. For this analysis, I optimized my workflow to efficiently access relevant physics information from AOD as well as NanoAOD. For this analysis, I have used CMSSW_10_6_34 for AOD to MiniAOD conversion and CMSSW_10_6_34 for MiniAOD to NanoAOD conversion. A flowchart depicting the data flow for Run 2 across various tiers is shown in 4.1.

4.2 Obtaining Taus at low momentum from AOD

Before developing any algorithm for identifying taus at low momentum, the first step was to set a benchmark using existing tau identification algorithms. This was done by using the DeepTau[5] algorithm of CMS. DeepTau is a multi-classifier that uses reconstructed tau candidates and provides various discrimination scores of taus against jets, electrons, and muons.

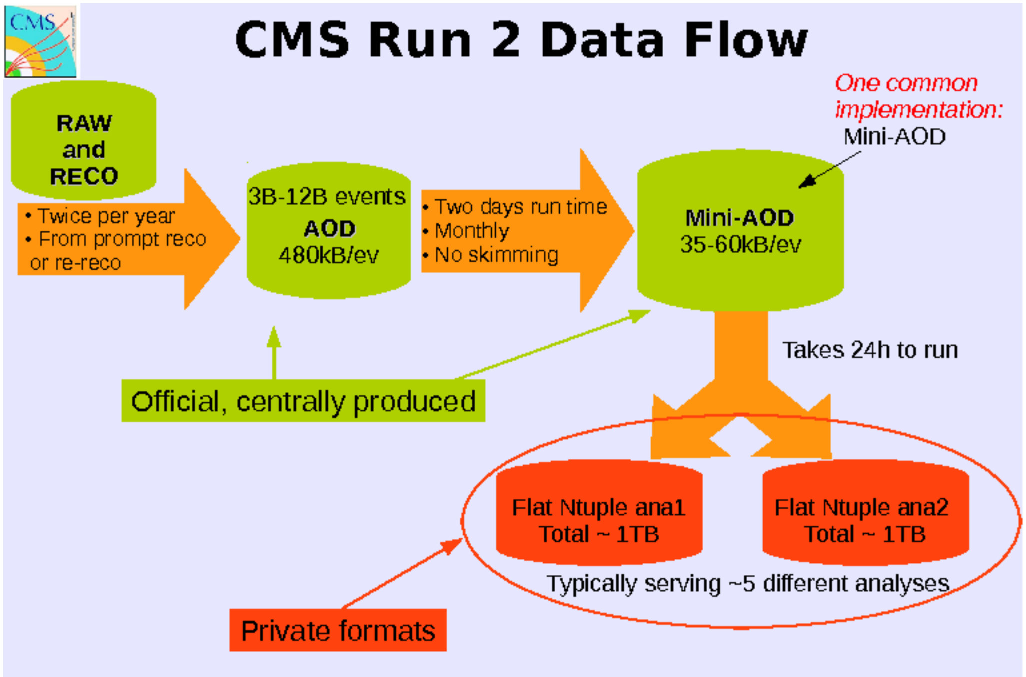


Figure 4.1: Data flow across different tiers for Run 2[26]

However, DeepTau uses MiniAOD level variables and provides the score at NanoAOD. This was a problem since neither MiniAOD nor NanoAOD contain information on taus below 18 GeV. This is done to reduce the storage information as low momentum taus are not used in most analyses.

This required me to find the module that converts AOD to MiniAOD and MiniAOD to NanoAOD and find the particular lines of code that impose this cut at 18 GeV. I created a local copy of the `RecoTauTag` and `PhysicsTools` packages. Using the help of CMS LXR[27], I located the particular set of cuts within thousands of lines of code and reduced the 18 GeV cut to 10 GeV. The same procedure was carried out for converting MiniAOD to NanoAOD. The entire procedure is described below.

1. Create local copy of `PhysicsTools` and `RecoTauTag` repositories using the `git cms-addpkg` command.
2. Change the cuts from 18 GeV to 10 GeV in:

   2.1 `RecoTauTag/RecoTau/python/PFTauPrimaryVertexProducer_cfi.py`

2.2 `PhysicsTools/PatAlgos/python/slimming/miniAOD_tools.py`

2.3 `PhysicsTools/NanoAOD/python/taus_cff.py`

2.4 `RecoTauTag/ImpactParameter/python/PFTau3ProngReco_cfi.py`

2.5 `RecoTauTag/RecoTau/src/PFTauPrimaryVertexProducerBase.cc`

2.6 `RecoTauTag/Configuration/python/tools/adaptToRunAtMiniAOD.py`

2.7 `PhysicsTools/NanoAOD/python/nanoDQM_cfi.py`

2.8 `PhysicsTools/PatAlgos/python/cleaningLayer1/tauCleaner_cfi.py`

2.9 `PhysicsTools/PatUtils/python/toolsrunMETCorrectionsAndUncertainties.py`

and do `scram b`.

3. To produce the configuration file for AOD to MiniAOD conversion, run the cmsDriver command:

```
$cmsDriver.py --python_filename *ConfigFileName.py* --eventcontent
MINIAODSIM --datatier MINIAODSIM --fileout file:*OutputMiniAODFilePath*.root
--conditions 106X_upgrade2018_realistic_v16_L1v1 --step PAT --geometry
DB:Extended --filein file:*InputAODFilePath*.root --era Run2_2018
--runUnscheduled --mc -n -1 --no_exec$
```

4. Run the command `cmsRun *ConfigFileName.py*` to produce the MiniAOD file with taus as low as 10 GeV transverse momentum.

5. Import the DeepTau module following the documentation[28] to have the DeepTau score in the NanoAOD to be produced in the next step.

6. To produce NanoAOD using the newly made custom MiniAOD, change the following set of cuts:

6.1 `PhysicsTools/NanoAOD/python/taus_cff.py`

6.2 `PhysicsTools/NanoAOD/python/genparticles_cff.py`

6.3 `PhysicsTools/NanoAOD/python/jets_cff.py`

6.4 `RecoTauTag/RecoTau/python/RecoTauCleanerPlugins.py`

and do `scram b`.

7. To produce the configuration file for MiniAOD to NanoAOD conversion, run the cmsDriver command:

   ```
   $cmsDriver.py *NanoConfigFileName.py* --filein file:*InputMiniAODFilePath*.root
   --fileout file:*OutputNanoAODFilePath*.root --mc --eventcontent NANOAODSIM
   --datatier NANOAODSIM --conditions auto:phase1_2017_realistic --step NANO
   --nThreads 2 --era Run2_2017 --no_exec -n -1$
   ```

8. Run the command `cmsRun *NanoConfigFileName.py*` to produce the NanoAOD file with taus as low as 10 GeV transverse momentum.

Using this procedure, one can obtain low momentum taus with p_{T} as low as 10 GeV at NanoAOD and its DeepTau score. Using this, we can set a benchmark, calculating reconstruction and identification(DeepTau) efficiency for taus between 10 and 20 GeV.

4.3 AOD to flat tree conversion

AOD has many more variables than MiniAOD for all physics objects, including taus. This is useful for a machine-learning algorithm, as AOD variables may contain more useful information to differentiate between taus and other objects than MiniAOD, which is what DeepTau is trained on. A flat ntuple or a flat tree is one that has no dependencies on CMSSW. AOD variables can be directly written out to a flat n-tuple, which can then be converted into a Python data frame using `uproot`[29]. Further machine-learning methods can then be applied to that set.

This led me to develop a pipeline for writing a flat n-tuple from any AOD file, writing out select branches with AOD-level variables. I wrote another filtering code to convert the flat n-tuple into a text file with all the relevant variables, which could be easily used for training and testing any machine learning algorithm.

Chapter 5

Methodology

Improving the yield of low p_{T} taus can be achieved by either improving the reconstruction or identification schemes at low p_{T}. The first step was to study how the existing generator level reconstruction and identification efficiencies depend on the p_{T} of the tau and its values at lower p_{T} ($10-20$ GeV). As per the procedure described in 4, I obtained a NanoAOD sample of taus as low as 10 GeV, with their corresponding DeepTau discrimination scores. The `/DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/RunIISummer20UL18RECO-106X_upgrade2018_realistic_v11_L1v1-v1/AODSIM` sample was used at AOD to obtain the custom sample of NanoAOD with low p_{T} taus. Figure 5.1 describes the dependence of gen reconstruction and identification efficiencies as a function of tau p_{T}. Here, generator-level reconstruction and identification efficiencies are defined as follows.

$$\text{Gen Reconstruction Efficiency} = \frac{\#\ \text{Tau Candidates matching pVisTau}}{\#\ \text{pVisTau in } p_{\mathrm{T}} \text{ range}} \tag{5.1}$$

$$\text{Gen Reco + ID Efficiency} = \frac{\#\ \text{Tau Candidates matching pVisTau \& ID}}{\#\ \text{pVisTau in } p_{\mathrm{T}} \text{ range}} \tag{5.2}$$

where pVisTau is a collection of generator-level visible taus, which means that it only contains the visible part of the decay (excluding neutrino four-vector). The pVisTau is a custom-made collection, prepared using the gen-level particle information, with p_{T} in the particular bin range, and $|\eta| < 2.3$. The identification requirement used is VTightVSJet, LooseVSe, LooseVSμ, which is the recommended DeepTau discrimination used for taus, as used in[2].

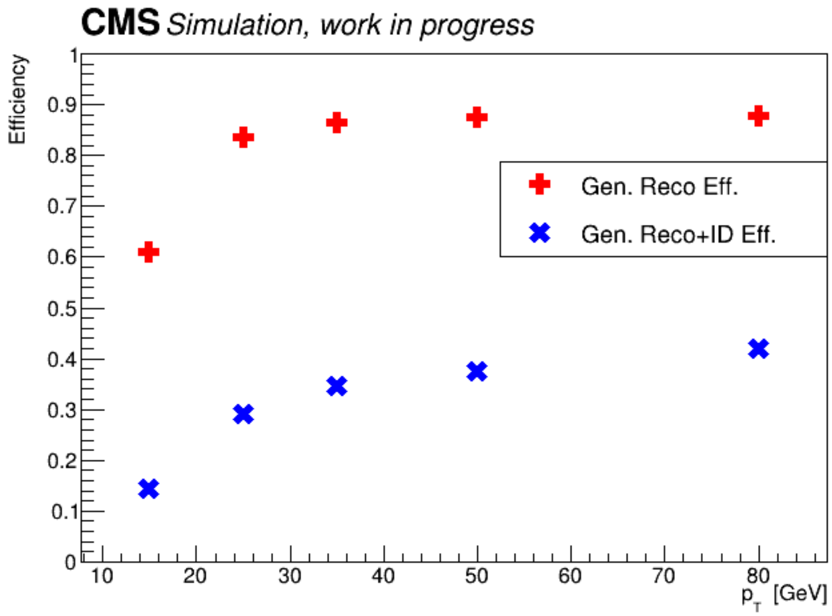
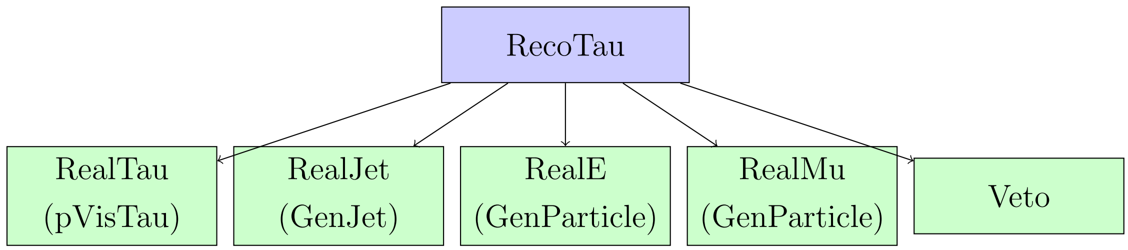


Figure 5.1: Generator level reconstruction and identification efficiency of taus as a function of p_{T}. The error bars of the points are much smaller than the points themselves.

As we can see from figure 5.1, the gen-level reconstruction efficiency is pretty high, even at low p_{T}. However, the gen-level Reco+ID efficiency is significantly lower than the gen-level reconstruction efficiency alone. This shows that improving identification schemes for taus, especially at low momentum, will be more beneficial than targeting reconstruction. Therefore, in this thesis, I attempt to improve the identification scheme for taus at low momentum, especially below 20 GeV.

5.1 Classification Scheme

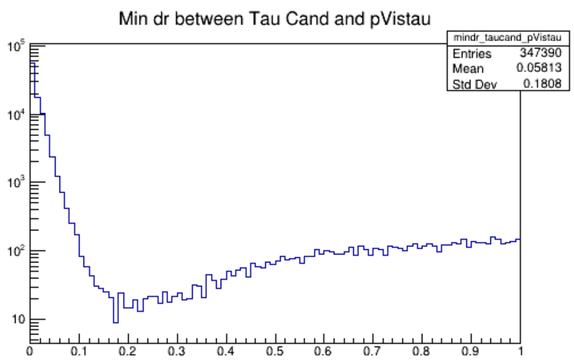
Instead of directly identifying taus, the easiest thing one can do is use existing identification algorithms for objects that can fake a tau (such as e/μ) and remove such objects from the reconstructed tau collection. But before doing this, it is important to study the nature of the objects reconstructed as a tau to see what fraction of them are really taus, as well as what is the composition of the different background objects faking a tau (jets/e/μ). In order to do this, I defined a classification scheme where reconstructed taus are matched to generator-level quantities to identify each reconstructed tau as a real tau(matching pVisTaus), real e(matching gen-electron), realMu(matching gen-muon) or realJet(matching gen-Jet). If it doesn't match with any of the generator objects, then it is classified as a veto object. The schematic for this classification and matching scheme above and below 20 GeV is shown in the schematic below.



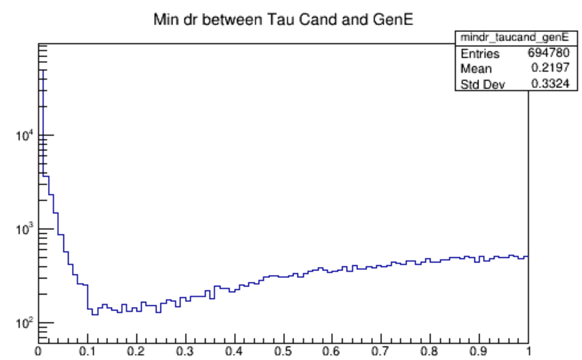
The following analysis is performed with taus of low p_{T} between $10 - 20$ GeV, as well as for taus above 20 GeV. A reconstructed tau with $p_{\mathrm{T}} \in 10 - 20$ (> 20) GeV, $|\eta| < 2.3$, and decay mode either one-prong or three-prong(dm < 3 or dm > 9) are selected. pVisTaus are generator-level hadronic taus constructed by adding four vectors of their visible decay products. pVisTaus are required to have $p_{\mathrm{T}} > 7(12)$ and $|\eta| < 2.3$. GenElectron and GenMuon collections are built from the GenParticle collection, with $p_{\mathrm{T}} > 7(12)$, $|\eta| < 2.4$, and status$=1$ objects. The GenJet collection is required to have $p_{\mathrm{T}} > 7(12)$ and $|\eta| < 2.4$. The minimum distance in $\eta - \phi$ space (dR) between each reconstructed tau and objects in each collection is calculated, and the reconstructed tau is classified into one of five classes as follows.

Take a reconstructed tau object and compute its minimum dR against a collection of different objects; pVisTaus, which are generator-level visible taus constructed from the GenParticle record; GenElectrons, which are generator-level electrons; GenMuons, which are generator-level muons; GenJets, that are generator-level AK4 jets. If the reconstructed tau object matches a pVisTau object within a dR<0.25, it is classified as a real tau. Instead, if it matches with a GenElectron object within a dR<0.20, it is classified as a real electron. Instead, if it matches with a GenMuon object within a dR<0.20, it is classified as a real muon. Instead, if it matches with a GenJet object within a dR<0.40, it is classified as a real jet. If it does not get a label assigned at the end of the process, the tau is called a veto tau.

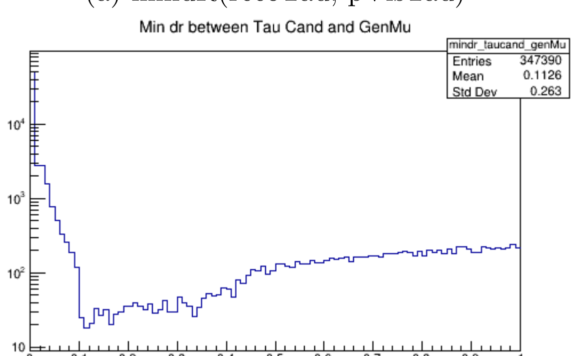
The dR values used in the procedure are obtained by plotting the minimum dR between each recoTau and objects in each collection. The plots for low p_{T} recoTaus are shown in Figure 5.2. Now that we have defined a classification scheme, we can study the composition of recoTau objects, both above and below 20 GeV. Figure 5.3 shows the composition of recoTau objects both above and below 20 GeV. Note that below 20 GeV, almost a third of all recoTaus are classified as veto! This means that they did not match with any pVisTau, GenJet, GenElectron, or GenMuon within $dR = 0.4$! The same fraction above 20 GeV is less than 2%. This indicates that at low momentum, there are a lot of recoTaus that do not match any of the obvious candidates that could potentially fake a tau.



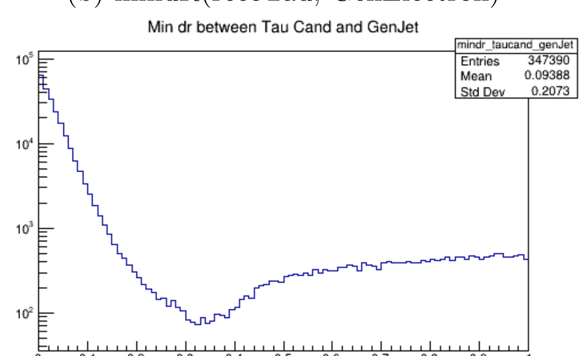
(a) mindR(recoTau, pVisTau)



(b) mindR(recoTau, GenElectron)

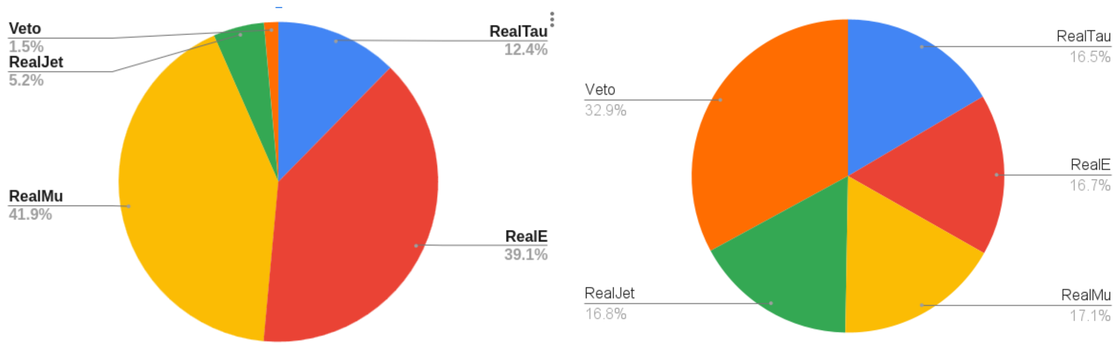


(c) mindR(recoTau, GenMuon)



(d) mindR(recoTau, GenJet)

Figure 5.2: Minimum dR plots for recoTaus vs various collections.



(a) Above 20 GeV (b) 10-20 GeV

Figure 5.3: Classified objects in the recoTau collection

5.2 Veto studies

Since the number of veto objects below 20 GeV is a large fraction of all objects, even if overlap removal were successful, it would be difficult to account for these veto objects. It is important to further understand what these veto objects are. Where are they coming from? Are they matching with some other GenParticle apart from generator level visible taus, electrons, muons, and jets? To study the nature of these veto objects, I plotted the PDGID of the closest GenParticle to the veto object within a dR cone of 0.6. This cone size is generously large to account for even very far away objects. Figure 5.4 shows that most($>$ 99%) of the veto objects do not match any GenParticle within dR $<$ 0.6 (filled as 0 in the histogram). No concrete source of these veto taus could be found, but they could possibly originate from inefficiencies in the detector simulation, where noise in the detector could possibly get reconstructed as a tau. These veto objects can also originate from out-of-time pileup, in which case no matching GenParticle would be found. It can also come from bookkeeping errors in the generator-level record, where certain particles are not listed. From 5.3, we see that about 80% of reconstructed taus above 20 GeV are electrons and muons, while the same fraction is only 35% for taus between 10-20 GeV. Since the proportion of electrons and muons is still considerably large, we can attempt to use electron and muon identification criteria to first select identified muons and electron collections and then remove any of these objects that may be listed in the reconstructed tau collection as well.

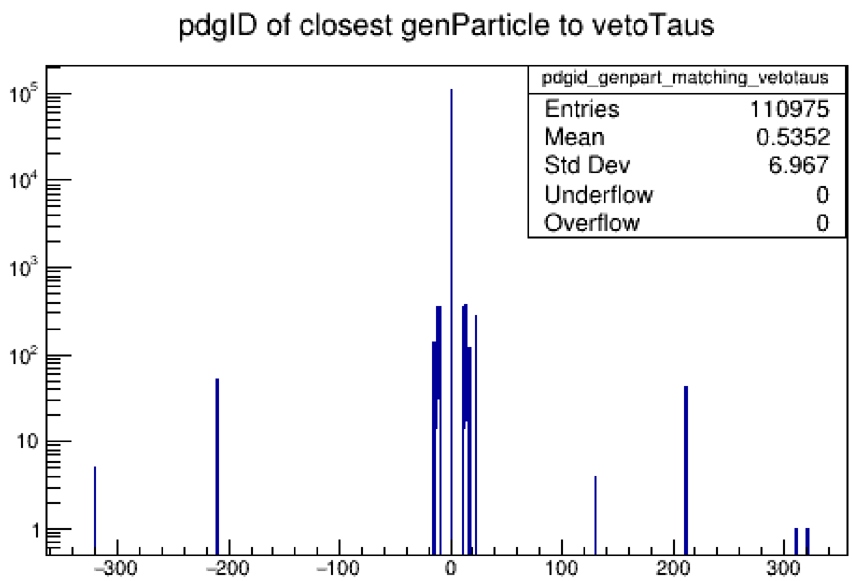


Figure 5.4: PDGID of closest GenParticle to recoTau within dR < 0.6

5.3 Overlap Removal

Reconstructed muon and electron collections are selected with $p_T > 7(10)$ GeV and $|\eta| < 2.4$. LooseID and cutBased> 0 identification requirements are applied on muons and electrons, respectively. Table 5.1 summarizes all the selections used in the construction of all collections so far.

Collections	Selections
goodElectron	$p_T > 7(10)$ GeV, $\|\eta\| < 2.4$, cutBased> 0
goodMuon	$p_T > 7(10)$ GeV, $\|\eta\| < 2.4$, LooseID
RecoTau	$p_T \in 10 - 20(> 20)$ GeV, $\|\eta\| < 2.3$, Decay Mode (< 3 or > 9)
Gen-Electron/Muon	$p_T > 7(12)$ GeV, $\|\eta\| < 2.4$, status$= 1$
GenJet	$p_T > 7(12)$ GeV, $\|\eta\| < 2.4$
pVisTau	$p_T > 7(12)$ GeV, $\|\eta\| < 2.3$, hadronic taus

Table 5.1: Summary of selections

We only overlap remove against electrons and muons, while the DeepTau standard ID selections include a cut on jet discrimination score as well. For fair comparison against

DeepTau, a jet score cut is also applied in addition to overlap removal. The working points of the existing DeepTau algorithm can also be modified to possibly obtain better discrimination for low p_{T} taus. I performed a signal-to-background comparison for various sets of cuts on the raw jet score, electron score, and muon score, as well as overlap removal and the standard DeepTau selections.

RecoTaus with jet-score < 0.1 and not matching any goodMuon or goodElectron within $dR < 0.2$ are matched to generator-level collections to classify them into five categories as per the classification scheme described in 5.1. The signal-to-background ratio for standard DeepTau cuts, overlap removal, as well as various other triplets of DeepTau discrimination scores are summarised in tables 5.2 for taus above 20 GeV, and in 5.3 for taus between 10-20 GeV.

Total #'s	Taus	Jets	Muon	Electron	Veto	Sig/Bkg
Total	$90,787$	$33,203$	$330,148$	$288,731$	9.520	0.14
(VTight, Loose, Loose)	$51,837$	$1,641$	34	$3,358$	565	9.26
$(0.1, 0.1, 0.3)$	$52,391$	$1,878$	23	551	655	16.86
$(0.1, 0.3, 0.3)$	$57,024$	$2,246$	36	$1,602$	814	12.14
$(0.1, 0.5, 0.5)$	$59,151$	$2,611$	53	$3,335$	921	8.55
OvRem(jet_score < 0.1)	$60,073$	$3,416$	179	$21,327$	$1,023$	2.32

Table 5.2: Comparison of DeepTau score cuts and Overlap Removal for taus above 20 GeV. A standard DeepTau working point VTightvsJet, LoosevsE, and LoosevsMu is also shown for comparison. The triplet of cuts is applied to Reco taus (jet_score, e_score, mu_score), and statistics in each category are listed. For overlap removal, jet_score cut of 0.1 is applied, and Reco taus are cleaned against identified muon and electron collections using dR matching.

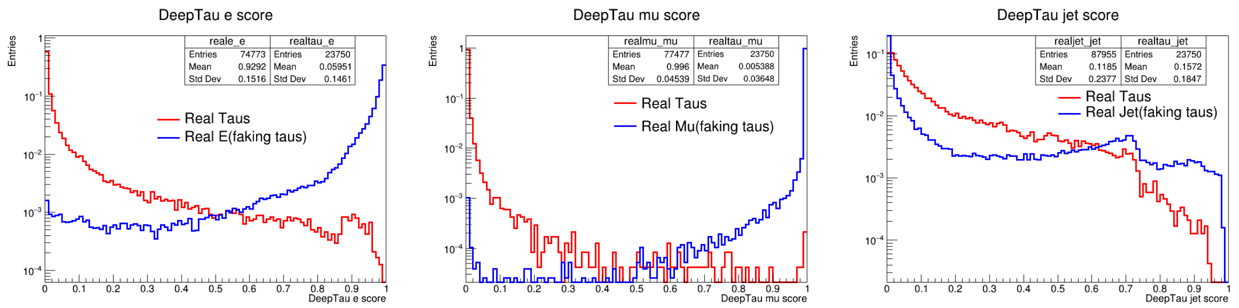
Total #'s	Taus	Jets	Muon	Electron	Veto	Sig/Bkg
Total	$51,871$	$49,375$	$57,453$	$54,305$	$98,828$	0.20
(VTight, Loose, Loose)	$17,430$	$1,504$	11	$1,277$	$5,647$	2.06
$(0.1, 0.1, 0.3)$	$17,350$	$1,675$	9	190	$6,072$	2.18
$(0.1, 0.3, 0.3)$	$20,594$	$2,305$	14	652	$8,744$	1.76
$(0.1, 0.5, 0.5)$	$22,292$	$2,787$	26	$1,482$	$11,164$	1.44
OvRem(jet_score < 0.1)	$24,034$	$3,541$	59	$9,223$	$15,010$	0.86

Table 5.3: Comparison of DeepTau score cuts and Overlap Removal for taus between $(10-20)$ GeV. A standard DeepTau working point VTightvsJet, LoosevsE, and LoosevsMu is also shown for comparison. The triplet of cuts is applied to Reco taus (jet_score, e_score, mu_score), and statistics in each category are listed. For overlap removal, jet_score cut of 0.1 is applied, and Reco taus are cleaned against identified muon and electron collections using dR matching.

As we can see from tables 5.2 and 5.3, overlap removal performs significantly worse than DeepTau working points and also some other triplets of Deeptau score based cuts, both above and below 20 GeV. This indicates that an overlap removal strategy for low p_{T} taus will not work. A better strategy may be to train a neural network to discriminate between these categories like DeepTau does.

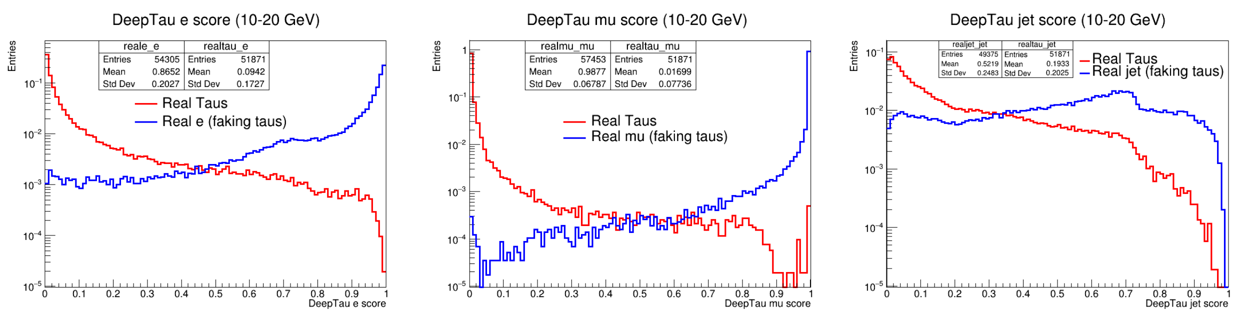
5.4 DeepTau at low p_{T}

Before training a neural network specifically for low p_{T} taus, I studied the performance of DeepTau for taus between 10-20 GeV using their raw neural network scores, as computed from equations 3.2 and 3.3. Note that this is the raw neural network score(y_i) before discrimination scores are calculated. Plots of raw DeepTau e-score, mu-score, and jet-score for Real taus and Real Electrons, Muons, and Jets, are shown in figures 5.5 and 5.6, respectively. Plots are shown, both for taus above 20 GeV and between 10-20 GeV. As expected, the e-score, mu-score, and jet-score of real taus peak on the left, while the same scores for real electrons, real muons, and real jets peak on the right.



(a) e-score (b) mu-score (c) jet-score

Figure 5.5: Overlay plot of e, mu and jet scores for real taus and e's, mu's and jets faking taus, for Reco taus with p_{T} above 20 GeV.



(a) e-score (b) mu-score (c) jet-score

Figure 5.6: Overlay plot of e, mu and jet scores for real taus and e's, mu's and jets faking taus, for Reco taus with p_{T} between $10-20$ GeV.

All the scores, e-score, mu-score, and jet-score as a discriminator, seem to become worse below 20 GeV. I also compared the DeepTau signal acceptance and background rejection for two sets of DeepTau working points; (VTightvsJet, LooseVsE, LooseVSMu) and (MediumvsJet, LooseVsE, LooseVSMu) for taus above 20 GeV as well as between 10-20 GeV. The numbers were obtained by applying these sets of cuts to reconstructed tau objects as defined in 5.1 and studying the remaining number of signal objects and background objects after the cut.

It can be seen from 5.7 that above 20 GeV, both the working points have better signal efficiency and background rejection than the points below 20 GeV. These points below 20 GeV can now be treated as a benchmark for comparison against any future identification schemes that may be developed.

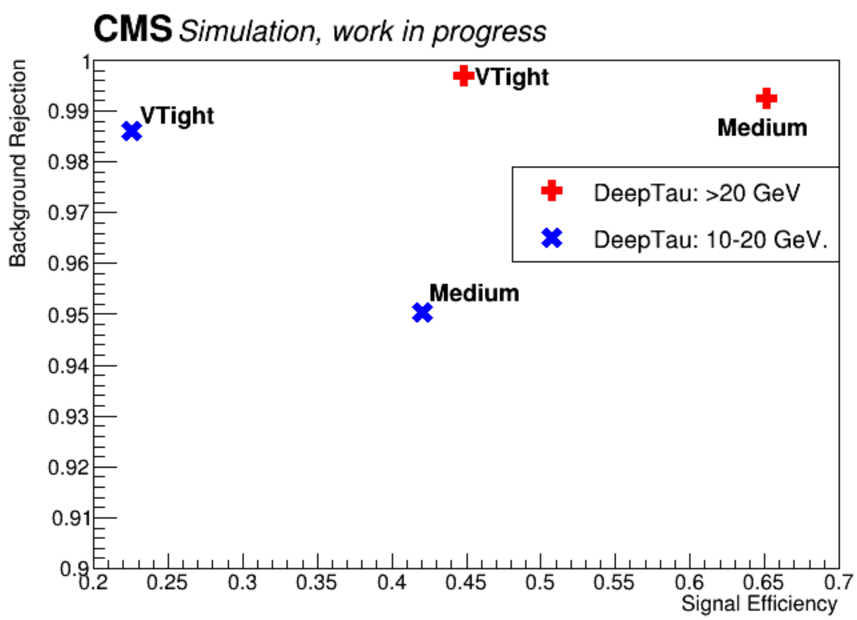


Figure 5.7: VTightVSjet, MediumVSjet working points above and below 20 GeV for DeepTau ID. LoosevsE and LoosevsMu are also applied as the standard triplet of cuts.

Chapter 6

Targeted Neural Networks

We have seen that overlap removal does not perform well against DeepTau using a very crude metric of signal-to-background. I used many variables of HPSpfTaus at AOD, and wrote them out to a flat-tree for the sample `/DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/RunIISummer20UL18RECO-106X_upgrade2018_realistic_v11_L1v1-v1/AODSIM`. This flat tree, containing object-wise information of all low p_T taus was used to train a neural network specifically for identifying low p_T taus. The flat tree was read into Python using the **uproot** package. The workflow is summarized in figure 6.1.

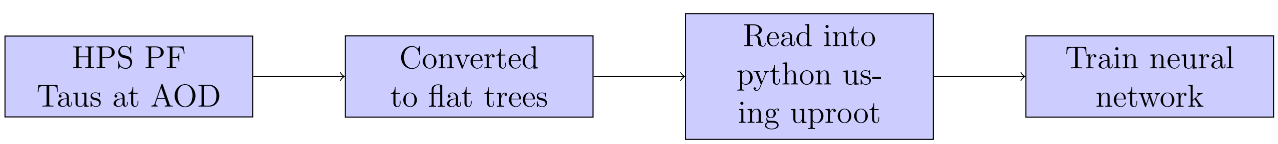


Figure 6.1: Flowchart with four rectangular boxes from left to right

6.1 Overview of Neural Networks

A neural network is a multivariate classifier that applies very elegant linear algebra to classify data. It consists of various nodes connected via edges carrying a weight and bias. It is essentially a large transformation matrix applied to a set of input variables such that a trained network can distinguish between two kinds of objects based on the output score. Typically, for a binary classifier, the singular output node takes any value between 0 and 1, 0 signifying one kind of object, and 1 signifying another object. A typical image of a neural network with input layers and hidden layers is shown in figure 6.2.

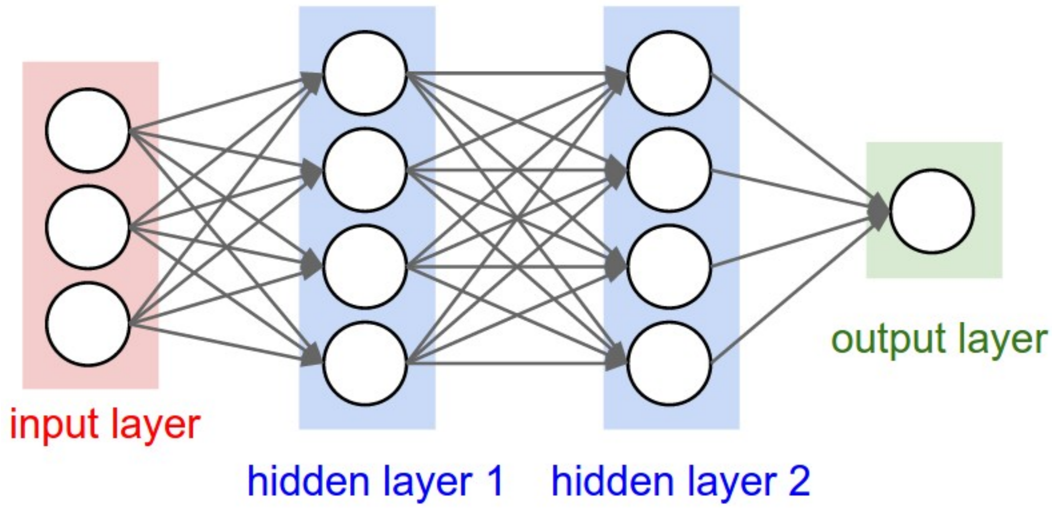


Figure 6.2: Example of a deep neural network with two hidden layers and a single output node[30].

The network is trained using backpropagation to minimize a loss function. The output value for each kind of object will typically be a distribution peaking at 0 and 1, respectively. Ideally, one object should peak at 1, while the other should peak at 0, but perfection is a myth. See figures 5.6 for instance. A binary classification made based on any cut chosen on the value of the output neuron will have some signal acceptance and background rejection. Signal acceptance is the fraction of signal events correctly identified by that cut, and background rejection is the fraction of non-signal events correctly rejected by the cut. The ideal neural network should then have signal efficiency = 1 and background rejection = 1. Signal acceptance is also called the true positive rate(TPR), while a different axis is sometimes used instead of background rejection, called false positive rate(FPR), which is just 1− background rejection. Typically, since the choice of cut is arbitrary, the performance of a neural network is visualized through a receiver operator characteristics(ROC) curve, where the area under the curve(AUC) somewhat signifies how good the network is. An AUC of 1 signifies perfect classification. Figure 6.3 shows an example of good and bad ROC curves with different values of AUC.

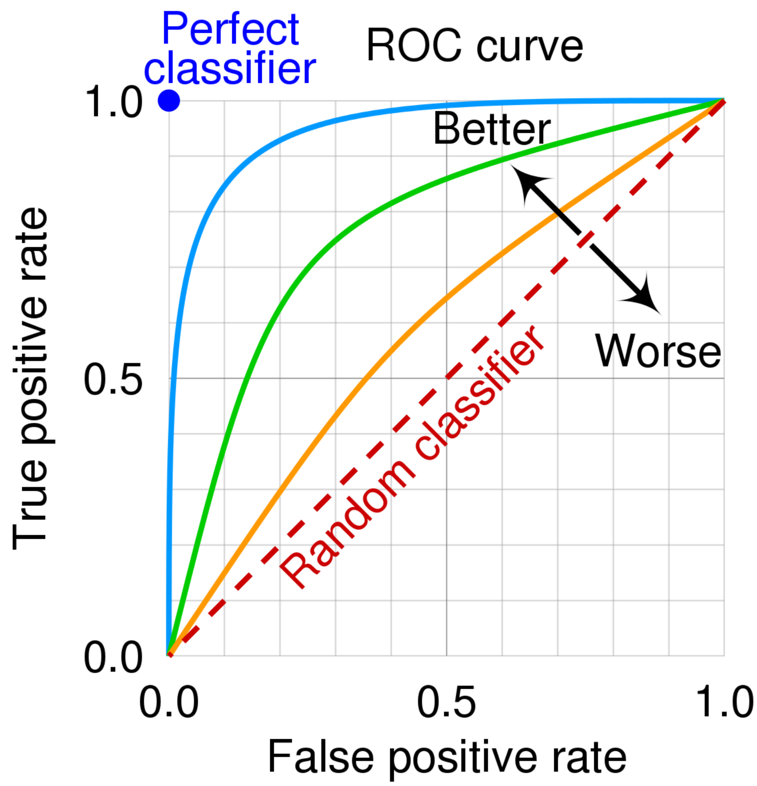


Figure 6.3: Examples of various ROC curves[31].

In figure 6.3, the blue curve performs better than the green curve since it has a larger area under curve. A random classifier would correctly or incorrectly identify objects with equal probability, so the ROC curve for a random classifier is a diagonal line, with AUC = 0.5.

6.2 Input variables

DeepTau uses over 200 variables and a CNN to train its network. We start small by finding many equivalent variables of DeepTau and a few other inspired variables. A total of just 28 variables have been used to train the neural network(NN). A list of these is shown below in table 6.1.

Variables	Linear Transform	Comments
centrality_fraction	$[0-5]$	Ratio of E_{T} sum of constituents in a $dR < 0.1$ cone to a $dR < 0.2$ cone around the tau direction.
leading_track_mom_fraction		Ratio of p_{T} of highest p_{T} object in the signal cone to E_{T} sum of constituents in a $dR < 0.2$ cone around the tau direction.
track_radius	$[0-5]$	p_{T} weighted average dR of all constituents.
max_track_sep_sig	$[0-5]$	Maximum dR between a constituent and the tau direction.
track_mass		Invariant mass of all charged constituents.
iso_03		Ratio of the sum of the momentum of constituents within $dR < 0.3$ to tau p_{T}
iso_05		Ratio of the sum of the momentum of constituents within $dR < 0.5$ to tau p_{T}
no_neutral_prongs		
no_charged_prongs		
no_signalCands	$[0-5]$	
no_signalPiZeroCandidates		
no_isolationCands	$[0-5]$	
leadChargedHadrCand_mass		
no_signalChargedHadrCands		
no_signalGammaCands	$[0-5]$	
no_isolationChargedHadrCands	$[0-5]$	
no_isolationGammaCands	$[0-5]$	
no_isolationPiZeroCandidates		
leadChargedHadrCand_pt_BY_Tau_pt		
leadCand_pt_BY_Tau_pt		

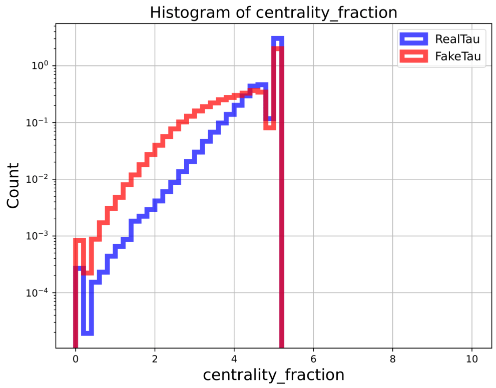
tau_isolationPFChargedHadr Cand-sPtSum_BY_Tau_pt	$[0-5]$	
tau_maximumHCALPFClusterEt _BY_Tau_pt		
tau_isolationPFGammaCandsEt Sum_BY_Tau_pt		
leadNeutralCand_pt_BY_Tau_pt	$[0-5]$	
leadPFChargedHadrCand_pt_BY _Tau_pt	$[0-5]$	
leadPFNeutralCand_pt_BY_Tau_pt		
leadPFCand_pt_BY_Tau_pt		
leadPFChargedHadrCandsigned Sipt_BY_Tau_pt	$[0-5]$	

Table 6.1: List of variables and transformations used for training

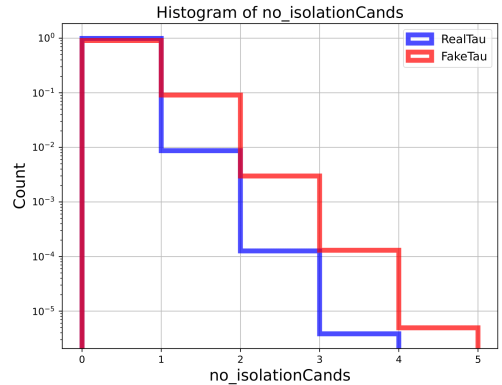
Table 6.1 describes all the 28 variables used for training taus at low p_{T}. For certain variables, a linear transformation based on their minimum and maximum values is applied to rescale the values to a number between $0-5$. This is a general scheme akin to normalization of input variables, applied to reduce the potential biasing of the network due to large input variables. Certain variables are inspired by [32], while other variables are constructed by considering certain DeepTau variables. Plots for some of these variables are shown in figure 6.4.

In 6.4, the blue curve shows the distribution of real taus, as per the classification scheme described in 5.1. The red curve is the combination of all other classified objects, that is, any reconstructed tau that was not matched to a generator-level visible tau. In contrast to DeepTau, where other non-tau objects have separate discriminators, we attempt a binary classification.

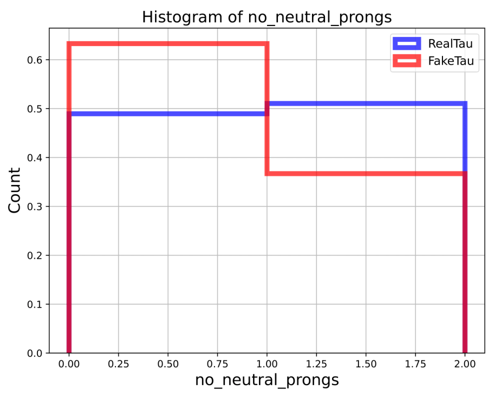
Another important piece of information to study is the correlation between all the input variables as well as their relative importance in the binary classification. This indicates the amount of repetitive information contained in the training variables. The correlation matrix with the label of real or fake is shown in figure 6.5. This is why the correlation along the



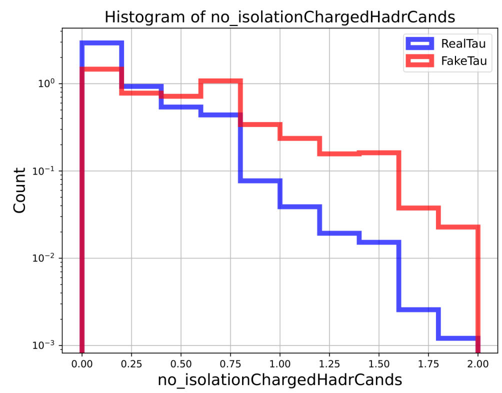
(a) Centrality Fraction



(b) No. of isolation candidates



(c) No. of neutral prongs



(d) No. of iso-charged hadron cands

Figure 6.4: Distributions of a few of the training variables shown for real and fake objects.

diagonal is one.

6.3 Network architecture and training

Reconstructed taus at low p_T from the `/DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/RunIISummer20UL18RECO-106X_upgrade2018_realistic_v11_L1v1-v1/AODSIM` sample, classified using the scheme described in 5.1 are used to train a deep neural network. A binary classifier is trained, where all real taus are assigned 1, while all other objects are assigned 0.

The 28 variables described in 6.1 are set as an input to a $256 \times 128 \times 64 \times 32 \times 8 \times 1$ node network. ReLu[33] activation is applied on all the hidden layers, while a sigmoid acti-

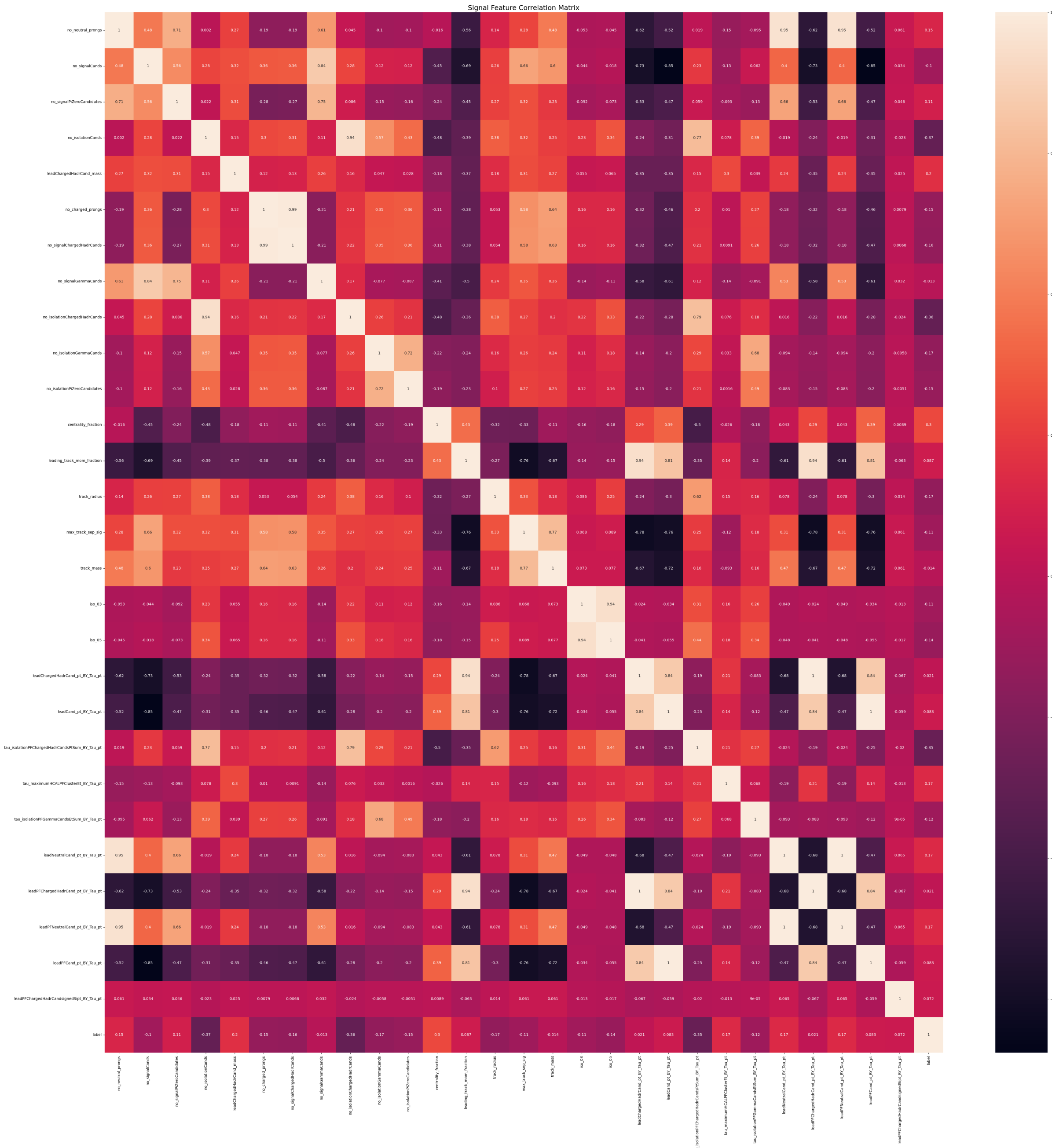


Figure 6.5: Correlation matrix of all input variables

vation is applied to the output node. The training is optimized with *Adam*[34], and the loss function is defined as binary cross-entropy. The network is trained for a total of 40 epochs. Variations of training batch size from 64 to 512, have been tested, and the optimal value was found to be 128. The AUC values from the ROC curve in each training sample are shown in table 6.2.

Batch Size	Testing AUC	Training AUC
64	0.9041	0.9162
128	0.9051	0.9170
256	0.9041	0.9157
512	0.9019	0.9117

Table 6.2: Training and testing AUC values for various batch sizes. Epochs = 40.

We choose batch size = 128 since that has the largest AUC for both training and testing.

6.4 Permutation invariance

The relative importance of each input variable is also tested using the permutation invariance method. Initially inspired by DeepTau, another variable was being used for training, the E/p_{T} of the tau object. However, permutation invariance with $n = 10$ showed E/p_{T} as the most significant feature [6.6], even though the distributions of the variable for real and fake objects do not look very different. This is concerning, as this may indicate some E or p_{T} dependence that has crept in. E is divided by p_{T} to account for this, but that assumes a linear relationship between the two objects, which may not be necessarily true. To test this, a previous version of the NN trained on a set of 31 variables including E/p_{T} is tested on two samples; taus between $10 - 15$ GeV and $15 - 20$ GeV. The ROC curves for both are shown in figure 6.7.

As we can see, the performance is significantly better for objects with p_{T} between $15 - 20$ GeV. Hence, the variable E/p_{T} was removed from the set of training variables. A few other variables were dropped from the input variable list based on their low feature importance from [6.6].

6.5 Smaller p_{T} window training

It is important that any neural network that we train, does not make a decision based on the p_{T} of the object. This is also the reason why p_{T} is not used as a training variable. The neural

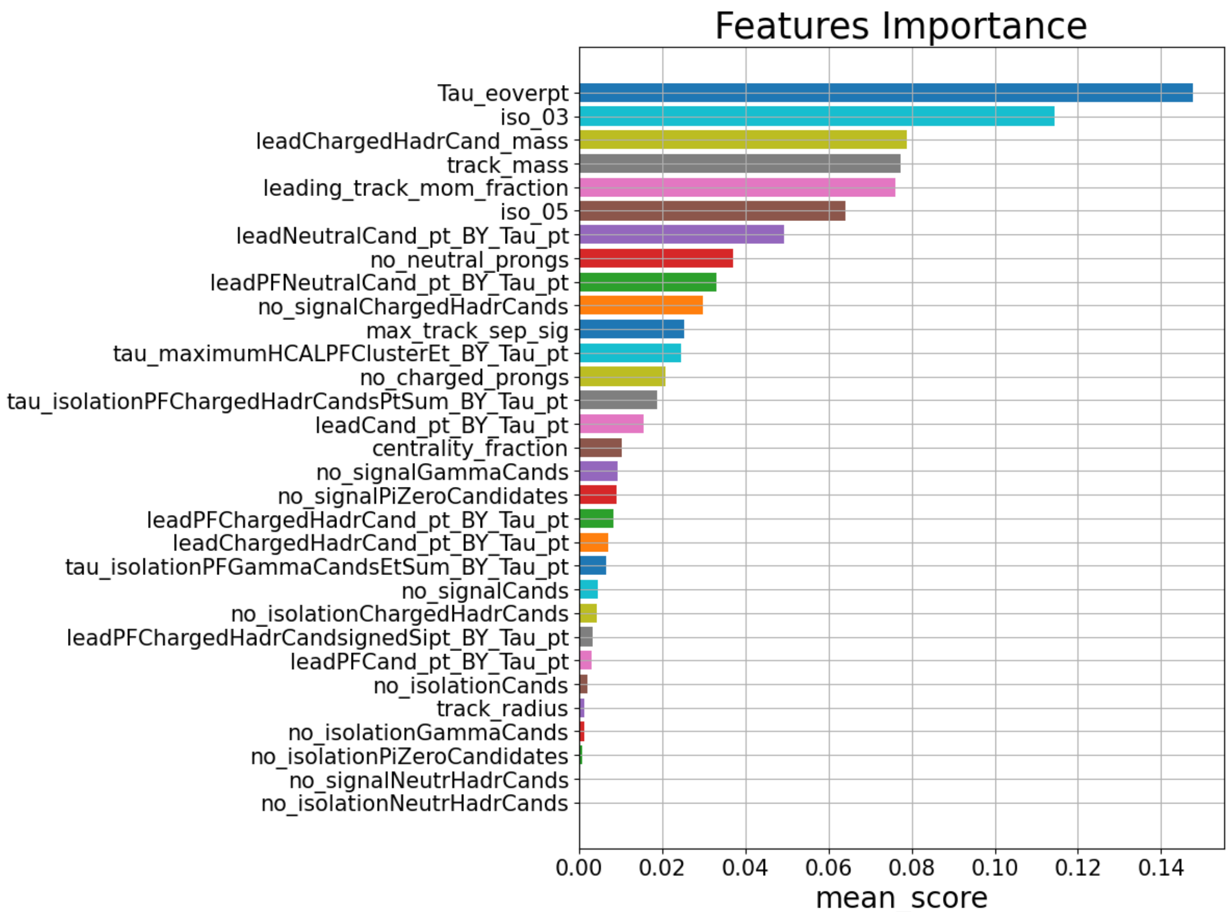
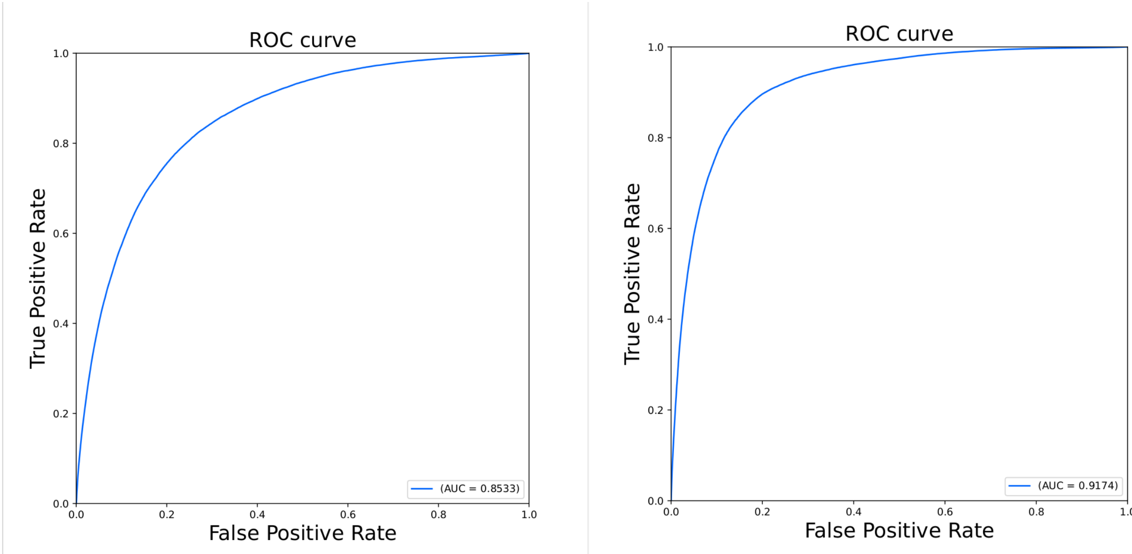


Figure 6.6: Feature importance distribution using permutation invariance.

network with the same architecture as mentioned in 6.3 is trained on a subset of objects, with p_{T} between $10-15$ GeV, and $15-20$ GeV. Both networks are trained with an equal mixture of real and fake objects, with $\tilde{1}32k$ real and fake tau candidates each for the $10-15$ GeV range, and $\tilde{1}29k$ real and fake candidates each for the $15-20$ GeV range. The AUC from the dedicated training in different p_{T} bins is compared to the AUC from the neural network trained on the entire $10-20$ GeV bin, but tested specifically on $10-15$ GeV, and $15-20$ GeV bin. The AUC values are listed below in table 6.3.

We can see from the data in 6.3 that even after dedicated training at finer p_{T} bins, there is only a very slight improvement in the AUC, so the neural network is frozen.



(a) Tested on $10 - 15$ GeV (b) Tested on $15 - 20$ GeV

Figure 6.7: ROC curves for NN tested on taus between $10 - 15$ and $15 - 20$ GeV.

p_{T} range (GeV)	Dedicated training in p_{T} range	Trained on full range but tested on dedicated range
$10 - 15$	0.880	0.853
$15 - 20$	0.925	0.917

Table 6.3: AUCs of the ROCs for various networks trained and tested on different p_{T} ranges

6.6 Comparison against DeepTau

Now that we have a network dedicated to identifying taus at low p_{T}, we can compare how it performs against DeepTau at low p_{T}. The collection of objects as defined in 5.1 are used to quantify the signal acceptance versus the background rejection for my neural network against the DeepTau performance as shown in figure 5.7. The signal efficiency of my network for the same value of background rejection of DeepTau is shown in figure 6.8 in addition to the previous figure 5.7.

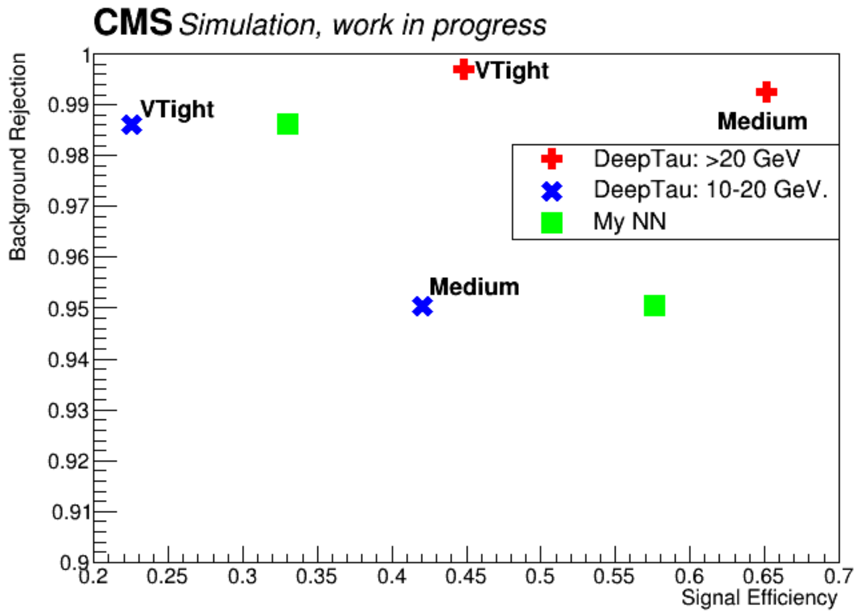


Figure 6.8: VTightVSjet, MediumVSjet working points above and below 20 GeV for DeepTau ID. LoosevsE and LoosevsMu are also applied as the standard triplet of cuts. Shown in green are working points of my neural network trained at low p_T at the same level of background rejection.

As we can see from figure 6.8, using dedicated training for taus between $10 - 20$ GeV, we can obtain better signal efficiencies for the same value of background rejection obtained from DeepTau. For the VTightvsJet working point, the signal efficiency increased from 22% to 33%, while for the MediumvsJet working point, the signal efficiency increased from 42% to 58%. This is a significant increase, as this clearly allows for better identification of low p_T taus.

Chapter 7

Acceptance studies

In 6.6, we have shown a proof of concept that low p_{T} taus can be identified better just by training a dedicated neural network between $10 - 20$ GeV. The immediate next step to do is to demonstrate an increase in acceptance for a particular signal sample. In this section, I attempt to calculate the additional acceptance gained by adding taus between $10 - 20$ GeV. For this specific analysis, we choose to use the following $H \rightarrow \tau\tau$ sample.

1. $H \rightarrow \tau\tau$
   `/GluGluHToTauTau_M125_TuneCP5_13TeV-powheg-pythia8/`
   `RunIISummer20UL18RECO-106X_upgrade2018_realistic_v11_L1v1-v3/AODSIM`

Taus between $10 - 20$ GeV have been obtained for the $H \rightarrow \tau\tau$ sample using the same workflow as defined in 4.

7.1 Final States

Two final states $\mu\tau_h$ and $e\tau_h$ are chosen to study the sample acceptance for a particular set of DeepTau cuts. The $H \rightarrow \tau\tau$ sample is a purely signal-like sample. Barring any initial state or final state radiation jets faking a muon and tau, most of the events in the $\mu\tau_h$ and

$e\tau_h$ final state should come from one of the taus decaying hadronically (τ_h) and the other tau decaying leptonically into μ or e. We will first compute the acceptance for taus with p_{T} above 20 GeV using the standard DeepTau algorithm to identify the taus, and then we will see how much additional acceptance is gained by using my algorithm to identify additional taus with p_{T} between $10 - 20$ GeV.

7.1.1 Event selection for DeepTau

Muons are selected with $p_{\mathrm{T}} > 25$ GeV, $|\eta| < 2.1$, RelIso04(relative isolation) < 0.15, $|d_{xy}| < 0.045$ and $|d_z| < 0.2$ and are required to pass looseID. Since AOD does not contain any identification variables, looseID for muons is not available. However, the equivalent of looseID is created at AOD using the definition of looseID[35]. Electrons are selected with $p_{\mathrm{T}} > 33$ GeV, $|\eta| < 2.1$, RelIso03(relative isolation) < 0.15, $|d_{xy}| < 0.045$ and $|d_z| < 0.2$. Since AOD does not contain the cutBased identification for electrons, a custom identification scheme is defined for electrons in order to maintain uniformity between selections at AOD and NanoAOD tiers. The cutBased variable at NanoAOD for identifying electrons is a combination of many variables such as $\sigma_{i\eta i\eta}$, h/e, etc[36]. A subset of these variables, which were available at both AOD and NanoAOD, was used to construct a custom identification scheme for electrons. Electrons are required to have $\sigma_{i\eta i\eta} < 0.0126$, $h/e < 0.05$, *pfRelIso_03_all*< 0.2, *#lost hits*≤ 2 and *conversion veto*$=$ True.

Reconstructed taus are selected with $p_{\mathrm{T}} > 20$ GeV, $|\eta| < 2.3$, $|d_z| < 0.2$. These reconstructed taus are also required to pass the DeepTau MediumvsJet working point. Additionally, the τ_h in the $e\tau_h$ channel must satisfy DeepTau TightvsE and VLoosevsMu, while the τ_h in the $\mu\tau_h$ channel must satisfy DeepTau VLoosevsE and TightvsMu working points.

The leading $e(\mu)$ and leading τ_h must have opposite sign and be separated by a $dR > 0.5$. If an event contains both a leading electron and a leading muon with opposite signs and is separated by $dR > 0.3$, then that event is vetoed, irrespective of whether it contains a τ_h or not. These sets of selections are summarized below in table 7.1.

Muons	$p_T > 25$ GeV, $\|\eta\| < 2.1$, RelIso04 < 0.15, $\|d_{xy}\| < 0.045$, $\|d_z\| < 0.2$, looseID = is a PF Muon and (is Global or Tracker Muon)
Electrons	$p_T > 33$ GeV, $\|\eta\| < 2.1$, RelIso03 < 0.15, $\|d_{xy}\| < 0.045$, $\|d_z\| < 0.2$, $\sigma_{i\eta i\eta} < 0.0126$, $h/e < 0.05$, *pfRelIso_03_all* < 0.2, #*lost hits* ≤ 2 and *conversion veto*= True
Taus	$p_T \geq 20$ GeV, $\|\eta\| < 2.3$, $\|d_z\| < 0.2$, MediumvsJet. Additionally, (VLoosevsMu + TightvsE) for $e\tau_h$ channel and (VLoosevsE + TightvsMu) for $\mu\tau_h$ channel.

Table 7.1: Selections on objects for populating final states

7.1.2 Event selection for neural network

The goal is to compute the gain in acceptance if low p_T taus for the $H \to \tau\tau$ sample is also considered. Low p_T taus(between $10 - 20$ GeV) are written out to a flat tree only if the event passes all the selections from table 7.1 except for the vsE and vsMu conditions for taus. Instead of requiring a discrimination score against electrons and muons, taus are cleaned against selected electrons and muons, requiring a $dR > 0.2$. This will allow us to see the amount of additional acceptance gained by adding low p_T taus as well.

7.1.3 Defining and computing acceptance

Using the selections described in 7.1, events are selected for each sample, and the number of events populating each channel $e\tau_h$ and $\mu\tau_h$ are counted. The acceptance for each channel is defined as:

$$\text{Acceptance } e\tau_h = \frac{\#\text{ of events in } e\tau_h \text{ channel}}{\#\text{ total events}} \tag{7.1}$$

$$\text{Acceptance } \mu\tau_h = \frac{\#\text{ of events in } \mu\tau_h \text{ channel}}{\#\text{ total events}} \tag{7.2}$$

The acceptance for each of the two channels using the selections described in 7.1 is computed for the $H \to \tau\tau$ sample using taus above 20 GeV only. The acceptance values are quoted below in the following table.

Final state/Sample	$e\tau_h$	$\mu\tau_h$
$H \to \tau\tau$	0.971%	1.936%

Table 7.2: Acceptance using DeepTau for taus above 20 GeV for $e\tau_h$ and $\mu\tau_h$

The next thing to calculate is the acceptance using low p_T taus. This analysis will be performed soon.

7.2 Signal to Background studies

We showed in section 7.1 that the signal acceptance for taus in the $H \to \tau\tau$ sample could be improved by additional selection of taus at low momentum. However, it is also important to study how the background increases for various samples when we accept taus at low momentum. For this, we use the following set of samples where we compare the signal-to-background between my neural network and DeepTau extended to low momentum.

1. $H \to \tau\tau$
   `/GluGluHToTauTau_M125_TuneCP5_13TeV-powheg-pythia8/`
   `RunIISummer20UL18RECO-106X_upgrade2018_realistic_v11_L1v1-v3/AODSIM`

2. WJetsToLNu
   `/WJetsToLNu_TuneCP5_13TeV-amcatnloFXFX-pythia8/`
   `RunIISummer20UL18RECO-106X_upgrade2018_realistic_v11_L1v1-v1/AODSIM`

3. DYJets$\to$LL
   `/DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/`
   `RunIISummer20UL18RECO-106X_upgrade2018_realistic_v11_L1v1-v1/AODSIM`

The WJets sample is a purely background-like sample. The W can decay leptonically into μ or e or hadronically producing jets, while another jet can fake a tau to populate the selected final states.

The DYJets→LL sample is trickier since it contains both signal-like and background-like events. If both the leptons in the final state are taus(DYJets→ $\tau\tau$), then the sample is similar to the $H \rightarrow \tau\tau$ sample, where all of the events are signal-like. Otherwise, if both the leptons in the final state are electrons or muons (DYJets→ ee or DYJets→ $\mu\mu$), then the sample is similar to the WJets sample, where all of the events are background-like.

The same set of final states as described in 7.1 with the same set of cuts as described in 7.1 is used, only now that instead of using taus above 20 GeV transverse momentum, we use taus between $10 - 20$ GeV. The goal is to obtain acceptance using DeepTau extended to low momentum, as well as well as using my neural network. The same definition for acceptance, as defined in 7.1.3, is used here as well. The acceptance for DeepTau applied for low momentum taus for various samples in both the final states is shown in table 7.3.

Sample/Final State	$e\tau_h$	$\mu\tau_h$
Signal:		
$H \rightarrow \tau\tau$	0.255%	0.325%
DYJets→ $\tau\tau$	0.087%	0.180%
Background:		
DYJets→ $ee/\mu\mu$	0.0167%	0.051%
WJetstoLNu	0.0057%	0.028%

Table 7.3: Acceptance for the set of cuts as defined in 7.1, now using taus with p_{T} between $10 - 20$ GeV, for various samples and both the $e\tau_h$ and $\mu\tau_h$ final states.

The next step is to obtain the same table as 7.3 using my neural network and compare the signal-to-background in both, which remains to be seen.

Chapter 8

Conclusions

In this thesis, I have explored the possibility of identifying taus with p_{T} between $(10-20)$ GeV for use in any generic analysis. Certain barriers within the CMS software framework are also identified, such as accessing taus at low momentum in NanoAOD requires one to start from AOD, while DeepTau runs on MiniAOD. This significantly hinders any attempt to even obtain these low momentum objects in the first place, let alone work with them and compare their properties to taus above 20 GeV. The choice to use DeepTau on MiniAOD, an intermediate, rather than one of the endpoints is curious and remains to be justified in light of the multi-tier data handling acrobatics performed in this thesis.

At the beginning of this thesis, there were many interesting questions that needed to be answered, such as "Why does CMS recommend using taus above 20 GeV?" "Are low momentum taus harder to detect and work with?" and "If yes, what is the bigger problem; reconstruction or identification?". We found that taus at low momentum are indeed harder to detect and work with for two reasons. Firstly, the background increases as a lot of soft objects are reconstructed as taus, and many of these objects do not match any generator-level object as seen in figure 5.3. Secondly, there is a substantial amount of multi-data-tier jugglery required to obtain taus at low momentum at NanoAOD, along with their DeepTau scores. The entire procedure, from obtaining the taus to obtaining their DeepTau scores, is described in 4.

The reconstruction efficiency of the HPS algorithm as well as the identification efficiency of DeepTau, is studied as a function of tau p_{T} as shown in figure 5.1. It is seen that even though both suffer at low momentum, there is a larger scope for improving identification since reconstruction efficiencies at low momentum are still high. A multi-layer neural network is trained specifically for taus at low momentum, and it is shown to improve the signal efficiency at the same value of background rejection for DeepTau extended below 20 GeV in figure 5.7. This serves as a proof of concept that taus at low momentum can be identified better with a targeted identification algorithm.

Improved identification translates to improved signal acceptance. A $H \to \tau\tau$ sample is used to quantify the additional acceptance gained by also using low momentum taus in section 7. However, this task remains to be performed along with the quantification of how much background increases when extending our selection cuts on tau momentum to 10 GeV. Increased acceptance at low momentum can enhance the yields of many standard model particles, such as the Higgs boson in the $H \to \tau\tau$ decay. It also improves our sensitivity to many models of BSM physics such as VLLs etc. In a sentence, this thesis explores the feasibility of having a generic tau identification algorithm at low momentum, and it positively turns out to be possible!

Bibliography

[1] CMS Tau POG. *Tau ID Recommendation for Run 2*. 2023. URL: https://twiki.cern.ch/twiki/bin/viewauth/CMS/TauIDRecommendationForRun2#Kinematic_tau_selection (visited on 12/13/2023).

[2] CMS Collaboration. "Inclusive nonresonant multilepton probes of new phenomena at $\sqrt{s} = 13$ TeV". In: *Phys. Rev. D* 105 (11 2022), p. 112007. DOI: 10.1103/PhysRevD.105.112007. URL: https://link.aps.org/doi/10.1103/PhysRevD.105.112007.

[3] CMS. *Workbook NanoAOD*. 2024. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookNanoAOD#Electrons (visited on 01/22/2024).

[4] CMS. *Workbook NanoAOD*. 2024. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookNanoAOD#Muons (visited on 01/22/2024).

[5] The CMS collaboration. "Identification of hadronic tau lepton decays using a deep neural network". In: *Journal of Instrumentation* 17.07 (2022), P07023. DOI: 10.1088/1748-0221/17/07/P07023. URL: https://dx.doi.org/10.1088/1748-0221/17/07/P07023.

[6] Assaf Shomer. "A Pedagogical explanation for the non-renormalizability of gravity". In: (Sept. 2007). arXiv: 0709.3555 [hep-th].

[7] Takaaki Kajita. "Discovery of neutrino oscillations". In: *Reports on Progress in Physics* 69.6 (2006), p. 1607. DOI: 10.1088/0034-4885/69/6/R01. URL: https://dx.doi.org/10.1088/0034-4885/69/6/R01.

[8] Kaustubh Agashe. *Naturalness and Supersymmetry*. 1999. arXiv: hep-ph/9908491 [hep-ph].

[9] Nilanjana Kumar and Stephen P. Martin. "Vectorlike leptons at the Large Hadron Collider". In: *Physical Review D* 92.11 (Dec. 2015). ISSN: 1550-2368. DOI: 10.1103/physrevd.92.115018. URL: http://dx.doi.org/10.1103/PhysRevD.92.115018.

[10] Prudhvi N. Bhattiprolu and Stephen P. Martin. "Prospects for vectorlike leptons at future proton-proton colliders". In: *Physical Review D* 100.1 (July 2019). ISSN: 2470-0029. DOI: 10.1103/physrevd.100.015033. URL: http://dx.doi.org/10.1103/PhysRevD.100.015033.

[11] Kazuo Fujikawa. "A Vector-Like Extension of the Standard Model". In: *Progress of Theoretical Physics* 92.6 (Dec. 1994), pp. 1149–1160. ISSN: 0033-068X. DOI: 10.1143/ptp/92.6.1149. eprint: https://academic.oup.com/ptp/article-pdf/92/6/1149/5188749/92-6-1149.pdf. URL: https://doi.org/10.1143/ptp/92.6.1149.

[12] Stephen P. Martin. "Extra vectorlike matter and the lightest Higgs scalar boson mass in low-energy supersymmetry". In: *Phys. Rev. D* 81 (3 2010), p. 035004. DOI: 10.1103/PhysRevD.81.035004. URL: https://link.aps.org/doi/10.1103/PhysRevD.81.035004.

[13] Peter W. Graham et al. "Little solution to the little hierarchy problem: A vectorlike generation". In: *Phys. Rev. D* 81 (5 2010), p. 055016. DOI: 10.1103/PhysRevD.81.055016. URL: https://link.aps.org/doi/10.1103/PhysRevD.81.055016.

[14] Kyoungchul Kong, seong chan Park, and Thomas Rizzo. "A vector-like fourth generation with a discrete symmetry from Split-UED". In: *Journal of High Energy Physics* 2010 (Apr. 2010). DOI: 10.1007/JHEP07(2010)059.

[15] ner, familyi=v., given=Ilja, giveni=I., ,, Svjetlana Fajfer, familyi=c., given=Ivana, giveni=I., ,, and. "Light vectorlike fermions in a minimal $SU(5)$ setup". In: *Phys. Rev. D* 89 (11 2014), p. 115004. DOI: 10.1103/PhysRevD.89.115004. URL: https://link.aps.org/doi/10.1103/PhysRevD.89.115004.

[16] James Halverson, Nicholas Orlofsky, and Aaron Pierce. "Vectorlike leptons as the tip of the dark matter iceberg". In: *Phys. Rev. D* 90 (1 2014), p. 015002. DOI: 10.1103/PhysRevD.90.015002. URL: https://link.aps.org/doi/10.1103/PhysRevD.90.015002.

[17] CDS CERN. *CMS Slice images*. 2018. URL: https://cds.cern.ch/record/2628641/files/?ln=en.

[18] The CMS Collaboration. "Particle-flow reconstruction and global event description with the CMS detector". In: *Journal of Instrumentation* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. URL: https://dx.doi.org/10.1088/1748-0221/12/10/P10003.

[19] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-kt jet clustering algorithm". In: *Journal of High Energy Physics* 2008.04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. URL: https://dx.doi.org/10.1088/1126-6708/2008/04/063.

[20] R. L. Workman and Others. "Review of Particle Physics". In: *PTEP* 2022 (2022), p. 083C01. DOI: 10.1093/ptep/ptac097.

[21] The CMS Collaboration. "Performance of reconstruction and identification of leptons decaying to hadrons and v in pp collisions at s=13 TeV". In: *Journal of Instrumentation* 13.10 (2018), P10005. DOI: 10.1088/1748-0221/13/10/P10005. URL: https://dx.doi.org/10.1088/1748-0221/13/10/P10005.

[22] SciPost. *The 15th International Workshop on Tau Lepton Physics (TAU2018)*. 2018. URL: https://scipost.org/SciPostPhysProc.1.

[23] CMS Twiki. *Data Tiers: Reconstructed (RECO) Data and Analysis Object Data (AOD)*. 2018. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookDataFormats#RecoData.

[24] CMS. *CMS Software Application Framework*. 2018. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSSWFramework.

[25] Rene Brun et al. *root-project/root: v6.18/02*. Version v6-18-02. June 2020. DOI: 10.5281/zenodo.3895860. URL: https://doi.org/10.5281/zenodo.3895860.

[26] CDS CERN. *CMS Run2 Data Flow*. 2018. URL: https://cds.cern.ch/record/2029414/plots.

[27] CMS. *CMSSW Software Cross-Reference*. URL: https://cmssdt.cern.ch/lxr/.

[28] CMS. *Software Guide for Tau Reconstruction*. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuidePFTauID#Rerunning_of_the_tau_ID_on_AOD_e.

[29] Jim Pivarski et al. *scikit-hep/uproot: 3.12.0*. Version 3.12.0. July 2020. DOI: 10.5281/zenodo.3952728. URL: https://doi.org/10.5281/zenodo.3952728.

[30] Medium. *Deep Dive in Recurrent Neural Networks for Binary Classification Project*. URL: https://medium.com/@jklein694/deep-dive-in-recurrent-neural-networks-for-binary-classification-project-cd15d89694da.

[31] Medium. *Understanding AUC - ROC Curve*. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[32] Steffen Schaepe. “Search for Supersymmetry in Tau Lepton Final States with the ATLAS Detector”. PhD thesis. Bonn U., 2015.

[33] Abien Fred Agarap. “Deep learning using rectified linear units (relu)”. In: *arXiv preprint arXiv:1803.08375* (2018).

[34] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[35] CMS Twiki. *Baseline muon selections for Run-II*. URL: https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideMuonIdRun2#Medium_Muon.

[36] CMS Twiki. *Cut Based Electron ID for Run 2*. URL: https://twiki.cern.ch/twiki/bin/view/CMS/CutBasedElectronIdentificationRun2#Recipe_for_regular_users_formats.