ISTANBUL TECHNICAL UNIVERSITY

FACULTY OF SCIENCE AND LETTERS

Physics Engeering Design II Report

MET Identification in TTbar events

Mert Dil

Department : Physics Engineering

ID : 090170120

Advisor : Prof. Dr. M. Altan Çakır

Fall 2022/2023

ABSTRACT

In recent years, GANs have gained increasing attention as generative models that are often superior to Variational Autoencoders and have shown very impressive results in image formation. While the Standard Model (SM) has shown great success in providing experimental predictions, it leaves some phenomena unexplained and is far from being a complete theory of fundamental interactions. It is therefore very interesting to integrate these new tools into our search for missing transverse energy. This work is dedicated to the use of a special type of Neural Network called WGAN-GP: WGAN-GP is focused on the search for missing transverse energy and kinematic distributions for charged leptons and jets. It aimed to use the method of missing transverse energy distributions to estimate and predict anomalies below the missing signal region. A correlation between Monte Carlo simulations and WGAN-GP signals was found in extracting a distribution of signals produced under missing energy.

ACKNOWLEDGEMENTS

I would like to give special thanks to my teacher Altan Ç. for supervising me and taking great care of my project. Thanks to him, I was able to work theoretically in a particle physics research lab for my engineering Design II. With this project, I learned a lot about particles, from how to analyze particle data from the Large Hadron Collider with machine learning algorithms. Every part of this project was fun and enlightening. In addition, I had taken many courses from him before. In each of them, he has greatly influenced the way I think and work, but most importantly, I met him in my first course, which laid the foundations for my graduate plans at this university. I thank him for awakening my interest and curiosity in physics. It is with joy that I complete my physics degree.

Mert Dil

CONTENTS

ABBREVIATIONS

$\mathbb{R}$	A symbol for the set of real numbers
η	Greek letter minuscule eta
ϕ	Greek letter minuscule phi
θ	Greek letter minuscule theta
E_T	Transverse Energy
GeV	Gigaelectron Volt
p_T	Transverse momentum
TeV	Teraelectron Volt
W(P_d, P_g)	Earth Mover Distance

INTRODUCTION

Large Hadron Collider (LHC) with CMS experiments has opened a new chapter in particle physics. A new window of opportunity opens for the search of direct evidence of physics beyond the Standard Model (SM) in a field that appears to be. The use of semi-supervision in Machine Learning (ML) for the search of physics beyond the SM has gained momentum in recent years. Machine Learning algorithms are trained on control samples that contain information regarding Missing Transverse Energy and in data samples that may contain new physics in addition to Missing Transverse Energy. Several anomalies have been identified in the production of leptons at the ATLAS and CMS experiments of the Large Hadron Collider(LHC). To date, these anomalies remain unexplained by MC tools that are becoming more and more accurate, while the effects are statistically compelling.[1].

The high accuracy of state-of-the-art Monte Carlo (MC) simulation software, typically based on the DELPHES3, has a high cost: MC simulation amounts to about one-half of the experiment computing budget and a large fraction of the available storage resources, the other half being largely used to process simulated and real data (event reconstruction). All these studies formalized the simulation task in terms of either image generation or analysis of specific high-level kinematic features.[2]

Dark Matter is one, of the several, phenomena that the SM fails to explain. On the other hand, missing transverse energy, an extension to the SM, is a good candidate to solve this one, and possibly other, yet unexplained phenomena. Missing transverse energy (MET) or a jet is used to probe possible dark matter (DM) production. Since DM travels through the Compact Muon Solenoid detector (CMS), it appears as MET in the resulting collision fragments. We study the dilepton final states. Searches are being conducted at possibly the largest data generation machine ever built, the LHC. With over tens of petabytes of data generated every year, different data analysis techniques are employed such as Wasserstein Generative Adversarial Network Gradient Penalty (WGAN-GP). On the other hand, Machine Learning offers novel data analysis tools, such as Neural Networks, which provide state-of-the-art performance in data analysis with increasing data with Monte Carlo simulations. Thus, it is very interesting to integrate these novel tools in our missing transverse energy searches. This work is dedicated to the use of a particular type of Neural Network called: WGAN-GP, in the search for missing transverse energy and kinematic distributions for charged leptons and jets. This work aims to use

the missing transverse energy distributions method to extrapolate and predict the anomalies below the missing signal region. To do so, two uncorrelated variables in the analysis are needed, and hence; a generative adversarial network is developed to decorrelate one of the physics variables from the network's output. We present WGAN-GP, an event particle based generative model that can be used to emulate MET identification simulation at the LHC.

The organization of this work is as follows. Section 2 presents some fundamental concepts including a brief introduction to Neural Networks. Sections 3 and section 4,5 present the methods and conclusions respectively.

FUNDEMENTAL CONCEPTS

2.1 Standard Model and Beyond Standard Model

All matter around us is made of elementary particles, the building blocks of matter. These particles occur in two basic types called quarks and leptons. Each group consists of six particles, which are related in pairs, or generations. The lightest and most stable particles make up the first generation, whereas the heavier and less stable particles belong to the second and third generations. All stable matter in the universe is made from particles that belong to the first generation; any heavier particles quickly decay to the next most stable level. The six quarks are paired in the three generations – the "up quark" and the "down quark" form the first generation, followed by the "charm quark" and "strange quark", then the "top quark" and "bottom (or beauty) quark". The six leptons are similarly arranged in three generations – the "electron" and the "electron neutrino", the "muon" and the "muon neutrino", and the "tau" and the "tau neutrino". The electron, the muon, and the tau all have an electric charge and a sizeable mass, whereas the neutrinos are electrically neutral and have very little mass. There are four fundamental forces at work in the universe: the strong force, the weak force, the electromagnetic force, and the gravitational force. Gravity is the weakest but it has an infinite range. Electromagnetic force also has an infinite range but it is many times stronger than gravity. The weak and strong forces are effective only over a very short range and dominate only at the level of subatomic particles. Despite its name, the weak force is much stronger than gravity but it is indeed the weakest of the other three. The strong force, as the name suggests, is the strongest of all four fundamental interactions.

Three of the fundamental forces result from the exchange of force-carrier particles, which belong to a broader group called "bosons". Particles of matter transfer

discrete amounts of energy by exchanging bosons with each other. Each fundamental force has its corresponding boson – the strong force is carried by the "gluon", the electromagnetic force is carried by the "photon", and the "W and Z bosons" are responsible for the weak force. Although not yet found, the "graviton" should be the corresponding force-carrying particle of gravity. Even though the Standard Model is currently the best description there is of the subatomic world, it does not explain the complete picture. The standard model falls short of being a theory of everything. These include the full theory of gravitation described by general relativity in all its details, the accelerating expansion of the universe (possibly as described by dark energy), and the dark matter particle with all the properties observed in observational cosmology.

BSM The Standard Model has many known limitations, the effective field theory describing interactions close to the TeV scale, new particles or dynamics that might arise in proton-proton collisions at the Large Hadron Collider (LHC), most of the available phase space for natural solutions to outstanding problems is excluded. If there is new physics, it is either heavy (i.e. beyond the reach of current research) or hidden (i.e. currently indistinguishable from standard model backgrounds). In a supersymmetric theory, the equations of force and the equations of matter are identical. In supersymmetry, every particle in one class will have an associated particle in the other class whose spin is half an integer number different, known as a superpartner. Strong theoretical advances are needed to understand the nature of the dark matter particle, which the standard model cannot fully explain.

New physics provides the opportunity to search for direct evidence. Anomalies in lepton production have been detected in the LHC's ATLAS and CMS experiments, including several final states: di-leptons with opposite sign, di-leptons with the same sign, and three leptons in the presence and absence of b-quarks. These final states can be observed at the corners of the phase space where different SM processes dominate. The anomalies cannot be explained by Monte Carlo (MC) tools, which have become more sensitive in recent years. The focus is on non-resonant searches in signatures. Generative modeling is considered to mimic the relevant separation of observables. In this research, GANs are used to mimic MC estimates.

2.2 Production of top quark

Events containing one or more top quarks produced with additional prompt leptons are used to search for new physics. In modern collider experiments at high energy,ttbar events are among the most copious signatures observed in the detectors.

When one top quark decays leptonically and the other hadronically, the signature is characterized by one lepton, missing transverse energy, and four jets, two of the originate from the fragmentation of b quarks. Moreover, at the LHC, about 50% of events have extra hard jets coming from initial or final state radiation. we focus on the mass of the hadronically-decaying top quark(In shown in figure 2.1). Their construction has been performed via Delphes using the detector configuration designed to mimic the performance of the CMS detector. As a particle produced in abundance at the LHC, top quarks are produced as top quark pairs. The most common is the production of a top–antitop pair via strong interactions. In a collision, a highly energetic gluon is created, which subsequently decays into a top and antitop.

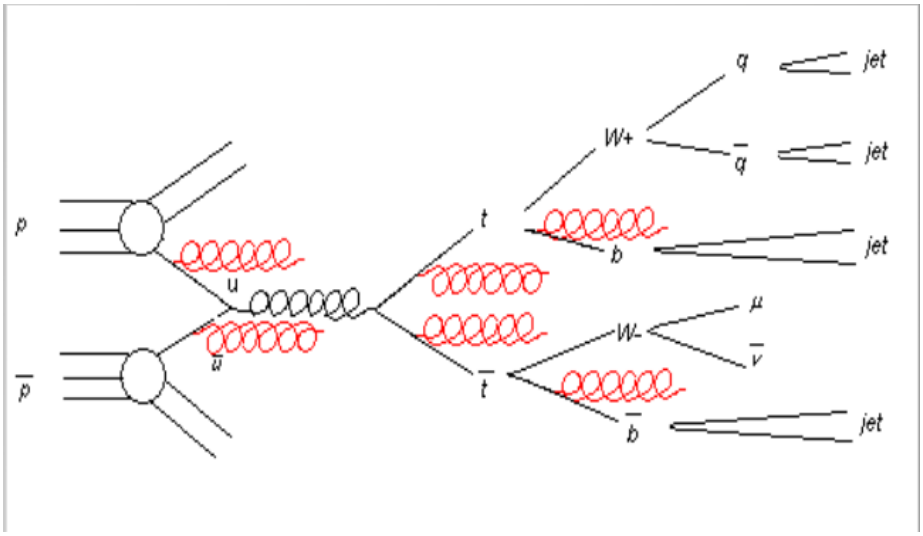


Figure 2.1: Top quark production.

2.3 Kinematics of elementary particles

The kinematics features are important parameters for classifying particles and defining certain properties. Kinematic variables have been playing an important role in collider phenomenology, as they expedite discoveries of new particles by separating signal events from unwanted background events and allow for measurements of particle properties such as masses, couplings, spins, etc. For the past 10 years, an enormous number of kinematic variables have been designed and proposed, primarily for the experiments at the Large Hadron Collider, allowing for a drastic reduction of high-dimensional experimental data to lower-dimensional observables, from which one can readily extract underlying features of phase space and develop better-optimized data-analysis strategies. As machine learning is nowadays percolating through many fields of particle physics including collider phenomenology, discussed

the interconnection and mutual complementarity of kinematic variables and machine learning techniques. discussed how the utilization of kinematic variables originally developed for colliders can be extended to other high-energy physics experiments including neutrino experiments[3]

In experimental particle physics, pseudorapidity, η, is a commonly used spatial coordinate describing the angle of a particle relative to the beam axis. It is defined as Eq.2.1

$\eta = 1/2.\ln(\tan(\theta/2),$ (2.1)

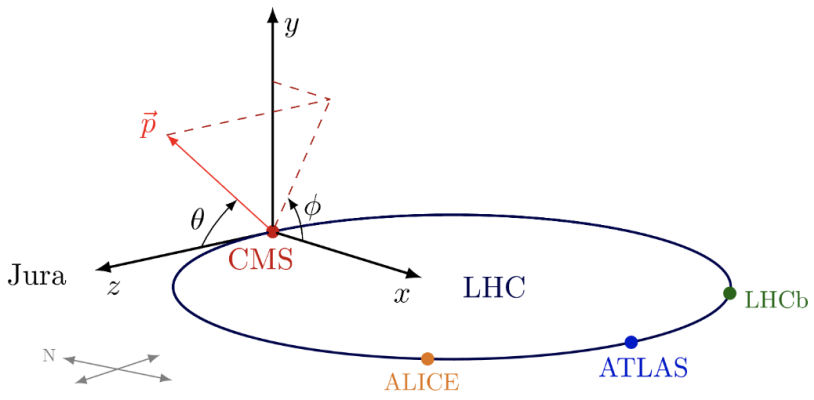


Figure 2.2: Experimental Cordinate System.

To understand Transvers Momentum, η, and ϕ about particle physics, one must first understand that the Compact Muon Solenoid (CMS) detector is a component of CERN's Large Hadron Collider (LHC). This is cylindrical, which is important to know to understand all of the components used to understand Transversal Momentum, Eta, and Phi. If one were to take a section of the CMS detector to analyze a collision of particles, it could be divided into two planes: the "transverse" xy-plane and the -plane. The xy-plane is either the bottom or the top of a cylinder, it depends on which way one chooses to view a collision. The x and y axes are the same as any average graph with the x being the horizontal axis and the y the vertical axis. The xz-plane includes the z-axis or beam-axis, which is the path the particles that collided followed through the cylinder, and the x-axis, which connects both planes. For the xz-plane, the x-axis is the horizontal axis and the z-axis is the vertical axis

and is perpendicular to the y-axis of the xy-plane. Transversal Momentum and Phi (also known as the Azimuthal or Scattering Angle and is measured in radians).At hadron colliders, a significant and unknown fraction of the beam energy in each event escapes down the beam pipe. Net momentum can only be constrained in the plane transverse to the beam z-axis!(In shown in figure 2.2).

$$\sum p_t(i) = 0 \quad (2.2)$$

Eq.2.2 is the net momentum, which must be zero in total ,

$$| p | = p_T . \cosh \eta \quad (2.3)$$

$| p |$is magnitude of momentum,

$$\begin{aligned} \mathrm{p}_x &= p_T \cos \varphi \\ p_y &= p_T \sin \varphi \\ p_z &= p_T \sin \eta \quad (2.4) \end{aligned}$$

p_x, p_y, p_zcomponents of momentum,

$$\mathrm{p}_T = \sqrt{p_x^2 + p_y^2} \quad (2.5)$$

p_T is transverse momentum with 2 vectors

$$\mathrm{E}_T = \frac{E}{\cosh \eta} \quad (2.6)$$

E_T is value of transverse energy

Accelerators at CERN boost particles to high energies before they are made to

collide inside detectors. The detectors gather clues about the particles – including their speed, mass, and charge – from which physicists can work out a particle's identity. The process requires accelerators, powerful electromagnets, and layer upon layer of complex subdetectors.Particles produced in collisions normally travel in straight lines, but in the presence of a magnetic field, their paths become curved. Electromagnets around particle detectors generate magnetic fields to exploit this effect. Physicists can calculate the momentum of a particle – a clue to its identity – from the curvature of its path: particles with high momentum travel in almost straight lines, whereas those with very low momentum move forward in tight spirals inside the detector. Tracking devices:Tracking devices reveal the paths of electrically charged particles as they pass through and interact with suitable substances. Most tracking devices do not make particle tracks directly visible but record tiny electrical signals that particles trigger as they move through the device. A computer program then reconstructs the recorded patterns of tracks. One type of particle, the muon, interacts very little with matter – it can travel through meters of dense material before being stopped. Muons, therefore, pass easily through the inner layers of a detector, which is why muon chambers – tracking devices specialized in detecting muons – usually make up the outermost layer of a detector. Calorimeters:A calorimeter measures the energy a particle loses as it passes through. It is usually designed to stop entirely or "absorb" most of the particles coming from a collision, forcing them to deposit all of their energy within the detector, thus measuring their full energy. Calorimeters have to perform two different tasks at the same time – stopping particles and measuring energy loss – so they usually consist of layers of different materials: a "passive" or "absorbing" high-density material – for example, lead – interleaved with an "active" medium such as plastic scintillators or liquid argon. Electromagnetic calorimeters measure the energy of electrons and photons as they interact with the electrically charged particles in matter. Hadronic calorimeters sample the energy of hadrons (particles containing quarks, such as protons and neutrons) as they interact with atomic nuclei. Calorimeters can stop most known particles except muons and neutrinos.

Particle-identification detectors: In addition to measuring a particle's momentum in tracking devices and its energy in calorimeters, physicists have further methods of narrowing down its identity. These methods all rely on measuring a particle's velocity, since this, in combination with the momentum measured in the tracking devices, helps to calculate a particle's mass and therefore its identity. Velocity can be measured using several methods. The simplest is to measure how much time it

takes for a particle to travel a certain distance, using precise time-of-flight detectors. Another method looks at how much a particle ionizes the matter that it passes through, as this is velocity-dependent and can be measured by tracking devices. If a charged particle travels faster than light through a given medium, it emits Cherenkov radiation at an angle that depends on its velocity. Alternatively, when a particle crosses the boundary between two electrical insulators with different resistances to electric currents, it emits transition radiation, the energy of which depends on the particle's velocity. Collating all these clues from different parts of the detector, physicists build up a snapshot of what was in the detector at the moment of a collision. The next step is to scour the collisions for unusual particles, or for results that do not fit current theories.

2.4 MET with ttbar events

In experimental particle physics, missing energy refers to the energy that is not detected in a particle detector, but is expected due to the laws of conservation of energy and conservation of momentum. Missing energy is carried by particles that do not interact with electromagnetic or strong forces and thus are not easily detectable most notable neutrinos.

Many beyond the standard model scenarios, including supersymmetry, predict events with large MET. The reconstruction of MET is very sensitive to particle momentum mismeasurements, particle misidentification, detector malfunctions, particles impinging on poorly instrumented regions of the detector, cosmic-ray particles, and beam-halo particles, which may result in artificial MET. Understanding the nature of dark matter (DM) is the focus of extensive research at collider- and astrophysics-based experiments. The most well-known signature for DM production at the LHC is the so-called "mono-X" topology, for which events are characterized by the presence of a high-momentum object (e.g. a jet in the case of a mono-jet signature) from initial-state radiation in combination with significant missing transverse energy (MET). If supersymmetric particles exist, they are very likely to be produced in collisions in the LHC. The heavy particles will decay into combinations of leptons (like electrons and muons) and quarks (which will cause sprays of particles called jets) as well as into neutralinos that will not decay any further. Therefore many neutralinos will pass through the CMS detector, without depositing any energy or leaving a trail.

The experimental challenge lies in the fact that in proton collisions, many other Standard Model particles with similar final decay products mimic the signal. These

collisions are called background. The most important backgrounds arise from collisions with a pair of W bosons in the final state, such as the Standard Model Higgs boson, the non-resonant Standard Model di-boson W+W- and the top quark pairs (almost 99% of the time each top quark decays to a W boson and a bottom quark). We have to dig deep inside the collected data to try to identify which events could have Dark Matter particles in them. We can only claim the observation of dark matter if there is a significant amount of signal events after all such backgrounds are removed.

Experimentally, interesting top quark pair collisions are selected by searching for the specific decay products of a top quark-antiquark pair. In the overwhelming majority of cases, top quarks decay into an energetic jet and a W boson, which in turn can decay into a lepton and a neutrino. Jets and leptons can be identified and measured with high precision by the CMS detector, while neutrinos escape undetected and reveal themselves as missing energy. The measured MET scale agrees with the expectations of the detector simulation, but the resolution is degraded by 10% in data. CMS has three different algorithms for calculating MET. Algorithms using tracker information have an improved resolution, and the use of a global particle-flow event reconstruction gives the best resolution [4].

We can classify the difficulties of experimental detection of missing energy in several ways: Energy Resolution, Multi-jet events Resolution, Electronic Noise, Clutter and Underlying Events, MET tails, High Magnetic field, Energy loss due to penetrations, Faulty calorimeter cells. In this study, we used machine learning algorithms to find potential signals because they were below the MET

2.5 Monte Carlo simulations

Monte Carlo simulations are conventionally used to setup the event selection while the real data will be revealed afterward independently MC events can be used to estimate the corresponding trials factor through a frequentist inference. However, MC events that are based on full detector simulations are resource intensive. The probability of false signals needs to be carefully estimated based on toy Monte Carlo (MC) studies. Unfortunately, MC samples based on a full detector simulation are CPU expensive.

The ATLAS and CMS experiments rely on Monte Carlo (MC) software for the simulation of events at the LHC as per the searches for new resonances. Therefore a need to address what can be referred to as the inverse problem in particle physics is of the essence, which would be addressing the possibility of whether the extraction of

information to build a new theory from the data is feasible. Several anomalies have been identified in the production of leptons at the ATLAS and CMS experiments of the LHC. These include several final states.

These final states appear in corners of the phase space where different SM processes dominate. To date, these anomalies remain unexplained by MC tools that are becoming more and more accurate, while the effects are statistically compelling. Machine learning can play a significant role to explore a deeper phase-space available at the LHC, where model dependence of the signal needs to be significantly reduced.

This can be achieved through the use of semi-supervision resting based on mixing samples. Side bands, signal-depleted corners of the phase space, and signal-enriched samples are defined. These are confronted with each other as two distinct samples, where the ML algorithm performs a classification task. The data in each of the samples is unlabelled, where prior knowledge of signal modeling is not necessary. The side-band brings insights from the SM background, either in the form of real data or simulated data, depending on the level of realism that the MC displays. An agreement between the MC and the WGAN-GP generated events is searched for the observables selected in the study

2.6 Machine learning with supervised learning

Supervised learning no prior knowledge of this underlying density is needed nor desired. Full supervision can often be aimed at solving a problem of classification. This is performed through generative or discriminative algorithms. Generative algorithms model how the data is generated, where the conditional probability P(x|y) is inferred[5]. By contrast, discriminative algorithms do not deal with how x is generated, but rather concentrate on modeling P(y|x) Logistic regressions are commonly used to achieve this goal. Semi-supervision is a hybrid where there are elements of both supervised and unsupervised learning. Labeled data is provided but not for all data sets. In this setup. For searchers for new physics at the LHC, semi-supervision can be applied, where the labeled sample corresponds to the background. The unlabelled sample would contain background in addition to an unknown admixture of signal from new BSM physics.

While semi-supervision, as set up here, has the advantage of not relying on a model for the signal it is necessary to constrain the phase space where the side band and the signal region are confronted with each other. While signature and topological requirements are driven by physics considerations, the search is not biased by the phenomenology of a model with a particular set of parameters. Situations where two

more corners of the phase space with different admixtures of the BSM signal. In any of the cases considered, no prior knowledge of the yield nor the model dependence of the BSM signal is required. In practice, as the task in hand is to efficiently classify between SM backgrounds and new BSM physics, it is essential that a good understanding of background modeling be provided through control samples where BSM signals are expected to be negligible. Machine learning can play a significant role to explore a deeper phase-space available at the LHC, where model dependence of the signal needs to be significantly reduced. This can be achieved through the use of semi-supervision resting based on mixing samples. Side bands, signal-depleted corners of the phase space, and signal-enriched samples are defined. The side-band brings insights from the SM background, either in the form of real data or simulated data, depending on the level of realism that the MC displays.[6]

Monte Carlo (MC) software for the simulation of events at the LHC as per the searches for new resonances. An inverse problem in particle physics is of the essence, which would be addressing the possibility of whether the extraction of information to build a new theory from the data is feasible. For this purpose, particle physicists went beyond the classical methods and introduced machine learning (ML) techniques, a component of artificial intelligence (AI), as effective aspects for analyzing complex and big data, with the hope of eliminating human intervention. With the previously celebrated successes in ML, specifically Deep Neural Networks (DNN), an unsupervised learning technique called Generative-Adversarial Networks (GANs) can synthesize fake-looking samples based on sets of un-labelled training examples. In practice, GANs are mostly used to generate photo-realistic images, medical imaging, and recently with applications in high energy physics for simulating detector responses. There are multiple black-box deep learning-based approaches to generative modeling such as Variational Autoencoders (VAE),Mixture Density Networks (MDN), and Generative Adversarial Networks (GAN)that can be used to reproduce the kinematic distributions that are obtained from MC simulations.These models have been considered complimentary to the already work in progress using GANs since they are taken as extensional building blocks to a GAN.

METHODS

3.1 Data preparation and processing

Delphes 3.0 simulation was used in this study. To preserve the tree structure of our data and to make it faster and more useful, we used the Root format and performed our analyses with Uproot. We ran programs called MadGraph and Pythia,

which emulate the hard scattering, display, and hadronization process. We created the pp$\Longrightarrow$LEPTONS, NEUTRINOS process in response to the detector with Delphes and continued our analyses . jets were used in the kinematic selection. pp collisions with MET origin were selected. Our top mass m_T=172.5 GeV was selected. Events was processed over 100000.

MadGraph5_aMC@NLO is a framework that aims at providing all the elements necessary for SM and BSM phenomenology, such as the computations of cross sections, the generation of hard events and their matching with event generators, and the use of a variety of tools relevant to event manipulation and analysis

PYTHIA is a program for the generation of high-energy physics events, i.e. for the description of collisions at high energies between elementary particles such as e^+, e^-, p and $\overline{p}$ in various combinations. This also includes heavy-ion collisions.

Delphes fast simulation version 3.0 was used. The purpose of Delphes is to allow the simulation of a multipurpose detector for phenomenological studies. The simulation includes a track propagation system embedded in the magnetic field, electromagnetic and hadron calorimeters, and a muon identification system. Physics objects that can be used for data analysis are then reconstructed from the simulated detector response. While some aspects of Delphes are specific to hadron colliders, it is flexible enough to be adapted to the needs of electron-positron collider experiments. Typical workflow diagram of a Delphes fast simulation. Event files from external Monte-Carlo generators are processed by a reader stage. Delphes provide leptons (electrons and muons), photons, and missing transverse energy. Accumulation includes charged particle propagation in a magnetic field ("tracking"), electromagnetic and hadronic calorimeters, and muon systems.

ROOT is a data processing framework born at CERN, at the heart of high-energy physics research. Every day, thousands of physicists use ROOT applications to analyze their data or perform simulations.

3.2 Modelling Approach

3.2.1 Related GAN work in HEP

The technique is first introduced by, as a class of unsupervised generative models that are deployed in adversarial settings of two network blocks that implicitly learn the underlying probability distribution of a given data set. After training, a GAN model provides a mechanism that helps to efficiently sample from the learned distri-

bution with no need for an explicit probability density function. From the original inception of GANs, a generator network (G_Q) and discriminator network (D_Q) are pitted against each other over the parameters θ and ϕ respectively. Here the generator takes input samples from a prior distribution $P_z(z)$ over some pre-defined latent variable $z \in \mathbb{R}$, where this is usually set as a uniform or normal distribution which it can serve as a source of variation for G. G learns a distribution $P_g(x)$ from the given training dataset $x \in \mathbb{R}^n$ that approximates the real data-generated distribution $P_d(x)$ with

$m \times n$. G(z) is therefore defining a mapping from P_z to the data space. D has the task of correctly distinguishing between real samples drawn from P_d and synthesized samples produced by G. The inputs to D can either be real or synthesized samples at a time, while its outputs a scalar value between 0 and 1 which are a representation of the probability of the inputs coming from P_d. In this sense, this constitutes a competitive two-player minimax game with players being D and G with the optimal solution being a Nash-equilibrium. In a case where G synthesizes samples that are the same as the real data distribution [5] (i.e. P_g=P_d) that point D becomes maximally confused and thus D(x) = D(G(z)) = 0.5.

The training objective of a vanilla GAN is given by Eq.3.1:

$$V_{\text{GAN}} = \min_{\theta} \max_{\phi} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_d(\mathbf{x})}\big[\log D_\phi(\mathbf{x})\big] + \mathbb{E}_{\mathbf{z}\sim\mathbb{P}_z}\big[\log(1 - D_\phi(G_\theta(\mathbf{z})))\big] \tag{3.1}$$

A WGAN differs from the vanilla-GAN in that it minimizes the Earth-Mover distance (also known as the Wasserstein distance) W(P_d, P_g) as an alternative distance measure for training a GAN in shown Eq.3.2. Informally it can be interpreted as the minimum amount of energy required to transform one probability mass over a distance to transform a distribution P_g into a target distribution P_d ;

$$W(\mathbb{P}_d, \mathbb{P}_g) = \inf_{\gamma\in\Pi(\mathbb{P}_d,\mathbb{P}_g)} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}\Big[\|x - y\|\Big] \tag{3.2}$$

The value function of the WGAN now becomesin shown in Eq.3.3:

$$V_{\text{WGAN}} = \min_{\theta} \max_{\phi \in \Omega} \left[\mathbb{E}_{\mathbf{x} \sim \mathbb{P}_d(\mathbf{x})} \left[D_\phi(\mathbf{x}) \right] - \mathbb{E}_{\mathbf{z} \sim \mathbb{P}_\mathbf{z}} \left[D_\phi(G_\theta(\mathbf{z})) \right] \right] \tag{3.3}$$

3.2.2 WGAN-GP architecture

The works on GANs became an appealing method within HEP, there are several applications ranging from images to non-image data applications. The earlier works within HEP are strictly on images, and that was short-lived since that part could not address some critical questions. One of the earliest non-image data GAN within HEP was around the years of 2018 for an unfolding task. The most dominating work and currently active area of research are mainly focused on Parton showers and event generation by Generative models in HEP are also explored in the modeling of muon interactions with dense targets, phase space integration, event subtraction, and unweighting.[7]

In this formalism D no longer plays a role of a classifier but rather a regressor, hence referred to as a critic. It is tasked to learn a function that approximates W(P_d, P_g).[5] Unlike in a vanilla-GAN, the WGAN has a critic that does not suffer from vanishing gradients. This is a major improvement since it supports cases where the distributions do not overlap. With the Minimizing Equation, concerning the parameter Q minimizes the learning function W(P_d, P_g). However, made a further improvement by proposing an implementation that constrains by adding a gradient penalty (GP) term to the critic's objective which penalizes the objective whenever the norm of the critic's gradients exceeds 1.

This further improves the value function (in shown in Eq.3.4) by adding a penalty term,making the function to be:

$$V_{\text{WGAN-GP}} = V_{\text{WGAN}} + \lambda \mathbb{E}_{\hat{\mathbf{x}} \sim \mathbf{P_d}(\hat{\mathbf{x}})} \left[(\|\nabla_{\hat{\mathbf{x}}} D_\phi(\hat{\mathbf{x}})\|_2 - 1)^2 \right] \tag{3.4}$$

RESULTS

4.1 MC Samples with Kinematic Variables

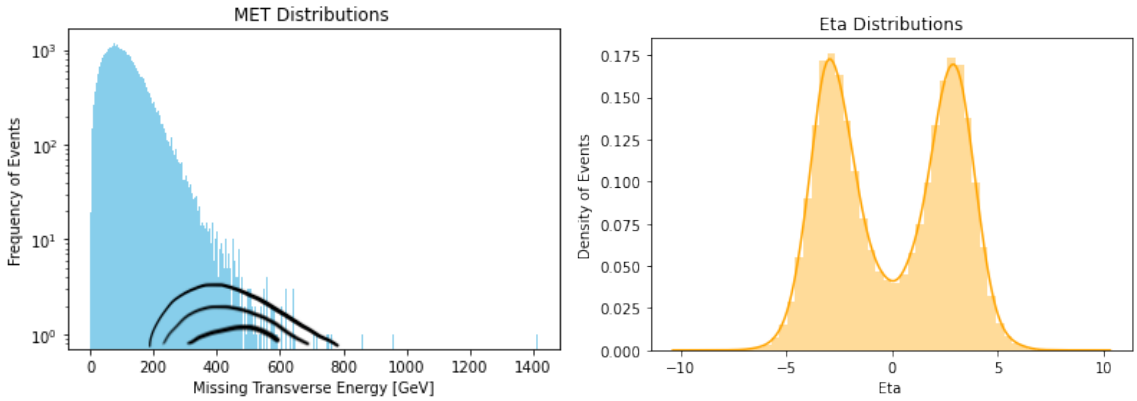
Kinematic variables have been playing an important role in collider phenomenology, as they expedite discoveries of new particles by separating signal events from unwanted background events and allow for measurements of particle properties such as masses, couplings, spins.

For the past 10 years, an enormous number of kinematic variables have been designed and proposed, primarily for the experiments at the Large Hadron Collider, allowing for a drastic reduction of high-dimensional experimental data to lower-dimensional observables, from which one can readily extract underlying features of phase space and develop better-optimized data-analysis strategies.

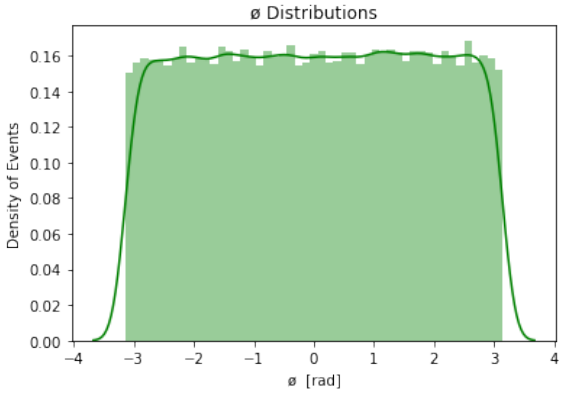
In this study, kinematic values, which are important in particle physics, were obtained by MC simulations. MET data are shown and anomalies are deepened through MET tails.

4.1.1 Missing Transverse Energy

In the last 10 years, a large number of kinematic variables have been designed and proposed, especially for experiments at the Large Hadron Collider, allowing a drastic reduction of high-dimensional experimental data to lower-dimensional observational variables that can easily extract key features (η,ϕ) of the phase space and develop better-optimized data analysis strategies(In shown figure 4.1 (b),(c)).We thought there might be anomalies where there might be a shift based on the presence of signals, as shown in the figure. We tried to deepen the anomalies on the Met tails because some anomaly signals are lost in the graph because they are below the missing energy values and we used generative algorithms to reconstruct these missing signals.(In shown in figure 4.1 (a))



(a) Missing Transverse Energy **(b)** Missing Transverse Eta distribution



(c) Missing Transverse Phi distribution

Figure 4.1: Monte Carlo Simulations For MET

4.1.2 Charged Leptons-Electron

An electron is a subatomic particle with a negative elementary electric charge.Electrons belong to the first generation of the lepton particle family.Since they have no known components or substructures, they are generally considered to be elementary particles. Monte Carlo simulations are simulated for testing with observable data, as shown in the figures 6.1.

4.1.3 Charged Leptons-Muon

A muon is an elementary particle similar to an electron, with an electric charge of -1 e and a spin of 1/2, but with a much larger mass. It is classified as a lepton. As with other leptons, the muon is not thought to be composed of simpler particles; that is, it is a fundamental particle. Like the Electron, the Muon has been used in observations because it is a charged lepton. It was simulated to test its compatibility

with real data in figures 6.2

4.1.4 Jet

A jet is a narrow cone of hadrons and other particles produced by the hadronization of a quark or gluon in a particle physics or heavy ion experiment. Particles that carry a color charge, such as quarks, cannot exist in a free state due to the limitation of quantum chromodynamics (QCD), which allows only colorless states. We separated the processing of Jet data from Charged Leptons and therefore created by MC simulations in shown figure 6.3.[8]

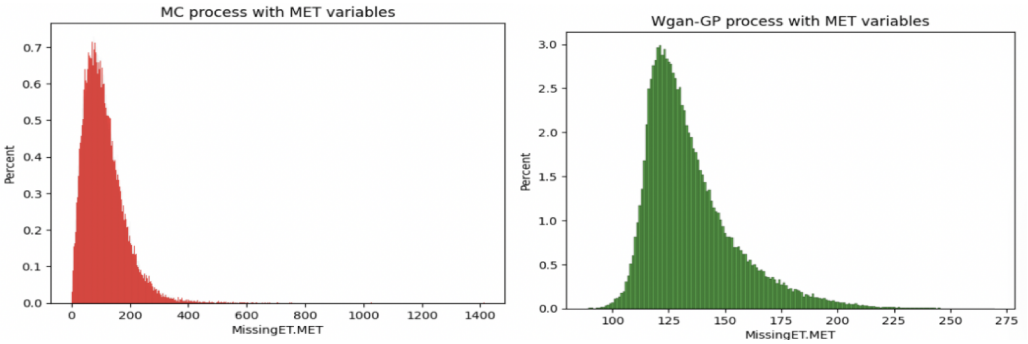
4.1.5 Scalar sum of all input particles momenta(H_T)

H_T is used as input to the final prediction. $H_\perp$, the scalar sum of the transverse moments of the charged lepton and jet. It is not expected to be very sensitive to details in the coupling schemes. in general it is used because the change in the overall sum is an important indicator in terms of control in shown in 6.4

4.2 WGAN-GP training

4.2.1 Missing Transverse Energy

The anomaly signals that disappeared under the text were tried to be reproduced with the generative algorithm WGAN-GP as shown in the figure 4.2. The density differences figure 4.3 here are related to the learning width of the algorithm. Higher training was tried but was not attempted because it was too time-consuming. The graphs peak at a close average value. Future work is being done on density variation to achieve better correlation in figure 6.6 and figure 6.8



(a) MC process with MET **(b)** WGAN-GP process with MET

Figure 4.2: Comparisons with MET between MC and WGAN-GP process)

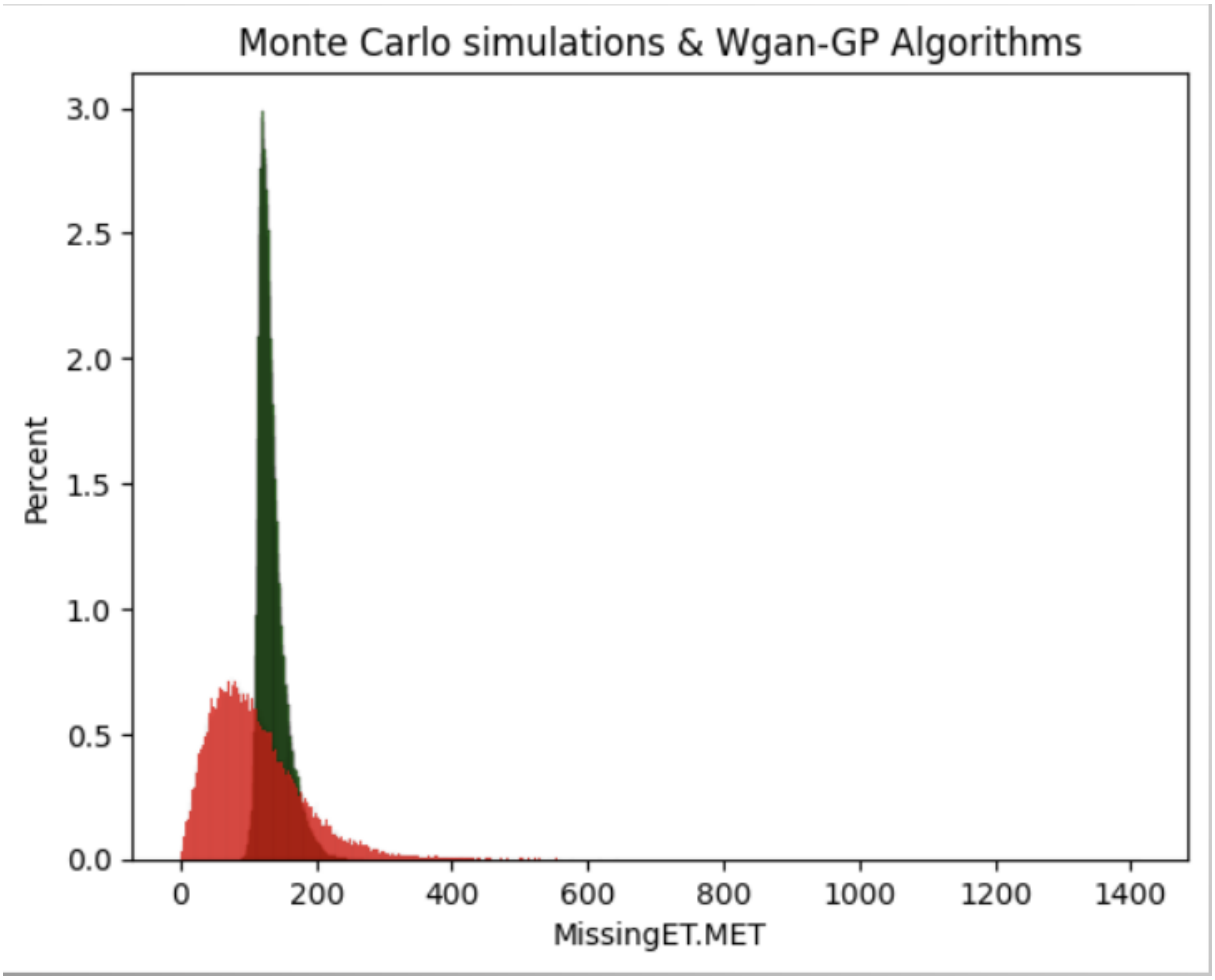


Figure 4.3: Monte Carlo Simulations and WGAN-GP training

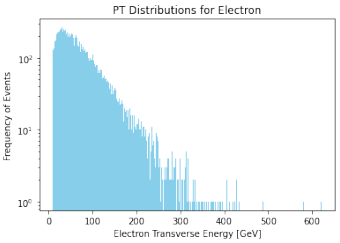
CONCLUSION

GANs are one of the most preferred semi-supervised methods, but it is not easy to generalize each setup to other tasks, each architecture tends to solve only a specific problem. Studies have been done between events generated by MC and WGAN-GP, which showed a density difference in Fig.4.3 on the two graphs due to the difficult integration of the GAN structure concerning events. The next step is to further strengthen the convergence by studying WGAN-GP [9] . An attempt is being made on Generative Adversarial Networks such as (GAN). Training time is a major constraint we face with the current setup of Keras with the Tensorflow backend, so it is appropriate that we move to PyTorch. PyTorch is an excellent alternative that can significantly increase our research productivity while scaling on GPUs, making it easier to execute new research ideas. It can reduce training iteration time in generative modeling from weeks to days.

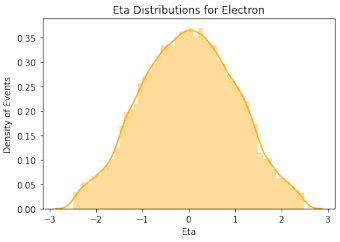
Optimizing the reconstruction of Wgan-GP and PT tails and MET tails. Creating the difference matrix from WGAN-GP and MC simulations will be used in future studies.

APPENDIX

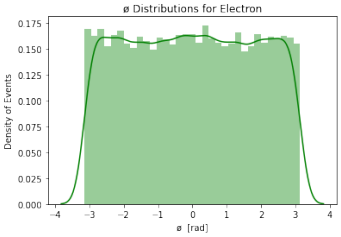
6.1 Monte Carlo Simulation Histograms



(a) Electron p_T

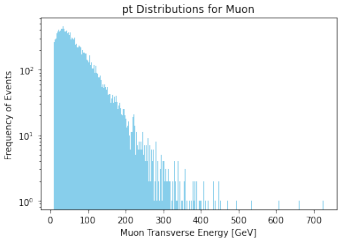


(b) Electron η distribution

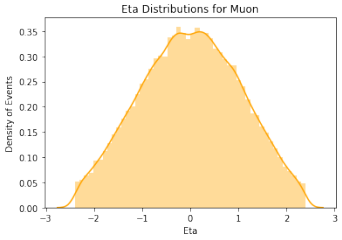


(c) Electron ϕ distribution

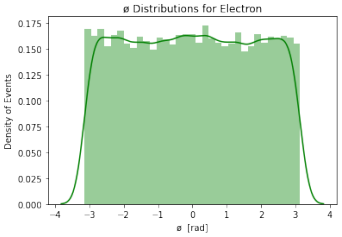
Figure 6.1: Monte Carlo Simulations For Electron



(a) Muon p_T

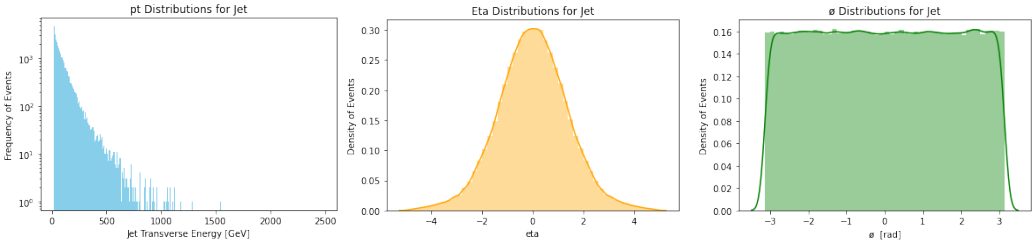


(b) Muon η distribution



(c) Muon ϕ distribution

Figure 6.2: Monte Carlo Simulations For Muon



(a) Jet p_T **(b)** Jet η distribution **(c)** Jet ϕ distribution

Figure 6.3: Monte Carlo Simulations For Jet

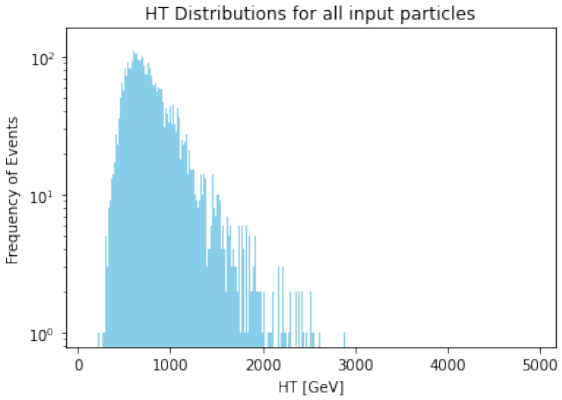


Figure 6.4: H_T distributions

6.2 WGAN-GP Training Histograms

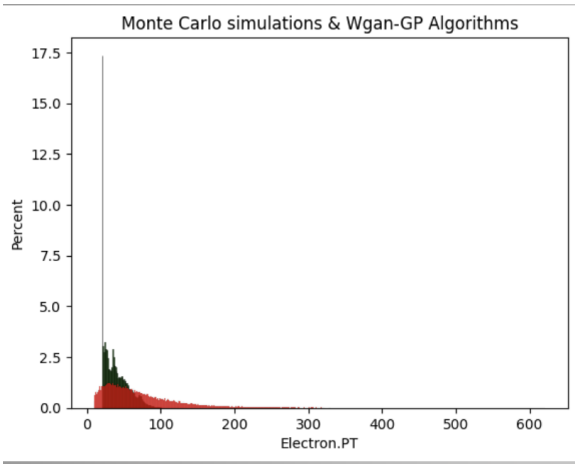
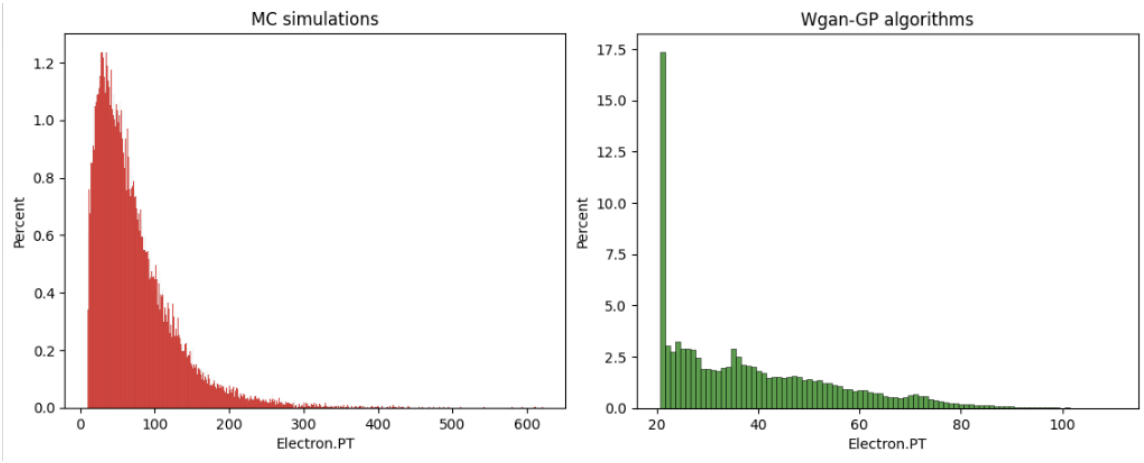
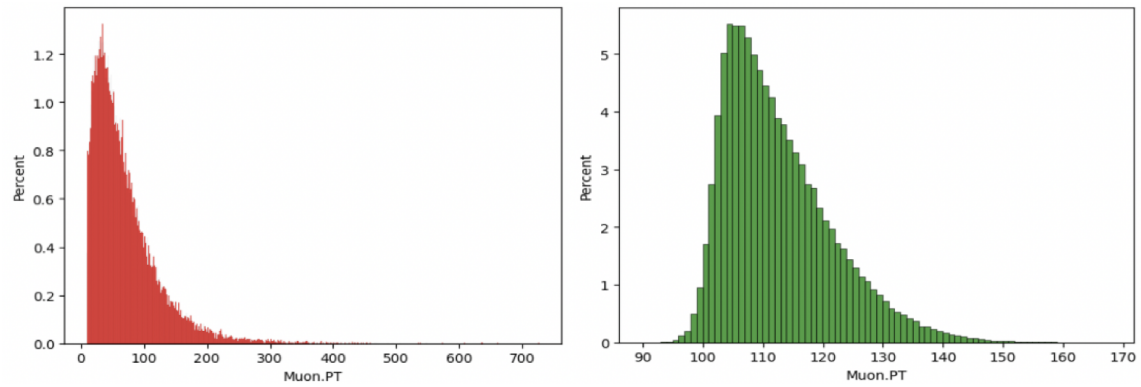


Figure 6.6: Comparisons density of Electron p_T between MC and WGAN-GP process



(a) MC process with Electron p_T **(b)** WGAN-GP samples with electron p_T

Figure 6.5: Comparisons with Electron p_T between MC and WGAN-GP process



(a) MC process with Muon p_T **(b)** WGAN-GP process with Muon p_T

Figure 6.7: Comparisons with Muon p_T between MC and WGAN-GP process

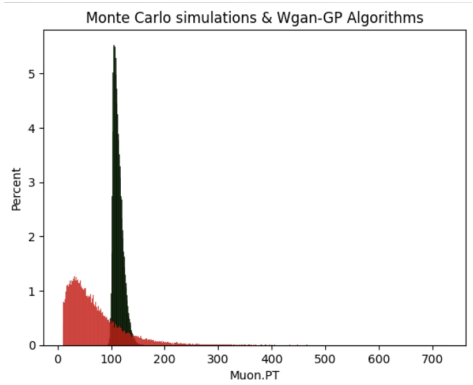


Figure 6.8: Comparisons density of Muon p_T between MC and WGAN-GP process

Bibliography

[1] Mona Anderssen. Performance of deep learning in searches for new physics phenomena in events with leptons and missing transverse energy with the atlas detector at the lhc, Jan 2021. URL https://www.duo.uio.no/handle/10852/82487.

[2] Jesus Arjona Martínez, Thong Q Nguyen, Maurizio Pierini, Maria Spiropulu, and Jean-Roch Vlimant. Particle generative adversarial networks for full-event simulation at the lhc and their application to pileup description. *Journal of Physics: Conference Series*, 1525(1):012081, 2020. doi: 10.1088/1742-6596/1525/1/012081.

[3] Ritter von Merkl, Schwanenberger, and Grohsjean. Ritter von merkl, k., Jan 1970. URL https://bib-pubdb1.desy.de/record/449738.

[4] The CMS collaboration. Missing transverse energy performance of the cms detector. *Journal of Instrumentation*, 6(09), 2011. doi: 10.1088/1748-0221/6/09/p09001.

[5] Thabang Lebese and Xifeng Ruan. The use of generative adversarial networks to characterise new physics in multi-lepton final states at the lhc, Feb 2022. URL https://arxiv.org/abs/2105.14933.

[6] Anja Butter, Tilman Plehn, and Ramon Winterhalder. How to gan lhc events, Nov 2019. URL https://arxiv.org/abs/1907.03764.

[7] Torben Lange. Applications of deep neural networks in a top quark mass measurement at the lhc, Jun 2018. URL https://cds.cern.ch/record/2621556.

[8] M. Binkley. Kinematic variables as ttbar discriminants in w + jets, Jan 1995. URL https://inis.iaea.org/search/search.aspx?orig_q=RN%3A28002772.

[9] Mert Dil. Mertdil/graduationproject: Met identification with wgan-gp algorithms. URL https://github.com/Mertdil/graduationproject.git.