DSC-PHYS-02/2024

Transverse momentum ($\mathbf{p_T}$) reconstruction at Large Hadron Collider using Machine Learning

Summer Research Internship Program - 2024,
Dyal Singh College (Univesity of Delhi), Lodhi Road, New Delhi - 110003
Is being submitted to: Internal Quality Assurance Cell (IQAC), DSC

By
Harshit Miglani, Radhika Pandey, Swayam Jha
Department of Physics, Dyal Singh College (University of Delhi)
Delhi-110003

Mentor: Naveen Gaur Department of Physics, Dyal Singh College (University of Delhi)
Delhi-110003, India

Dedication

This report is dedicated to the memories of our brilliant student Swayam Jha. Swayam was part of this summer project but unfortunately left us on 19 July 2024.

Acknowledgement

This summer project was undertaken in the Dyal Singh College (University of Delhi), Lodhi Road, New Delhi-110003 during the period June-August 2024. We thank the IQAC cell of the college for the support. We would like to thank Dr. Thallapalli Tejaswini (IIT Delhi) for help and discussions.

Harshit Miglani

Radhika

Harshit Miglani Radhika Pandey Swayam Jha

Naveen Gaur

Disclaimer: We do not claim the originality of the work

Contents

Chapter 1

Introduction

Proton-proton collisions in present times are very significant processes in particle physics. They are studied at high-energy particle accelerators like the Large Hadron Collider (LHC) at the European Organization for Nuclear Research (abbreviated as CERN) [1]. In these collisions, two protons are accelerated to near the speed of light and made to collide head-on. The immense energy from these collisions breaks the protons apart, producing a variety of subatomic particles and possibly giving us a glimpse of the underlying physics. LHC provides a new energy frontier to not only study what is known as the Standard Model of Particle Physics but also give glimpses of physics beyond the standard model. The study of these collisions is done via many of the kinematical variables of the produced particles. Some basic kinematical variables in the context of high-energy colliders are rapidity, azimuthal angles, transverse momentum (p_T), invariant mass etc. In this report, we try to build a machine learning model to predict such crucial parameters providing insights into how the universe works and providing a glimpse of the high energy state of the early universe.

1.1 Motivation

As a physicist, we are always interested in finding out how the universe started, the basic nature of the fundamental forces governing our existence, the fundamental particles their behavior etc. In the quest to understand the universe, physicists are often stuck with unresolved phenomena like dark matter. Many experiments have been conducted to improve our understanding and at present a new high-energy frontier in the name of LHC is currently operational at CERN, Geneva, Switzerland. In this experiment, two proton beams are collided in opposite directions

[1] https://home.cern/

at the center of mass energy of about 13/14 TeV. The results of these collisions are studied at the many detectors that are placed at various strategic locations. The detectors that are supposed to observe the results of these collisions are CMS, ATLAS, ALICE, and LHCb. In this work, we have tried to develop and use a Machine Learning model that can help in reconstructing some of the kinematical variables at the di-electron (electron-positron pair) final state of LHC.

1.2 Project Scope

We have considered the data of the following process:

$$p + p \to e^+e^- + X \tag{1.1}$$

where X could be any stuff produced along with the electron-positron pair. We have considered 100,000 events in the above process for our analysis. These events have the four-momenta of electron and position. These were used to reconstruct p_T and invariant mass (M).

1.3 Overview of Collision Parameters

Energy-momentum-mass relationship of a particle of rest mass M is given as :

$$E^2 = |\vec{p}|^2c^2 + M^2c^4 \tag{1.2}$$

where c is the velocity of light, $\vec{p}$ is the three momentum of the particle. In this work we will be using *natural system* of units where $c = 1, \hbar = 1$ where $\hbar = \frac{h}{2\pi}$. In the natural system of units, the above relation becomes

$$E^2 = |\vec{p}|^2 + M^2 \tag{1.3}$$

The collision involves several key parameters that must be understood to make accurate deductions from the experiment.

1. $\mathbf{p_x}$: x-component of the 3-momentum of the particle.

2. $\mathbf{p_y}$: y-component of the 3-momentum of the particle.

3. $\mathbf{p_z}$: z-component of the 3-momentum of the particle.

4. $\mathbf{p_T}$: Transverse momentum (p_T) of particle is defined as:

$$p_T = \sqrt{p_x^2 + p_y^2} \tag{1.4}$$

5. **Rapidity (η)** : The definition of rapidity is

$$y = \frac{1}{2} ln \left(\frac{E + p_z}{E - p_z} \right) \tag{1.5}$$

   It is difficult to measure rapidity in a collider environment like LHC and hence one defines another quantity derived from rapidity and is known as pseudo-rapidity

$$\eta = -\frac{1}{2} ln \left(tan \frac{\theta}{2} \right) \tag{1.6}$$

   where θ is the angle made by particle trajectory from beam-pipe. Some points about rapidity

   - For highly relativistic particles $y \approx \eta$. As we are dealing with highly relativistic particles hence we will use η and call it as rapidity.
   - When $\theta = 90$ (the particle is coming out perpendicular to the beam pipe), $\eta = 0$.
   - When $\theta = 0$ (the particle is along the beam pipe), $\eta = \infty$
   - Rapidity is an additive quantity.

6. **Azimuthal angle (ϕ)**: Angle made by the projection of net momentum for the x-axis

$$\phi = tan^{-1} \left(\frac{p_y}{p_x} \right) \tag{1.7}$$

7. **Invariant mass (M)**: The system's invariant mass is independent of the frame of reference.

$$M^2 = E_{total}^2 - P_{total}^2 \tag{1.8}$$

   where P_{total} is the total momentum of the system of two particles. The equation, in verbose form, is as follows:

$$M^2 = (E_1 + E_2)^2 - (p_1 + p_2)^2 \tag{1.9}$$

   where, p_1 is the net momentum of the first particle, and p_2 is the net momentum of the second particle.

1.4 Rapidity in High-Energy Collisions: Why It Matters

Rapidity is a crucial concept in the description of the motion of nearly light-speed particles in high-energy physics. The rapidity can be defined as :

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right)$$

where E is the energy of the particle and p_z is the z-component of the momentum of the particle. To simply describe it, the choice of frame of reference affects how time, distance, and momentum are measured and the relevance of rapidity in high-energy collisions primarily lies in its ability to describe particle motion in such a way that it is independent of the observer's frame of reference. This is crucial for accurately interpreting collision data, as the motion of these nearly light-speed particles must be analyzed in a manner that is unaffected by the observer's frame. To sum up, rapidity is a Lorentz invariant quantity; hence, the rapidity calculations can be done in any frame of reference.

The problem with the definition of rapidity as given in eqn 1.5 involves the total energy and z-component of the particle's momentum. These quantities are difficult to measure in a collider environment. The reason for this is that the initial state particles (protons here) are traveling along the beam pipe that is considered to be the z-axis and it is not possible to put detectors along the beam pipe. This makes it difficult to measure any quantity (including momentum) along the beam pipe. Hence it is not possible to measure p_z thereby creating problems in the estimation of exact energy. This is also the reason why it is difficult to measure rapidity as defined in eqn 1.5 as the estimation of rapidity involves E and p_z. Thus the concept of pseudo-rapidity is introduced.

1.5 Pseudorapidity

Rapidity is a very useful concept for which we need both the Energy of the particle and its momentum components. As explained in the previous section rapidity being a Lorentz invariant quantity is a useful kinematical variable but still, it is difficult to estimate in a collider environment. In colliders, the particles involved are highly relativistic *i.e.* particles moving at a speed close to the speed of light and hence the expression of rapidity eqn. 1.5 is modified to give pseudo-rapidity given in eqn 1.6.

Chapter 2

Methodology

In the section 2.2 the results of event-based distributions of some of the kinematical variables are presented. In section ?? we have used Machine Learning (ML) techniques to predict the distributions of some of the variables presented in section 2.2.

2.1 Event based data Analysis

It is important to know the distribution of the physical parameters in an experiment, to identify outliers and provide a rough idea of what to expect, ensuring more accurate and reliable results. In every run, several subatomic particles are produced, we have considered a dataset where an electron-positron pair is present.

The analysis of the event file was done using `Python` and some of the libraries of Python like `Pandas, seaborn` were used for the analysis and presentation of the events. All the programs and data files used for the analysis are given in public domain [1].

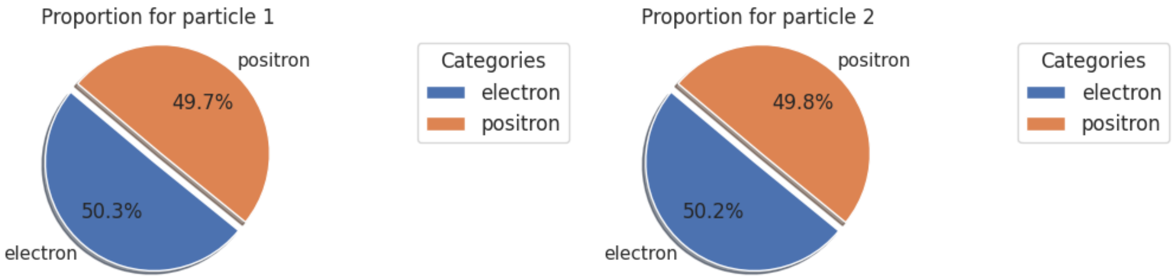


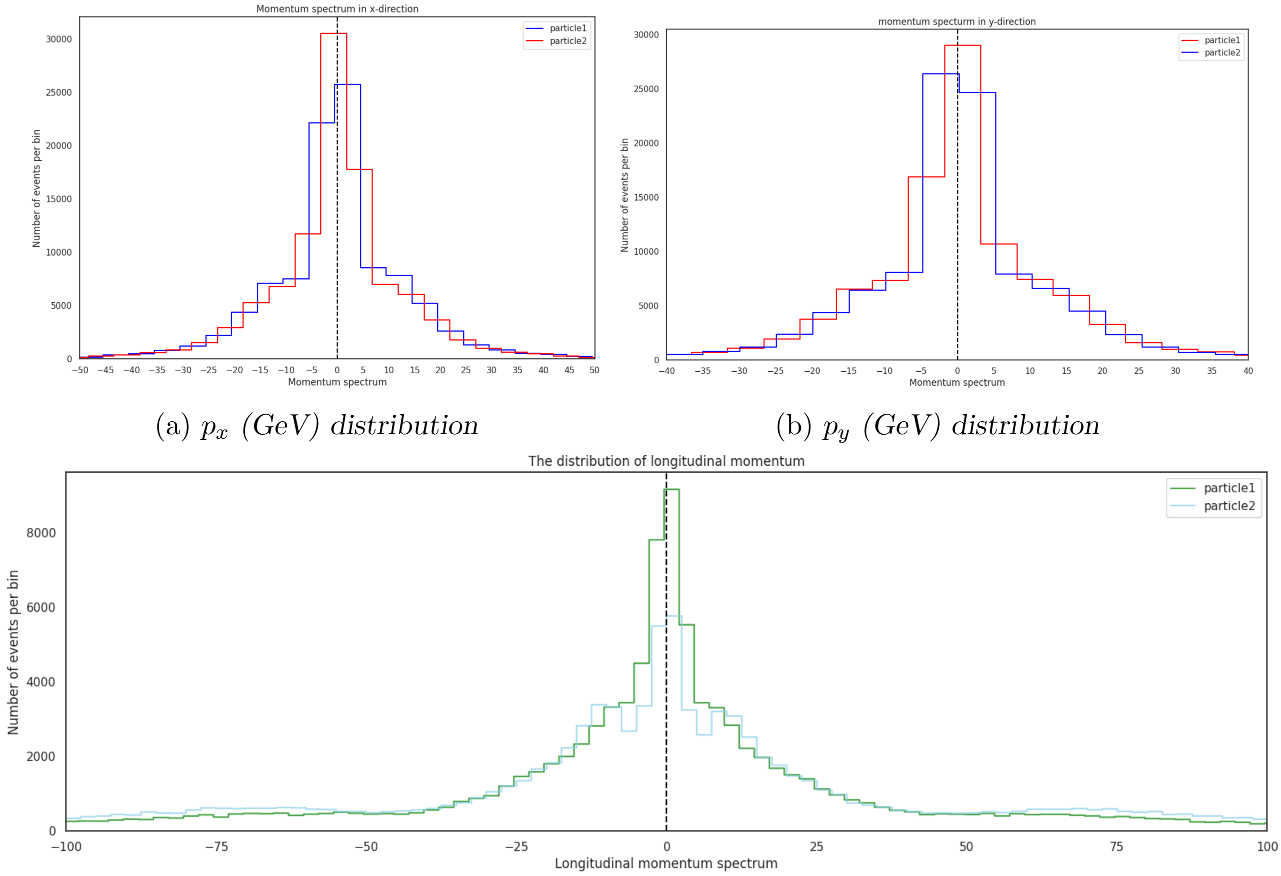
Figure 2.1: Proportion of electron and positron in the data.

In Fig 2.1 we have shown the pi-plot of the proportionate representation of electrons and

[1] https://github.com/navgaur/SP-ML-PT-2024/tree/main

positrons in particles labeled 1 and 2. As can be seen, approximately electrons and positrons exist with 50% probability *i.e.* nearly equal distribution of e^+ and e^- are present.

Using the dataset of 100,000 events having e^+ and e^- we have tried to plot some of the histograms of the kinematical variables. In fig 2.2a, 2.2b, 2.2c we have plotted the histogram of x,y, z components of the momentum of e^+ and e^-.



(a) p_x *(GeV) distribution*

(b) p_y *(GeV) distribution*

(c) p_z *(GeV) distribution*

As expected the symmetry of the problem suggests that the distribution is symmetrical around the origin. The distributions of both electrons and positrons are similar.

In Fig 2.3 we plotted both particles' p_T (in GeV) distribution. The mathematical expression of p_T is given in eqn 1.4. We again wish to note that transverse momentum is one of the most important kinematical variables in collider experiments as it is difficult to measure the z-component of the momentum hence for the physics analysis, one uses transverse momentum. The p_T distributions of both the particles appear to be similar. It is to be noted that p_T by definition is a positive quantity.

In Fig. 2.4 we have plotted the (pseudo)rapidity (η) and azimuthal angel (ϕ) distributions. As can be seen, the rapidity distributions peak for larger values of η. Note larger values of η mean

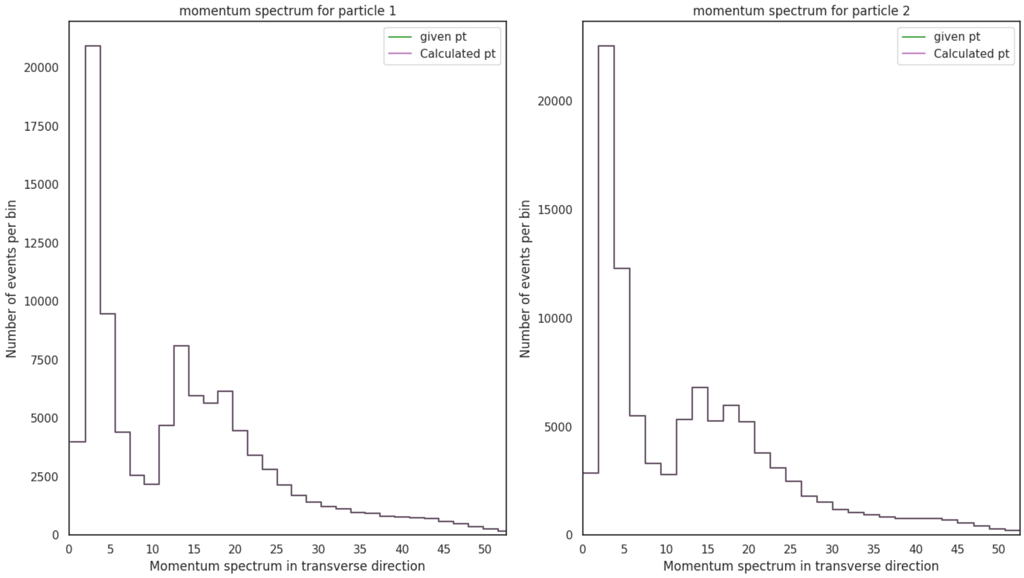


Figure 2.3: *p_T (GeV) distribution for the particle 1 (left panel) and particle 2 (right panel)*

the particles are coming in the direction transverse to the beam direction.

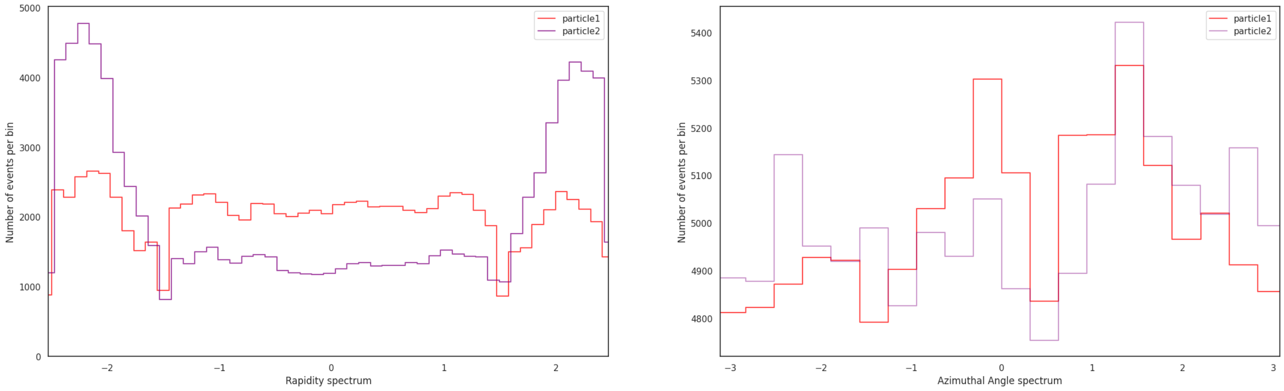


Figure 2.4: Rapidity (η) on left panel and Azimuthal angle (ϕ) on right panel distributions

The energy distribution of the particles is shown in Fig 2.5.

In Fig 2.6 the invariant mass distribution of di-electron (electron-positron pair) has been plotted. This distribution is plotted from the invariant mass given in the event file. However, in Fig 2.7 we have plotted the invariant mass of the electron-position pair using the eqn 1.9. The invariant mass distribution shows the well-known s-channel fall. Interestingly the distribution shows a peak/hump around 90 GeV. This is indicative that the electron-position pair was produced via an intermediatory neutral particle of mass about 90 GeV via the process like $pp \to Z(\to e^+e^-)X$. It is well known that this particle is a Z-boson that has a mass of 91.2 GeV. The existence of this boson was predicted in the Standard Model of particle physics and

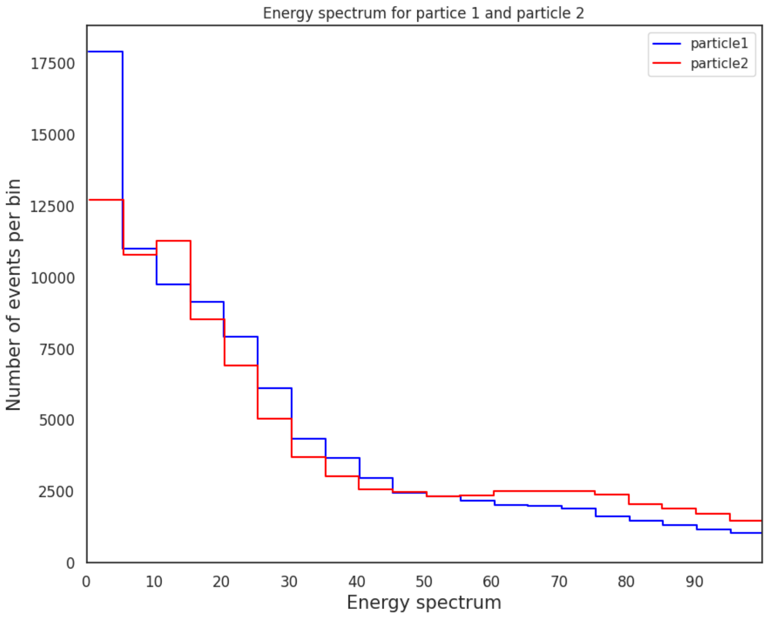


Figure 2.5: Energy distribution of the particles

was finally discovered at UA1 and UA2 collaborations of Super proton anti-proton collider ($Sp\bar{p}S$) which was a modification of the Super Proton Collider (SPS) at CERN in 1983 [1]. Carlo Rubbia headed these experiments and within a few months of the announcement of the discovery was awarded the Nobel Prize in Physics in 1984 [2]. This indicates the importance of the neutral Z-boson in our understanding of fundamental physics. An excellent writeup on the issue can be found in CERN monthly magazine named CERN courier [3].

2.2 Prediction Using Machine Learning

In this study, we aimed to model the transverse momentum ($\mathbf{p_T}$) of two particles based on their momentum components in the x and y directions ($\mathbf{p_x}$ and $\mathbf{p_y}$). Here we employ a machine learning approach using the Random Forest Regressor (RFR) algorithm. Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and reduce the risk of overfitting. In this study, we used Random Forest Regressor due to its capability to handle non-linear relationships and its robustness to noise in the data.

2.2.1 Data Preparation

The dataset used in this study consisted of the momentum components $\mathbf{p_x}$ and $\mathbf{p_y}$ for two particles and their corresponding transverse momentum $\mathbf{p_T}$. We divided the dataset into training

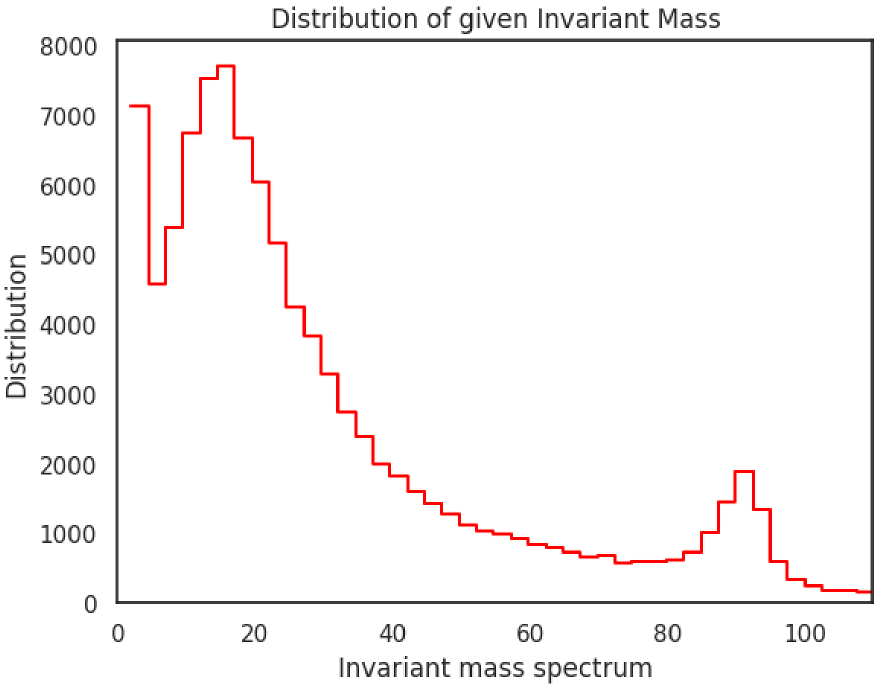


Figure 2.6: Invariant mass of the particles

and testing sets. The training set was used to fit the model, while the testing set was used to evaluate its performance.

For particle 1, the input features were $\mathbf{p_x^1}$ and $\mathbf{p_y^1}$ the target variable was $\mathbf{p_T^1}$. Similarly, for particle 2, the input features were $\mathbf{p_x^2}$ and $\mathbf{p_y^2}$ the target variable was $\mathbf{p_T^2}$.

2.2.2 Model Training and Evaluation

Two separate Random Forest models were trained — one for each particle. The models were initialized with default parameters, which typically include 100 decision trees, a maximum depth that allows full growth of the trees unless otherwise limited by the data, and other hyper-parameters that control the randomness and tree structure. The models were trained using the input features and target variables for each respective particle. After training, we predicted $\mathbf{p_T}$ for the training set to evaluate the model's performance. The performance of the model was assessed using the R^2 score, which measures how well the predictions match the actual values. The R^2 score ranges from 0 to 1, with 1 indicating perfect predictions.

For particle 1, the model achieved an R^2 score of 0.9969, indicating a strong correlation between the predicted and actual $\mathbf{p_T^1}$ values For particle 2, the model achieved an R^2 score of, 0.9955 which similarly indicates a high level of accuracy.

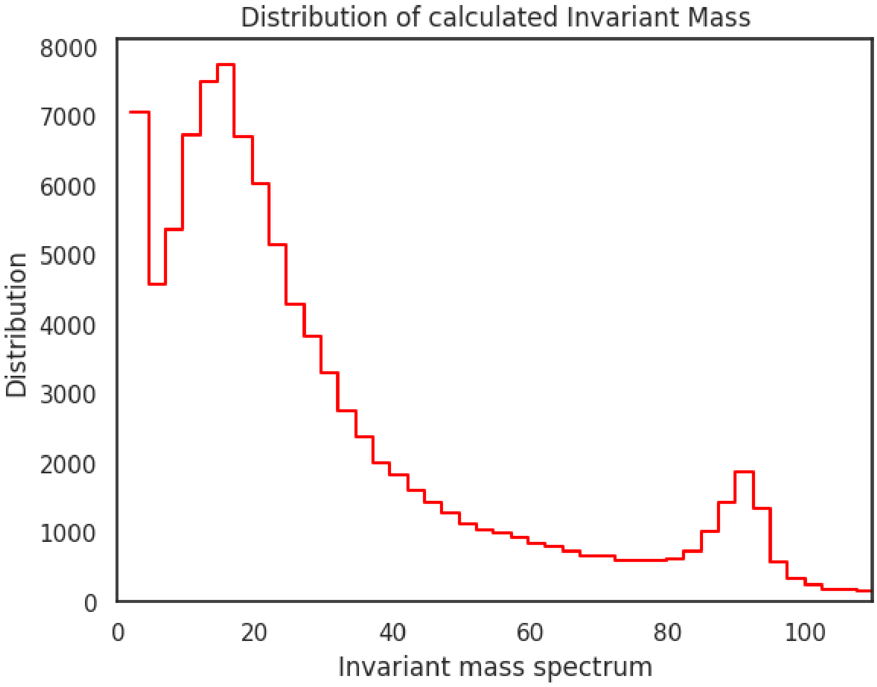


Figure 2.7: *Invariant mass*

Chapter 3

Results

We have compared the distributions of the modeled $\mathbf{p_T}$ values against the original $\mathbf{p_T}$ values. Histograms were generated for both particles, showing the modeled and original transverse momentum distributions. The results of the same are shown in Fig 3.1. As can be seen from the figures for both the particles, the modeled distributions closely matched the original distributions, indicating that the Random Forest model successfully captured the relationship between $\mathbf{p_x}$, $\mathbf{p_y}$ and $\mathbf{p_T}$.

The data files and the programs used in this work are placed in the public domain [4] This document is prepared using $\LaTeX\,2_\varepsilon$documentation system on the Overleaf [5] platform.

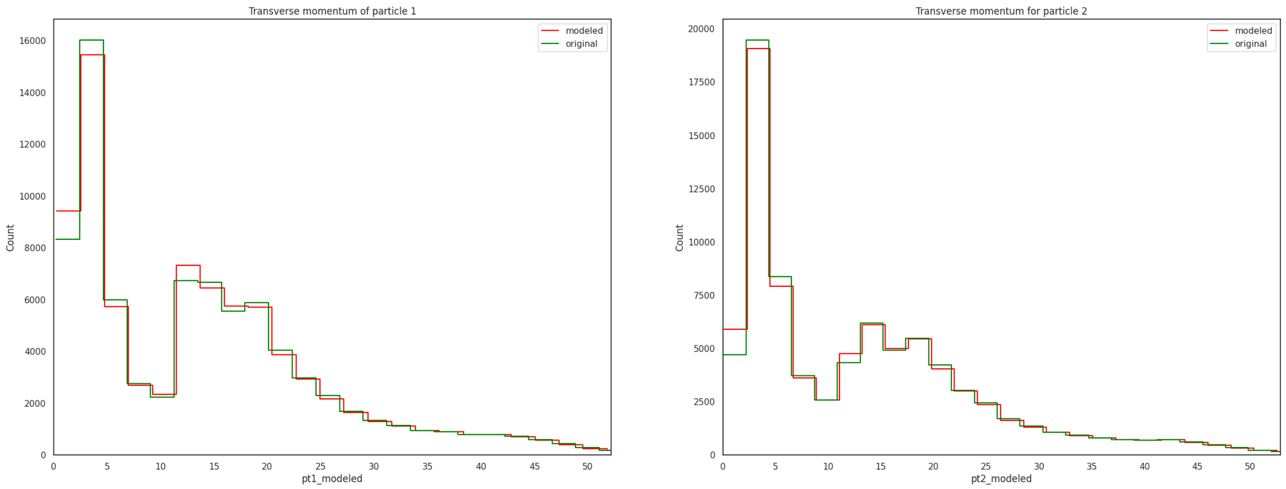


Figure 3.1: *Predicted vs Original* $\mathbf{p_T}$ *distributions for particle 1 and 2*

Bibliography

[1] https://home.cern/science/physics/z-boson

[2] https://www.nobelprize.org/prizes/physics/1984/summary/

[3] https://cerncourier.com/a/when-cern-saw-the-end-of-the-alphabet/

[4] https://github.com/navgaur/SP-ML-PT-2024/tree/main

[5] https://www.overleaf.com