CERN Summer Student Report

Improving Transverse Invariant Mass Resolution for Dark Photon Searches Using Machine Learning

By:
Adnan Alwahedi
(UAEU/CERN)
Supervisor:
Dr. Mohamed Belfkir (UAEU)

Acknowledgements

I would like to express my gratitude to the European Organization for Nuclear Research (CERN) for giving me the opportunity to intern at CERN, and providing their facilities, teaching, and knowledge for me to use and learn from.

I would like to also thank my supervisor Dr. Mohamed Belfkir for teaching and guiding me on my project. I have learned a lot from Dr. Mohamed as his knowledge has been beneficial in my work, as it has enhanced my knowledge in both physics and programming skills. As well I want to thank Prof. Salah Nasri for his contribution in my journey since he has helped me and taught me before coming for the program. I would like to express my gratitude as well to my university for informing and preparing me for this program, by teaching me the necessary courses to be ready.

Abstract

There have been huge contributions to the search for Dark Matter (DM) candidates at the Large Harden Colliders (LHC) through the Higgs boson decay. This is because the Higgs portal can give us huge access to visible and dark interactions. The DM photon as known as (γ_d) is a theoretical particle that can be created in the Higgs boson decays H$\rightarrow$ $\gamma\gamma_d$. We take into consideration the production of the Higgs boson in the LHC through the gluons fusion (ggF) which has the largest production rate. So our goal in our project is to improve the results of finding the transverse invariant mass (mT) resolution for Dark Photon by using machine learning.

Contents

Chapter 1

Introduction

The existence of dark matter is most definitely certain since there is cosmological evidence to show it. One of the main ways to look into dark matter is when dark particles have an interaction with other dark particles. The second interaction is when dark particles interacts and develops into visible sector particles. Last interaction is when normal particles interacts and develop into dark particles, as shown in Figure 1.1

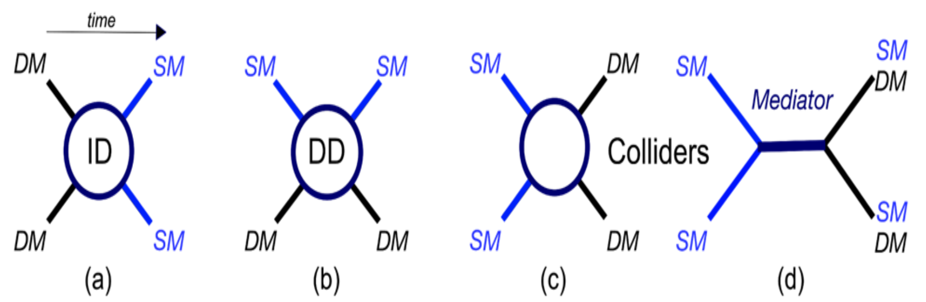

Figure 1.1: Schematic illustration of dark matter production modes [4]

The First two models can not be investigated inside the colliders, therefore the concentration is on the third interaction that happens. As it is known the Higgs boson is what gives particles their masses and since dark matter is huge, we can understand there might be a link between dark matter and the Higgs boson. Therefore that's why we might consider the use of the Higgs portal to gain access to the sector of dark matter.

The majority of searches about dark was concentrating on recognizing a weakly interacting massive particle, but till now there has been no success in finding such a particle. As a result of that, studies now focus on a different candidate and between those studied candidates have been the dark photon

(γ_d) [2]

In this project, we will be focusing on improving the Transverse Invariant mass (mT) resolution for Dark Photon searches by using a Machine Learning model. With this machine learning model we will try to make a model that gives us results as close to the truth value of mT and will go into detail in what type of machine learning model we used and the various data we used in our model.

Chapter 2

Experimental Setup

2.1 The CERN-Large Hadron Collider (LHC)

The Large Hardon Collider (LHC) is the biggest and the most powerful particle accelerator created in the world, the LHC is located between both the Swiss and French borders. The collider was built by the European Organization for Nuclear Research (CERN) in 2008 and the first beam collision was in 2010 [3]

The LHC is made of 27 kilometres of underground superconducting magnets with many accelerators that are built inside of it to expand the particle energy. The LHC main use is to study high-energy particles. Inside the collider, the two beams of hardons which are usually protons move almost as the speed of light and they move in opposite direction. Calling attention to Einstein equation ($E = mc^2$), the collision of these energized hardons develops in conversion of energy to mass and therefore new particles are created. We then study these produced particles to be able to understand our universe much better. [3]

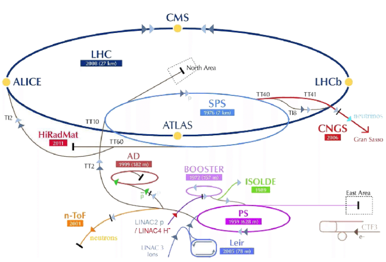

Figure 2.1: (LHC diagram) [10]

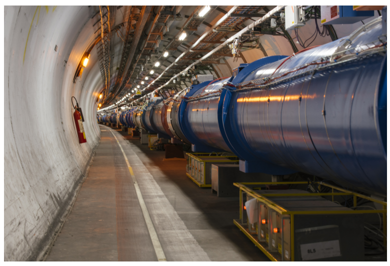

Figure 2.2: (LHC pipe photo) [12]

2.2 The ATLAS Experiment at LHC

There are a total of eight different experiments in the LHC using various detectors. There are four main experiments which are ATLAS, CMS, LHCb, and ALICE. Both ALICE and LHCb experiments focuses on detecting and studying the Quark-Gluon Plasma(QGP) and b-hadrons. CMS and ATLAS have many different usage and experiments that they focus on, where they focus on an enormous range of physics phenomena.

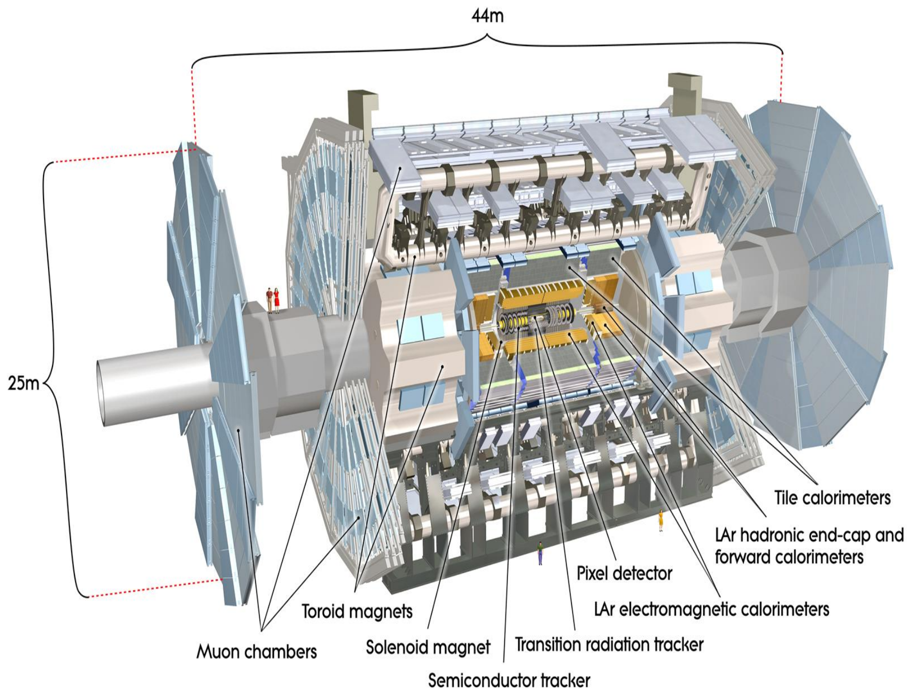

Figure 2.3: The layers of ATLAS detector [1]

The ATLAS detector has a diameter of 25 m and os 46 m long where it is put 100 m underground. It also Weight approximately around 7000 tons.[14]

Atlas is divided into four sub-detectors, where every single one is developed and enhanced to measure and reconstruct a certain set of particles. These are composed into the following, the inner detector (ID), the moun spectrometer (MS) and electromagnetic and hadronic calorimetrs (ECAL and HCAL).

The closest layer to the collision point is the ID and its aim is to reconstruct the direction of the charged particles and compute their momentum. In addition, a 2T magnetic field is used to bend the trajectory of the charged particles.

The ATLAS calorimeter system is meant to supply a accurate energy meaurment and position reconstruction for electromagnetic (EM) particles (electrons, photons) and jets (hadrons). The hermiticity of the calorimeter also let it measure the missing transverse energy and shows the separation of electron and photons from hadrons and jets. The calorimeter is made of two calorimeters: The hadronic calorimeter (HCal) and the electromagnetic calorimeter (ECal). Both are sampling calorimeter, with layers of heavy absorber material and an active material in which an ionization signal is made. The ECal is the first sub-detector after the ID. It is maximised for the energy reconstruction of photons and electrons by triggering EM showers and measuring its properties. At high energy colliders, gluons and quarks shred into fragments of particles known as jets. The HCal finishes the measurement of the jet energy started by the ECAL

Mouns are slightly ionizing particles because of their high mass, thus they are not stopped by the calorimeter system. The Moun spectrometer is the furthermost part of the ATLAS detector intended to spot the moun leaving the calorimeter and calculate their momentum. A huge superconducting air-cored toroid magnet is used to bend mouns directions. We can see the full picture of ATLAS various layers of detectors as shown in Figure2.3.

2.3 Triggers

The main purpose of the Trigger system is to run the the signals created in the detectors and speedily choose events that has only the needed and exciting information. The events are saved and then are accessible for future analysis. The system uses a simple criteria and based on the type of events being studied by scientists, various types of triggers can be made and tested out. Even though the system might result in losing or cutting some data, it is needed because of the restriction of the storage space. The rate of the Event generation in the LHC is around 40 MHz and it is cut into almost to 10 $10Hz$ [15].

ATLAS uses two trigger levels [7]: L1 trigger (L1T) and High Level Trigger (HLT). The L1T, which is like a built-in part of each detector, does an important job by filtering out events that don't really have much useful physics information. Its main goal is to bring down the number of events to at least 75 KHz. After this first filtering, the events that pass through the L1T are then sent to another system called the High-Level Trigger (HLT), which works with software.At the HLT stage, the detector's different parts are quickly analyzed, and then the events are reconstructed in a way that's similar to how it's done offline (when we're not in a hurry). This whole process takes around 0.2 seconds. The HLT produces an output of events with a frequency of about 1 KHz.When dealing with situations where there are a lot of events being created, we might need to keep only a fraction of them to fit within our storage limits for offline analysis later on.

2.3.1 Run-2 ATLAS trigger menu

In the second run or Run-2, the trigger is made of about 1500 triggers chain for selecting events. The selection process is based on the physics signatures, these can have presence of isolated and/or energetic photon, leptons, jets, or also missing energy [7]

2.3.2 Missing Transverse Energy (MET)

The MET are particles that do not interact with the detector material, therefor the detectors can't find them, like for example neutrinos. The particles leave the detector with no trace of their existence. Someone might ask how do invisible particles can be traced even though by their meaning they can not be traced or detected. The way to detect them is by indirect methods. The technique to use to detect them is by calculating the imbalance of momentum from the visible particles. The beams in the LHC travel perpendicularly to the beams axis, meaning that the transverse energy is 0. So by applying the conservation of momentum rule, the transverse momentum of all the particles will be equal to 0 when you add them up together after the collision. The events that has invisible particles will have a huge transverse missing momentum in the planes, due to the missing part. This is known as the missing transverse momentum E^T_{miss}. The LHC looks for dark matter and keeps an eye to collisions with huge value of E^T_{miss}.

2.4 Higgs Boson production at LHC

It is known that particles get their mass through the interaction with the Higgs boson field and since this guides to the idea that there is relation between the large dark matter particles and the Higgs boson. This idea can be put to test by looking for events where the Higgs boson decay through an invsible channel. This means we need to look for events where this is some missing energy which can be an invisible dark particle.

Inside the LHC, the most common Higgs boson production are the gloub gloun fusion (ggF), vector boson fusion (VBF) and the Higgs boson strahlung or Z-associated production (ZH) [5]

2.5 Transverse mass

Taking into consideration the H$\rightarrow \gamma\gamma_d$ decay channel,the mass of the Higgs boson can reconstructed in the transverse plane which is the transverse mass mT. Thus it can be defined as the following [11]

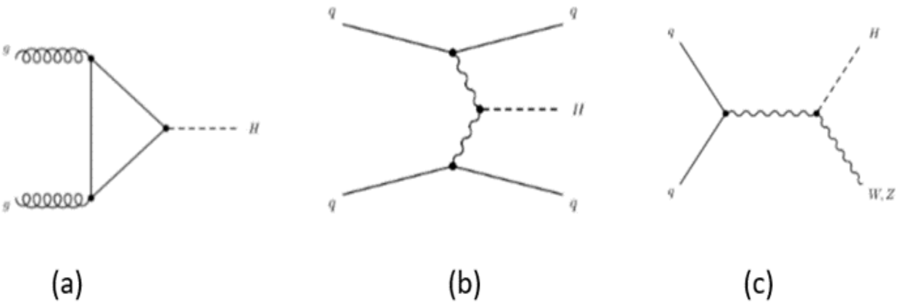

Figure 2.4: Feynman diagrams of the Higgs boson production modes: (a) ggF (b) VBF (c) ZH [8]

$$mT = \sqrt{2 \times P_T \times Met \times (1 - cos(\phi_{Photon} - \phi_{Met}))} \tag{2.1}$$

2.6 Significance

When we're looking at an event we've observed, we say it's "statistically significant" if it's really unlikely that the event just happened randomly. In other words, if the chances of the event being a coincidence are very low. This is determined by something called the "p-value," which tells us how likely it is that the event is just by chance, we take the following equation to be our significance[13]:

$$Z = \frac{s}{\sqrt{s+b}} \tag{2.2}$$

where s and b are the number of signal and background events.

Chapter 3

Machine Learning Model Setup

3.1 What is Machine learning?

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that revolves around using data and algorithms to mimic the way humans learn. Over time, it gradually improves its accuracy by continuously learning from new information [6].

3.2 Machine learning model used

The Machine learning model we used for our project is called the Support Vector Regression (SVR). SVR is a machine learning model mainly used for data classification on continuous data. It is also regression function that is generalized by Support Vector Machine [9]. Before diving into SVR will go briefly into what linear regression is. Linear regression is when we try to draw a line through some points on a graph in a way that the line is as close as possible to all those points. The idea is to reduce the difference between the actual points and our line, kind of like finding the best fit.Now, SVR builds upon this, but it's a bit more flexible [9]. What makes SVR good to use in our project is because In SVR, we allow a certain amount of error around the line we're drawing. Think of it as a buffer zone around the line where we tolerate some mistakes. This buffer zone is like a "no-penalty" area. Unlike regular linear regression where every point counts, SVR is a bit more forgiving. If a point falls within this buffer zone, it's like the model gives it a thumbs-up and doesn't penalize it much. But if a point is outside this buffer zone, that's when SVR starts caring. It looks at how far the point is from the line and takes that into account. Creating the puzzle in SVR is by forming this special zone around the line, and anything within it is safe from punishment. Anything outside gets some attention based on how far it is from the line. This forms an interesting puzzle of how to find the best line while considering this buffer zone. We use some fancy math and techniques to make sure our line fits well while not punishing the points inside that buffer too much. The main goal of SVR is to come up with this super-smart line that fits the data well but doesn't sweat the small stuff outside the buffer zone. By doing this, SVR is like a cool problem solver when it comes to predicting things that don't necessarily follow a straight line. To put this SVR plan into action, we use Python. We get our data all set up, scale it (to make it easier for the computer to handle), and then use various functions to create our SVR model. We play around with our data, see how well the SVR model fits, and make predictions. And with that, we've got a tool that can predict what we need in a pretty flexible way.

Chapter 4

Analysis, Results, and Discussions

4.1 Analysis Plots

Before starting our SVR we need to look at some plots to determine and get an idea of what we are dealing with. So first thing we did is plot both the photon Transverse momentum P_T and the missing transverse energy and comparing them to their truth value.

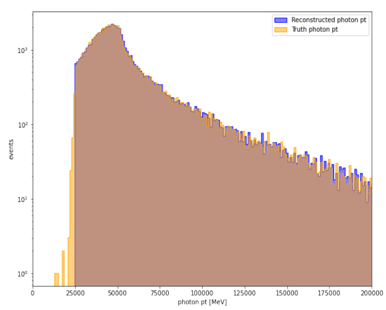

Figure 4.1: (Photon pT MeV)

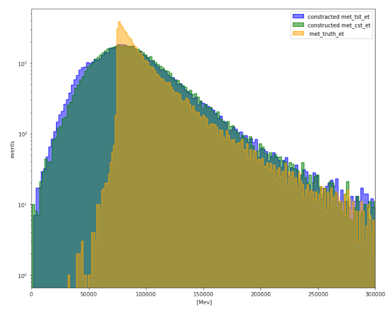

Figure 4.2: (truth, met(cst) and met(tst) plots)

We can see that both the P_T of the constructed and the truth value are very close to each other, which their value is very close to 50000 MeV. While for the Met the values seems to be different between the constructed Met and the truth Met and we see the peak to be around 75000 MeV With this information we can compute the transverse invariant mass by using the equation we have in 2.1 where as we mentioned will calculate P_T, Met, and ϕ direction for both the P_T and Met.

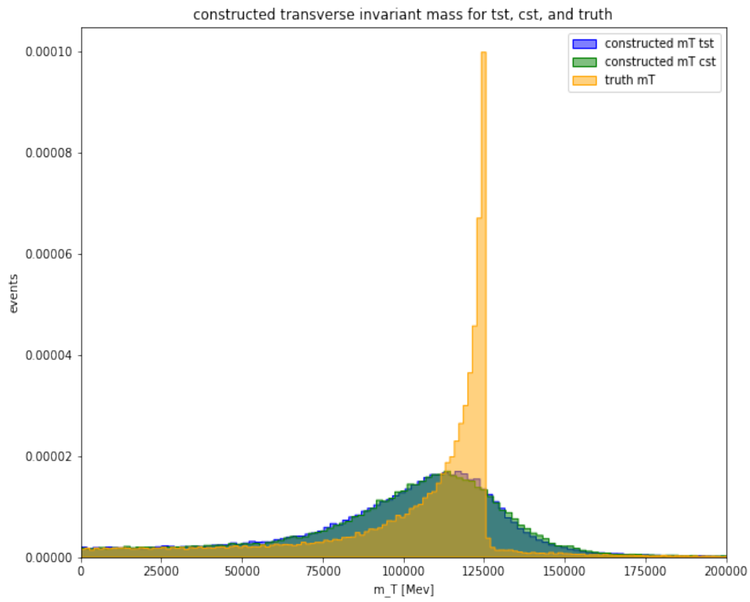

Figure 4.3: mT resolution for truth, constructed cst, and tst

We can see that with this plot the constructed value of the mT is far from our truth mT value, so our goal now is to make the SVR model to have a better fit that is closer to the truth value. In addition we can observe from the plot that the distributions are peaking around 125 GeV of the Higgs mass.

4.2 SVR Plots

4.2.1 Jet dependency

Before starting our SVR training we will have to take some consideration about certain values in our data. One of them being if our mT will have a jet dependency which will affect our results. So we will plot first number of jets per event to see where jets occur more.

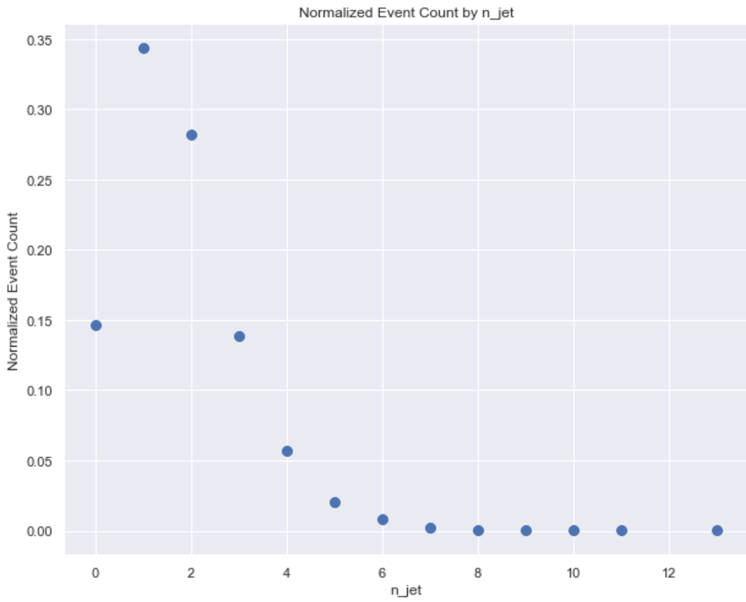

Figure 4.4: number of jet vs events

From the plot we can see that around 15% of events have 0 jets, while most of events have at least 1 or two jets. In addition, we can see that jet occur mostly in 0 jet and 1 jet more specifically, while 2 and more jets seems to have less jets. So in order to solve the issue we will combine all the jets from 2 and more for our next plot. To determine that there is dependency between mT and jets will have to plot the resolution of the mT where we will except to get an Gaussian which will confirm if there is a dependency.

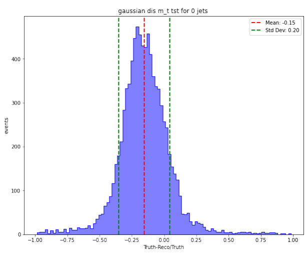

Figure 4.5: gauss dis for 0 jet and mT

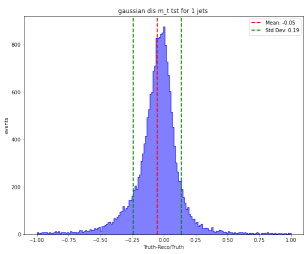

Figure 4.6: gauss dis for 1 jet and mT

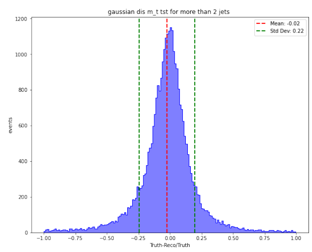

Figure 4.7: gauss dis for more than 2 jet and mT

As we can see there is a huge dependency on the jet so we can assume that that when we do our SVR model and include the jets that the predicted to our truth value.

4.2.2 First SVR Model Test

Our main inputs in the first SVR model will be the Met et, phi, and the sumet.In addition with the photon transverse momentum and the photon in the phi direction, these inputs are used for when we have no jets. These inputs are important since we use them to calculate the mT. The inputs for when we have number of jets will be similar with just the addition of number of jet in the input. The SVR parameters used in this model is known as kernel (RBF) function which is used mostly for non linear regression tasks.

So for our first SVR model will be focusing only for tst values since they gave us the best results for our prediction. So, in this process in the SVR model we will look into training two models one with number of jet and one without number of jet so we can compare the results together and see if there is any improvement in the results and if they are close to the truth value. So after we trained both the models we got the following:

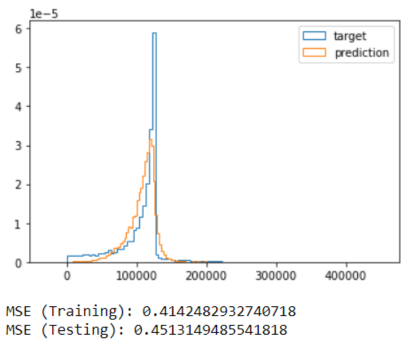

Figure 4.8: mT ML tst with jet

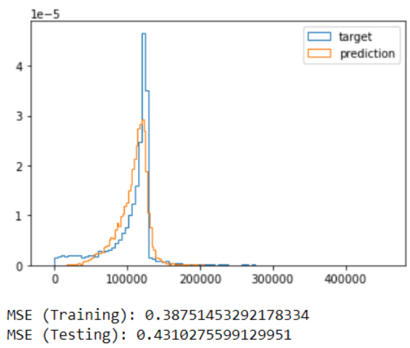

Figure 4.9: mT ML tst with no jet

Unfortunately, from the results we got were not great since we can see that our mean square error is high for both plot and in addition, the plots are not close to our truth value. Even with the number of jets the plots do not show any improvement, this might have been because the mT resolution is mostly dominated by more than 2 jets events, so statistically this will play no role in our model that much. For our next step to try to improve our resolution will try to add the background to see if there is any improvement in our results.

4.2.3 Second SVR Model Test

Before doing our second test, we will plot the first model with background to see what kind of graph we get, where we except to get a smooth curve with no peaks, the following graph:

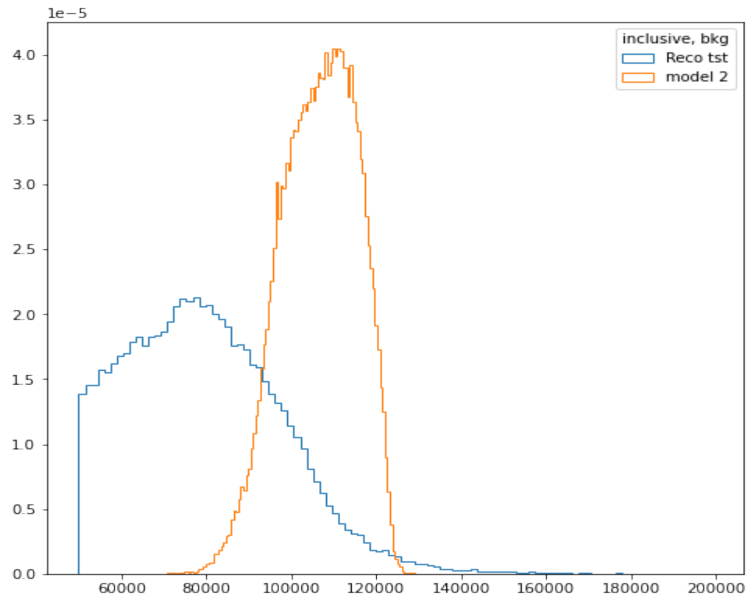

Figure 4.10: First model on Background

As we can see there is a huge peak instead of smooth line we wanted, we can conclude that in the first model there are a lot of interference from things inside the data, where it affects ML model while it learns from it.

For our second test where we are required to better our resolution mT, We will consider the background in making our mT model. We will as well take into consideration the signal, this is because when we usually apply only the background on our model it will create for us an artificial peak which is not something we want as it happened in 4.10. With the addition of the signal it will help us produce a nice peak which will be around 125 GeV. So in conclusion we will train the model with both the background and the signal together.

After we trained our model and scaled the input and prediction data, and plotted it we get the following mT plot where we plotted only the signals events for the SVR with background, SVR with no background, the reconstructed mT, and the truth mT. Also to mention we started the plot from 50 GeV so our plot because the results below 50 GeV do not matter to our main result, so the following plot:

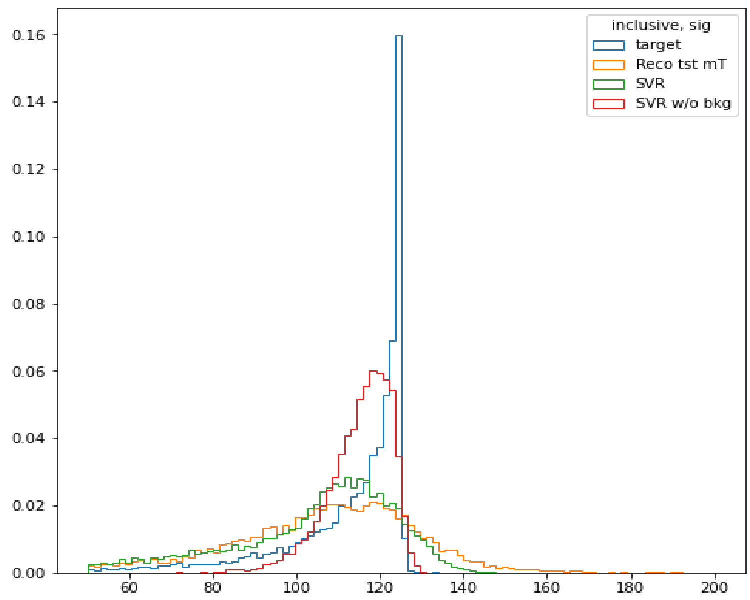

Figure 4.11: Inclusive sig

As we can see our SVR plot with background is doing much better than the reconstructed one, since it's been trained with the background and signal. While the one with no background had much better peak, this is because when you add background into your training model it lowers the performance of the model.

After that we plotted the mT with only the background, if we plot this we except our plot to have a smooth curve similar to our truth value, with no peak between 120 to 130 GeV. So the following graph we get is:

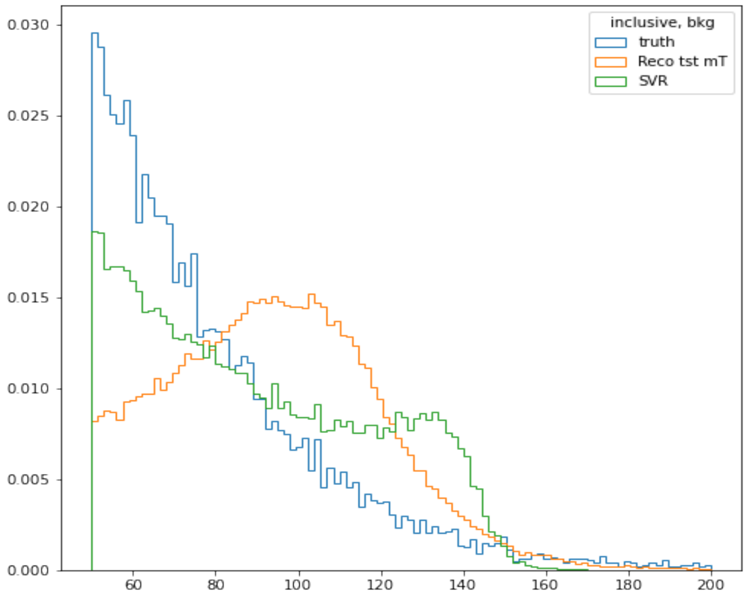

Figure 4.12: Inclusive bkg

As expected out mT model got a smooth curve till 120 GeV similar to the truth value, even though there is a peak at 140 GeV we don't have to worry about it because it reads something else in the data that do not contribute anything to the information we need.

Moving on we will plot and calculate the significance before and after the regression. This will help us understand how much better our data is before and after. For calculating the significance we will use the follwing formula mentioned in chapter 2 2.2, where in order to calculate the significance before the regression we will keep the range of the mT reconstructed to be from 80 GeV to 140 GeV, as this will help us calculate the number of signals. To count the number of signals we simply multiply the filtered range with the event weight which is $eventweight = 0.777546$ which will give us the number of signals. For the background will use only use the weight. This goes the same as well for the after regression but the only difference is the filtered range is going to be for mT predicted. So the following plot we get and result:

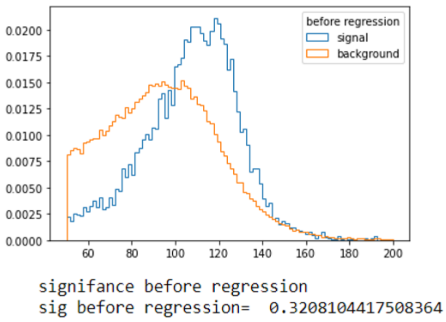

Figure 4.13: Before regression

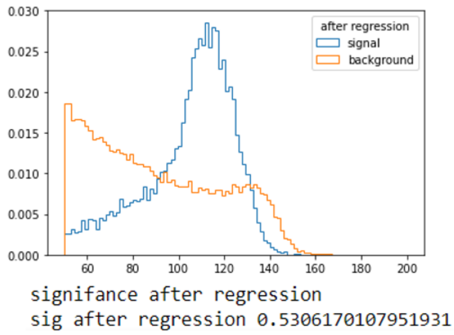

Figure 4.14: After regression

As expected we see an increased improvement on the significance after the regression, which means that the mT plot we did by the SVR model has been improved much more than the first test and it is much closer to the truth value.

Chapter 5

Conclusion

In conclusion, in this project we were able to introduce and explain what type Machine Learning model we used for our project where it helped us improve the resolution of the mT to be much closer as possible to our truth mT value. In, addition we were able graph all the required plots needed and understand all the results which helps us conclude that our mT resolution has improved. We still can do a lot to improve our results much better by using different type of methods but the only thing we need is time with us to be able to perform it.

Bibliography

[1] *ATLAS detector at the LHC*. URL: http://opendata.atlas.cern/release/2020/documentation/atlas/experiment.html.

[2] Hugues Beauchesne and Cheng-Wei Chiang. “Measuring properties of a dark photon from semi-invisible decay of the Higgs boson”. In: *Journal of High Energy Physics* 2022.4 (2022), pp. 1–21.

[3] Cern. *The Large Hadron Collider*. URL: https://home.cern/science/accelerators/large-hadron-collider.

[4] ATLAS Collaboration et al. *Atlas feature: Dark matter*. Tech. rep. 2019.

[5] LHC Higgs Working Group. *Higgs cross sections and decay branching ratios*. URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWG.

[6] IBM. *What is Machine Learning?* 2023. URL: https://www.ibm.com/topics/machine-learning.

[7] on behalf of the ATLAS Collaboration Javier Montejo Berlingen. “Triggering in the ATLAS experiment”. In: *40th International Conference on High Energy physics - ICHEP2020* (2020).

[8] Nguyn Anh Kỳ and Nguyn Thi Hng Vân. “Latest results on the Higgs boson discovery and investigation at the ATLAS-LHC”. In: *Journal of Physics: Conference Series*. Vol. 627. 1. IOP Publishing. 2015, p. 012011.

[9] Bejamin Naibei. *Getting Started with Support Vector Regression in Python*. 2022. URL: https://www.section.io/engineering-education/support-vector-regression-in-python/.

[10] Yuancun Nie et al. “Numerical simulations of energy deposition caused by 50 MeV—50 TeV proton beams in copper and graphite targets”. In: *Physical Review Accelerators and Beams* 20 (Aug. 2017). DOI: 10.1103/PhysRevAccelBeams.20.081001.

[11] KA Olive et al. “KINEMATICS”. In: *C:* 9 (2014), pp. 508–512.

[12] Corinne Pralavorio. *Record luminosity: well done LHC*. 2017. URL: https://home.cern/news/news/accelerators/record-luminosity-well-done-lhc.

[13] *Statistical Significance — Brilliant Math Science Wiki*. URL: https://brilliant.org/wiki/statistical-significance/#:~:text=An%20observed%20event%20is%20considered.

[14] *The ATLAS Detector*. URL: https://atlas.cern/Discover/Detector.

[15] Ramon Cid Manzano Xabier Cid Vidal. *LHC Trigger :Taking a closer look at LHC*. URL: https://www.lhc-closer.es/taking_a_closer_look_at_lhc/0.lhc_trigger#:~:text.