CERN ONLINE SUMMER STUDENT 2021

Study of semi-visible jet using machine learning method

Supervision:

Deepak Kar
Sukanya Sinha
Tasnuva Chowdhury

Augustin Tribolet

August 20, 2021

Contents

1 Introduction

There are several evidences that the theory of the Standard Model has to be extended in a more general theory. One of the most important is the dark matter problem. About 30% of the total amount of matter in the universe remains unknown and invisible to all the current detectors. None of the particles in the SM can explain the state of the dark matter. The dark matter candidate should be a stable and neutral particle which does not or extremely rarely interact with the SM particles.

The Large Hadron Collider at CERN is strongly involved in this issue. The search for a neutral and non-interacting dark matter candidate in the detector is expected to be seen as missing transverse energy. Recently, the study of the Hidden Valley theory has been investigated. This theory predicts the existence of the so-called semi-visible jet that could be seen in LHC data.

The application of machine learning techniques in LHC analysis offers a new way of looking for dark matter. In this project, the theory of the Hidden Valley is investigated with the used of machine learning algorithm. The goal is to produce images of the semi-visible jet and see if they are different from usual jets from the SM using Convolutional Neural Network.

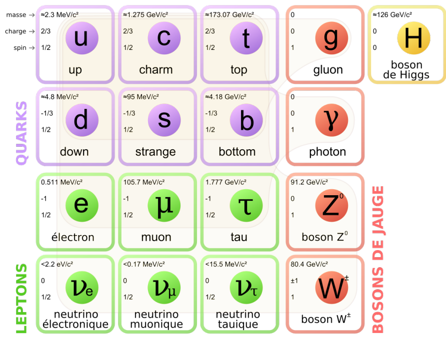


Figure 1: The Standard Model

2 Dark matter searches

2.1 The Hidden Valley theory

The idea of the Hidden valley is to assume the existence of a new SU(2) gauge symmetry group independently of the SM gauge group and constituting the dark sector. In this model, the dark sector is strongly interacting and the two sectors have the possibility to communicate through portal interactions (figure 2). There are several possibilities for the dynamics within the DM sector.

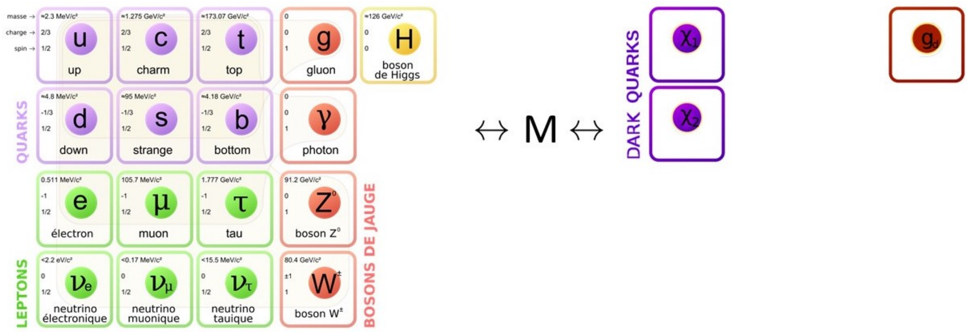


Figure 2: The Hidden valley (1)

In this work, the strongly coupling hidden sector is assumed to be composed of dark quarks with a coupling constant $\alpha_d = \frac{g_d^2}{4\pi}$ with g_d is the strength coupling constant. The dark quarks can form bound states, ie. dark hadrons. Like in the SM, a confinement scale is assumed in which the strong coupling constant defined perturbatively diverges. As a consequence, the dark hadrons production is done for energy smaller than the confinement scale Λ_d. The quarks with $pT > \Lambda_d$ will hadronize and produce a shower of dark hadrons exactly in the same way as in the SM.

As in the SM, the dark fermions can take two different states and are designed by $\chi_a = \chi_{1,2}$. The dark Lagrangian can be written in the same way as the QCD Lagrangian in term of the dark quarks and the dark gluon field strength tensor ($F^d_{\mu\nu}$):

$$\mathcal{L}_{QCD} = -\frac{1}{4}G^{\mu\nu}G_{\mu\nu} - \overline{\psi}(i\not{D} - m)\psi \tag{1}$$

$$\mathcal{L}_{dark} = -\frac{1}{4}F^d_{\mu\nu}F^{d\,\mu\nu} - \overline{\chi}_a(i\not{D} - M_{d,a})\chi_a \tag{2}$$

The first term represents the kinetic term of the dark gluon and the second term is the fermion Dirac Lagrangian with $M_{d,a}$ the mass of the dark fermion. This similarity between the two Lagrangian reflects the fact that the dark sector will behaves exactly in the same way as QCD in the SM. All the QCD physics is then well refined inside the dark sector, especially the hadronisation and the formation of jet.

The theory assumed also a portal between the two sectors. The mediators of the Hidden Valley allow the communication of the visible and the dark sector. The interaction between the SM and the DM sector can be seen as an effective theory in which the interaction is modelled by a contact operator. This means that the mediator is integrated out only at high energy. At higher energy, the UV completion occurs which means that it becomes possible to describe the interaction in a more general way with the exchange of the mediator. The figure 3 illustrates the different possibilities for the UV completion. In this work, the case of the t-channel is investigated. It corresponds to the collision of two quarks trough the mediation of the ϕ mediator. The result is the production of two dark hadrons corresponding to two dark jets. The search of at least two semi-visible jets in the data will be necessary in order to access this process.

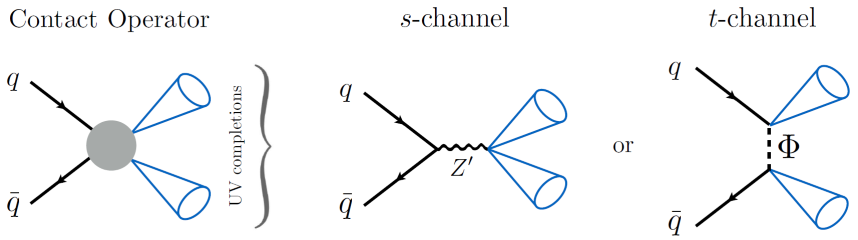


Figure 3: Different portals possible (2)

In this effective theory, the Lagrangian for the t-channel reduces to an effective Lagrangian of dimension 6 where c_{ijab} encode the flavor possible structure and Λ is the characteristic scale of the contact operator (beyond this scale, the mediator have to be integrated out):

$$\mathcal{L}_{contact} \supset \frac{c_{ijab}}{\Lambda^2}(\bar{q}_i\gamma^\mu q_j)(\overline{\chi}_a\gamma_\mu\chi_b) \tag{3}$$

2.2 Semi-Visible jet

Starting from the proton of the LHC, two visible quarks can go to the dark sector through a mediator into two dark hadrons. These two quarks will hadronize and form bound state. The figure 4 represents these interaction for the s-channel.

There are two different dark quark flavours. The two different dark quarks flavours can form 6 different hadrons Π^+, Π^-, Π^0 and ρ_+, ρ_-, ρ_0. The dark sector must by definition constitutes the dark matter. As a consequence, it is necessary that some of these bound states remain stable and do not decay into the SM sector in order to explain the current state of the DM problem. The stability of these dark hadrons depends on the gauge symmetry group of the hidden sector. Only the hadron ρ_0 is unstable and decays to the visible sector. It conducts to a jet formed of visible quarks (coming from the unstable dark hadrons) and of dark quarks (coming from the stable dark hadrons). These kind of jets are called semi-visible jets.

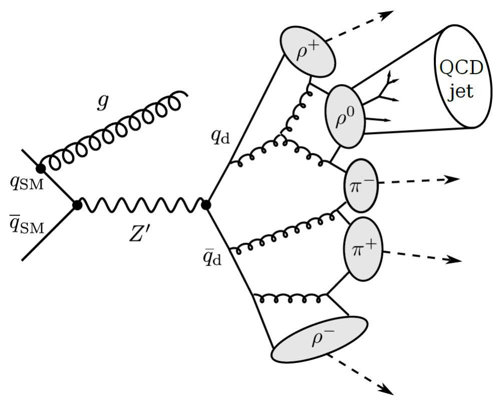


Figure 4: Dark shower from the s-channel (3)

In the figure 5, the different cases are shown depending on the stability of the dark hadrons. The semi-visible jet (in the centre) are characterised by a visible and dark component. In the case where all the dark hadrons are unstable, the jet decay totally in the visible sector becoming a QCD jet. In the case where all the dark hadrons are stable, the jet stays in the dark sector. The dark hadrons are not directly detected and are inferred as missing transverse energy (MET).

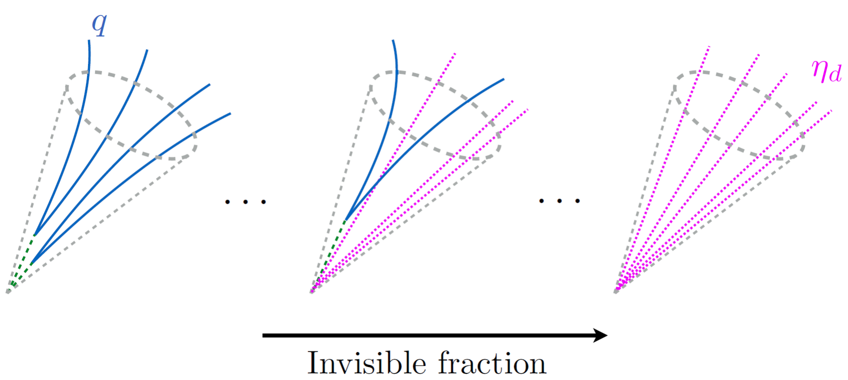


Figure 5: Dark sector parton shower (2)

A way to quantify these jets is them to compare the number of stable dark hadrons over the number of total dark hadrons which is introduce trough the parameter r_{inv} :

$$r_{inv} = \frac{\#\ of\ stable\ dark\ hadrons}{\#\ of\ dark\ hadrons} \tag{4}$$

If all the dark hadrons decay in the visible sector, then $r_{inv} \to 0$. And if all the dark hadrons are stable, then $r_{inv} \to 1$. For the semi-visible jet, r_{inv} is comprised between 0 and 1.

2.3 Signature

Looking for the t-channel, two jets will be emitted in opposite directions. Different possibilities can occur depending on the parameter r_{inv} of the jets (figure 6).

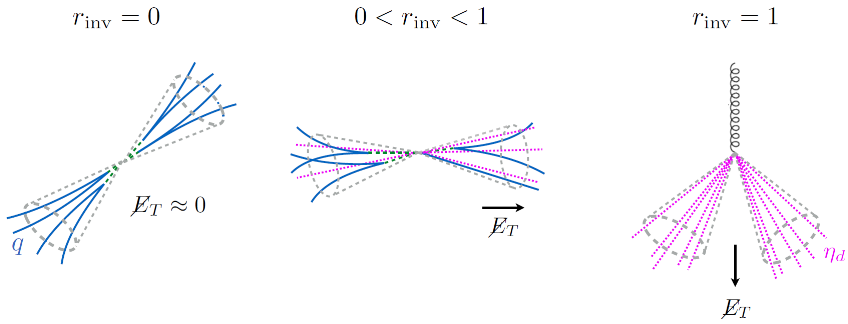


Figure 6: CharacteriDark sector parton shower (2)

By definition, all the semi-visible jets will contain missing transverse energy coming from the dark particle. If the two jets have not the same composition and so the same r_{inv}, one jet will have less pT (subleading jet) than the other (leading jet). This difference of pT compared to the leading jet will be compensated in the dark sector. As a consequence the subleading jet will have more MET than the leading jet. This gives a unique signature for semi-visible jet, MET will always be aligned with the subleading jet.

Only unbalanced semi-visible jets can have a non zero MET. In general, the two jets will have the same missing transverse energy and the total cancelled out. These semi-visible jets look like SM jet events and are as a consequence not interesting. Indeed, pure SM jet do not contain any MET in general (MET from neutrinos are coming from the decay of W, Z or top bosons) and if the SM jet contains it, there is no correlation between the jet and MET.

Note that when $r_{inv} = 0$, all the dark quarks decays in the visible sector and there is no missing energy. In the other extreme case, $r_{inv} = 1$, there is no missing energy except if an Initial-State-Radiation is emitted before the collision so that the missing energy can be inferred from the jet produced by the ISR (figure 6).

3 Data analysis

3.1 Generation of the data

The signal of this dark sector is simulated using the Hidden Valley module of Pythia8 for the generation of the dark shower and for the hadronization. Then it is necessary to simulate the response of the detector to these events. This is done using the DELPHES detector simulator. However, to study the properties of the dark events, the passage trough DELPHES can be skipped.

3.2 Basics properties

As seen below, the data for the signal need to be constituted of two semi-visible jets. Because of the fluctuations and the second order process, other jets will be emitted in addition. A pT cut is then required to not consider jets with a pT too low to be detected.

In order to see the prediction of the theory for the signature of the semi-visible jet, it is useful to plot basic properties of the signal events. The figure 7 show the distribution of missing transverse energy and its azimuthal angle distribution on the figure 8.

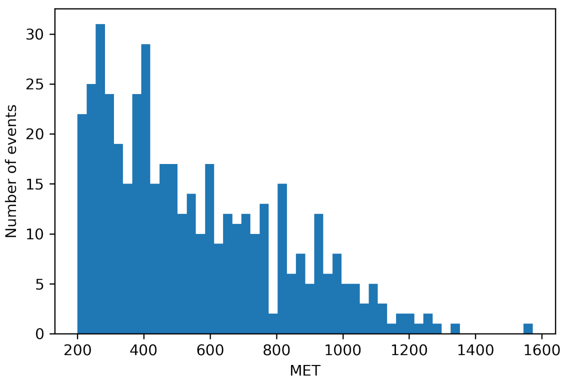


Figure 7: Distribution of MET

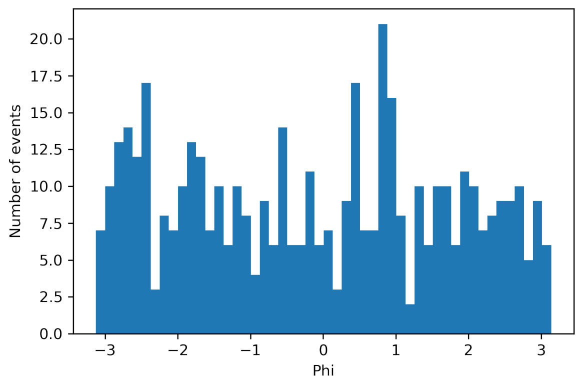


Figure 8: Distribution of phi of MET

The azimuthal angle between the leading jet and MET (figure 10) is picked at pi which is well in agreement with the prediction. The leading jet is emitted opposite to the subleading jet containing MET.

The subleading jet (figure 11) is aligned with MET as predicted (figure 12).

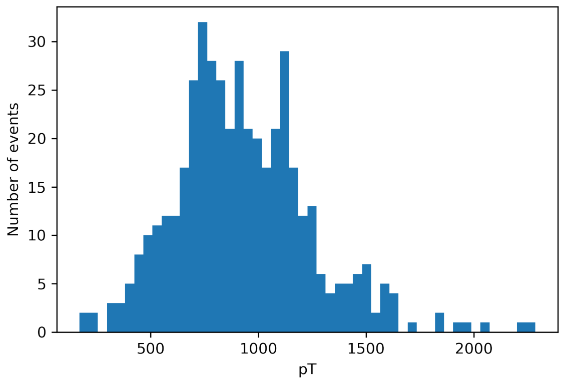


Figure 9: Distribution of the leading jet pT

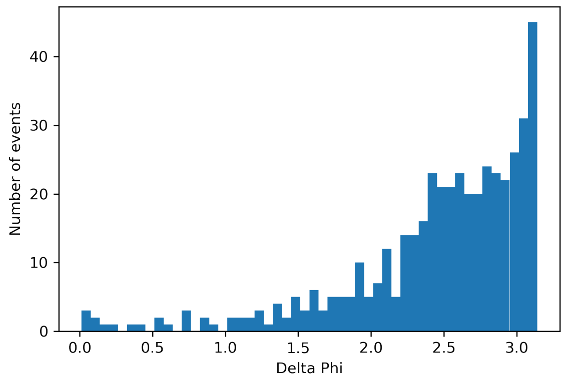


Figure 10: Distribution of delta phi between MET and the leading jet

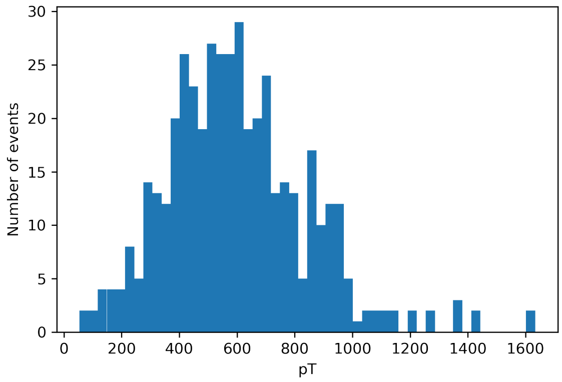


Figure 11: Distribution of the subleading jet pT

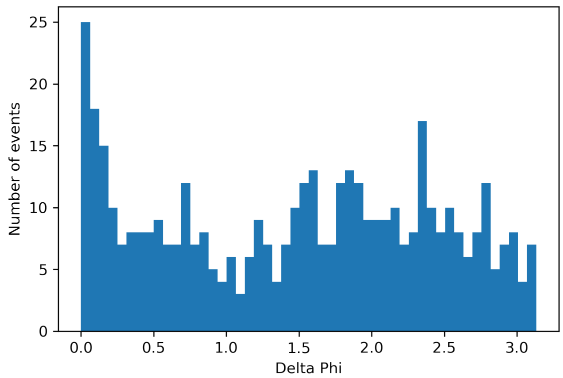


Figure 12: Distribution of delta phi between MET and the subleading jet

3.3 Comparaison

The background events are constituted of events with multijets (at least two jets) coming from SM interaction. Jets events coming from QCD interactions are made of the hadronization of quarks and gluons (colour charged particles). The machine learning algorithm will be used to distinguish them from the semi-visible jet.

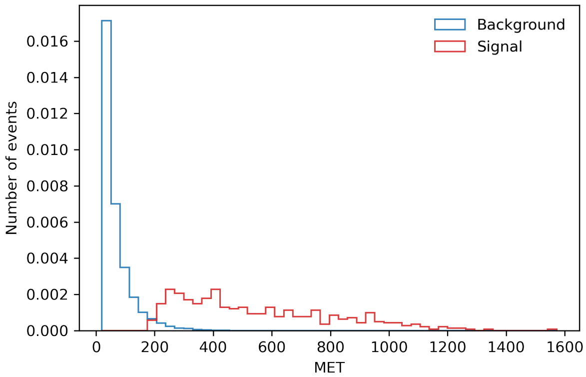


Figure 13: Distribution of MET

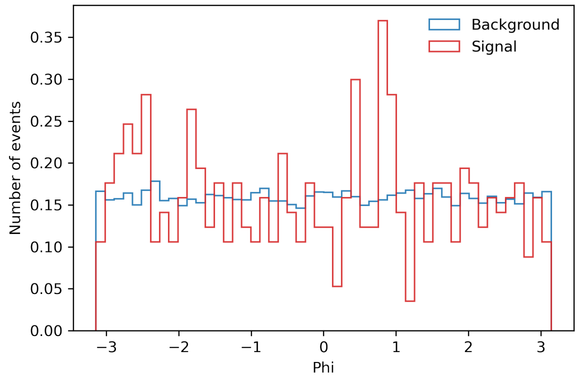


Figure 14: Distribution of phi of MET

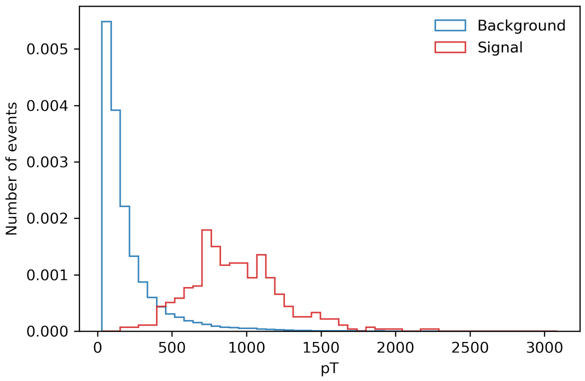


Figure 15: Distribution of the leading jet pT

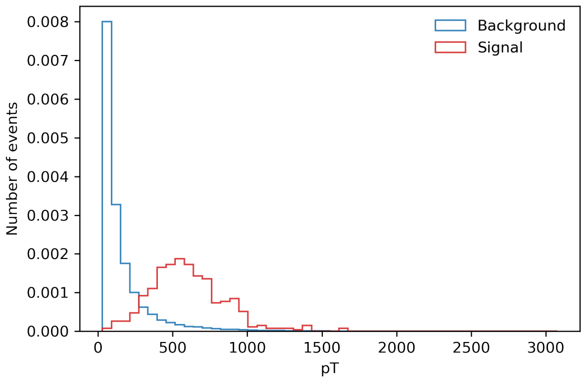


Figure 16: Distribution of the subleading jet pT

4 Machine learning

4.1 Images formation

Semi-visible jets and SM jets are very similar which explain why they do not have been observed yet. It is therefore necessary to investigate the problem through machine learning. Supervised machine learning algorithms are used when the data are labelled. This means that the algorithm can have the direct correct answer during the training part. In this project, the algorithm start from simulated data from the Hidden Valley module from Pythia. The machine learning algorithm is so supervised.

Note that it will not be possible to apply an unsupervised learning with semi-visible jet directly to data from the LHC. Indeed, alignments of MET with one jet inside the data occur in SM events. They are classified as miss reconstruction of the jet due to some miss measurement of the detector and are thrown away. This show even more the difficulty of finding semi-visible jet. It justified the used of machine learning algorithm and the used of supervised learning on simulated data first.

The machine learning part will be applied to images of jets. The jets images consist of a distribution of the transverse momentum of each constituent of the jets into the azimuthal angle (y axis) and the pseudorapidity (x axis) plane. This gives an 2D image with transverse energy deposition as pixel intensity.

Before applying the machine learning algorithm on these images, it is necessary to prepossess the images. This part makes sure that all the images have the same representation. It will allow the algorithms to work more efficiently.

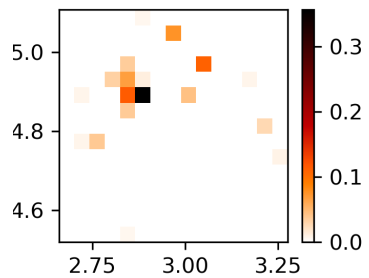
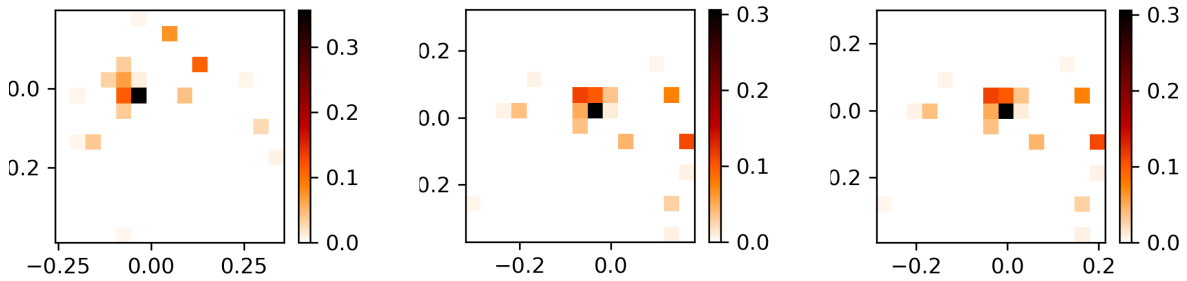


Figure 17: Jets image: BG00 event

4.2 Preprocessing

The preprossessing part is done in three steps. The first step is a displacement of the centre of the jet to the coordinate (0,0). The centre of the jet is calculated taking into account all the transverse energy of each jet. This step does not change the distribution of transverse energy of the constituents.



(a) Step 1 (b) Step 2 (c) Step 3

Figure 18: Preprocessing steps

The second step is a rotation in the phase plane. The two constituents with the highest transverse energy must be parallel to the phi axis. The rotation is done in such a way that the second highest particle is on the negative side of the highest particle. On figure 18b, the result of the rotation has redistributed the constituents. This is because the grid is rotated inside the phase plane which lead to a redistribution of all the constituents (figure 19).

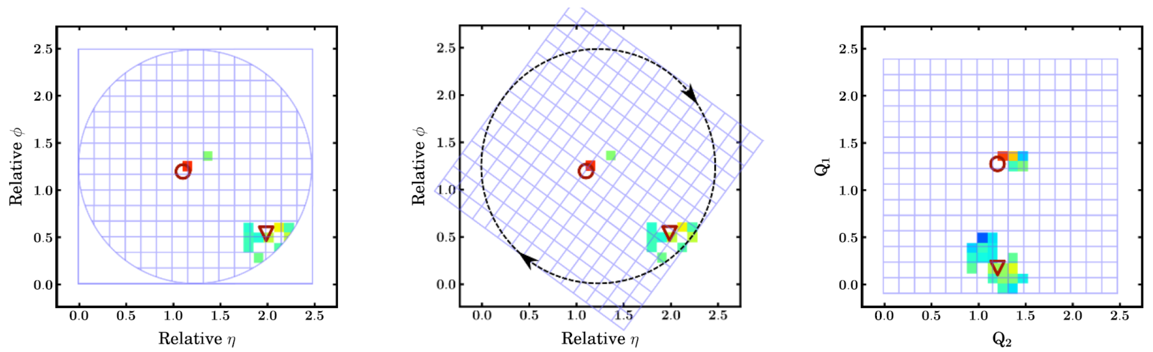


Figure 19: Rotation of the grid (4)

The third step places the origin on the highest transverse energy particle (figure 18c).

All the images has been plotted with a normalisation factor. This makes sure that the comparison for the machine learning part will be made on the same kind of intensities for each image. In this project, the pT of each constituents is divided by the sum of all the constituents inside the jet. The result of the step 3 without normalisation is shown in the figure 20.

In order to see the effect of the preprocessing part, it is useful to plot the average images for all the different steps. The result of the preprocessing is shown on figure 21d for the background events and on figure 22d for the signal events. The goal is clearly to have a systematic pattern in all the images, a vertical behaviour.

The images need to be produced in colour and in black and white without any colour-bar. Indeed the CNN will learn what is present in the images and no important information are contained in the axis and the colour-bar. The black and white images are used to train easily the machine learning part. The next figures show the final images. For

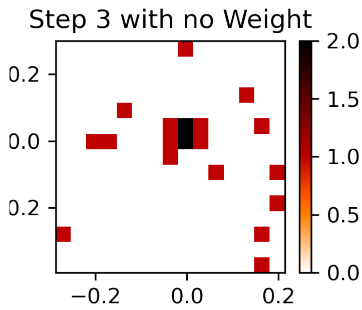
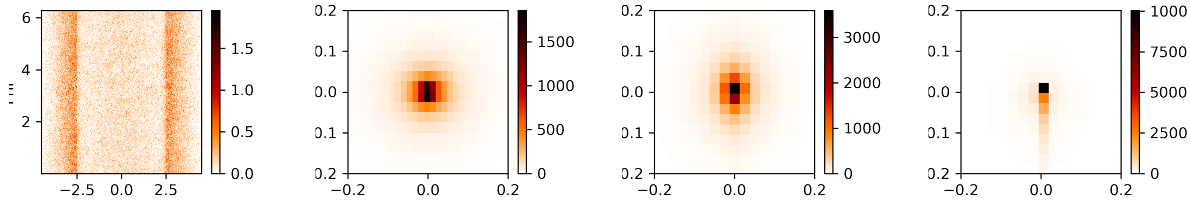
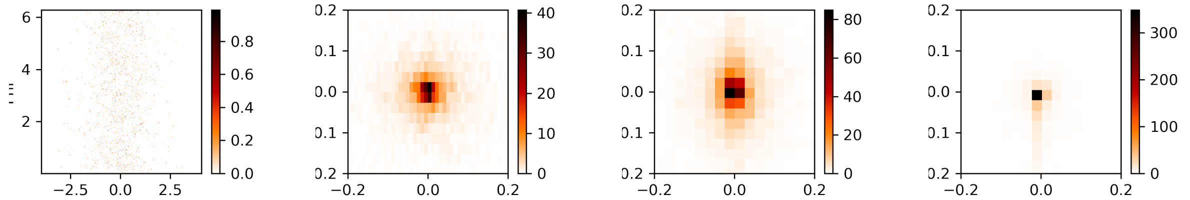


Figure 20: Preprocessing: Step 3 without normalisation



(a) Step 0 (b) Step 1 (c) Step 2 (d) Step 3

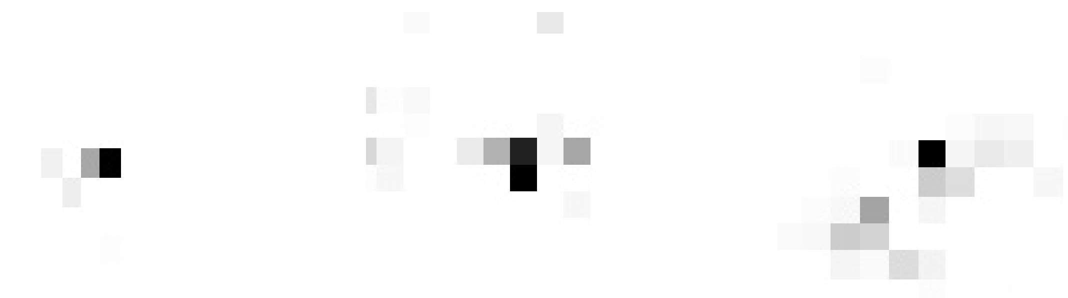
Figure 21: Average image for background images



(a) Step 0 (b) Step 1 (c) Step 2 (d) Step 3

Figure 22: Average image for signal images

the signal events, 917 images have been produced. The background events have to have the same pT range as the pT of the signal events (the range chosen is 300 to 700 GeV) and lead to around 2000 images. The sets of images have to be constituted of the same amount of background and signal images. In total, 1834 images are used for the machine learning.



(a) SIG00 (b) SIG03 (c) SIG05

Figure 23: Final images

4.3 Convolutional Neural Network

Convolutional Neural Network allowed the classification of images. This type of neural network is divided into two parts: the convolution and the classification. The neural network will find some features in semi-visible jets using convolution operation. The CNN will then classify and label all the images in input into background or signal jet according to these features. The goal is to build the best model so that semi-visible can be distinct from SM jets.

The images are separated into 80% train and 20% validation set. The model runs on two classes: background class and signal class. The model used is shown on figure 24.

```
Model: "sequential_41"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_128 (Conv2D)          (None, 15, 15, 16)        448
_________________________________________________________________
max_pooling2d_117 (MaxPoolin (None, 7, 7, 16)          0
_________________________________________________________________
conv2d_129 (Conv2D)          (None, 7, 7, 32)          4640
_________________________________________________________________
max_pooling2d_118 (MaxPoolin (None, 3, 3, 32)          0
_________________________________________________________________
conv2d_130 (Conv2D)          (None, 3, 3, 64)          18496
_________________________________________________________________
max_pooling2d_119 (MaxPoolin (None, 1, 1, 64)          0
_________________________________________________________________
dropout_40 (Dropout)         (None, 1, 1, 64)          0
_________________________________________________________________
flatten_41 (Flatten)         (None, 64)                0
_________________________________________________________________
dense_82 (Dense)             (None, 128)               8320
_________________________________________________________________
dense_83 (Dense)             (None, 2)                 258
=================================================================
Total params: 32,162
Trainable params: 32,162
Non-trainable params: 0
_________________________________________________________________
```

Figure 24: Sequential model

The principle of the convolution operation is shown on figure 25. A filter, the kernel, is displaced from the right top corner of the image to the left bottom corner applying systematically the operation. This enhances some features in the image.

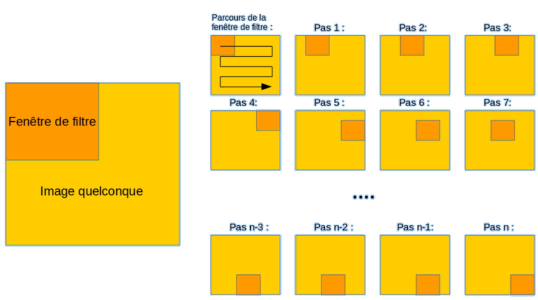


Figure 25: Principle of the convolution operation on images (4)

After a layer of convolution, a layer of pooling is used. This layer is used to reduce the size of the images keeping the most important information find by the layer of convolution. The figure 26 shows the effect of the max-pooling layer.

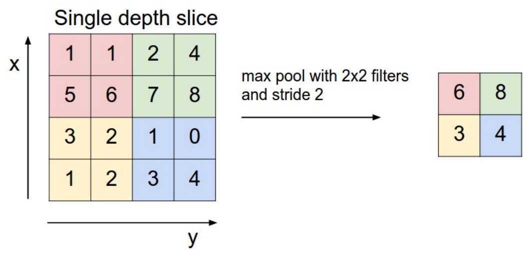


Figure 26: Maxpooling convolution (4)

Overfitting is a challenging problem encounter in ML. The parameters are fitted for this specific set of training data and does not work on the validation data. In order words, the model is not generalisable. A dropout layer can be introduced in the model. It will randomly dropout some output of the model in the training part. Doing so, overfitting can be decreased.

The layer flatten is used to make an 1-dimension map of the features for the dense layer. This last one initialise the connected network and the classification of images.

The result of this model is shown on figure 27 and 28. The validation accuracy goes down and the validation loss goes up for the last iterations. The model needs to produce the inverse behaviour. The problem encounter is overfitting.

Figure 27: Loss of the model

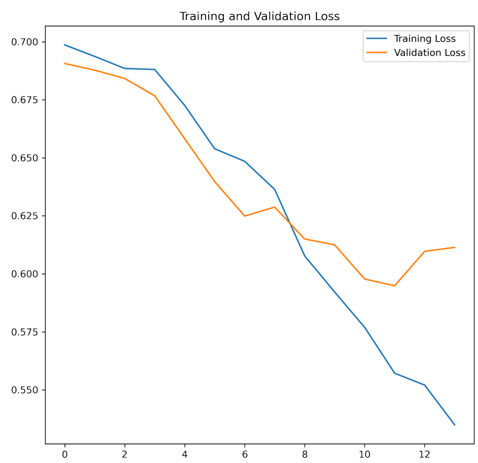


Figure 28: Accuracy of the model

The model used here is a binary classifier. A way to evaluate a binary model is to plot the Receiver Operating Characteristic curve (ROC curve). The ROC curve represents the true positive rate as a function of the false positive rate. In this model, the true positive rate is defined as labelling a background image as background and the negative rate as labelling a background image as signal. The model has to be found in such a way to have a ROC curve as close as possible to 1 in the true positive rate part.

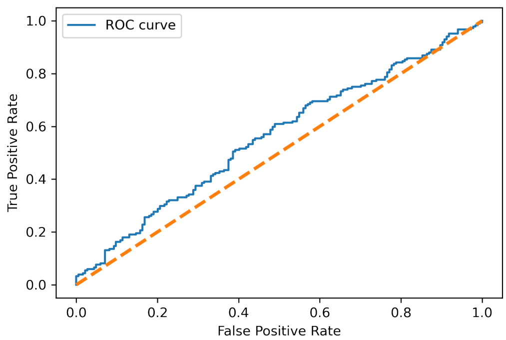


Figure 29: ROC Curve

This model needs to be run on several iterations. An epoch is an iteration on the whole model. If the model runs on too few epochs, it will reach a local minimum and not the global minimum of the derivative of the loss function. In order to find the global minimum, the model need to be run on at least 50 epochs. Doing so, the result of the current model is even worse. This model is a consequence not good and need to be improved. As the project is going to its end, the model could not be investigated in more details. An ideal result is shown on figure 30 and 31. The accuracy increase and loss decrease for the validation set.

Figure 30: Accuracy of the model

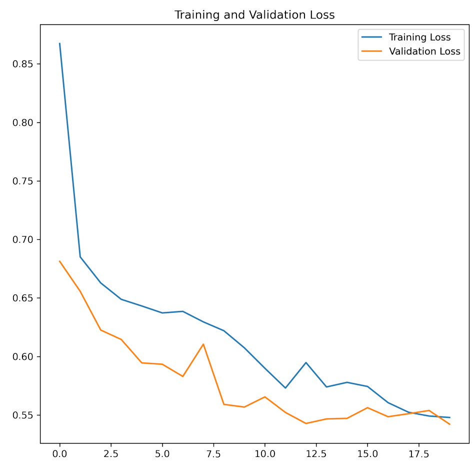


Figure 31: Loss of the modell

5 Conclusion

In this project, the dark matter problem has been investigated. The Hidden Valley theory predicts the existence of the so-called semi-visible jet. Semi-visible jets are composed of visible and invisible hadrons. The alignment of MET with the subleading jet gives a unique signature and offers a new way to look for dark matter. However the difficulty to find out such signature from SM jets impose the used of machine learning method.

Jets images are images of all the constituents in the the azimuthal and pseudorapidity plane. These images have been produced for semi-visible jets and for usual jets. In order to make the learning process more efficient, all of the images have been preprocessed. In this part, all the images are manipulated to have a systematic behaviour in the images.

Finally, these images have been used for the Convolution Neural Network. The model that have been built is not a good model and need to be improved. The ROC curve of the model has to be as close as possible to 1 in the y axis. This condition is necessary in order to apply the model on real data from the LHC and hope to find traces of semi-visible jets. This work has to be followed in future searches.

Acknowledgements

Thank to CERN to have organised this summer student 2021 in this pandemic context and for offering such unique opportunity to students from all around the world. I would like to thank my supervisors Deepak Kar, Sukanya Sinha and Tasnuva Chowdhury to have supported me in this work. This project has been really interesting and inspiring, many thanks to you Deepak, Sukanya and Tasnuva.

I want also to thank the administration team for making this summer student as close as possible to an on-site internship and for their interaction with the students. Many thanks to Dominique Lambert and Michele Lucente for their trust and their reference letter, making this dream possible.

References

[1] Colin Fallon. Searching for Dark Matter with Semi-Visible Jets at CMS - Workshop, August 2019.

[2] Timothy Cohen, Mariangela Lisanti, Hou Keong Lou, and Siddharth Mishra-Sharma. LHC Searches for Dark Sector Showers. J. High Energ. Phys., 2017(11):196, November 2017. arXiv: 1707.05326. doi:10.1007/JHEP11(2017)196.

[3] Elias Bernreuther, Felix Kahlhoefer, Michael Krämer, and Patrick Tunney. Strongly interacting dark sectors in the early Universe and at the LHC through a simplified portal. J. High Energ. Phys., 2020(1):162, January 2020. arXiv: 1907.04346. doi:10.1007/JHEP01(2020)162.

[4] Josh Cogan, Michael Kagan, Emanuel Strauss, and Ariel Schwarztman. Jet-images: computer vision inspired techniques for jet tagging. J. High Energ. Phys., 2015(2):118, February 2015. doi:10.1007/JHEP02(2015)118.

[5] Luke de Oliveira, Michael Kagan, Lester Mackey, Benjamin Nachman, and Ariel Schwartzman. Jet-Images - Deep Learning Edition. J. High Energ. Phys., 2016(7):69, July 2016. arXiv: 1511.05190. doi:10.1007/JHEP07(2016)069.

[6] Deepak Kar. Experimental Particle Physics - Understanding the measurements and searches at the Large Hadron Collider. IOP Expanding Physics, August 2019. doi:10.1088/2053-2563/ab1be6.

[7] Sinha Sukanya. Dark matter search using semi-visible jets - HEPP Workshop, January 2020.

[8] Deepak Kar and Sukanya Sinha. Exploring Jet Substructure in Semi-visible jets. SciPost Phys., 10(4):084, April 2021. arXiv: 2007.11597. doi:10.21468/SciPostPhys.10.4.084.

[9] Torbjorn Sjostrand. Introduction to Pythias HV - Workshop, November 2019.

[10] G. Kasieczka, T. Plehn, A. Butter, K. Cranmer, D. Debnath, B. M. Dillon, M. Fairbairn, D. A. Faroughy, W. Fedorko, C. Gay, L. Gouskos, J. F. Kamenik, P. T. Komiske, S. Leiss, A. Lister, S. Macaluso, E. M. Metodiev, L. Moore, B. Nachman, K. Nordstrom, J. Pearkes, H. Qu, Y. Rath, M. Rieger, D. Shih, J. M. Thompson, and S. Varma. The Machine Learning Landscape of Top Taggers. SciPost Phys., 7(1):014, July 2019. arXiv: 1902.09914. doi:10.21468/SciPostPhys.7.1.014.

[11] Classification des images | TensorFlow Core. URL: https://www.tensorflow.org/tutorials/images/classification?hl=fr.

[12] Convolutional neural network, June 2020. URL: https://datascientest.com/convolutional-neural-network.