Physics at the Large Hadron Collider: A Data Analytic Approach

Krish Vijayan (2019PH10634)
Ishaan Watts (2019PH10629)

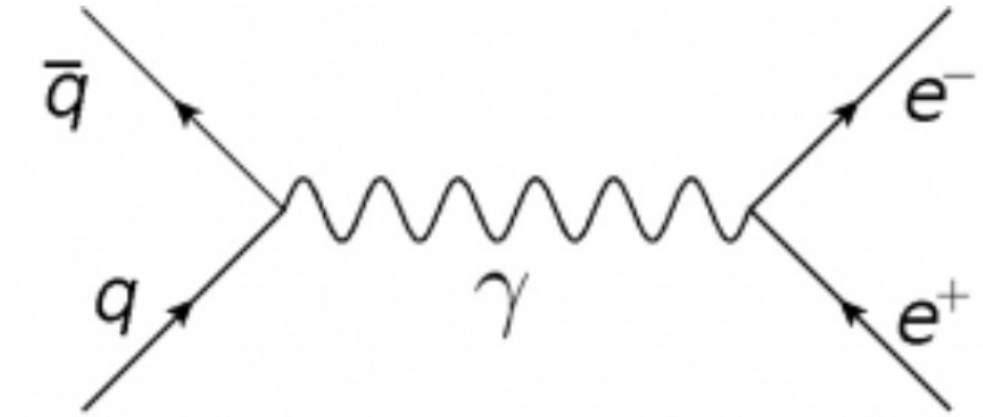
Advisor: Prof. Abhishek Muralidhar Iyer

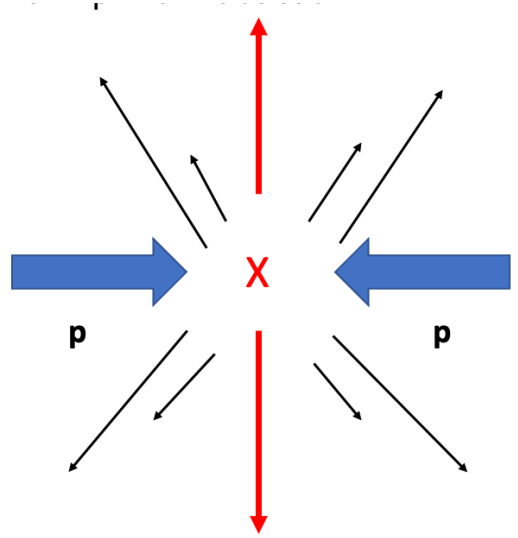
- Large Hadron Collider
- Problem Statement
- Generation of Dataset
- Recap
- Model Architecture
- Results and Discussion
- Conclusion

Large Hadron Collider

Typical Process in the LHC

A typical process in the Large Hadron Collider which contributes to the dataset used in this project is the collision of a quark (q) and an anti-quark (q′) to become a photon (γ) which further decays into an electron (e-) and a positron (e+) in the final state.





Final states are four-vectors

Problem Statement

Anomaly Detection

- Basic idea of LHC is to compare 2 datasets. How can one reject one dataset in favor of another.
- The discovery of an expected signal, such as the Higgs Boson, is associated with a p-value. The trained model should ideally predict whether a given event is an anomaly with a high probability based on features such as the energy and momenta of different kinds of particles.

Generation of Dataset

Variables

In order to reject one dataset in favour of the other, the difference in pseudo-rapidities of the final state particles is considered to be a distinguishing feature:

$$\eta = -ln(tan(\frac{\theta}{2})). \qquad \Delta\eta = \eta_1 - \eta_1'$$

Another important metric which is relevant to the classification is the Signal Discovery Significance (Z), given by:

$$Z = \sqrt{\sum_{i=1}^{N}(2(s_i + b_i)log(1 + \frac{s_i}{b_i}) - 2s_i)}$$

(summed over all the bins, s_i and b_i are the expected numbers for signal and background events in the i^{th} bin)

Using the mass and momentum, the invariant mass is defined as:

$$m = \frac{\sqrt{E^2 - p^2c^2}}{c^2}$$

Dataset

Three datasets are used for different tasks, which are:

1. Signal Discovery Significance (Z): Background and a signal (bump) with 46,611 and 4,600 values of $\Delta\eta$.
2. Neural Network: Classification of 122,854 (Dataset A) and 133,619 (Dataset B) values of $\Delta\eta$.
3. Auto-encoders: Anomaly detection using 5 samples with values of invariant mass.

Data-set	Number of Values
Signal with peak at 0.5TeV	3982
Signal with peak at 1TeV	4409
Signal with peak at 2TeV	4600
Signal with peak at 3TeV	4602
Background Distribution	46611

Table 1: Invariant Mass Measurement Data-sets

For last task, more training data was also generated artificially by adding +-5% noise to the background bin values.

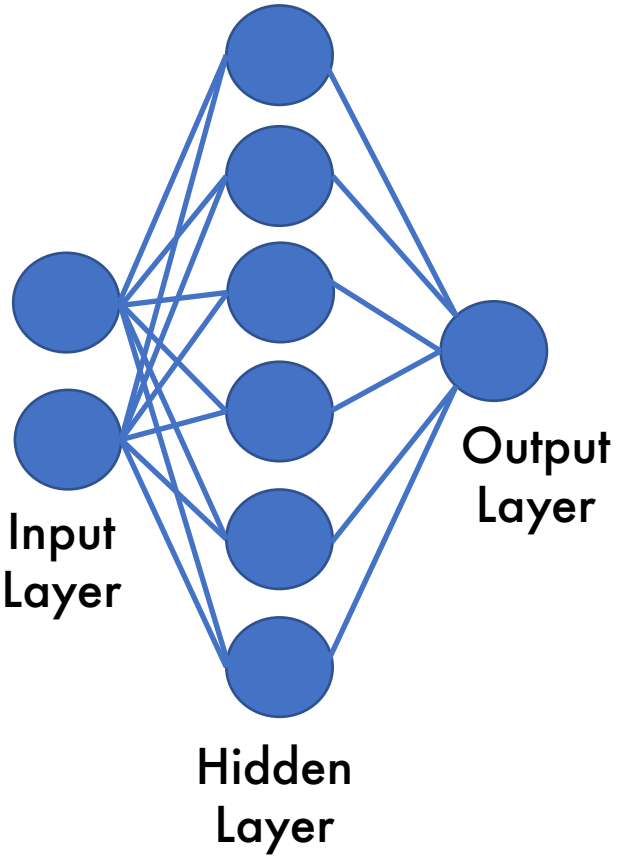
https://inspirehep.net/literature/1607893

Recap

Neural Network

Used an artificial neural network (ANN) to separate two distributions with the following parameters:

Parameter	Value
Loss Function	Binary Cross-Entropy
Activation	ReLU + Sigmoid
Optimizer	Adam, SGD, RMSProp
Epochs	200
Hidden Units	5 - 128
Learning Rate	0.001 – 0.5



$$ReLU(z) = max(0, z).$$

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

$$L(y, p) = -\sum_{i=1}^{m} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

Results

Z value in Task 1: 2183.79.

Optimizing ANN using Optuna in Task 2:

Final Parameter	Value
Loss Function	Binary Cross-Entropy
Activation	ReLU + Sigmoid
Optimizer	Adam
Epochs	200
Hidden Units	112
Learning Rate	0.036

Results	Value
Training Accuracy	96.97%
Test Accuracy	96.83%
ROC	99.73

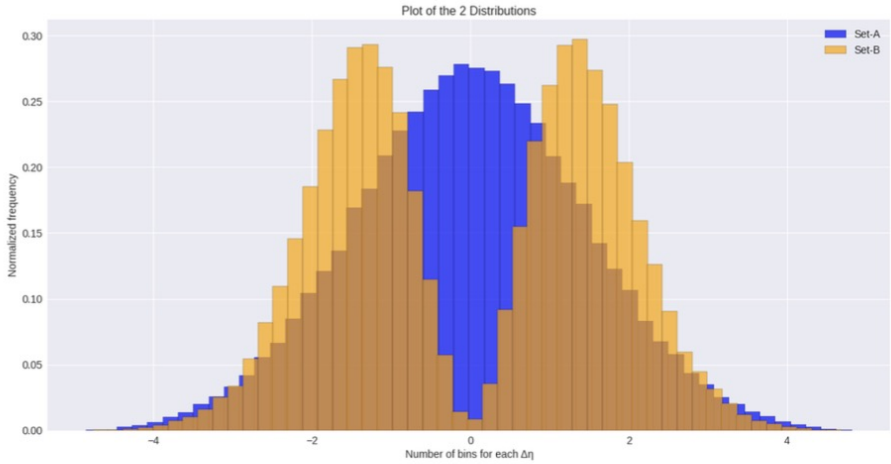


Figure 6: Histogram of the range of values of $\Delta\eta$ vs the frequency, after normalization

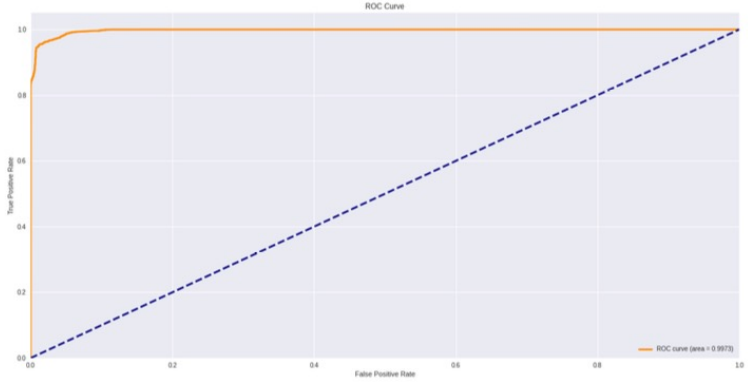


Figure 8: Plot of the *True Positive Rate (TPR)* vs the *False Positive Rate (FPR)*

Model Architecture

Auto-Encoder: Motivation

Auto-encoders: Self-supervised neural networks in which input and output are identical. Compress the input into a code with fewer dimensions and then reconstruct the output based on this representation.

Encoder, code, and decoder are the three components that make up an auto-encoder. The encoder compresses the input and generates the code, while the decoder reconstructs the input using only the code.

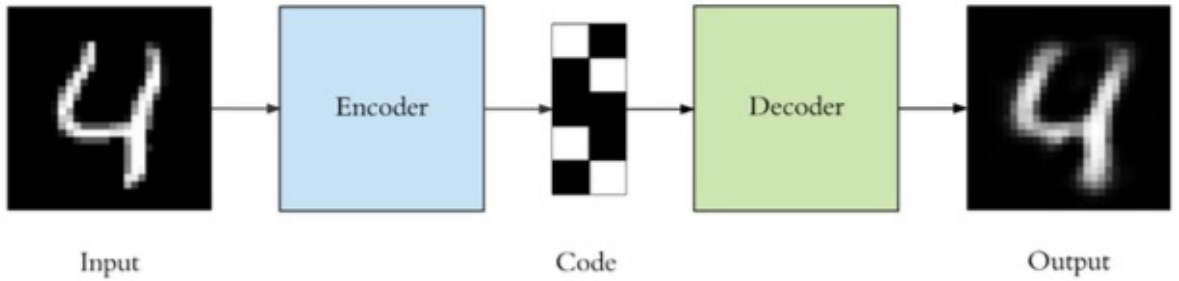
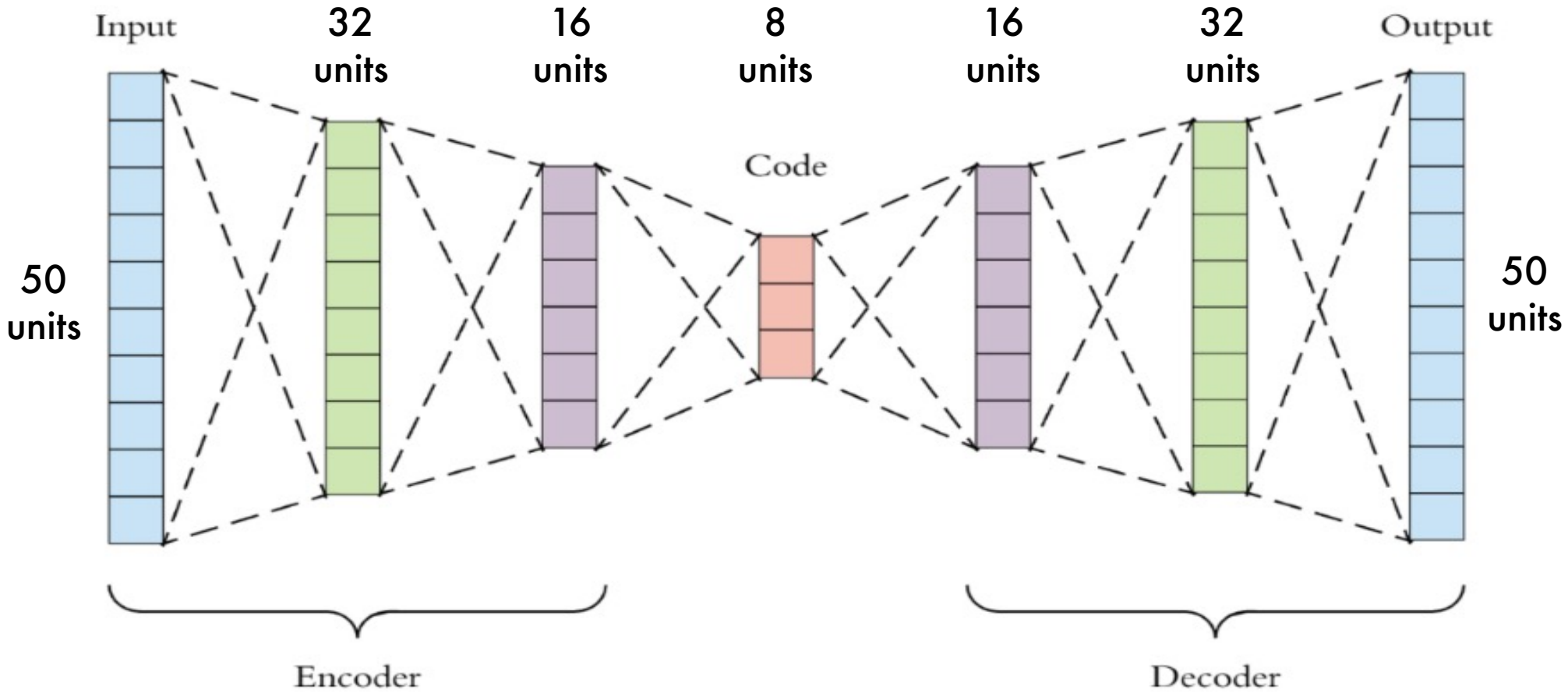


Figure 3: Auto-encoder Architecture

Auto-Encoder: Architecture

The model is trained using the following architecture:



Auto-Encoder: Architecture

Each layer also has a ReLU activation function. The training takes place over 200 epochs, with a learning rate of 10^{-3} and Adam Optimizer.

The loss function used is Mean Squared Error (MSE) given by:

$$L(y) = -\frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{50}(x_{ij} - \hat{x}_{ij})^2$$

where x_{ij} is the j^{th} bin value of original input and $\hat{x}_{ij}$ is the j^{th} bin value of reconstructed output averaged over m training examples.

Results and Discussion

Distribution Plots

For each of the signals in Table 1, the normalised histograms were plotted by grouping the values into 50 bins against the background, and the following plot was obtained:

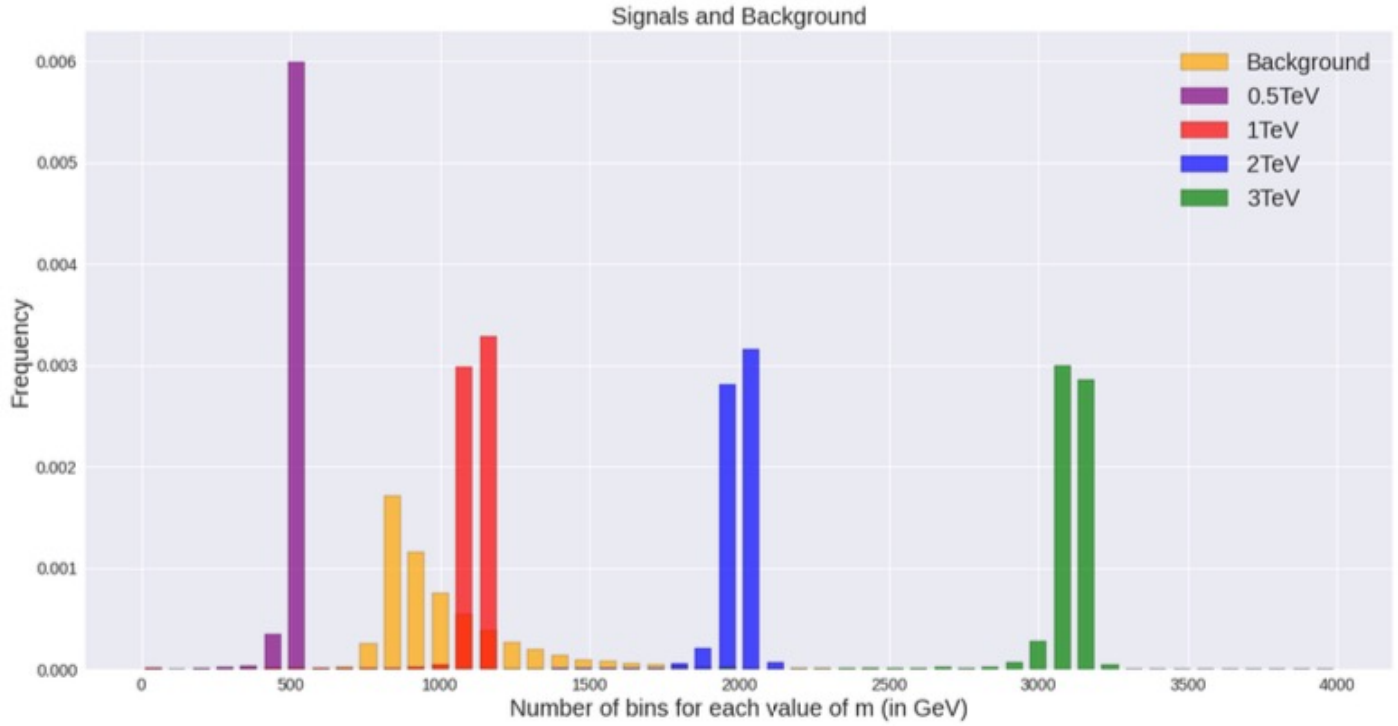


Figure 9: Signals with peaks at 0.5,1,2,3 TeV with background

Background reconstruction

Trained on the input background and the new distributions generated by adding random noise to the background. The auto-encoder learnt the reconstruction for the given data, without exactly learning the exact values of the input (which would just be an identity function).

The reconstruction MSE loss for the original background is 7.138e-09.

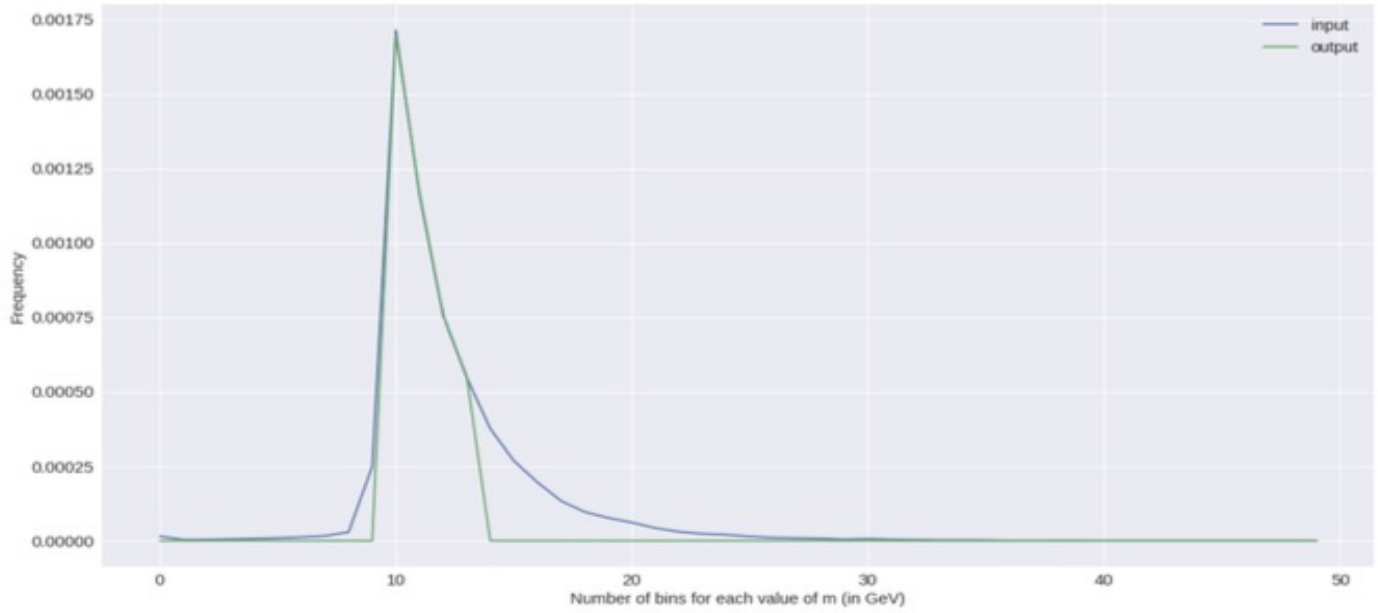


Figure 11: Background (Input) and its Reconstruction (Output)

Signal Reconstruction

Now, we reconstruct the signal distributions by feeding them as input to the model. Since the model has only learned to reconstruct the background, it is not able to reconstruct the signal distributions.

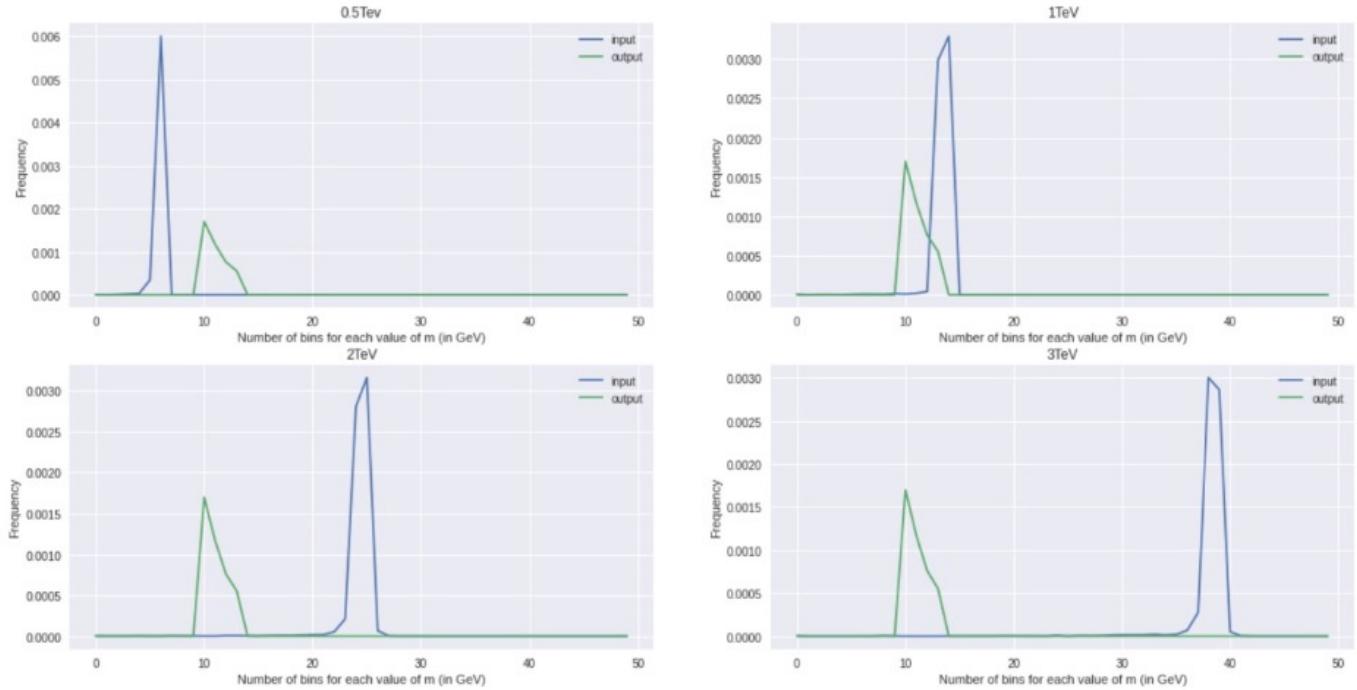


Figure 12: Signals (Input) and their reconstruction (Output)

Input	Loss
0.5TeV	8.183e-07
1TeV	4.273e-07
2TeV	4.596e-07
3TeV	4.472e-07
Background	7.138e-09

Table 3: Loss for various inputs

Re-construction losses for signals are greater than the reconstruction loss of background by a magnitude of 100.

Conclusion

Conclusion

1. A high value of Z indicates that for a sufficiently large background we can differentiate between signal and background curves.

2. We are able to distinguish between distributions with high accuracy in a Supervised Learning task using Artificial Neural Networks.

3. Using unsupervised learning methods, it is feasible to identify anomalies in the data, allowing us to identify if a z-boson was formed in the process by learning the distributions.

All the programs can be found in the PYD411 Google Colab Notebook

Thank You

QnA