**DAML Report-2**

**Files in submition:**

* Preprocess\_factory (analysis-1) (Jupyter- Notebook)
* Network\_factory ((analysis-1) (Jupyter- Notebook)
* Analysis\_1 (Jupyter- Notebook)
* Preprocess\_factory\_2 (Analysis-2) (jupyter-Notebook)
* Network\_factory (analysis-2) (Jupyter- Notebook)
* Analysis\_2 (Jupyter- Notebook)
* 9 Trained Model. (Turnitin is not accepting files above 100mb)
* I can provide preprocessed csv file, but there is same problem. (although preprocess\_factory takes 30 seconds to process file)

**Chronology:**

1. Preprocess\_factory (analysis-1)
2. Network\_factory ((analysis-1)
3. Analysis\_1
4. Preprocess\_factory\_2
5. Network\_factory (analysis-2)
6. Analysis\_2

**Preprocessing:**

In Preprocess\_factory\_1 notebook takes (csv file as input and out put processed csv file)

Data is preprocessod and made appropirate for the network, work is documented in notebook itself. The same notebook is used to preprocess all the data, So exactly loded file in notebook factory is irrrelevant. Also after first analysis, there seems some random noise in analysis, so in preprocess\_factory 2 data is further cleaned so redule the noise.

**Network:**

Network\_factory is very basic notebook. (BackgrownTrain.csv as input and trained model as output)

Which will create model in demand model. I tried various models in brainstroming phase, But finally concluded to go with (adams, mean-square error), Unfortunately turnitit is not accepting my model.h5 file. But this note book is very basic in details.

Streatgy is to analyse at least 6 trained models to see which one fits our need. I have selected 6 models with latent space varying from 5, 8, 10, 15, 20 and 30.

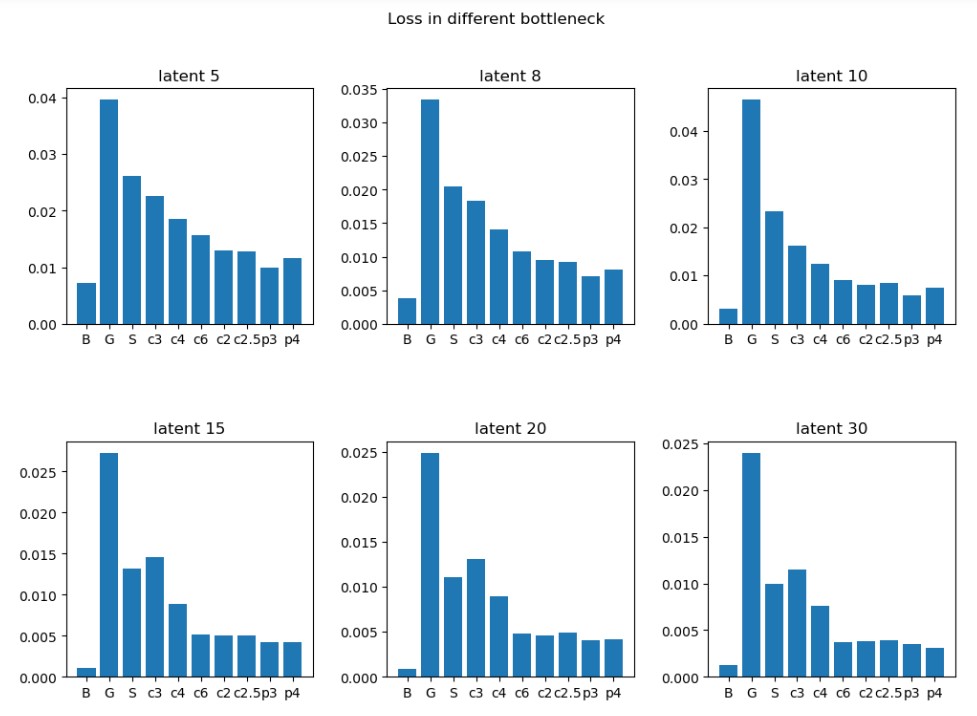
Training of model seems not very helpful in our ultimate aim, so I restrain from saving history of traning, Basicicy all models are trained with at least 50 epocs.

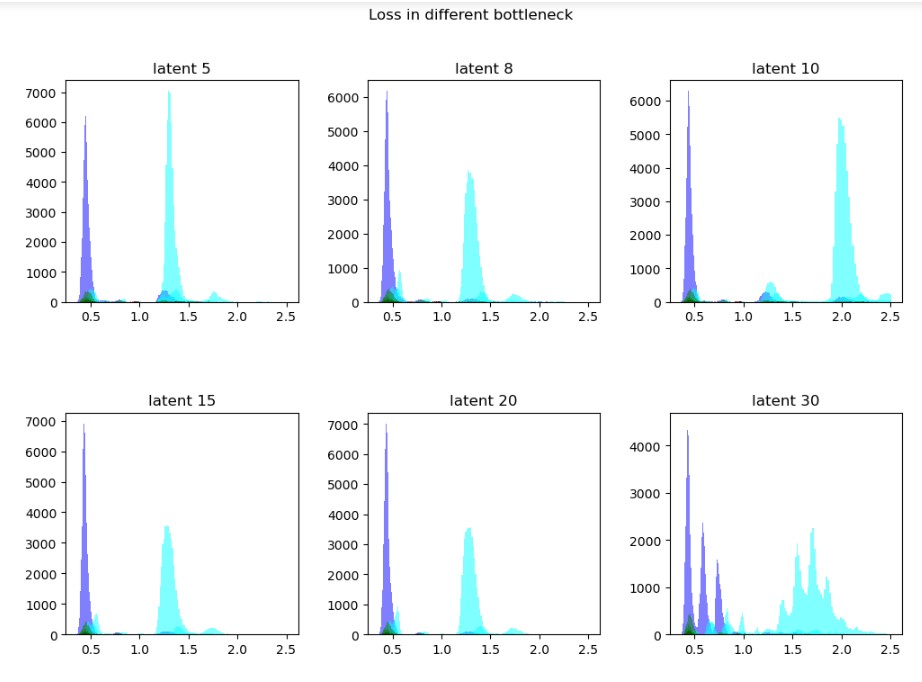
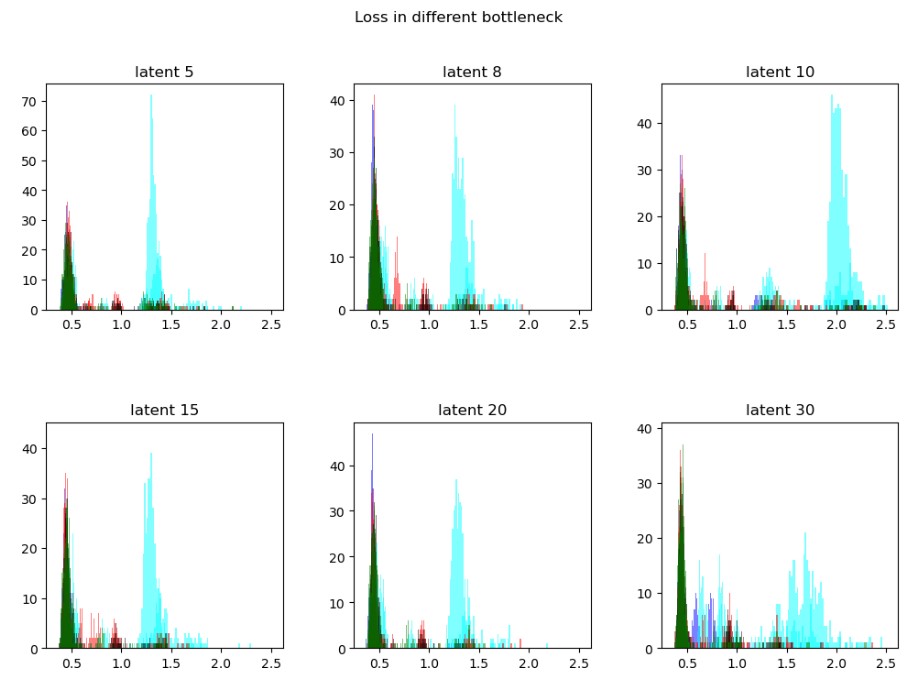
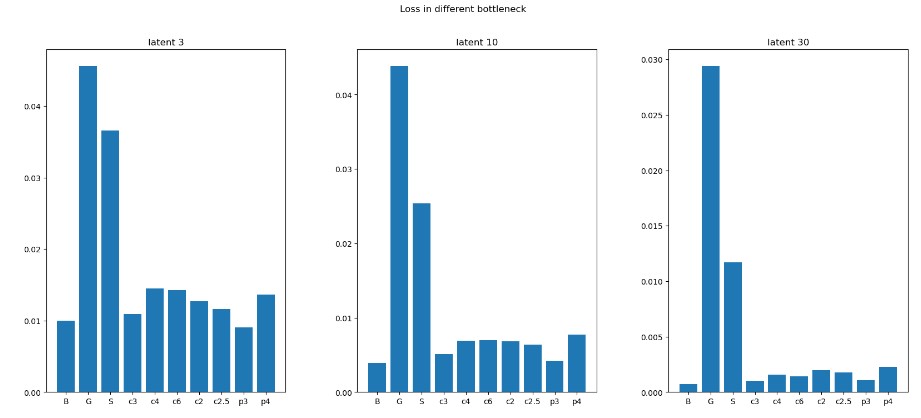
**Analysis:**

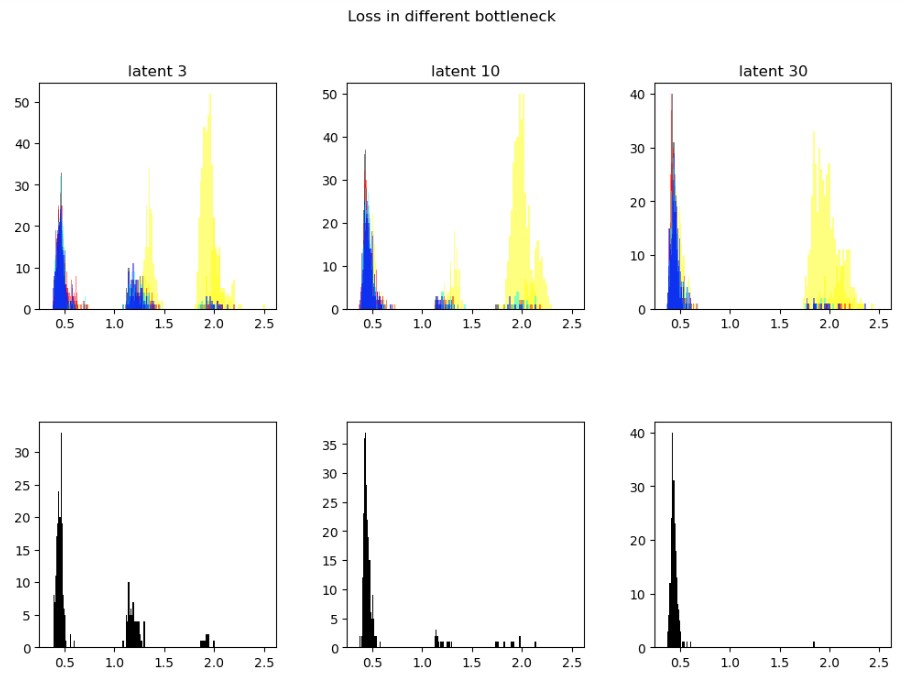
After two above factory (Notebooks) have provided, preprocess data and trained models. Analysis Notebook Analyse the performance of all the networks on all the data. Aim is to find a way to filter out a anomly events.

**Summary:**

* Finding Mean square error of each model on each data sets, Comparing results side by side.



* Finding distribution of loss of each permutaion.
* Normalising the distribution.
* Seeing to much Noise, Reprocessing data in analysis 2.
* Finding the spectrum agin on reprocessed Data.



**Conclusion:**

1. With Data preprocessing technique-1, Its more convinient to find overall error in data, While preprocess technique 2 is more convinient in signaling anomly events.
2. It depends on the relative strenth of BSM on the effictiveness of signal, And some models are more effective to different signal, So Its not possible to find single anomly score.
3. None of the models is working on **Gluino and RPV SUSY** Data set, Reson being this is the data set where all events have more than 8 particles(mostly). But Traning dataset hardly capture physics of those.
4. As the Latent space dimention decrese, the error spectrum divides into three gausian(see last figure) But despite sharpning of backgrown gausian, BSM gaussing is not sharpning with dimentions decrese, Which indicate that in analysis 2 with hours of traning we can clearly find anomly events.
5. As latent Space Dimentions are increased, error specturm tends to sharpen, But so does error spectrum of BSM, thus making it hard to single out BSM events. So In my conclusion, autoencoder with latent dimention 7-14 will be best to factor out BSM events.

**What I could have done if given time**

* Organisesation of work.
* More traning of network.
* Counting events in non bacgrown spectrum to give some anomly score.(although I have defined function)
* Find alpha and beta values.
* Use convolutional networks to work on Gluino dataset.

-Ritesh

S2456148