DISCRIMINATION OF QUARK AND GLUON INITIATED JETS USING MULTIVARIATE ANALYSIS

Project report submitted to Central University of Karnataka as a part of completion of

Master of Science
(Physics)

by
ATHULYA C K

2025

Department of Physics
School of Physical Sciences
Central University of Karnataka
Kadaganchi, Kalaburagi-585367
India

DECLARATION

I hereby declare that the work reported in this project report titled "Discrimination of Quarks and Gluon initiated jets using Multivariate Analysis" submitted to Central University of Karnataka, Kalaburagi, India is an authentic record of my work carried out under the supervision of Dr. Jyothsna Rani Komaragiri. It is further certified that this work is not plagiarised and has not been submitted for the award of any other degree/diploma of this University or any other institution. I further attest that this work is original and that I am fully responsible for the content of this thesis.

Athulya C K

(Enrollment No.: 23PPHYSO10)

Place: Central University of Karnataka

Date:

CERTIFICATE

This is to certify that the work reported in the project report titled **"DISCRIMINATION OF QUARK AND GLUON INITIATED JETS USING MULTIVARIATE ANALYSIS"** submitted by ATHULYA C K (Enrolment No.: 23PPHYS010) to Central University of Karnataka, Kalaburagi, India is a bona fide record of her original work carried out under my supervision.

Dr. Jyothsna Rani Komaragiri
Thesis Supervisor
Associate Professor
Center for High Energy Physics (CHEP)
Indian Institute of Science (IISc), Bangalore,
Signature:

Dr. Bharat Kumar (Head of the Department)
Associate Professor
Department of Physics
Central University of Karnataka
Signature:

Dr. Suchismita Sahoo (Head of the Department)
Local Mentor Assistant Professor
Department of Physics
Central University of Karnataka
Signature:

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Dr. Jyothsna Rani Komaragiri, Associate Professor in the Center for High Energy Physics (CHEP) at Indian Institute of Science (IISc), Bangalore, for allowing me to work under her guidance. The immense support and guidance during this period have been a great help for me. I would also like to thank Mr. Kaushik Gupta for assisting me during my early period. I am also thankful for the support I have received from my local mentor, Dr. Suchismita Sahoo, Assistant Professor, Central University of Karnataka and also for giving me the opportunity to work with Dr. Jyothsna Rani Komaragiri. I would like to acknowledge everyone who have helped me during this period.

Abstract

In this project we use Multivariate Analysis, which is a statistical method used to analyze the simulated data from CMS opendata website, to discriminate light quark jets from overwhelming gluon jets. The work focuses on the differences in quarks and gluon initiated jets, preprocessing and modeling the data and evaluating the model performance.

Contents

INTRODUCTION 4

EXPERIMENT THEORY 6

A BRIEF INTRODUCTION TO THE STANDARD MODEL 6

THE LARGE HADRON COLLIDER (LHC) 7

CMS DETECTOR 8

Electromagnetic Calorimeter (ECAL) 9

Hadronic Calorimeter (HCAL) 10

Superconducting Solenoid Magnet 10

JETS 11

MULTIVARIATE ANALYSIS IN HIGH ENERGY PHYSICS 13

Supervised learning 13

Unsupervised learning 14

THE QUARK-GLUON DISCRIMINATOR 15

NEURAL NETWORK 15

Overfitting and Underfitting 16

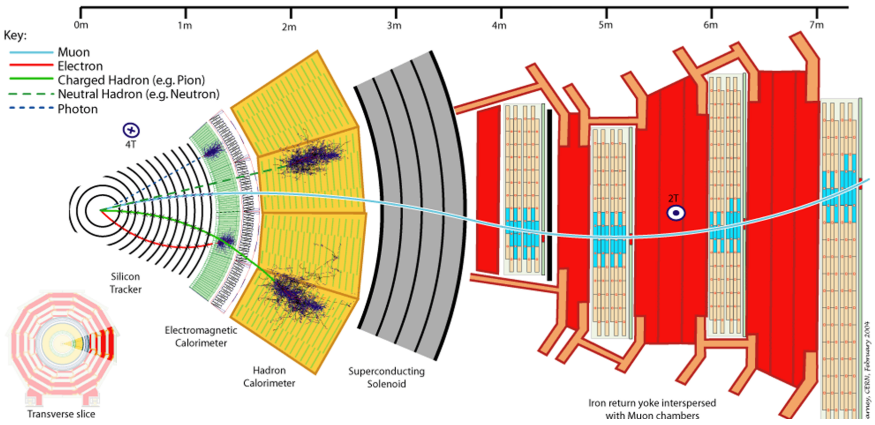
Hyperparameters 16

Forward and Backward Propagation 17

Receiver Operating Characteristic curve (ROC) and Loss function Curve 18

DATA PREPROCESSING 19
Selection of variables 19
Data Standardization 21
Handling Class Imbalance 22
Train-Test Split 22
DISCRIMINATOR 22
Model 1: Neural Network from Scratch 23
Model 2 : Keras-Based Neural Network 23
RESULTS 24
Model 1: Neural Network from Scratch 24
Model 2 : Keras-Based Neural Network 26
ROC Curve for different p_T and η regions 29
ROC Curve for all the variables 29
CONCLUSION 30

INTRODUCTION



A Transverse view of the CMS Detector showing the interaction of particles within the detector (https://doi.org/10.48550/arXiv.1706.04965)

Particle accelerators and detectors have been under constant evolution to unravel the fabrics of the universe. The Compact Muon Solenoid (CMS) at the Large Hadron Collider, plays a dominant role in this matter. CMS operates in complicated environments, capturing and reconstructing the swarm of particles created during the high-energy proton-proton collision. The fragmentation of quarks and gluons leads to the formation of jets, collimated spray of particles, which carry the key information about the physics at parton level. Distinguishing whether the jet originated from quarks or gluons is crucial for understanding the physics beyond the standard model.

Quantum Chromodynamics (QCD) is the root of jet physics. Since quarks and gluons cannot be observed directly due to color confinement, understanding jet substructure and energy distribution will give us new insights in the dynamics of QCD. Quark jets tend to be narrower with few but energetic particles, whereas gluon jets are much more broader with higher multiplicity and softer fragmentation. These differences arise due to the difference in color charge between quarks and gluons - Gluons have more color charge than quarks and hence it leads to more significant radiations.In many analyses, the dominant background

arises from gluon-initiated jets, while the signals of interest often produce quark-initiated jets. Since many rare phenomena have smaller cross-sections, isolating the signal from this overwhelming gluon background has become a challenge in high-energy physics.[1]

Early efforts in discriminating quarks and gluon jets was done by using single variable analysis. Exploiting the differences between quark and gluon jets, a number of observables were proposed to enhance the separation. This included jet multiplicity, energy depositions and angular distribution. However, the complexity of QCD radiation and high luminocity at the LHC requires more sophisticated method which can combine multiple variables for the separation. This led to the implementation of Multivariate Analysis (MVA), which uses the correlation between these variables to enhance the performance of the classifier.

MVA is used to combine multiple observables to identify subtle patterns which otherwise go unnoticed in single variable analysis methods and hence provide a better distinction between quarks and gluon initiated jets.[2] Early multivariate approaches included linear discriminants and simple boosted decision trees, but the complexity of data has opened the way for advanced machine learning (ML) methods, such as neural networks and deep learning architectures like convolutional and graph neural networks, which have been used to develop powerful classifiers that distinguish quark jets from gluon jets with great accuracy. MVA also makes it easy to analyze datasets which have high dimensions. The algorithm is trained on simulated datasets which gives us the optimal combinations of observables which maximize the separation between quark and gluon jets. Once trained, the model can be used on experimental data for signal to background discrimination. By interpreting the features in MVA models, researchers gain insights into the underlying QCD processes, such as the mechanisms governing parton showering and hadronization.

In the subsequent chapters of this thesis, we will focus more on the experimental set up of CMS detector, an overview of jet physics, and the differences between quarks and gluon jets. We will then focus in details about the Multivariate Analysis and then finally about the quark-gluon discriminator build using neural networking and talk about how we train, validate an test the model.

EXPERIMENT THEORY

A BRIEF INTRODUCTION TO THE STANDARD MODEL

The Standard Model consists of fundamental particles, which are considered to be the building blocks of the universe. By fundamental one means they are no further divisible with no internal structure. The Standard Model is based on the gauge symmetry($SU(3)_C * SU(2)_L * U(1)_Y$, and all the known fundamental particles are classified into fermions, gauge bosons and the scalar higgs boson[3]. Fermions consists of quarks and leptons which have half-integer spins($\frac{1}{2}$). The leptons fall into three generations;

Generation	Lepton	Q	L_e	$L_(\mu)$	$L_(\tau)$
First	e	-1	1	0	0
	ν_e	0	1	0	0
Second	μ	-1	0	1	0
	$\nu_(\mu)$	0	0	1	0
Third	τ	-1	0	0	1
	$\nu_(\tau)$	0	0	0	1

There are also antileptons with opposite signs for the quantum numbers, and hence there are in total 12 leptons.

The quarks, which are also spin$-\frac{1}{2}$ fermions, are classified into three generations. Each quark consists of three colors – Red, Green and Blue. Since all naturally occurring particles are colorless, we cannot find an individual quark. Quarks also have antiparticles, the anti-quarks with opposite quantum numbers of quarks, and hence there are 36 of them all.

Generation	Quark	Q
First	u	$\frac{2}{3}$
	d	$\frac{-1}{3}$
Second	c	$\frac{2}{3}$
	s	$\frac{-1}{3}$
Third	t	$\frac{2}{3}$
	b	$\frac{-1}{3}$

Gauge bosons (gluons, photons W and Z bosons) are interaction mediators with spin-1. There are 8 gluons in the Standard Model, and like quarks they also carry color charge and hence do not exist as isolated particles. They are detected inside hadrons or as colorless combinations with other gluons(glue balls). Gluons are responsible for the strong interaction. It is of very short range and gets weaker as the subatomic particles move closer and becomes stronger as they are farther apart. Photons are mediators of the electromagnetic interaction. It governs the interaction between charged particles. The weak force is mediated by W and Z bosons. It is responsible for radioactive decay and also acts in a very short range.

Force	Mediator
Electromagnetic	Photon(γ)
Weak	Intermediate vector bosons($\pm W, Z^0$)
Strong	Gluons

Despite the success, the Standard Model fails to give an explanation on many topics such as dark matter, baryon asymmetry of the universe, hierarchy problems, and neutrino oscillations. The Standard Model also does not include gravity and it does not explain spacetime at the quantum level. Ongoing experiments aim to provide explanation to these issues and extend physics beyond the Standard Model.

THE LARGE HADRON COLLIDER (LHC)

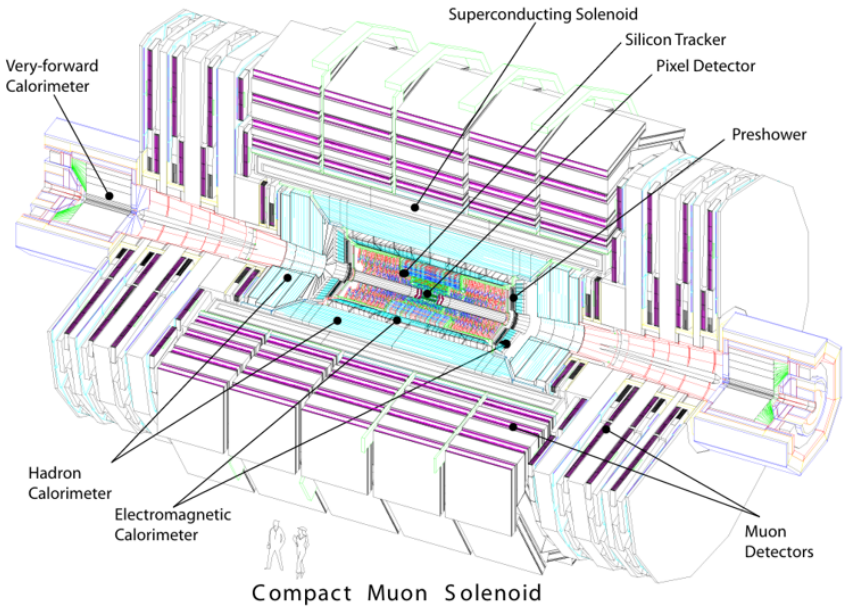
The LHC, situated at the European Organization for Nuclear Research, CERN, near Geneva, Switzerland, is the worlds most powerful accelerator. It is 27 km long and is also the world's largest particle accelerator. It accelerates proton beam to 7TeV and yields a center-of-mass energy of 14TeV with a design luminosity of $10^{34} cm^2 s^1$. [4]

The primary goal of LHC is to study the electroweak symmetry breaking and Higgs mechanism. It also focuses on supersymmetry, grand unification, and extra dimensions that could address the physics beyond the Standard Model. LHC's heavy-ion collision (lead-lead

collision) recreates the quark-gluon plasma, which is thought to have existed during the Big Bang. This allows for the exploration of parton dynamics under extreme conditions.

The proton beams travel in opposite directions and are guided by superconducting magnets. There are four collison points where different detectors reside to detect different phenomena. ATLAS and CMS are the two general purpose detectors, while others have more specialized research. For handling the large amount of data, there are infrastructures that connect computing centers in many countries to utilize maximum amount of data. A trigger system is used to filter out the phenomena that are of physics interest and to reduce the data size.

CMS DETECTOR



A perspective view of CMS Detector[5]

The CMS detector is situated at one of the four collision points, 100m underground at the LHC and has the most powerful solenoid magnet ever made. It weighs 12,500t and is 21.6m long and 14.6m in diameter. It has a cylinder shape with several concentric layers of components. CMS acts as a camera and can take 3D 'photographs' of particle collisions up

to 40 million times each second. It can detect almost all stable particles that have formed after the collision, and from the data acquired, the collisions can be further studied. The detector can also detect muons very accurately.[5]

The solenoid magnet produces a magnetic field of 4 tesla, which bends the particle emerging from the collision point. This helps to identify the charge and momentum of the particle. The tracks of these particles are identified by their electromagnetic interaction with the silicon Tracker. The information about its energy is collected from two calorimeters - the Electromagnetic Calorimeter(ECAL) in the inner layer which measures the electrons and photons and the Hadron Calorimeter(HCAL) at the outer layer which detects the hadrons. However, the muons are not stopped by the calorimeters, so they are detected by sub-detectors.

Calorimeters have two jobs; one is to stop the particle and the other is to measure the energy loss. Particles mass is calculated by measuring its velocity and momentum and hence it can be identified. There are two ways in which velocity can be detected. Using time-of-flight detectors or by looking at how much a particle ionizes the material that it passed through since it is velocity dependent. If the particle travels faster than the speed of light in that medium, it will emit Cherenkov radiation at an angle that depends on its velocity.

Electromagnetic Calorimeter (ECAL)

ECAL is designed to measure the energy and position of electrons and photons. When electrons and photons pass through it, the lead tungstate crystal will scintillate and produce light proportional to the particles energy. In the case of electrons the process starts with bremsstrahlung, and for photons it is pair production leading to showers. The shower stops when individual energy falls below a critical energy and ionization dominates. These secondary particles deposit energy in the crystals. The scintillator lights are then converted to electrical signals by the photodetectors, which are glued to the back of the crystal. The signal is then amplified and sent for analysis.

Hadronic Calorimeter (HCAL)

HCAL measures the energy of hadrons. Hadrons interact with the nuclei of the absorber material (brass) and initiates hadronic shower. The shower contains secondary hadrons and photons which is due to the immediate decay of pions. The shower is usually broader than the ones in ECAL. Plastic scintillators emit light cause of the interaction of charged particles, and this light is carried through a photomultiplier tube. The amount of light is summed up into a 'tower' which is considered as the measure of particles energy. HCAL provides indirect measurement of the non-interacting particles neutrinos or potential dark matter candidates. HCAL ensures the capture of all the particles produced during the collision, and hence an imbalance in energy or momentum can be deduced as an 'invisible particle'.

Superconducting Solenoid Magnet

The superconducting solenoid magnet of CMS detector is its central component. The solenoid produces strong magnetic fields which causes the particle inside the detectors to bend due to Lorentz force. The magnetic field produced is 3.8 Tesla and the magnet itself is large enough to enclose the calorimeters. The coil is made up of niobium-titanium, which is a superconducting material and operates at extremely low temperatures. Muons which have high penetrating power are detected here in the gas-ionization chamber.

JETS

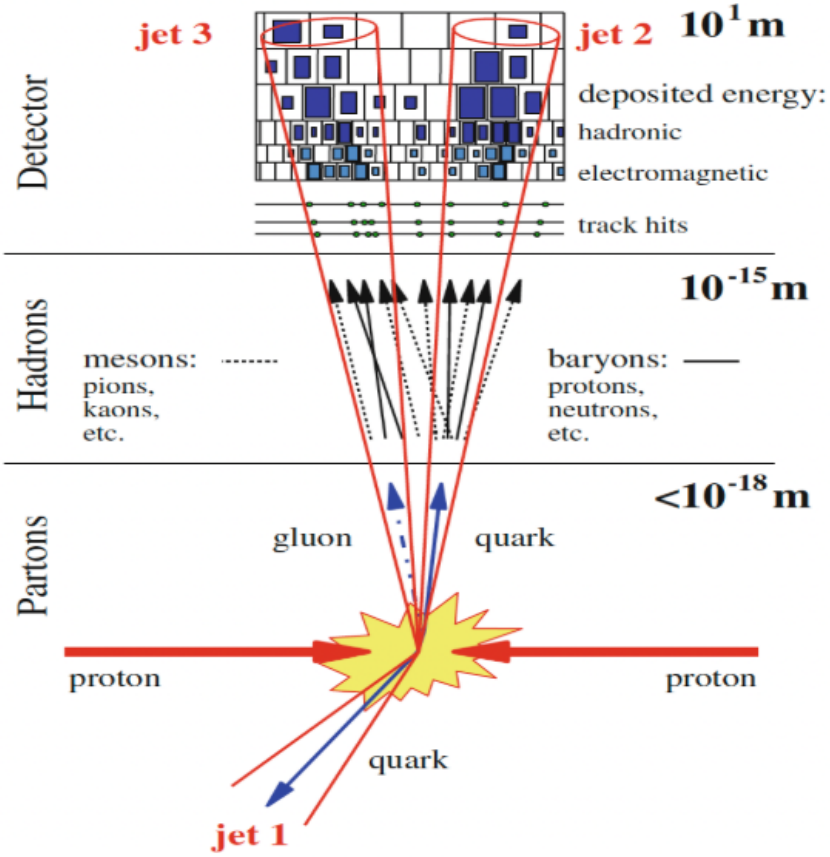
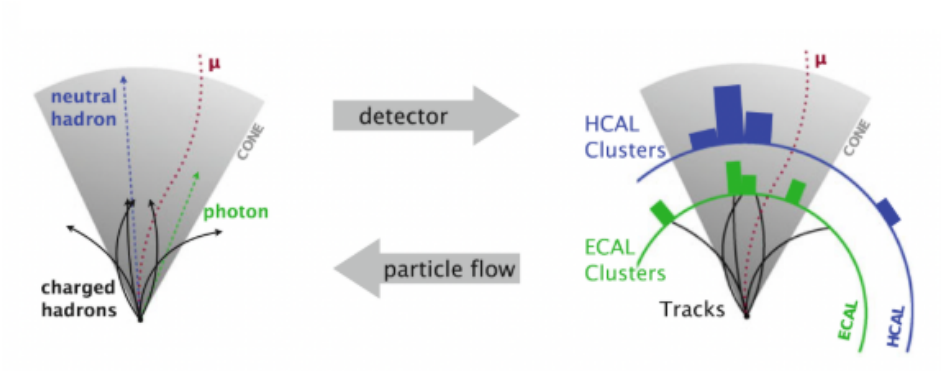


Illustration of a jet [6]

Jets are clusters of hadrons formed during high-energy particle collisions due to the fragmentation of quarks and gluons. Quarks and gluons are fundamental particles with a color charge. According to QCD, all colored particles must combine color-neutral hadrons due to confinement. This is a reason why we cannot observe a single quark or a gluon. Quarks contain a single color charge, while gluons contain two; one color and one anti-color charge. Parton showers are formed during the proton-proton collision where additional gluons or quarks and anti-quark pairs are radiated.[7]

Quarks are only free at extremely small distances. This is due to the unusual behavior of a strong force. Since gluons have more color charge they interact with itself, and hence when the quarks are separated, the gluons which are the mediators of the strong force will interact with themselves creating more quark and anti-quark pairs. It also produces more gluons,

which leads to intense radiation. As everything cools down, quarks and gluons combine to form color neutral hadrons by a process called hadronization. The spray of hadrons that approximately move in the same direction are called jets.



Schematic association of sub-detector measurements to physical particle candidate using PF technique of CMS (Illustration courtesy, F.Pandolfi)

Jets appear as a cluster of energy deposits in calorimeters. However, the energy deposited is not directly proportional to the particle's energy since it loses some part of its energy during hadronization and interaction with the nuclei in the calorimeters. To overcome this, a multiplicative factor known as the jet energy correction factor is applied to the raw energy found from calorimeters. Individual particles are reconstructed and clustered into jets by particle flow algorithms after combining information from all the sub-detectors. The particles are clustered on the basis of their relative distance in momentum space which produces conical, stable jets. This gives us an idea of the parton state.

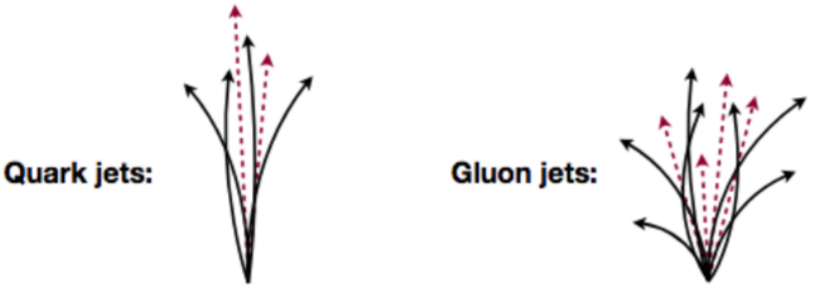


Illustration to show the differences between quark and gluon jet(Discrimination of quark and gluon jets at 13TeV in the CMS experiment- Jyoti Babbar)

Quark-initiated jets and gluon-initiated jets have different characterizations because of their differences in color charge. Hence, the jet originated from gluons is much more complex. Compared to a quark-initiated jet, gluon jets are less collimated and have soft fragmentation. The particle multiplicity is higher for gluon jets and their energy is distributed among many particles.

MULTIVARIATE ANALYSIS IN HIGH ENERGY PHYSICS

CMS is built to capture almost all the data that are acquired during the proton-proton collisions at LHC. With luminosity reaching $10^{34} cm^2 s^1$ it becomes hard to track down rare processes such as Higgs boson decay, or other beyond Standard Model phenomena with very small cross-section in an overwhelming background. Traditionally used cut-based method could not recognize the complex relations between the features obtained. Here is where machine learning, particularly multivariate analysis comes into picture. MVA is a statistical method that is used to analyze data with multiple variables. MVA helps us organize our data, find the correlation between the variables, reduce the dimensionality of the data, and make it overall easy to analyze. MVA improves the signal-to-background discrimination by combining information from many observables that are otherwise be obscured.

There are two types of multivariate analysis - supervised and unsupervised.

Supervised learning

The model is trained on data that is labeled or data that are already known (e.g., Monte-Carlo simulation). Features are the input variables used by the model to make predictions and target is the output variable or the output your model needs to predict. In supervised learning, features and targets are explicitly given. Examples for supervised learning are;

- Logistic regression - Classify models into 0 or 1 using the sigmoid function.
- Boosted Decision Tree - Multiple weak models are used to build a strong one. Errors are calculated for each weak model and subtracted from the final model.

- Standard Vector Model - Maximize the margin between the variables ,where negative variables fall under negative hyperplane and positive variables under positive hyperplane.
- Neural Network - A model that is inspired by our brain. It consists of layers with interconnected nodes called neurons that take input from the previous ones.

Unsupervised learning

There will be no prior knowledge about the features and the target. The model will try to classify by finding the structures and patterns within the data. Examples of unsupervised learning are;

- Principle Component Analysis(PCA) - Used for feature selection. Helps finding important variables for data processing.
- Clustering - Based on similarity, the data points are grouped into clusters.

THE QUARK-GLUON DISCRIMINATOR

A quark-gluon discriminator is a binary classifier that tells you whether the jet originated from light quarks or not by exploiting the differences in characterization of jets originated from quarks and gluons. The tagger treats light quarks as the signal gluons as the background and gives a clear distinction between them. This gives more exposure to rare events such as vector-boson fusion, supersymmetry searches, and the study of QCD.[8]

NEURAL NETWORK

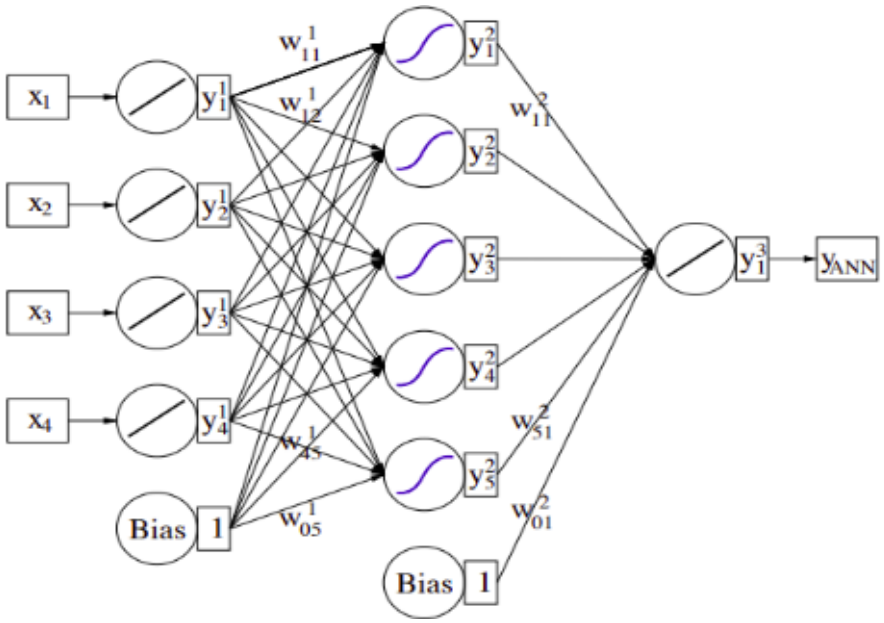


Illustration of a Neural Network with one hidden layer[9]

For the discriminator, we have used Neural Networking. It has an input layer where the features are applied, a hidden layer with activation functions, and an output layer that classifies the data. Every connection between the neurons has a weight associated with it and every neuron has an associated bias. Weights determine the strength of each connection. Weights for each neurons are found during the training phase by minimizing the error on the output of the neural network. Unlike weights which are connected to the input, biases are added to the neuron's output and provide more flexibility to the network. If x is the input and w and b are the weight and bias, the output of the neuron is given by,

$$y = wx + b$$

[10]

Overfitting and Underfitting

A model should be evaluated using test data and if it performs poorly, then the model may be overfitting or underfitting. Underfitting occurs when the model is too simple to capture the complexity of the data. Using very few neurons or very strong dropouts can lead to this. Overfitting occurs when the model becomes too complex and memorizes even the noises in the training data. Imbalanced dataset, model with too many neurons, and parameters can lead to overfitting. To avoid both of these situations, we tune the hyperparameters. By monitoring training and validation performance and adjusting the hyperparameters accordingly, one can find an optimum fit for their data.

Hyperparameters

- Learning Rate(LR) - Size at which the weights and biases are updated.
- Batch size - Number of samples processed before updating the weights. Smaller batch size(16-64) can introduce noise, while larger batch size(256 - 1024+) gives you a smoother gradient but can lead to overfitting.
- Number of epochs - Defines how many times the model goes through the entire data set. Too many epoch can lead to overfitting.
- Activation function - Not all dataset have linear connection. Activation function introduces non-linearity which helps the model to learn complex patterns. Some of the common choices are,
  - ReLU - max(0,x). No negative values.
  - Leaky ReLU - max(0,x) + a(min(0,a)). Small negative values.
  - Sigmoid - Used in the output layer for binary classification. $\frac{1}{1+e^{-x}}$
- Optimizer - Controls the updates of weights and biases. Popular choices are SGD, Adam and RMSProp.

- Dropout - Prevents overfitting by randomly deactivating neurons.
- L1/L2 Regularization - Prevents overfitting by adding penalties to large weights.

Forward and Backward Propagation

The data is passed through the network from the input layer, through the hidden layers, and finally to the output layer. During each propagation, for every neuron, weights are multiplied from each connection and then summed together along with its associated bias. This is then passed through the activation function which introduces non-linearity.

After each forward propagation, the models performance is evaluated by a loss function. The loss function calculates the error in the prediction. Gradients with respect to each weights and biases are calculated, and then the weights and biases will be updated by the optimizer. In this way, the network tries to decrease the loss. The rate at which the update should take place is determined by the learning rate. The loss function used for binary classification is Binary cross-entropy and it is defined as;

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right]$$

where,

- N - no of observables
- y_i - represents whether the $i^(th)$ observation is o or 1
- p_i - represents the predicted value of the $i^(th)$ observation

This process of forward propagation, finding the loss and updating the weights is repeated for many iterations, and over the time, the loss decreases and model becomes much more accurate.

Receiver Operating Characteristic curve (ROC) and Loss function Curve

ROC curve shows the performance of the model as the threshold changes. In the case of jets, jets whose value is predicted to be greater than the threshold are quarks, and those that are less than the threshold are gluons. It plots the true positive rate (TPR) vs the false positive rate (FPR) at different classification thresholds. For the quark-gluon discriminator, the terms are as follows;

- TP (True Positive): correctly predicted quark jet
- FP (False Positive): gluon jet incorrectly predicted as quark
- TN (True Negative): correctly predicted gluon jet
- FN (False Negative): quark jet incorrectly predicted as gluon

The x-axis of the ROC curve represents the False Positive Rate (FPR), which is the fraction of gluon jets that are misclassified as quarks.

$$FPR = \frac{FP}{FP+TN}$$

The y-axis represents the True Positive Rate (TPR); the fraction of quark jets that are correctly classified.

$$TPR = \frac{TP}{TP+FN}$$

Low thresholds (0.2) can lead to most of the jets being classified as quarks; hence high TPR. For high threshold (0.8) only few quark jets will be correctly classified and hence FPR will be high.

The area under the ROC curve gives the accuracy of the classifier. If it is closer to one, the model is a perfect classifier. If it is closer to 0.5, the classifier is randomly guessing.

The training and validation loss curve determines whether your model is underfitting or overfitting. Training loss is the error in the prediction of training data set, while validation loss is the error in prediction of unknown data set. The training and validation loss has to

gradually decrease with the epoch. If validation loss increases or diverges from the training loss, it is an indication to overfitting. If the losses are very thigh and decreases very slowly, the model might be underfitting. It might also indicate that the learning rate is small.

DATA PREPROCESSING

Data preprocessing is a crucial step in data analysis, especially in high energy physics, where raw data are transformed into a clean and structured format. This makes it much easier to handle complex information during analysis and also makes it reliable to train the model. Preprocessing includes feature selection, feature engineering, data scaling, handling class imbalance, and finally splitting the data set for training and testing.

Selection of variables

Feature selection involves the identification of observables which are relevant and can effectively discriminate between quarks and gluons. Here we exploit the differences between the quarks and gluon- initiated jets. [11]

- Multiplicity -

  Multiplicity is the total number of PFCandidates reconstructed within the jet. Since gluon jets have a higher color charge, they radiate more and are expected to have greater values of multiplicities with respect to quark jets for a given p_T scale. The multiplicity of both charged and neutral particles is taken into account for the analysis.

- Jet shape -

  Jets originated from gluon hadronization is much more wider compared to the ones originated from quarks. The shape of a jet can be approximated to an ellipse. If M is a

2*2 matrix constructed by the following elements:

$$M_{11} = \sum_i p_{T,i}^2 \, \Delta\eta_i^2$$
$$M_{22} = \sum_i p_{T,i}^2 \, \Delta\phi_i^2$$
$$M_{12} = M_{21} = -\sum_i p_{T,i}^2 \, \Delta\eta_i \, \Delta\phi_i$$

where,

- $p_{T,i}$ - transverse momentum of the $i^(th)$ constituent
- $\Delta\eta_i$ - pseudorapidity distance between each constituent and their average direction
- $\Delta\phi_i$ - azimuthal distances between each constituent and their average direction
- $p_{T,i}^2$ - -weighted direction of the jet constituents in $\eta - \phi$ space.

Then the major(σ_1 and minor (σ_2 axes of the jet are given by,

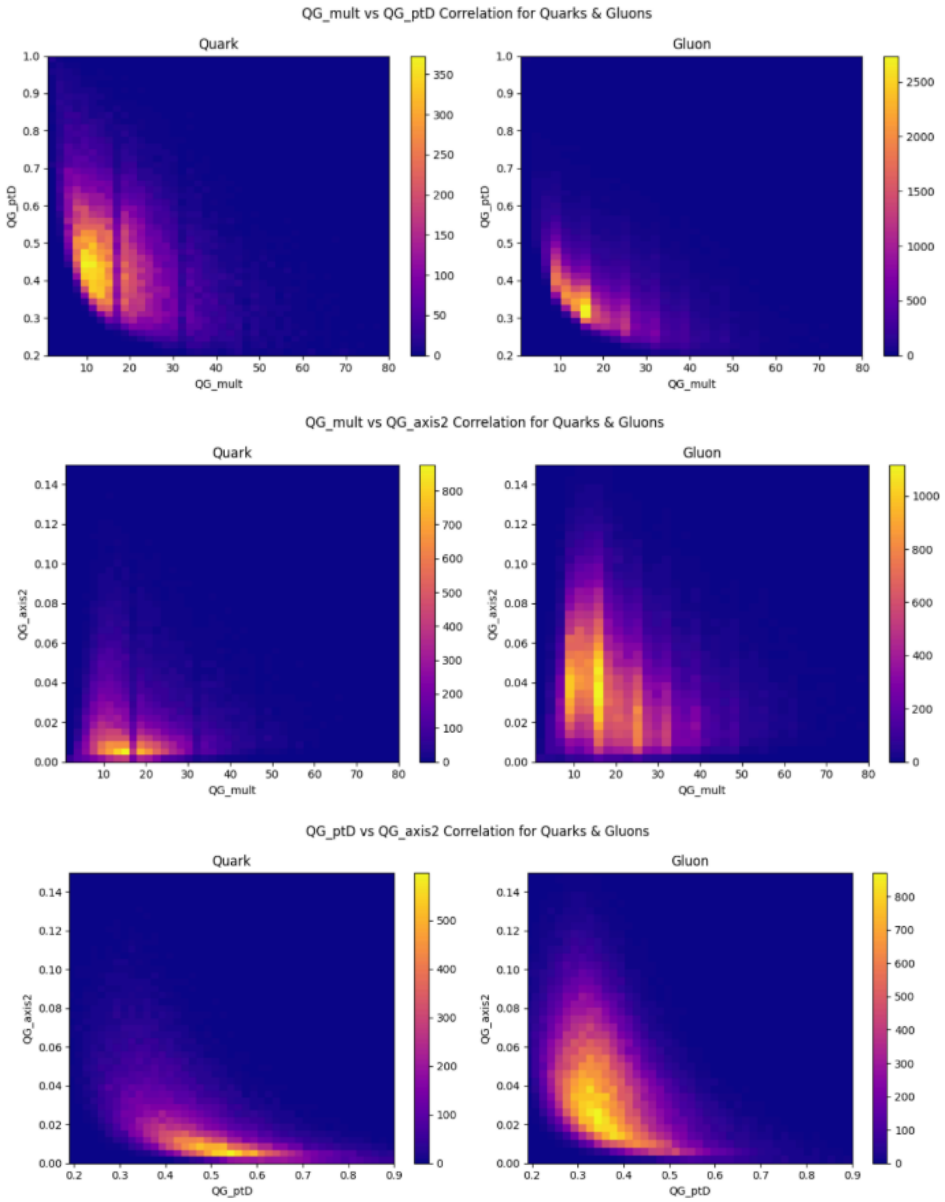
$$\sigma_1 = \left(\frac{\lambda_1}{\sum_i p_{T,i}^2} \right)^{1/2}$$
$$\sigma_2 = \left(\frac{\lambda_2}{\sum_i p_{T,i}^2} \right)^{1/2}$$

- Fragmentation function -

  Gluon jets have soft fragmentation. The energy of the jet is spread over a lot of constituent. Meanwhile quark jets have harder fragmentation leading to hard constituents which carry significant part of their energy. This can be showed by;

$$p_T D = \frac{\sqrt{\sum_i p_{T,i}^2}}{\sum_i p_{T,i}}$$

  $p_T D$ becomes 1 for jets made of only one particle, it tends to zero as the particle number increases.



Two-dimensional heat map showing the correlation between different variables chosen for the analysis

In addition to using raw observables, new features were constructed using combinations of existing variables considering how different jet properties work.

Data Standardization

The data is standardized to a common format or range by rescaling each features that is selected, to have a mean of zero and standard deviation of one. This prevents the model from being biased towards features with larger scales. The standardization is applied using;

$X' = \frac{X-\mu}{\sigma}$

where μ is the mean and σ is the standard deviation.

Handling Class Imbalance

Not every data has a class with a 50:50 ratio of quarks and gluons. A common issue is gluon jets outnumbering quark jets. Class imbalance causes the model to be biased towards the majority class resulting in false accuracy. Since majority of the class are gluons, the accuracy will mislead us into believing that the model works fine. Class imbalance can be tackled by either undersampling gluons or by oversampling quarks. Another method is to add more weight to quarks.

Since undersampling can lead to the loss of valuable information, SMOTE-Tomek was used to oversample the minority class. SMOTE (Synthetic Minority Oversampling Technique) algorithm finds nearest neighbors of a data point and synthesizes a new data point between the other two ones, while Tomek links are used to remove ambiguous samples that are near the class boundary. Isolation-based techniques are used to filter out noise. This creates samples that are not exact duplicates and we get a good balanced data set.

Train-Test Split

A model is valuated by how well it performs in classifying unknown data. To validate the model, we split our data set into training and testing data sets. This is done using the train-test-split function in Sklearn. Typically, the ration of training data (x-train, y-train) to testing data (x-test, y-test) is 8:2.

DISCRIMINATOR

Once the data have been preprocessed we train the model that is built using neural network. We map the features onto a single output, either one or zero, which we interpret as the likelihood that the jet was initiated by quarks. The model is built in two ways; a neural

network from scratch using just numpy and python, and the other one using TensorFlow 2.17 and Keras. The model architecture for both the models are shown below.

Model 1: Neural Network from Scratch

- Layers:
  - $Dense(13 \rightarrow 256) + ReLU + Dropout(0.2)$
  - $Dense(256 \rightarrow 128) + ReLU + Dropout(0.2)$
  - $Dense(128 \rightarrow 1) + Sigmoid$
- Loss: Binary cross-entropy
- Optimizer: Customized Adam ($lr = 0.005, \beta_1 = 0.09, \beta_2 = 0.999, \epsilon = 10^{-8}$)
- Epochs: 1000, with no early stopping

Model 2 : Keras-Based Neural Network

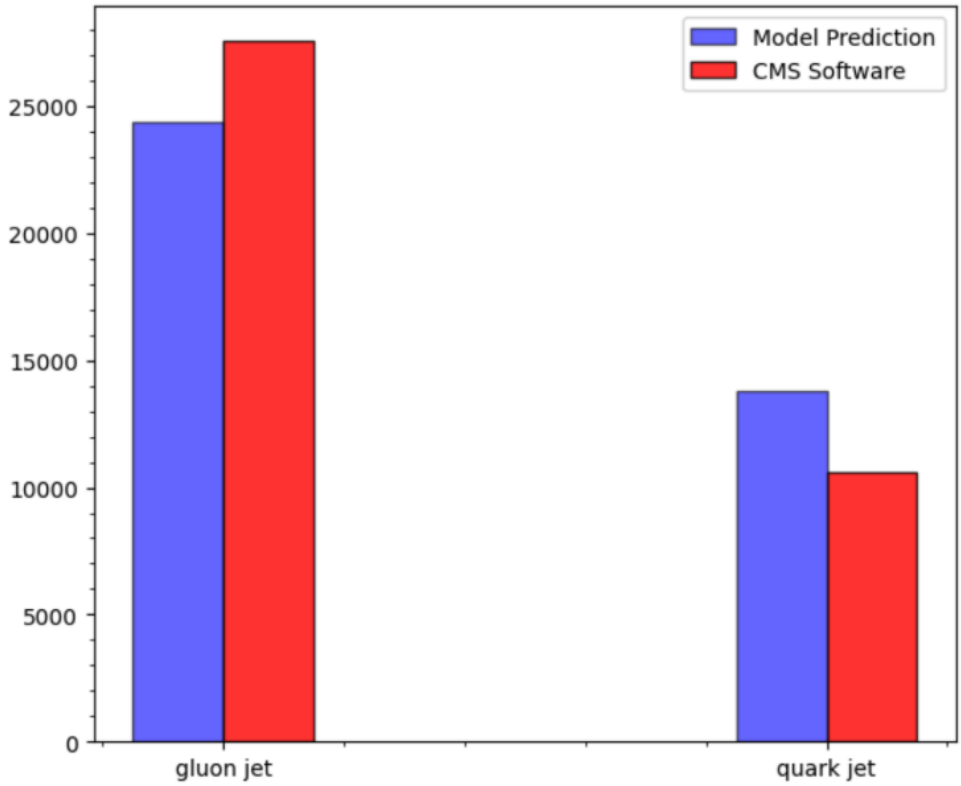
- Layers:
  - $Dense(13 \rightarrow 128) + ReLU + L2(1e-4) + Batchnorm + Dropout(0.40)$
  - $Dense(128 \rightarrow 128) + ReLU + L2(1e-4) + Batchnorm + Dropout(0.30)$
  - $Dense(128 \rightarrow 64) + ReLU + L2(1e-4) + Batchnorm + Dropout(0.30)$
  - $Dense(64 \rightarrow 1) + Sigmoid$
- Loss: Binary cross-entropy
- Optimizer: Adam $lr = 1e-3$
- Callbacks:
  - EarlyStopping (patience= 5)
  - ReduceLROnPlateau (factor 0.5, patience 3)
  - ModelCheckpoint (on validation loss)

- Epochs: up to 50 (stopped at 35)

RESULTS

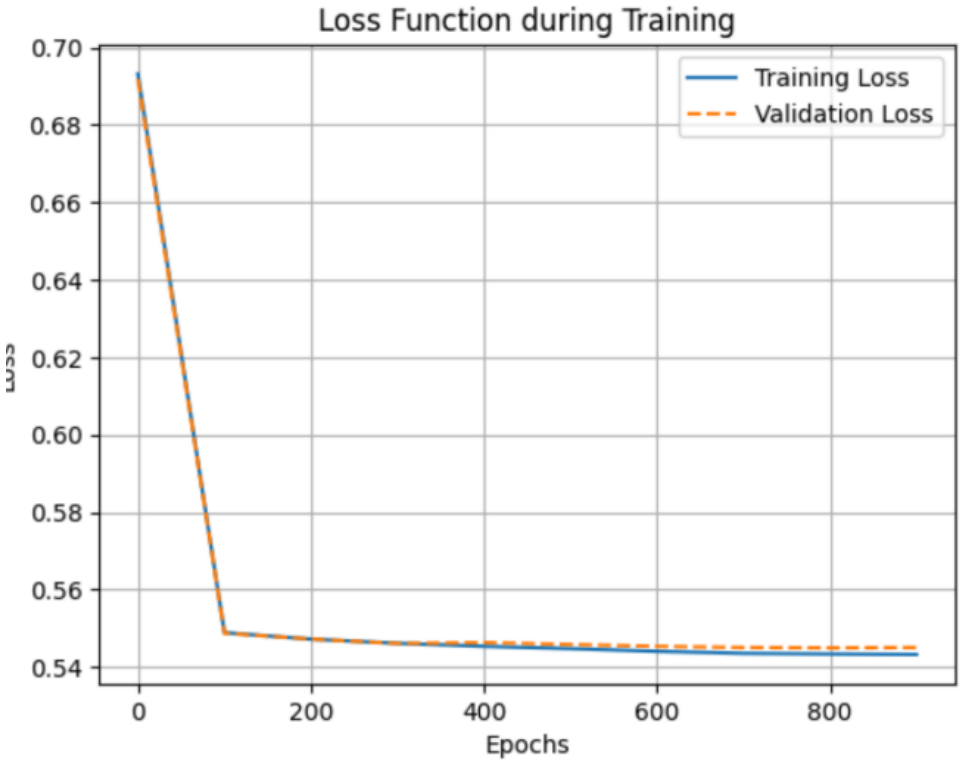
Model 1: Neural Network from Scratch

The number of gluon and quark jets tagged by numpy and python based model (blue) compared to the CMS simulation (red) is shown below. The network slightly under-calls gluons and over-calls quarks.



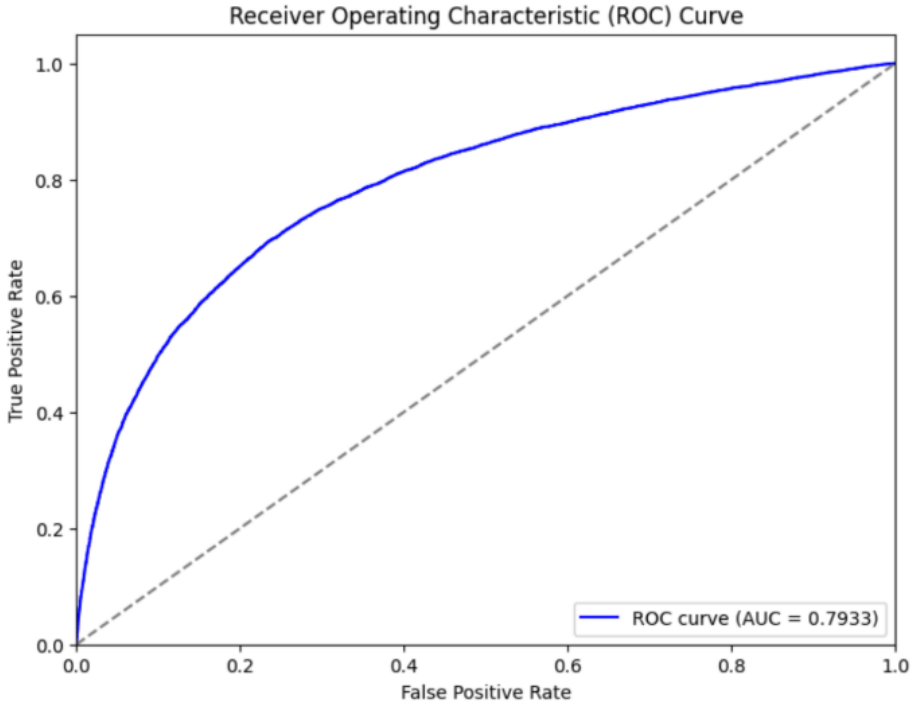
Comparison of Class Counts: Model 1 vs CMS data

Both the training and validation loss curve shows a sharp drop and then plateau near 0.54 - 0.55. There is almost no gap between the two curves, indicating that there is no overfitting. The curves are both smooth because there are 1000 epochs and also because of the simple architecture of the network, it might not pick up all the fluctuation.



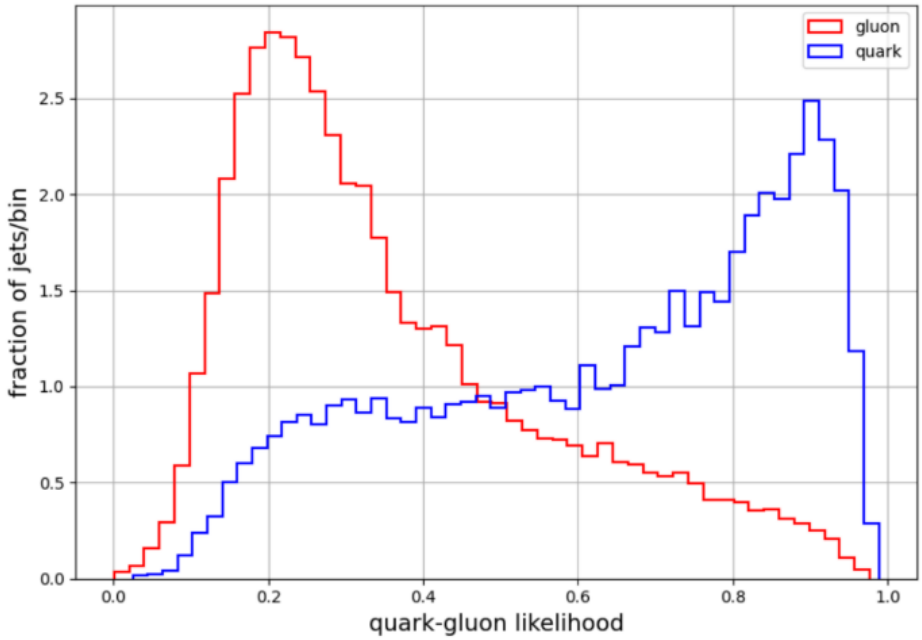
Binary-cross-entropy loss on the training set (blue) and validation set (orange) vs epoch for Model 1

The area under the curve is shown to be 0.79. This indicates that if a random jet is picked, the network correctly classifies the quark jet 80 percentage of the time.



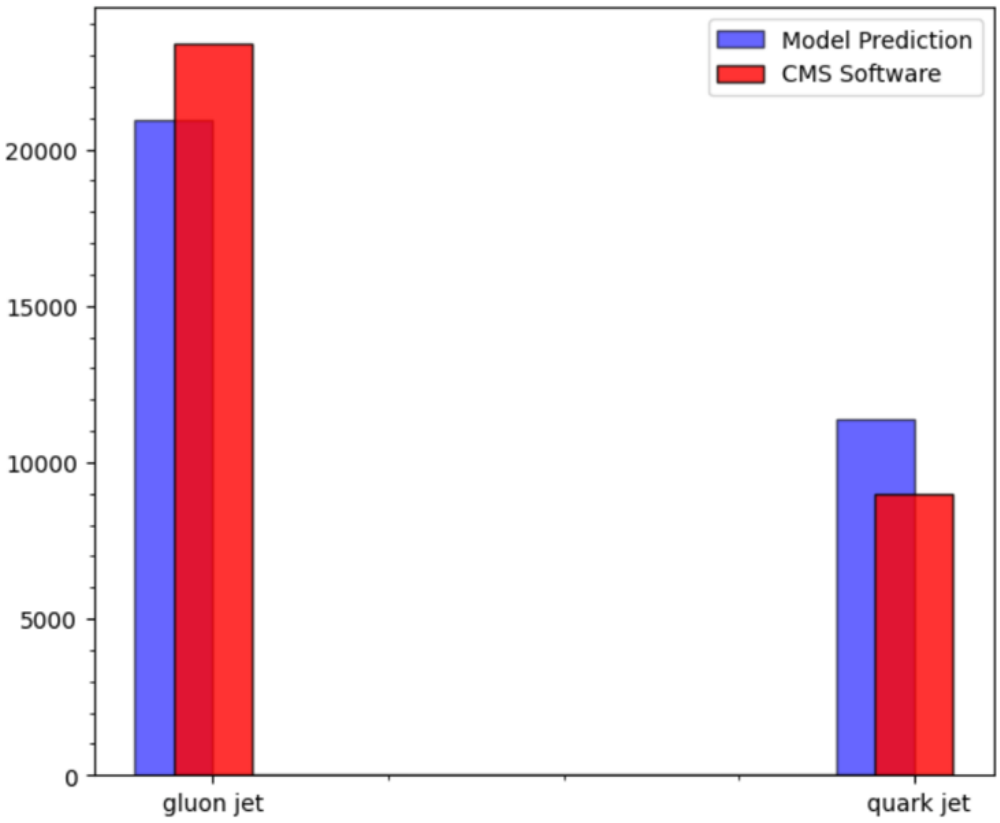
ROC curve for python and numpy based model.

Quark jets cluster near scores of 0.8–1.0, while gluon jets peak around 0.2. There is an overlap between the distributions around 0.4 - 0.6.



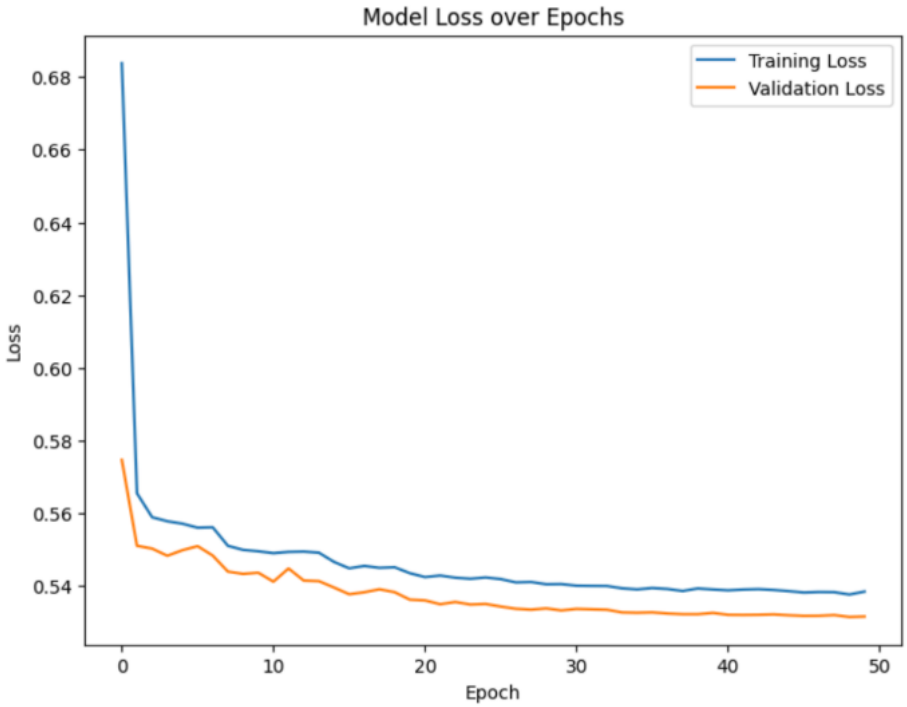
Normalized distribution for quark(blue) and gluon(red) initiated jets from Model 1

Model 2 : Keras-Based Neural Network



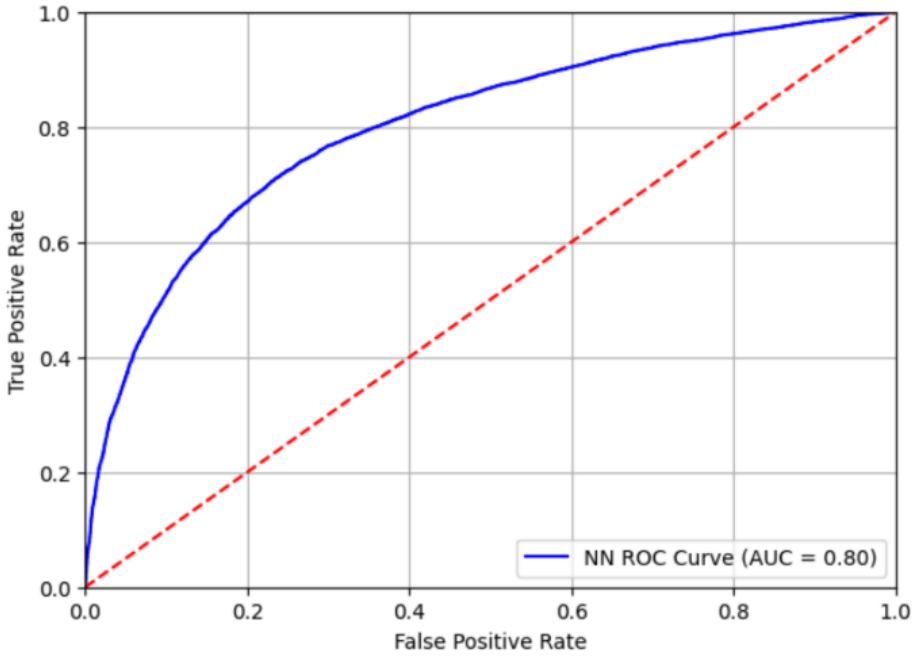
Comparison of Class Counts: Model 2 vs CMS data

There is a slight separation between the two curves, but there is still no sign of overfitting. The not-so-smooth curves are due to a small number of epochs and the model calculating the loss for each epoch rather than averaging it.



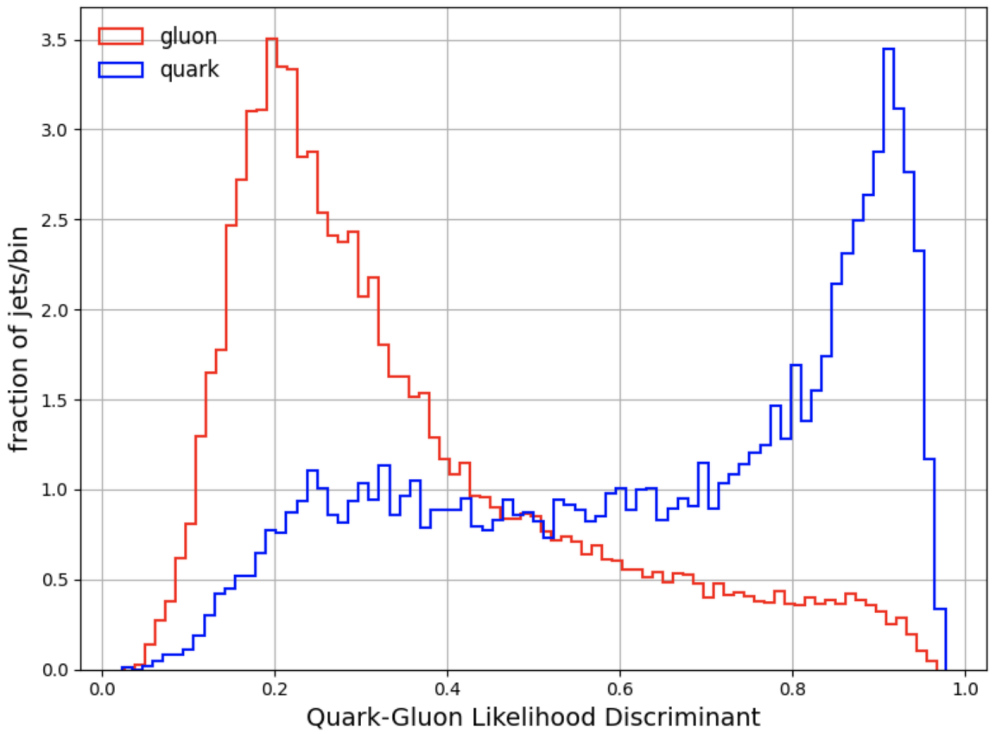
Binary-cross-entropy loss on the training set (blue) and validation set (orange) vs epoch for Model 2

The AUC is 80 percent. The steep slope of the curve even at low FPR shows that the model performs well.



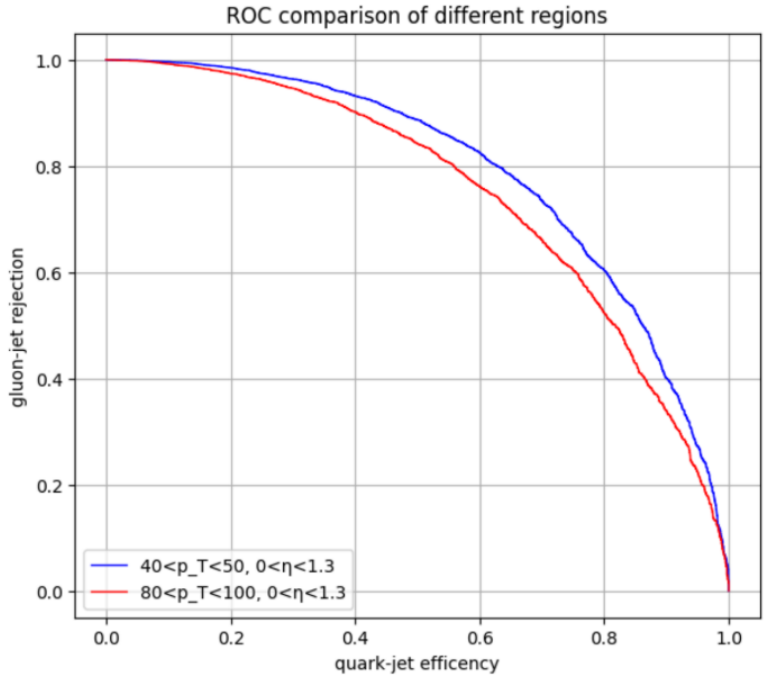
ROC curve for Keras - based model

The separation between quark and gluon initiated jets is much more distinct for this model.



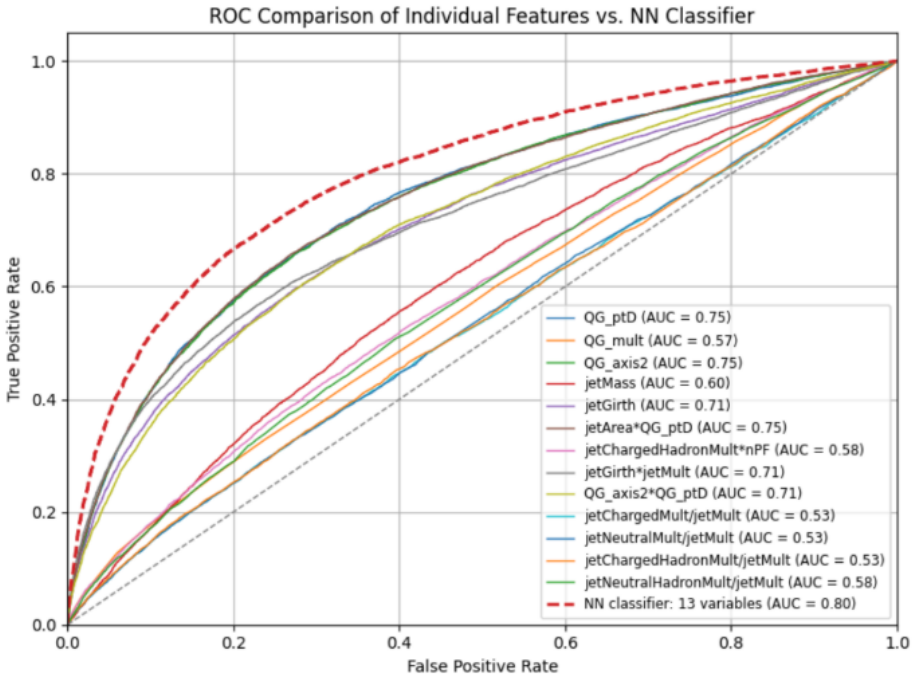
Normalized distribution for quark(blue) and gluon(red) initiated jets from Model 2

ROC Curve for different p_T and η regions



The model performs better in the lower transverse momentum region since it covers more area than the higher transverse momentum curve.

ROC Curve for all the variables



ROC curve for all the features including the final neural network output is shown.

CONCLUSION

In this project, we built a classifier which differentiates between quarks and gluon initiated jets from high-energy proton-proton collision. Formation and differences between quarks and gluon jets were studied, and physics-motivated input variables were selected for training and evaluating the model. The data obtained from the Monte Carlo simulation were processed and standardized. The class imbalance was handled by SMOTE-Tomek. After that it was split into training and validating set.

After preprocessing, two neural networks, one built using numpy and python, and the other built using Keras and TensorFlow were used to fit the data. Performance was assessed using binary cross-entropy and Receiver Operating Characteristic curves. The Keras-based model was much faster, and achieved an accuracy of 0.80. By evaluating the validation curve, we can see that there was no overfitting. The likelihood distribution confirmed the model's ability to discriminate jets correctly.

The project provided practical experience in handling collider data and learning various techniques in machine learning. Unlike the traditional cut-based method, Multivariate Analysis makes data analysis much more promising. More advanced architectures, further optimization of hyperparameters can lead to an improved model for a perfect analysis.

Bibliography

[1] Andrew J. Larkoski and Eric M. Metodiev. A theory of quark vs. gluon discrimination. *Journal of High Energy Physics*, 2019(10):14, 2019.

[2] Pushpalatha C. Bhat. Advanced Analysis Methods in High-Energy Physics. *AIP Conf. Proc.*, 583(1):22–30, 2002.

[3] R. L. Workman et al. Review of Particle Physics. *PTEP*, 2022:083C01, 2022.

[4] Lyndon Evans and Philip Bryant. Lhc machine. *JINST*, 3:S08001, 2008.

[5] CMS Collaboration. The cms experiment at the cern lhc. *JINST*, 3:S08004, 2008.

[6] Tancredi Carli, Klaus Rabbertz, and Steffen Schumann. *Studies of Quantum Chromodynamics at the LHC*, pages 139–194. Springer International Publishing, Cham, 2015.

[7] R. Keith Ellis, W. James Stirling, and B. R. Webber. *QCD and Collider Physics*. Cambridge University Press, Cambridge, UK, 1996.

[8] Jason Gallicchio and Matthew D. Schwartz. Quark and gluon tagging at the lhc. *Physical Review Letters*, 107(17), October 2011.

[9] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, M. Backes, T. Carli, O. Cohen, A. Christov, D. Dannheim, K. Danielowski, S. Henrot-Versille, M. Jachowski, K. Kraszewski, A. Krasznahorkay Jr., M. Kruk, Y. Mahalalel,

R. Ospanov, X. Prudent, A. Robert, D. Schouten, F. Tegenfeldt, A. Voigt, K. Voss, M. Wolter, and A. Zemla. Tmva - toolkit for multivariate data analysis, 2009.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

[11] Performance of quark/gluon discrimination in 8 TeV pp data. 2013.