Multivariate analysis to discriminate quark and gluon initiated jets in proton-proton collisions using CMS Open data

A dissertation submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Physics

by

Shivam Vaid

Roll No: 2023PGPHPH22

Under the guidance of

Prof. Jyothsna Rani Komaragiri

Centre for High Energy Physics (CHEP)

Indian Institute of Science (IISc), Bangalore

Submitted to

Prof. Rajeev Ranjan

Department of Physics

National Institute of Technology, Jamshedpur

May 2025

National Institute of Technology, Jamshedpur

Department of Physics

Certificate

This is to certify that the thesis entitled **"Multivariate Analysis to Discriminate Quark and Gluon Initiated Jets in Proton-Proton Collisions Using CMS Open Data"** submitted by **Shivam Vaid** in partial fulfillment of the requirements for the degree of **Master of Science in Physics** is a bonafide record of original work carried out by him under our supervision and guidance.

The work presented in this thesis has not been submitted elsewhere for the award of any degree or diploma.

Prof. Rajeev Ranjan
Associate Professor
Department of Physics
National Institute of Technology,
Jamshedpur

Prof. Jyothsna Rani Komaragiri
Associate Professor
Centre for High Energy Physics (CHEP)
Indian Institute of Science (IISc),
Bangalore

Declaration

I hereby declare that the work presented in this thesis entitled **"Multivariate Analysis to Discriminate Quark and Gluon Initiated Jets in Proton-Proton Collisions Using CMS Open Data"** is the result of my own work carried out at the Indian Institute of Science, Bangalore, during the fourth semester of my M.Sc. in Physics. This work was done under the guidance of external guide **Prof. Jyothsna Rani Komaragiri**, Centre for High Energy Physics (CHEP), Indian Institute of Science (IISc), Bangalore and internal guide **Prof. Rajeev Ranjan**, Department of Physics, National Institute of Technology, Jamshedpur.

This thesis has not been submitted, either in part or in full, for the award of any other degree or diploma at any other institution.

Shivam Vaid

Date: 2023PGPHPH22

Acknowledgement

I would like to express my sincere gratitude to my supervisor, **Prof. Jyothsna Rani Komaragiri**, Center for High Energy Physics (CHEP), Indian Institute of Science (IISc), Bangalore, for his invaluable guidance, support, and encouragement throughout the course of this project.

I am deeply thankful to the faculty of the Department of Physics, National Institute of Technology, Jamshedpur. In particular, I would like to thank **Prof. Rajeev Ranjan** for his support and encouragement, and our Head of Department, **Prof. Ujjwal Laha**, for fostering a positive academic environment. I also extend my gratitude to all the teachers and lab staff for their continuous help and inspiration during my studies.

My heartfelt thanks to the CMS Open Data team for making real collider data publicly accessible, which made this research possible.

My deepest thanks go to my family for their unconditional love, patience, and unwavering support in every step of my journey. I am equally grateful to my friends for their constant motivation, understanding, and companionship through both challenges and successes.

Above all, I thank God for giving me the strength, perseverance, and peace of mind to complete this work.

Shivam Vaid

Date: 2023PGPHPH22

Abstract

This thesis presents a study of quark-gluon jet discrimination in proton-proton collisions using CMS Open Data. Jets from the AK4 collection are analyzed with 11 substructure variables. A Multi-Layer Perceptron (MLP) classifier is trained using TMVA in C++, with all steps executed inside a Docker container for reproducibility.

The model achieves an AUC of 0.837, outperforming likelihood-based methods. Performance is evaluated through ROC curves and statistical metrics, and results are interpreted through the lens of QCD, highlighting physical differences in jet radiation patterns.

Contents

Chapter 1

Introduction

Motivation: The Large Hadron Collider (LHC) at CERN enables the study of fundamental forces by colliding protons at extremely high energies. These collisions produce high-energy quarks and gluons, which subsequently hadronize into collimated sprays of particles known as jets.

Importance of Quark-Gluon Discrimination: Distinguishing between quark-initiated and gluon-initiated jets is essential for a wide range of physics analyses. It plays a critical role in precision tests of the Standard Model, the search for new particles, and the study of the strong interaction dynamics governed by Quantum Chromodynamics (QCD).

Jet Substructure Differences: Jets originating from quarks and gluons exhibit different substructures due to their distinct color charges and radiation patterns. The color factor for quarks is $C_F = \frac{4}{3}$, while for gluons it is $C_A = 3$. As a result, gluons radiate more strongly, producing broader jets with higher particle multiplicity compared to quark jets.

Objective of This Study:

- To extract and preprocess relevant jet substructure variables from CMS Open Data.
- To implement a Multilayer Perceptron (MLP) classifier using the TMVA framework in C++.
- To evaluate the classifier performance using receiver operating characteristic (ROC) curves and likelihood-based discriminants.
- To interpret the results obtained within the theoretical framework of QCD, focusing on the effects of color factors and parton radiation patterns.

Previous Work: Quark-gluon discrimination has long been a subject of interest in high-energy physics. Early research was based on simple observables such as jet multiplicity

and width. With advancements in experimental techniques and detector precision, more sophisticated substructure variables and multivariate analysis methods have emerged.

Machine Learning in Jet Physics: The integration of machine learning techniques, particularly neural networks and decision trees, has significantly improved the ability to classify jets based on their substructure characteristics, enabling deeper insights into parton-level processes.

Use of CMS Open Data: The availability of publicly accessible CMS Open Data has opened new opportunities for independent researchers to perform high-quality analyses, reproduce official results, and develop innovative techniques within the high-energy physics community.

Challenges:

- The overlapping nature of quark and gluon jet substructures, especially at high transverse momenta.
- Detector effects such as resolution limitations and pile-up contamination.
- Dependence of classifier performance on jet energy scales, detector configurations, and data-taking conditions.

Scope and Limitations: This work is based on a supervised machine learning approach utilizing eleven carefully selected jet substructure variables from CMS Open Data, processed using a Multilayer Perceptron model within the TMVA framework. Although limited to these variables and this model, the methodology can be extended to include additional observables and advanced classification algorithms. The theoretical interpretation of the results is grounded in perturbative QCD and standard hadronization models.

Overview: This introduction provides the necessary context and motivation for the study, describing its significance, objectives, challenges, and methodology, thus setting the stage for the theoretical, computational, and analytical discussions that follow.

Chapter 2

Theoretical Background

What is QCD (Quantum Chromodynamics):
Quantum Chromodynamics (QCD) is the theory that explains the strong interaction — the force that holds quarks together inside protons, neutrons, and other hadrons. In QCD, quarks interact by exchanging particles called gluons. Both quarks and gluons carry a type of charge known as *color charge*, which is different from electric charge in electromagnetism.

The basic QCD Lagrangian is written as:

$$\mathcal{L}_{QCD} = \bar{\psi}_i(i\gamma^\mu D_\mu - m)\psi_i - \frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu}$$

Here: - ψ_i is the quark field for flavor i.
- m is the mass of the quark.
- D_μ is the covariant derivative that includes gluon interactions.
- $F^a_{\mu\nu}$ is the gluon field strength tensor.

Color Factors and Radiation Strength:
In QCD, the amount of radiation emitted by quarks and gluons depends on their color factors. For quarks, the color factor is:

$$C_F = \frac{4}{3}$$

For gluons, it is:

$$C_A = 3$$

Since $C_A > C_F$, gluons radiate more strongly than quarks. This makes gluon jets wider and with more particles inside them compared to quark jets.

Jet Formation:

When high-energy quarks and gluons are produced in a collision (like in the LHC), they cannot exist freely due to a property called *color confinement*. Instead, they produce many other particles through a process called *parton showering*, where they radiate more gluons and create quark-antiquark pairs.

This showering continues until the energy drops low enough that hadronization occurs — this means the partons convert into color-neutral hadrons. These hadrons form a collimated spray of particles known as a **jet**.

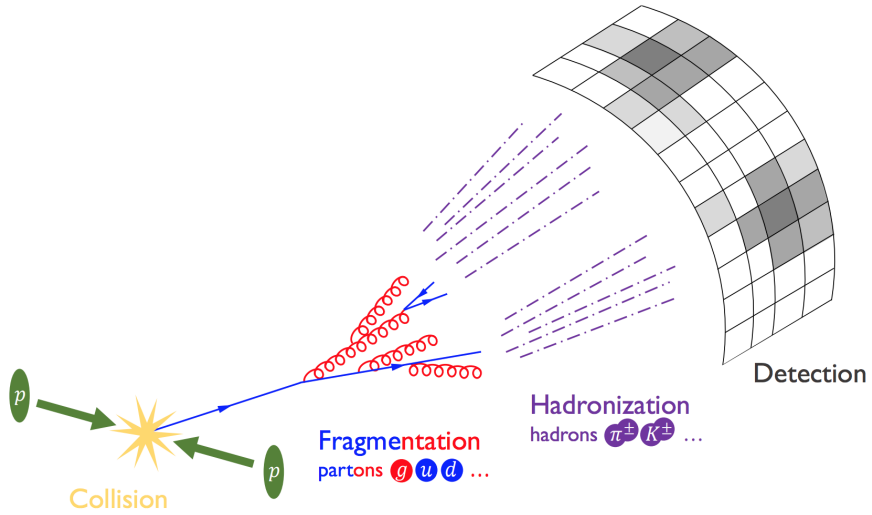

Figure 2.1: Illustration of a proton-proton collision producing quarks and gluons, followed by parton showering and hadronization into jets.

Differences in Quark and Gluon Jets:

Because gluons radiate more:

- Gluon jets have higher *multiplicity* (more particles).
- Gluon jets are *broader*.
- Gluon jets tend to have a slightly larger *jet mass*.

This makes it possible to tell quark jets and gluon jets apart by studying the internal structure of the jet, called **jet substructure**.

Important Jet Variables:

Some useful measurable quantities inside a jet are:

- **Multiplicity** (N): Number of particles inside the jet.
- **Jet Mass** (M_{jet}): The invariant mass of all the particles in a jet.

$$M_{jet} = \sqrt{\left(\sum_i E_i\right)^2 - \left|\sum_i \vec{p}_i\right|^2}$$

- **Girth (or Jet Width)**:

$$g = \frac{\sum_i p_{T,i} \Delta R_i}{\sum_i p_{T,i}}$$

where $p_{T,i}$ is the transverse momentum of particle i, and ΔR_i is its distance from the jet axis in (η, ϕ) space.

- **Jet Area**: The geometric area occupied by a jet in the detector.

- **ptD**:

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

This measures the spread in the transverse momentum of particles inside the jet.

Parton Showers and Fragmentation:

The evolution of a quark or gluon after production is described by the *DGLAP equations*, which govern how partons branch into other partons:

$$\frac{\partial f(x, Q^2)}{\partial \ln Q^2} = \frac{\alpha_s(Q^2)}{2\pi} \sum_j \int_x^1 \frac{dz}{z} P_{ji}(z) f\left(\frac{x}{z}, Q^2\right)$$

where: - $f(x, Q^2)$ is the parton distribution function.

- $P_{ji}(z)$ is the splitting function.

- $\alpha_s(Q^2)$ is the strong coupling constant.

Machine Learning in Jet Physics:

Since quark and gluon jets overlap significantly in these substructure variables, simple cut-based methods are not enough. Machine learning algorithms like **Multilayer Perceptrons (MLPs)** are used to classify jets by combining many variables together in a non-linear way.

An MLP is a type of neural network with:

- An input layer (for substructure variables)
- One or more hidden layers (with activation functions)
- An output layer (giving the probability of the jet being a quark or gluon)

Classifier Evaluation:

The performance of a classifier is measured using:

- **ROC Curve**: Plots True Positive Rate (TPR) vs False Positive Rate (FPR).
- **AUC (Area Under Curve)**: Larger values indicate better classification.
- **Likelihood Ratio**:

$$\mathcal{L} = \frac{P(\text{quark jet})}{P(\text{quark jet}) + P(\text{gluon jet})}$$

CMS Detector and Open Data:

The CMS (Compact Muon Solenoid) detector is a large, general-purpose detector at the LHC. It records the particles produced in collisions using:

- Tracker: For charged particle tracks.
- Calorimeters: For measuring energy.
- Muon Chambers: For detecting muons.

CMS has released a part of its data to the public under its **Open Data** policy, allowing researchers worldwide to perform independent studies.

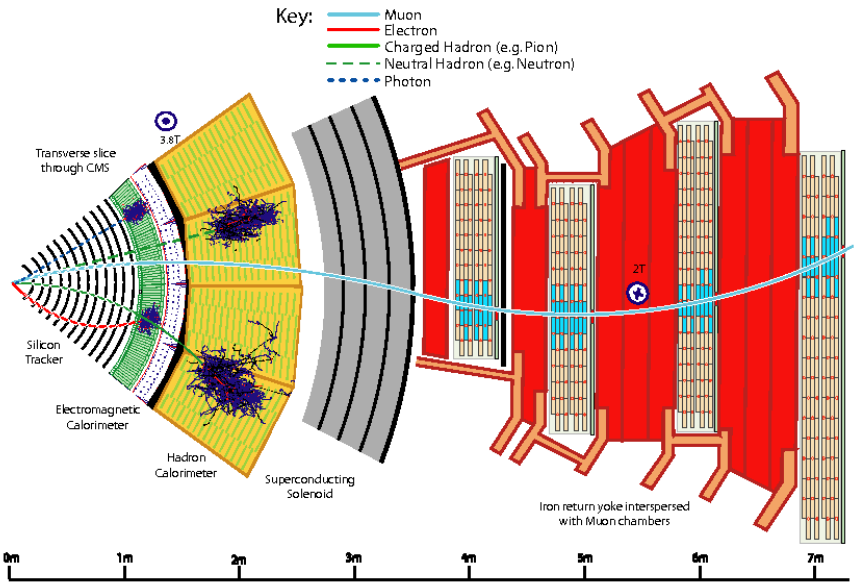

Figure 2.2: Cross-section of the CMS detector showing different parts relevant for jet detection.

Chapter 3

Data and Tools

3.1 Data Sample

This analysis is based on simulated Monte Carlo (MC) samples corresponding to proton-proton (pp) collisions at a center-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$. The dataset is publicly available through the **CMS Open Data** initiative, providing high-quality data and documentation to the broader scientific community.

The specific dataset used is:

- **Dataset:** `RunIISummer16 MiniAOD` simulation
- **File:** `JetNtuple_RunIISummer16_13TeV_MC_17.root`
- **Tree:** `/AK4jets/jetTree`

Jets are reconstructed using the anti-k_T clustering algorithm with a distance parameter of $R = 0.4$ (commonly referred to as AK4 jets). Each jet is associated with generator-level information, allowing identification of its origin.

In this analysis:

- Jets matched to light-flavored quarks (u, d, s) are selected using:

$$\texttt{isPhysUDS} == 1$$

- Jets matched to gluons are selected using:

$$\texttt{isPhysG} == 1$$

No additional kinematic cuts (such as on transverse momentum or pseudorapidity) are applied. Only generator-level information is used for jet labeling, ensuring a clean and unbiased training sample for classification.

3.2 Input Variables

A total of eleven jet substructure variables are selected as input features for the multivariate classification task. These variables are chosen based on their sensitivity to differences in quark and gluon jet properties, such as jet width, particle multiplicity, and radiation patterns.

The selected input variables are:

1. `QG_ptD`: Jet p_T-D variable (momentum dispersion inside the jet)
2. `QG_axis2`: Second principal axis (width) of the jet transverse profile
3. `QG_mult`: Jet particle multiplicity (number of constituents)
4. `jetMass`: Mass of the jet
5. `jetGirth`: Girth variable (radial energy profile)
6. `jetArea`: Effective area of the jet in (η, ϕ) space
7. `jetChargedHadronMult`: Number of charged hadrons within the jet
8. `jetNeutralHadronMult`: Number of neutral hadrons within the jet
9. `jetChargedMult`: Total number of charged particles
10. `jetNeutralMult`: Total number of neutral particles
11. `jetMult`: Total number of particles (charged + neutral)

These variables are used without any transformations unless otherwise specified during training. Normalization and preprocessing are internally handled by TMVA during classifier training.

3.3 Software Tools and Environment

The analysis is performed entirely using C++ and the ROOT framework. The main tools employed are:

- **ROOT** (version 6.xx): A software framework widely used in high-energy physics for data processing, histogramming, and statistical analysis.
- **TMVA** (Toolkit for Multivariate Data Analysis): A ROOT-integrated library providing tools for training and evaluating machine learning models such as decision trees, neural networks, and likelihood classifiers.
- **Docker**: Containerization technology used to create a reproducible and portable computational environment. All software dependencies are encapsulated within Docker images, ensuring consistency across different machines.

All data handling, classifier training, evaluation, and plotting are implemented through custom C++ ROOT macros. No external Python libraries are used in this work.

3.4 Jet Labeling Strategy

The correct labeling of jets as either quark-initiated or gluon-initiated is critical for supervised training. Instead of relying on reconstructed kinematic properties, generator-level information is used directly:

- Jets with `isPhysUDS` flag set to 1 are labeled as **quark jets**.
- Jets with `isPhysG` flag set to 1 are labeled as **gluon jets**.

This labeling ensures that the classifier learns from genuine quark and gluon jets, minimizing contamination and mislabeling effects.

Chapter 4

Methodology

4.1 Multivariate Analysis (MVA)

Multivariate analysis (MVA) is a powerful statistical tool used to analyze data with multiple variables simultaneously. In high-energy physics, MVA techniques are particularly useful for distinguishing between different classes, such as signal and background events. Instead of using a single variable for classification, MVA methods take advantage of the entire set of features (input variables) to improve classification accuracy.

Two commonly used MVA techniques in particle physics are the likelihood-based method and artificial neural networks, specifically the Multilayer Perceptron (MLP). Both methods aim to classify data into categories (e.g., quarks vs. gluons) based on the input variables that describe the events.

4.1.1 Likelihood-based Method

The likelihood-based method calculates the probability that a given event belongs to a particular class (signal or background) by modeling the probability distribution of the input variables for each class. This approach relies on Bayes' theorem and the assumption that the input variables are conditionally independent given the class. The likelihood for an event with input variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is given by:

$$\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

Where:

- $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the vector of input variables.

- $p(x_i|\theta)$ is the probability density function (PDF) of the i-th variable given model parameters θ.
- The product runs over all input variables.

For classification, we compute the likelihood for both classes (signal and background), and the class with the higher likelihood is chosen as the predicted class. The decision rule can be written as:

$$C = \arg\max_{C} \mathcal{L}_C(\mathbf{x}; \theta)$$

Where C represents the class (signal or background) and $\mathcal{L}_C$ is the likelihood for class C. This method assumes the independence of input features and uses the likelihood ratio to discriminate between classes.

4.1.2 Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) is a type of artificial neural network (ANN) used for classification tasks. Unlike the likelihood-based method, which is based on probabilistic models, MLP is a machine learning approach that learns a non-linear mapping between the input features and the output class.

MLP Architecture

The MLP consists of multiple layers: an input layer, one or more hidden layers, and an output layer. The input layer receives the features (input variables), the hidden layers process the information with non-linear activation functions, and the output layer produces a class prediction. Each layer is connected by weights and biases that are updated during the training process.

The output of each hidden layer $\mathbf{h}_i$ is calculated as:

$$\mathbf{h}_i = \sigma(\mathbf{W}_i \cdot \mathbf{x} + \mathbf{b}_i)$$

Where:

- $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid activation function.
- $\mathbf{W}_i$ is the weight matrix for the i-th layer.
- $\mathbf{b}_i$ is the bias term for the i-th layer.

- $\mathbf{x}$ is the input vector (features).

The final output y is computed using the sigmoid function again in the output layer:

$$y = \sigma(\mathbf{W}_2 \cdot \mathbf{h}_3 + \mathbf{b}_2)$$

Where:

- $\mathbf{h}_3$ is the output of the third hidden layer.
- $\mathbf{W}_2$ is the weight matrix for the output layer.
- $\mathbf{b}_2$ is the bias term for the output layer.

Training and Loss Function

The MLP is trained using backpropagation (BP) to minimize the loss function. The backpropagation algorithm adjusts the weights based on the gradient of the loss function with respect to the weights. The weights are updated using the following rule:

$$\mathbf{W}_i \leftarrow \mathbf{W}_i - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}_i}$$

Where:

- η is the learning rate.
- $\mathcal{L}$ is the loss function, which is the cross-entropy loss in this case.

The cross-entropy loss function is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Where:

- y_i is the true label for the i-th sample.
- $\hat{y}_i$ is the predicted probability for the class.

MLP Configuration

The configuration of the MLP used in this study shown in (Figure 4.1).

The specific choices in this configuration are:

```
dataloader->PrepareTrainingAndTestTree(
    signalCut, backgroundCut,
    "SplitMode=Random:"
    "NormMode=EqualNumEvents:"
    "!V"
);

factory->BookMethod(
    dataloader, TMVA::Types::kMLP, "MLP",
    "!H:"
    "!V:"
    "VarTransform=N:"
    "NeuronType=sigmoid:"
    "NCycles=300:"
    "HiddenLayers=N+10,N+5,N:"
    "TrainingMethod=BP:"
    "EstimatorType=CE:"
    "EpochMonitoring=true"
);
```

Figure 4.1: MLP Configuration

- **VarTransform=N**: No transformation of the input variables.
- **NeuronType=sigmoid**: Sigmoid activation function is used in the hidden layers.
- **NCycles=300**: The network is trained for 300 cycles (iterations) to ensure convergence.
- **HiddenLayers=N+10, N+5, N**: Three hidden layers with varying numbers of neurons ($N + 10$, $N + 5$, and N).
- **TrainingMethod=BP**: Backpropagation (BP) is used as the training method.
- **EstimatorType=CE**: Cross-entropy loss function is used.
- **EpochMonitoring=true**: Training progress is monitored for each epoch.

Why This Configuration?

The chosen configuration strikes a balance between model complexity and the ability to generalize. The three hidden layers with varying sizes allow the network to learn non-linear patterns while controlling overfitting. The use of the sigmoid activation function ensures smooth outputs that can be interpreted as probabilities for binary classification tasks. The 300 training cycles ensure that the network has enough time to learn the underlying patterns in the data.

This MLP configuration is suitable for the task of quark-gluon discrimination, as it provides a flexible, non-linear model capable of learning complex relationships between the input variables and the class labels.

Chapter 5

Results and Analysis

This chapter presents the performance evaluation and analysis of various classifiers developed for quark-gluon discrimination using jet substructure variables. First, we summarize the performance comparisons across different models. Then, we provide a detailed study focusing specifically on the final MLP model trained with 11 input variables, including its training diagnostics, classification metrics, input variable studies, and physical interpretation.

5.1 Overview of Classifier Comparisons

Several classifiers were trained and compared using different sets of input variables:

- MLP using only 3 CMS-inspired variables: `QG_ptD`, `QG_axis2`, `QG_mult`
- MLP using the remaining 8 additional substructure variables: `jetMass`, `jetGirth`, `jetArea`, `jetChargedHadronMult`, `jetNeutralHadronMult`, `jetChargedMult`, `jetNeutralMult`, `jetMult`
- CMS-style likelihood classifier with 3 variables: `QG_ptD`, `QG_axis2`, `QG_mult`
- Full MLP model with all 11 input variables: `QG_ptD`, `QG_axis2`, `QG_mult`, `jetMass`, `jetGirth`, `jetArea`, `jetChargedHadronMult`, `jetNeutralHadronMult`, `jetChargedMult`, `jetNeutralMult`, `jetMult`

5.1.1 Training Dataset and Configuration

The dataset used consisted of 64819 quark jet events and 64819 gluon jet events. Classifier training was performed using TMVA with the following configuration:

- **Split Mode:** Random

- **Normalization Mode:** Equal number of signal and background events
- **MLP Architecture:** Three hidden layers with $N+10$, $N+5$, and N neurons respectively
- **Activation Function:** Sigmoid
- **Loss Function:** Cross-Entropy
- **Training Method:** Backpropagation

5.1.2 Performance Summary

The area under the ROC curve (AUC) for each classifier is summarized in Table 5.1.

Classifier	**AUC**
CMS Likelihood (3 variables)	0.787
MLP (3 variables)	0.808
MLP (8 variables)	0.819
MLP (11 variables)	0.837

Table 5.1: Comparison of AUC values for different classifiers.

(ROC curves of all four classifiers together, Fig. 5.1)

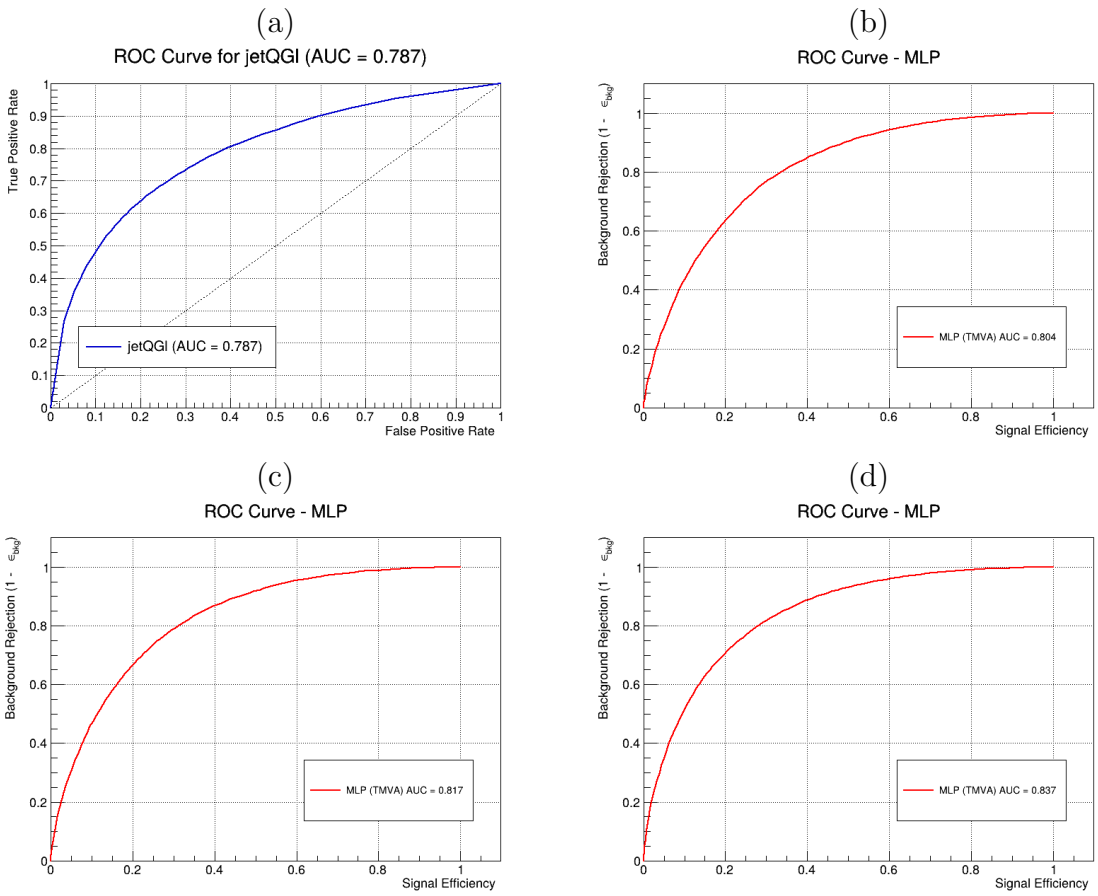

Figure 5.1: ROC curves of different classifiers: (a) CMS likelihood, (b) 3-variable MLP, (c) 8-variable MLP, and (d) 11-variable MLP.

The performance improvement when moving from 3 variables to a full 11-variable model highlights the importance of using additional jet substructure information for quark-gluon separation.

5.1.3 Discriminator Output: Response of Input Variables through MLP

After training the Multilayer Perceptron (MLP) model using the selected 11 input variables, we further analyzed the behavior of each input variable with respect to the model's discriminator output. For each input variable, we plotted the MLP response (i.e., the output node value) as a function of that variable. These plots provide insight into how each input contributes to the final classification decision made by the MLP.

For this study:

- The trained MLP model was applied to an independent test sample.
- The MLP output score was plotted as a function of each of the 11 variables individually.
- Separate curves are drawn for **quark jets (red)** and **gluon jets (blue)** to highlight differences in model response for each class.

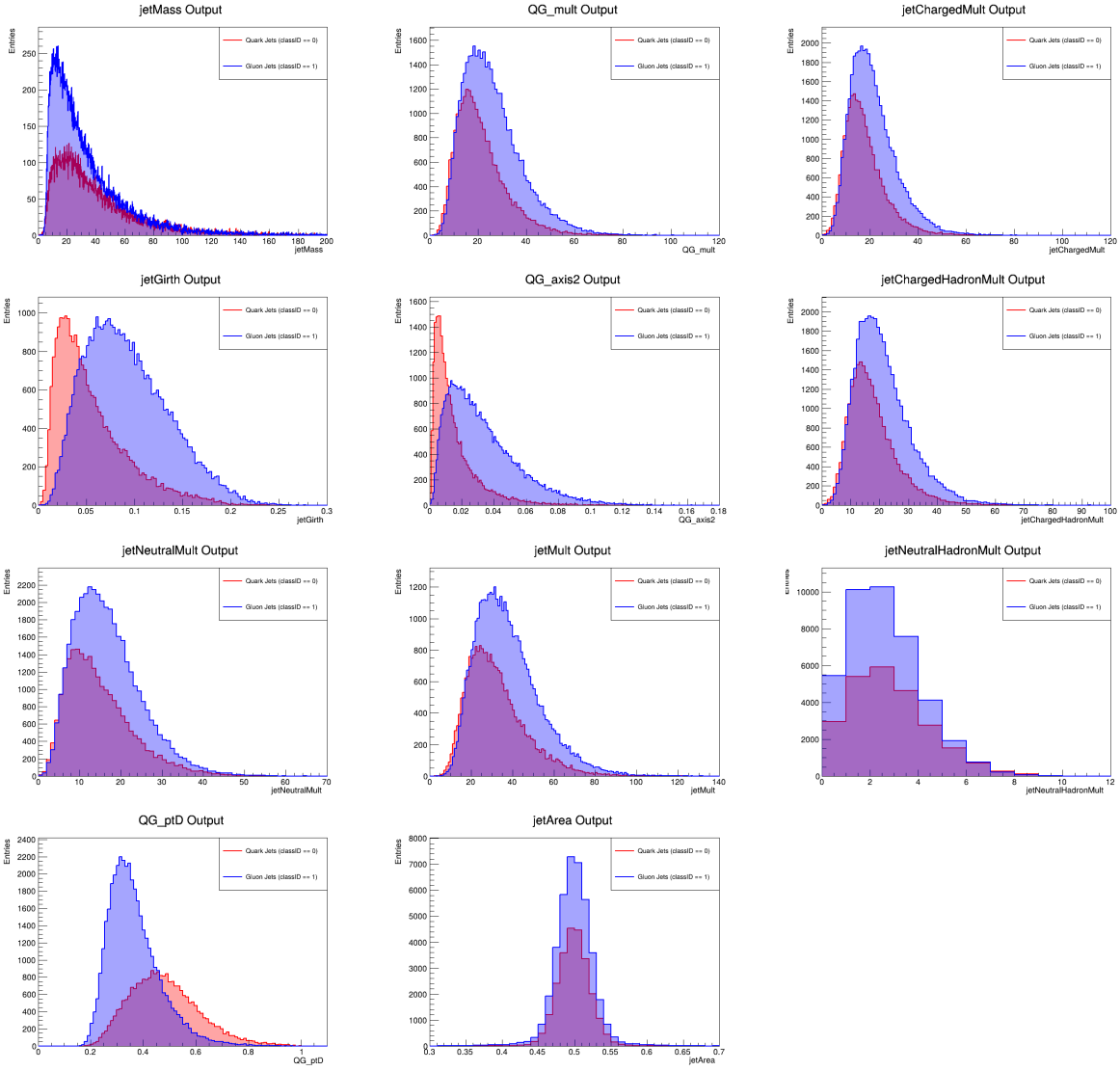

Figure 5.2: MLP output for each variable individually.

These plots help to visualize the **sensitivity** of the MLP output to each variable individually. A strong separation between quark and gluon responses indicates that the MLP has learned to exploit that feature effectively.

5.1.4 Feature-by-Feature ROC Curves with CMS Likelihood Comparison

To better understand the individual discriminating power of each input variable, we computed the Receiver Operating Characteristic (ROC) curve for each variable separately. In addition, we included the ROC curve for the CMS likelihood-based discriminator for comparison.

For each input variable:

- The ROC curve is constructed by using only that single variable to perform quark-gluon classification.
- The Area Under the Curve (AUC) is computed to quantify the performance.

Additionally, the CMS likelihood discriminator is included for direct comparison.

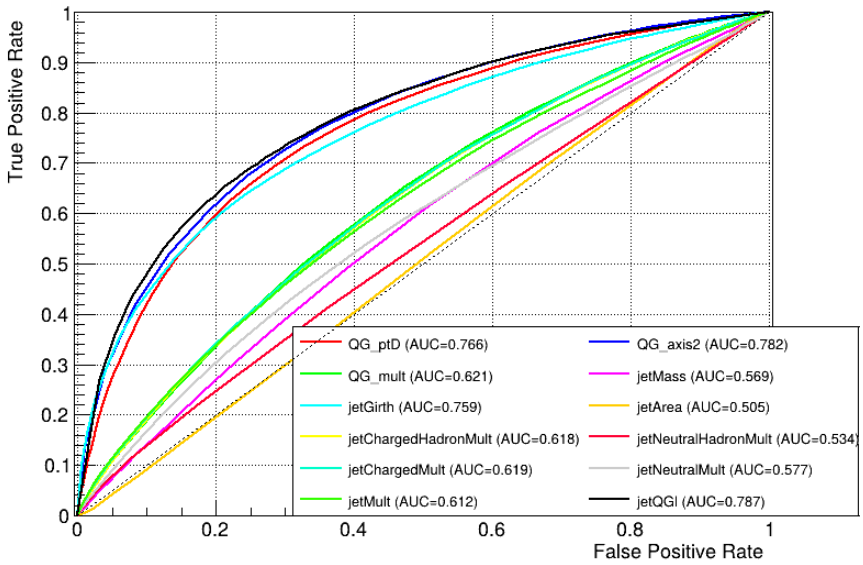

Figure 5.3: ROC curves of features and CMS Likelihood

The ROC curves clearly show that while some variables such as `QG_ptD`, `QG_axis2`, and `QG_mult` individually have strong discriminating power, combining them within a multivariate model such as the MLP leads to significantly improved performance.

5.2 Detailed Study of 11-Variable MLP Classifier

In this section, we focus on a detailed evaluation of the final MLP model trained with 11 normalized input variables.

5.2.1 Training Summary

The MLP model was trained on a balanced dataset of 64819 events each for signal (quark jets) and background (gluon jets). The training was completed in 398 seconds. The training configuration included 11 normalized input variables, listed with their statistical properties in Table 5.2.

Table 5.2: Normalized Input Variables: Mean, RMS, Minimum, and Maximum values for training sample.

Variable	**Mean**	**RMS**	**Min**	**Max**
QG_ptD	-0.153	0.260	-1.000	1.000
QG_axis2	-0.704	0.252	-1.000	1.000
QG_mult	-0.591	0.205	-1.000	1.000
jetMass	-0.873	0.112	-1.000	1.000
jetGirth	-0.484	0.317	-1.000	1.000
jetArea	-0.035	0.087	-1.000	1.000
jetChargedHadronMult	-0.653	0.169	-1.000	1.000
jetNeutralHadronMult	-0.589	0.293	-1.000	1.000
jetChargedMult	-0.650	0.170	-1.000	1.000
jetNeutralMult	-0.549	0.240	-1.000	1.000
jetMult	-0.555	0.215	-1.000	1.000

5.2.2 Input Variable Importance

The MLP ranked the input variables based on their contribution to the classification task. As shown in Table 5.3, `jetMass` emerged as the most influential variable, followed by `QG_mult`, `jetChargedMult`, and `jetGirth`. The variable `jetArea` contributed the least.

5.2.3 Classifier Response and Output Distributions

Figure 5.4 displays the MLP classifier output for signal (quark) and background (gluon) jets. The distributions are sharply peaked towards 1 and 0 respectively, indicating strong classification ability.

Table 5.3: Ranking of input variables by the MLP classifier. Higher values of importance indicate greater relevance to classification.

Rank	Variable	Importance
1	jetMass	385.2
2	QG_mult	61.29
3	jetChargedMult	22.27
4	jetGirth	21.63
5	QG_axis2	21.37
6	jetChargedHadronMult	20.54
7	jetNeutralMult	10.94
8	jetMult	8.35
9	jetNeutralHadronMult	6.33
10	QG_ptD	2.42
11	jetArea	0.36

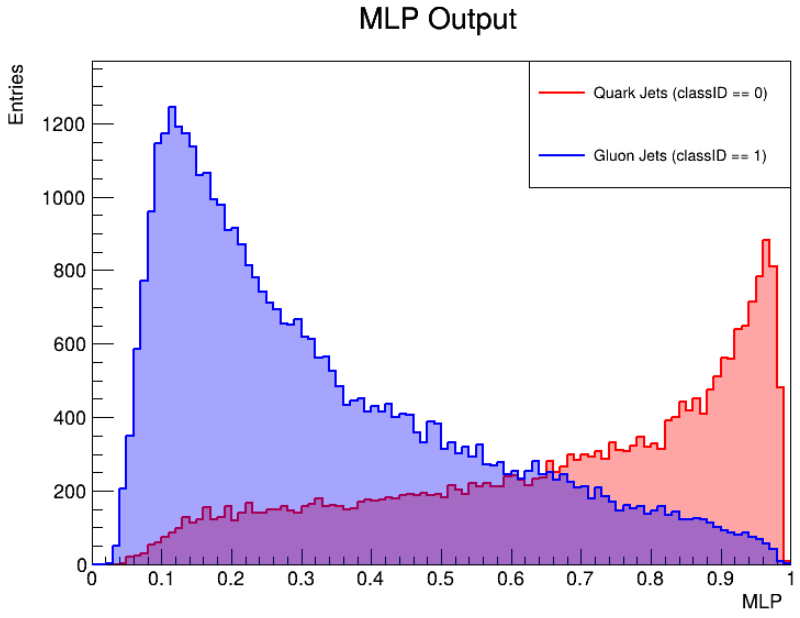

Figure 5.4: Classifier output distribution for signal and background jets.

5.2.4 Performance Metrics at Working Points

Classification performance at a working point corresponding to a 50% signal efficiency was evaluated. Confusion matrix elements were used to compute standard metrics:

$$\text{Efficiency (Recall)} = \frac{TP}{TP + FN} = 74.22\% \tag{5.1}$$

$$\text{Purity (Precision)} = \frac{TP}{TP + FP} = 66.83\% \tag{5.2}$$

$$\text{F1 Score} = 70.33\% \tag{5.3}$$

$$\text{Accuracy} = 76.53\% \tag{5.4}$$

5.2.5 Correlation Matrix of Input Variables

To study inter-variable dependencies, the correlation matrix of input features was computed (Figure 5.5). Low correlations among top-ranked variables confirm their complementarity.

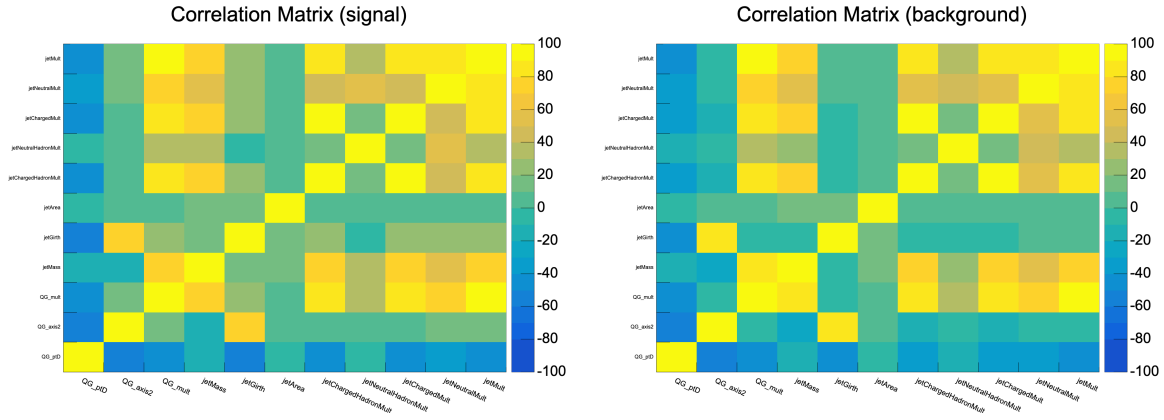

Figure 5.5: Correlation matrix for the 11 input variables used in the MLP: (a) Quark jets signal, (b) Gluon jets background.

5.2.6 2D Distributions of Important Variable Pairs

Joint 2D distributions reveal clear separation trends between quark and gluon jets:

- `QG_axis2` vs. `QG_ptD`
- `jetGirth` vs. `QG_ptD`
- `jetGirth` vs. `QG_axis2`

(2D histograms side-by-side showing quark and gluon densities.)

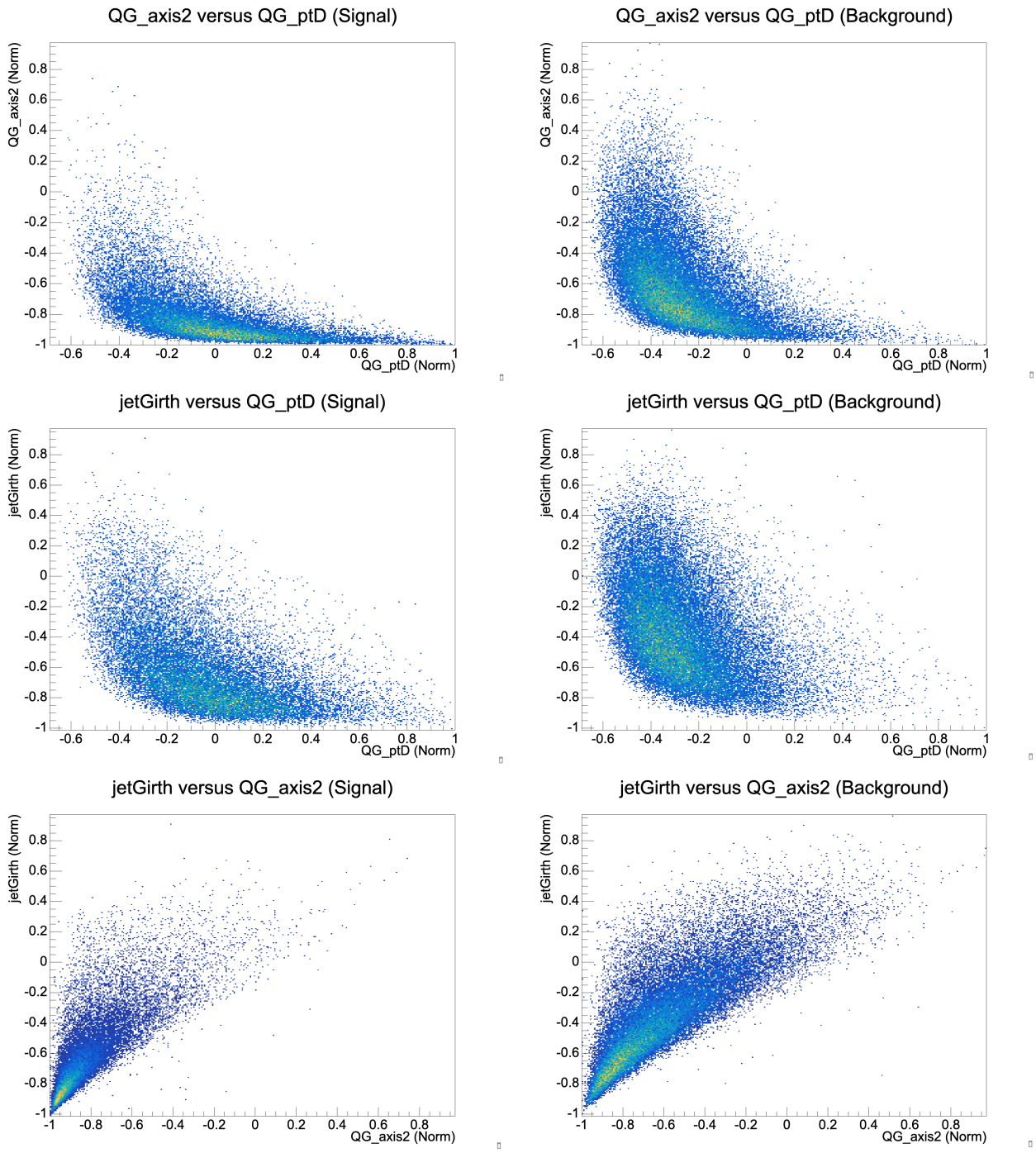

Figure 5.6: 2D distributions: Left - Quark —— Right - Gluon

5.2.7 Physical Interpretation of Input Variables

The rankings and separation power observed can be physically interpreted as follows.

- Gluons radiate more, leading to higher jet multiplicity, shown in Fig. 5.8.

- Gluon jets are broader, shown in Fig. 5.7 at (2,2) position as jetGirth and heavier, shown in Fig. 5.7 at (2,1) position as jetMass which are due to higher color charge.
- Variables such as jet width and multiplicity capture the broader radiation pattern typical of gluons.

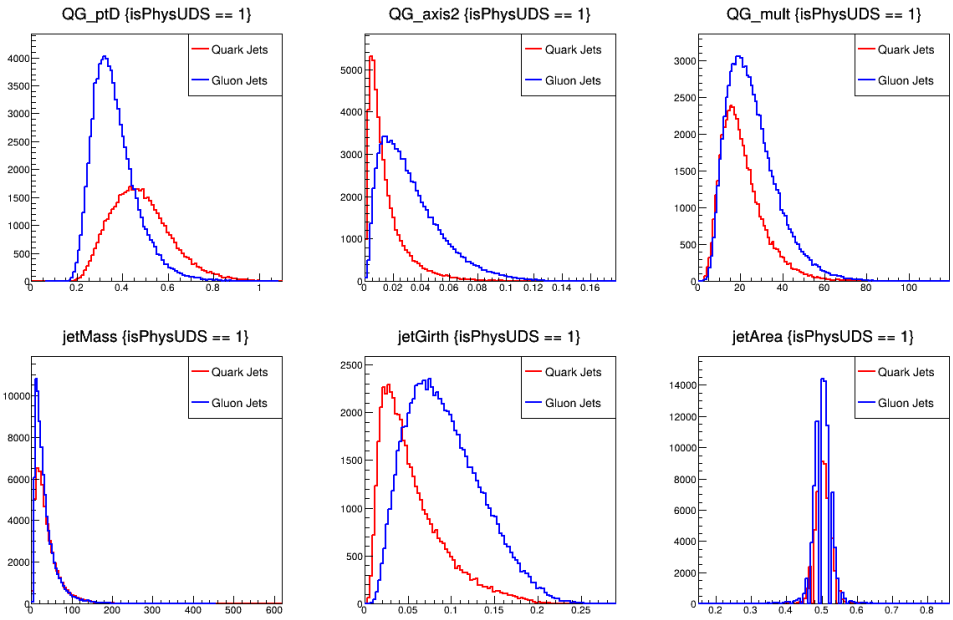

Figure 5.7: Input Variables (6 variables).

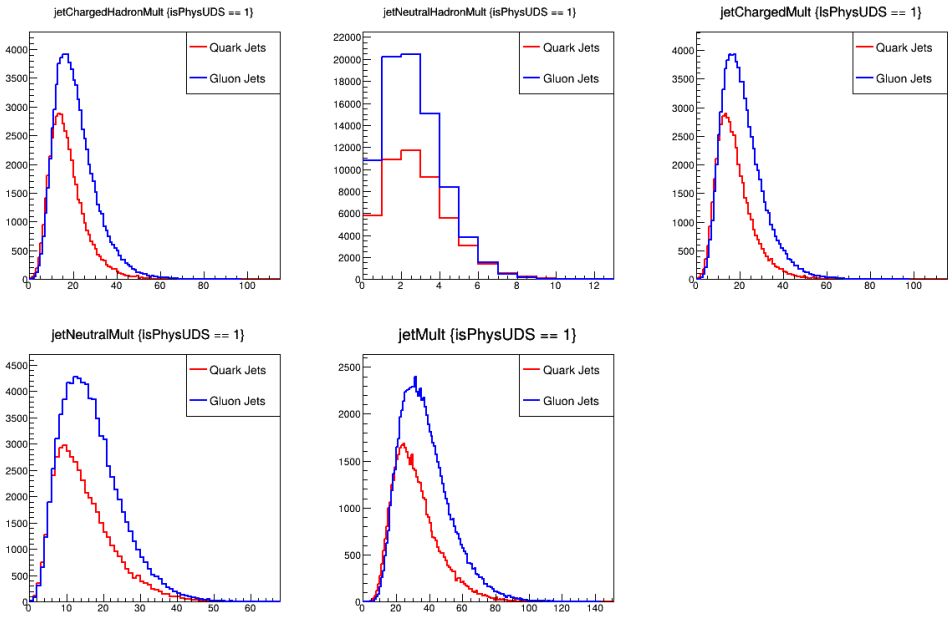

Figure 5.8: Input Variables (Jet Multiplicity)

5.3 Summary

The use of 11 input variables significantly improved quark-gluon separation performance compared to the standard CMS likelihood method. The detailed evaluation confirms the importance of incorporating a broader set of jet substructure observables. Future work can involve exploring even deeper networks or using adversarial methods to further improve generalization across different phase-space regions.

Chapter 6

Discussion

Variable Importance vs Individual Discriminating Power

In the course of this analysis, variable importance rankings obtained from the trained MLP model were compared with the individual discriminating power (as measured by the AUC value) of each input variable. A key observation is that **the order of variable importance in the MLP does not directly match the order based on individual AUC values**. This outcome is natural and can be explained through several interconnected effects:

- **Correlation Effects:**
  Many input variables exhibit mutual correlations. When two variables are correlated, the information provided by one can partially or wholly overlap with the other. As a result, once a correlated variable is included, the marginal contribution of the second variable to the multivariate classifier decreases, even if that second variable alone had strong discrimination power.

- **Information Redundancy:**
  Variables that show good separation individually might not provide significant additional information when combined with other already-available inputs. The MLP, being optimized on the full variable set, can deprioritize redundant variables in favor of inputs offering genuinely new or complementary information.

- **Nonlinear Combinations in MLP:**
  Neural networks like MLPs are capable of learning complex, nonlinear interactions between variables. Consequently, a variable that appears weak when evaluated alone may become highly valuable when combined nonlinearly with others. This allows the

MLP to exploit hidden features in the joint distribution of inputs that are not visible in individual AUC evaluations.

- **Training Dynamics and Feature Utilization:**
  During training, the MLP adjusts internal weights based on overall classification performance. Certain variables may receive larger weight adjustments depending on their ease of optimization and their synergy with other inputs. This adaptive behavior can lead to importance rankings that differ from simple one-variable performance measures.

In summary, **individual AUC values reflect the discriminating power of variables in isolation**, whereas **MLP variable importance reflects their utility within the multivariate context**. The difference between the two rankings highlights the complex interplay of correlations, redundancies, and nonlinear interactions among the input variables.

Variable Importance and QCD Interpretation

The eleven variables used for quark-gluon discrimination capture various physical properties of jet structure, driven by fundamental aspects of Quantum Chromodynamics (QCD). The larger color factor for gluons ($C_A = 3$) compared to quarks ($C_F = \frac{4}{3}$) leads to broader, softer, and more populated jets. This underlies the variable importance rankings observed.

- `jetMass` (385.2): Measures the invariant mass of the jet constituents. Gluon jets, being broader and containing more radiation, tend to have larger masses. The extremely high importance of this variable reflects its strong connection to color radiation differences in QCD.

- `QG_mult` (61.29): Counts the number of jet constituents. The higher radiation rate of gluon jets results in significantly higher multiplicity, making this a strong discriminator.

- `jetChargedMult` (22.27): Counts all charged particles in the jet. Charged multiplicity follows the same trend as total multiplicity, providing complementary discrimination power.

- `jetGirth` (21.63): Measures the radial spread of transverse momentum in the jet. Gluon jets are typically wider, and girth captures this effect independently of mass.

- `jetChargedHadronMult` (20.54): Counts the number of charged hadrons. As gluon jets fragment more, this variable also reflects the increased multiplicity characteristic.

- `QG_axis2` (21.63): Describes the second moment of the energy distribution (width) around the jet axis. A wider width is expected for gluon jets due to their larger color factor.

- `jetNeutralMult` (10.94): Counts neutral particles in the jet. Although less powerful than charged multiplicities, it adds information about the full particle content.

- `jetMult` (8.35): Total multiplicity, combining charged and neutral constituents. It reinforces the information provided by individual charged and neutral counts.

- `jetNeutralHadronMult` (6.33): Counts neutral hadrons such as neutrons and K_L^0. Its moderate importance supports the overall particle count difference between quark and gluon jets.

- `QG_ptD` (2.42): Measures the spread of momentum among constituents. Gluon jets typically have softer constituents, leading to lower ptD values. However, its low importance suggests that other variables already capture the key discriminating information.

- `jetArea` (0.36): Represents the effective spatial area of the jet. As jet area is more influenced by the clustering algorithm than fundamental QCD properties, it contributes minimally to discrimination.

In conclusion, variables linked to jet mass, particle multiplicity, and energy spread dominate the discrimination power, consistent with QCD expectations regarding gluon and quark jet properties. Variables less directly tied to radiation patterns, such as `jetArea`, are of limited utility.

Variable	Importance	QCD Interpretation
jetMass	385.2	Larger jet mass for gluons due to higher radiation and broader jets.
QG_mult	61.29	Higher constituent multiplicity in gluon jets from stronger color factor.
jetChargedMult	22.27	Charged particle multiplicity; gluon jets fragment into more charged particles.
jetGirth	21.63	Radial energy spread; gluon jets are wider than quark jets.
jetChargedHadronMult	20.54	Charged hadron multiplicity; gluons produce more hadrons through fragmentation.
QG_axis2	21.63	Jet width; gluons radiate more widely around the axis.
jetNeutralMult	10.94	Neutral particle multiplicity; gluon jets produce more neutrals.
jetMult	8.35	Total multiplicity; confirms multiplicity trends favoring gluons.
jetNeutralHadronMult	6.33	Neutral hadron multiplicity; provides additional soft radiation information.
QG_ptD	2.42	Momentum dispersion among constituents; gluon jets have softer fragmentation.
jetArea	0.36	Spatial jet size; weakly related to fundamental quark-gluon differences.

Table 6.1: Summary of variable importance and QCD interpretation for quark-gluon discrimination.

Chapter 7

Conclusion

In this chapter, a detailed study was conducted on the performance of eleven jet substructure variables for the task of quark-gluon discrimination. These variables were selected to capture distinct aspects of jet internal structure, including mass, particle multiplicity, radial energy distribution, and momentum dispersion. Using machine learning techniques, specifically a multilayer perceptron (MLP) classifier, we evaluated the importance of each variable during the classification task.

The measured variable importances, summarized in Table 6.1, indicate a clear hierarchy in discriminating power. Variables related to jet mass (`jetMass`), particle multiplicities (`QG_mult`, `jetChargedMult`, `jetChargedHadronMult`, `jetNeutralMult`, `jetMult`), and radial energy spread (`jetGirth`, `QG_axis2`) exhibit the highest importances. Conversely, variables such as `QG_ptD` and `jetArea` show relatively lower importance, suggesting their more limited role in separating quark- and gluon-initiated jets.

This observed trend aligns closely with fundamental Quantum Chromodynamics (QCD) expectations. In QCD, gluons possess a larger color charge ($C_A = 3$) compared to quarks ($C_F = 4/3$), leading to a stronger coupling to the strong interaction field. As a result, gluon jets are characterized by broader angular spread, larger particle multiplicities, softer fragmentation, and higher jet masses compared to quark jets. These theoretical expectations are well captured by the behavior of the measured variables:

- **Jet mass (`jetMass`)** reflects the overall energy spread and virtuality within the jet. Gluon jets, being more radiative, naturally exhibit larger jet masses.
- **Particle multiplicity variables** (`QG_mult`, `jetChargedMult`, `jetChargedHadronMult`, `jetNeutralMult`, `jetMult`) directly quantify the number of final-state constituents, which is higher for gluons.

- **Radial energy distribution variables** such as `jetGirth` and `QG_axis2` are sensitive to the transverse spread of radiation around the jet axis, capturing the broader structure of gluon jets.

- **Momentum dispersion** (`QG_ptD`) provides complementary information, as gluon jets tend to have more evenly spread, lower-p_T constituents, resulting in a softer fragmentation pattern.

- **Jet area** (`jetArea`), while geometrically related to the size of the reconstructed jet, does not show strong discriminating power since it is largely determined by external jet clustering parameters rather than intrinsic QCD radiation properties.

The dominance of jet mass and multiplicity-related observables suggests that the internal complexity and fragmentation pattern of jets are the primary drivers of quark-gluon discrimination performance. This finding is consistent with prior experimental studies and theoretical models, such as those implemented in parton shower Monte Carlo simulations.

It is noteworthy that despite their relatively lower individual importances, variables like `QG_ptD` and `jetArea` may still play a supporting role in improving classifier performance through multivariate correlations. In machine learning contexts, even low-importance features can enhance decision boundaries when combined appropriately with more powerful variables.

The systematic agreement between the variable importance rankings and QCD-driven physical expectations strengthens confidence in both the interpretability and robustness of the quark-gluon tagging framework developed in this work. It emphasizes that machine learning models, when trained carefully, are capable not only of achieving strong classification performance but also of uncovering underlying physical structures within complex datasets.

In summary, the detailed analysis presented in this chapter establishes a clear physical narrative for jet substructure-based quark-gluon discrimination. It highlights the critical role of mass, multiplicity, and radiation width observables, firmly rooted in QCD dynamics, and sets the stage for further optimization and deployment of advanced jet tagging strategies in high-energy collider experiments.

Appendix A

References

1. CMS Open Data
Sample with jet properties for jet-flavor and other jet-related ML studies
Available at: https://opendata.cern.ch/record/12100

2. TMVA
https://tmva.sourceforge.net/citeTMVA.html

3. David Tong: Lectures on Quantum Field Theory
Available at: https://www.damtp.cam.ac.uk/user/tong/qft.html

4. Multilayer Perceptrons in Machine Learning: A Comprehensive Guide
Available at: https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning

5. Software Tools
The ROOT Team. ROOT – Data Analysis Framework. CERN. https://root.cern
Docker Inc. Docker — Portable Containers for Software. https://www.docker.com

6. Additional Computational Tools and Scripts
Docker Image: https://hub.docker.com/r/rootproject/root

7. Jet Reconstruction and Particle Flow Philipp Schieferdecker (KIT)
Available at: https://indico.desy.de/event/3084/contributions/64996/attachments/42087/51952/2010-09-29.desy.jets.pdf

8. Metodiev, E. M. (Figure 2.1)
Jet Formation and Substructure. (2021).
Available at: https://www.ericmetodiev.com/post/jetformation/

9. Davis, Siona Ruth (CERN). (Figure 2.2)
An Interactive Slice of the CMS Detector. CERN Document Server (2020).
Available at: https://cds.cern.ch/record/2205172/