HiLL PUBLISHING

Theoretical Physics and Quantum Mechanics, 2026, 3(1), 1-7
https://www.hillpublisher.com/Journals/tpqm/
ISSN Online: 3066-5744

A Comparison of Machine Learning Classifiers for Quark and Gluon Jet Tagging Using CMS Open Data

Dylan Zhang

Del Norte High School, San Diego 92127, USA.

How to cite this paper: Dylan Zhang. (2026). A Comparison of Machine Learning Classifiers for Quark and Gluon Jet Tagging Using CMS Open Data. *Theoretical Physics and Quantum Mechanics*, 3(1), 1-7.
DOI: 10.26855/tpqm.2026.06.001

Received: September 7, 2026
Accepted: October 15, 2026
Published: November 3, 2026

***Corresponding author:** Dylan Zhang, Del Norte High School, San Diego 92127, USA.

© 2026 by the author(s).
This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) license, which permits non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and is not modified or adapted.
https://creativecommons.org/licenses/by-nc-nd/4.0/

Abstract

Separating jets initiated by quarks from those initiated by gluons has been a persistent problem in particle physics. Because the two types of jets are only different in the color charge of their initiating parton, the differences between them may not always exist on the scale of each individual jet. No singular variable is able to cleanly and consistently differentiate between the two types of jets, so practical taggers all include many variables. We analyze multiple machine learning methods to differentiate jets originating from a quark/gluon, including approaches ranging from the unsupervised k-means clustering and linear logistic regression to random forests, multilayer perceptrons, and a constituent-level transformer, noting the accuracy of each method by comparing resultant quark and gluon tags with Monte Carlo truth information. We show that a simple classifier can, while having a modest number of jet/constituent variables and a relatively small dataset, perform about as well as a far more complex model in certain setups. We also highlighted the fact that, despite testing 16 different transformer setups, all of them resulted in roughly similar accuracy/AUC values (AUCs of about 0.75 to 0.77). The strongest conventional classifiers also achieved AUC values similar to this, so increasing architectural complexity did not lead to a significant gain in discrimination ability in this case. This may be a result of the relatively limited dimensionality and number of variables of the available inputs, motivating future studies with larger datasets and a greater number of variables.

Keywords

Quark–gluon jet tagging; machine learning; jet substructure

1. Introduction

Particle colliders, such as the Large Hadron Collider at CERN, produce jets originating from quarks and gluons. As the original parton can't be directly observed from a jet, the differences between them may not always exist on the scale of each individual jet; however, on a larger scale, quark jets tend to be narrower and lower multiplicity, while gluon jets, on the other hand, tend to be broader and higher multiplicity [1-3]. However, no variable can cleanly separate the two, so any practical tagger has to combine many variables.

Being able to distinguish if a jet is from a quark or gluon is useful because different Standard Model processes and potential new physics signals can contain different proportions of quark and gluon jets. Depending on what's being studied, quark enriched or gluon enriched samples might be needed, so a reliable quark/gluon tagger can help separate the two jet populations and improve collider analyses and therefore make the search for new physics more efficient or easier [1-3].

This work compares models of substantially different complexity on simulated CMS Open Data [4] to identify which variables and model architectures yield the most reliable discrimination between the two types of jets. The study includes unsupervised K-means clustering, logistic regression, random forests, three feed-forward neural networks, and a constituent-level transformer in order to determine whether increasing the complexity of the model used provides a clear benefit for this dataset and input representation.

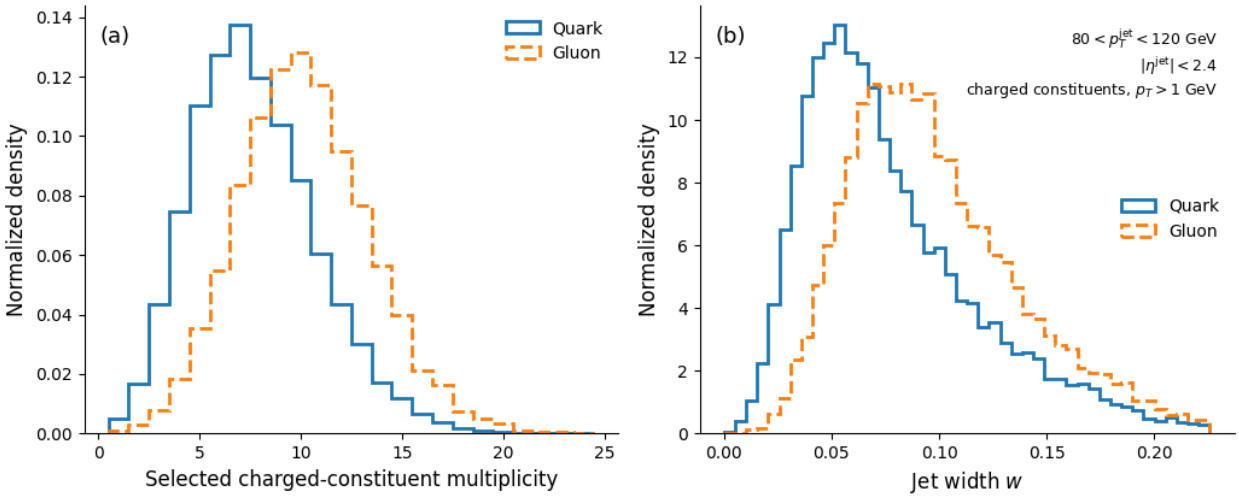

Figure 1. Normalized distributions of (a) selected charged constituent multiplicity and (b) jet width for Monte Carlo quark and gluon labeled jets. The different but overlapping distributions illustrate the statistical differences between quark and gluon initiated jets while also showing why no single variable provides complete separation.

2. Dataset and preprocessing

The simulated dataset used in this study is from the CMS Open Data Portal, and is a simulated QCD dataset corresponding to 2016 proton-proton collision conditions at a center-of-mass energy of

$$\sqrt{s} = 13\,TeV. \tag{1}$$

The sample was generated with Pythia 8 and is available in the CERN Open Data website [4]. Monte Carlo parton-flavor information from the dataset was used as the classification target. As per the Monte Carlo numbering scheme, quark jets were labeled with codes 1 through 8, antiquark jets were labeled with codes -1 through -8, and gluon jets were labeled with 21 [5]. Jets without these numbers were not included in the study. Also excluded were jets with leading transverse momentum

$$p_{T,i} < 1\,GeV \tag{2}$$

First, the following variables were observed from the data for jets and constituents: transverse momentum dispersion (p_T^D), constituent multiplicity, N_{95}, and jet mass. The transverse momentum dispersion was calculated with the formula

$$p_T^D = \frac{\sqrt{\Sigma_i {p_{T,i}}^2}}{\Sigma_i p_{T,i}}. \tag{3}$$

Also calculated were the radial displacement of each constituent from the jet axis, jet width, and leading constituent transverse momentum fraction, defined as

$$\Delta R_i = \sqrt{(\Delta\eta_i)^2 + (\Delta\phi_i)^2}, \tag{4}$$

$$w = \frac{\Sigma_i p_{T,i}\Delta R_i}{\Sigma_i p_{T,i}}, \tag{5}$$

and

$$f_{lead} = \frac{max_i p_{T,i}}{\Sigma_i p_{T,i}}, \tag{6}$$

respectively. Charged multiplicity, neutral multiplicity, mean constituent p_T, and maximum constituent p_T were also computed.

3. Classification methods

An unsupervised k-means baseline was used to test whether these variables alone allowed for an unsupervised method to separate the two populations consistently. The model was fit to 2 clusters with 50 random initializations, and the cluster with a larger fraction of gluon jets was labeled as the gluon-like cluster and vice versa. Supervised logistic regression, random forest, and feed-forward neural network classifiers were then trained. Logistic regression served as the simplest supervised baseline. The random forest classifier consisted of 300 decision trees with maximum tree depth 10 and a minimum of 10 samples per leaf. The feed-forward neural network compared three different feed-forward multilayer architectures. The architectures, (32, 32), (64, 64), and (10, 10, 10, 10), represent the number of neurons in each successive hidden layer: two 32-neuron layers, two 64-neuron layers, and four 10-neuron layers, respectively. These choices allowed for comparing architectures with moderate width, increased width, and increased depth, all while keeping the overall models relatively small.

Lastly, the transformer differed from the previous architectures in that it represented each constituent by the following quantities:

$$[\log(1+p_T), \Delta\eta, \Delta\phi, \log(1+m), q, \frac{p_T}{\Sigma_j p_{T,j}}]$$

where $\Delta\eta$ and $\Delta\phi$ are measured relative to the jet axis and q represents the constituent charge. This change in variables was made for the more flexible model of the transformer to see if it could learn useful substructure directly from the particles inside each jet instead of relying only on summary variables such as multiplicity or jet width. Constituents were ordered by decreasing p_T.

Then, a grid search tested

$$d_{model} \in \{32, 64\},$$
$$n_{heads} \in \{2, 4\},$$
$$n_{layers} \in \{1, 2\},$$

and

$$N_{constituents} \in \{30, 50\},$$

meaning that d_{model} could be either 32 or 64, n_{heads} could be either 2 or 4, and so on. Every possible combination of these choices was trained, totalling to 16 different transformer configurations.

For optimization of the transformers, 30% of the data was used as a test set, 59.5% was for training, and 10.5% was for validation.

4. Evaluation

The performance of each architecture was compared and quantified with the area under the receiver operating characteristic curve (ROC AUC). However, classification accuracy was also measured. The reason the ROC AUC was preferred to the classification accuracy is because the accuracy is based on a classification threshold and on class composition, while the ROC AUC evaluates ranking performance over all thresholds. As they measure different things, the model with the greatest classification accuracy might not have the highest ROC AUC, and vice versa.

The current analysis lacks statistical confidence intervals for point estimates, so differences in ROC AUC that are very small, like a few 10^{-3}, should not be taken as necessarily statistically significant.

5. Results

5.1 Comparison of classification methods

All supervised architectures achieved higher ROC AUC values than the unsupervised K-means baseline, as expected. K-means reached an AUC value of 0.7051, compared with AUC values ranging from 0.7582 to 0.7715 for the supervised models. Among the supervised methods, the performance and AUC values seemed to be very close, and for some, like random forest and feed-forward 64-64, nearly identical.

Table 1. Summary of the test accuracies and ROC AUC values for each method, sorted by ROC AUC values from high to low

Method	Test Accuracy	Test ROC AUC
Random forest	0.7292	0.7715
Feed-forward NN 64-64	0.7315	0.7714
Feed-forward NN 32-32	0.7329	0.7702
Constituent transformer	0.7377	0.7694
Feed-forward NN 10-10-10-10	0.7308	0.7669
Logistic regression	0.7194	0.7582
K-means	0.6804	0.7051

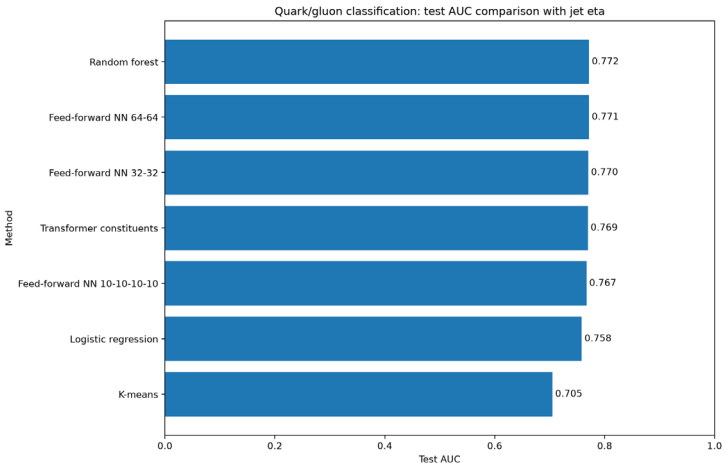

Figure 2. Test ROC AUC for the seven approaches. Differences among the random forest, feed-forward networks, and transformer are small compared with the difference between the supervised methods and K-means.

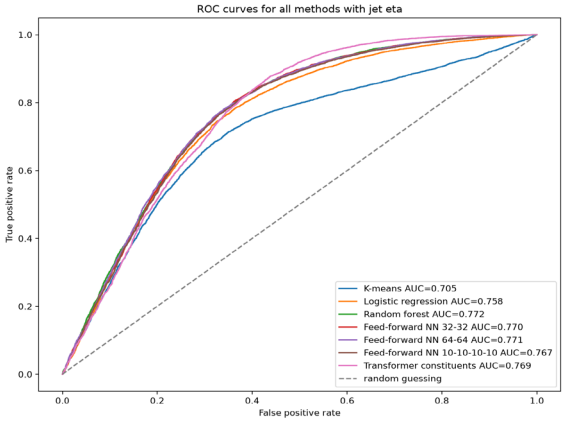

Figure 3. ROC curves for the seven classification approaches. The dashed line represents random classification.

While the transformer's AUC value was only the 4th highest out of all the methods, it had the highest accuracy value. At the default decision threshold, the transformer's confusion matrix was

$$\begin{pmatrix} 3577 & 3373 \\ 842 & 8279 \end{pmatrix},$$

where the rows represent true quark and gluon labels and the columns represent predicted quark and gluon labels. This highlights that the transformer identified approximately 90.8% of gluon-labeled jets but only 51.5% of quark-labeled jets at the default threshold.

5.2 Transformer hyperparameter scan

Table 2. Summary of the four hyperparameters varied in the transformer scan, totalling to 16 total transformer architectures

Hyperparameter	Values tested
Embedding dimension d_{model}	32, 64
Attention heads n_{heads}	2, 4
Encoder layers n_{layers}	1, 2
Maximum constituents $N_{constituents}$	30, 50

The transformer hyperparameter scan showed nearly no variation between the 16 transformer architectures, with validation AUC values only ranging from 0.7518 to 0.7552.

The small AUC range suggests that changing the edited hyperparameters within the considered ranges did not lead to a large change in performance for the transformer.

The configuration selected according to highest validation ROC AUC was

$$d_{model} = 64, n_{heads} = 2, n_{layers} = 1, N_{constituents} = 30,$$

with a validation AUC of 0.75518.

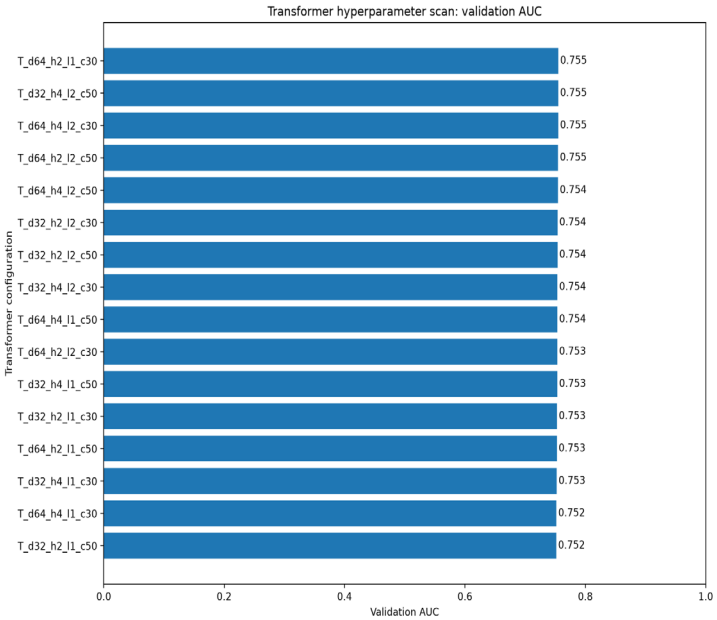

Figure 4. Validation ROC AUC for the 16 transformer configurations. The architecture with the greatest validation AUC was selected for the final transformer.

6. Discussion

The main and immediate takeaway from the results is that even very different model classes achieved similar levels of supervised discrimination of jet types. A random forest using engineered jet-level variables, relatively small feed-forward networks, and a constituent transformer all reached ROC AUC values of around 0.77, suggesting that increasing the model complexity didn't lead to a substantial increase in the level of the supervised discrimination in the current data representation and training setup. One possible explanation is that the amount of information available to the models was limited, since more complex architectures usually have more opportunity to perform better than simple architectures when they have higher dimensional inputs and more training data.

While the results of this study showed no major difference in the performance of different model classes, this does not mean that transformers or other deep learning models aren't able to help improve quark/gluon tagging in any circumstance. More sophisticated constituent-based architectures have had strong performance on jet-tagging benchmarks [6-8]. The result indicates that, in this case and in this setup, increasing the model complexity didn't lead to a substantial increase in the level of the supervised discrimination, though this may be simply because of the relatively small number of input variables used.

There are also several limitations to the study that could be addressed in future research: for example, the lack of statistical uncertainty measurements for the AUC values prevents definitively stating that one model outperforms another model. Future research should calculate these uncertainties. Similarly, a larger set of variables and a larger dataset could benefit future research.

7. Conclusion

This work analyzed multiple machine learning methods to differentiate jets originating from a quark/gluon, including approaches ranging from the unsupervised k-means clustering and linear logistic regression to random forests, multilayer perceptrons, and a constituent-level transformer, noting the accuracy of each method by comparing resultant quark and gluon tags with Monte Carlo truth information.

K-means produced a test ROC AUC value of 0.7051, while the other, supervised classifiers produced AUC values ranging between 0.7582 and 0.7715. The random forest architecture produced the greatest AUC value of 0.7715, while the constituent transformer produced the greatest test accuracy value of 0.7377, with an AUC of 0.7694.

Testing 16 different transformer configurations showed that they all produced roughly similar AUC values, with validation AUC values ranging from 0.7518 to 0.7552. The small AUC range suggested that changing embedding dimension, attention-head count, encoder depth, or maximum constituent count within the considered ranges did not lead to a large change in performance for the transformer. The selected transformer with highest validation AUC used a 64-dimensional representation, two attention heads, one encoder layer, and at most 30 constituents.

The similar AUC values of the multiple different supervised models that were tested suggest that, under the tested experimental conditions, increasing architectural complexity alone doesn't guarantee better discrimination between gluon and quark jets.

The comparison can act as a benchmark of how much classification performance is affected by engineered jet-substructure variables versus constituent-level learning. The study illustrates that the choice of how data is represented and the evaluation procedure can be at least as important as the complexity of the machine-learning architecture itself.

References

[1] Gallicchio J, Schwartz MD. Quark and gluon tagging at the LHC. Phys Rev Lett. 2011;107:172001.

[2] Tumasyan A, Adam W, et al. Study of quark and gluon jet substructure in Z+jet and dijet events from pp collisions. J High Energ Phys. 2022;2022:188.

[3] Larkoski AJ, Metodiev EM. A theory of quark vs. gluon discrimination. J High Energ Phys. 2019;2019:14.

[4] CMS Collaboration. Simulated dataset QCD_Pt_80to170_bcToE_TuneCP5_13TeV_pythia8 in MINIAODSIM format for 2016 collision data. CERN Open Data Portal. 2024. DOI:10.7483/OPENDATA.CMS.5NJW.EBHC

[5] Krauss F, Navas S, Richardson P, Sjöstrand T. Monte Carlo particle numbering scheme. In: Zyla PA, et al., editors. Review of Particle Physics. Prog Theor Exp Phys. 2020;2020:083C01.

[6] Komiske PT, Metodiev EM, Thaler J. Energy flow networks: deep sets for particle jets. J High Energ Phys. 2019;2019:121.

[7] Qu H, Gouskos L. Jet tagging via particle clouds. Phys Rev D. 2020;101:056019.

[8] Qu H, Li C, Qian S. Particle transformer for jet tagging. In: Proceedings of the 39th International Conference on Machine Learning. Vol. 162. PMLR; 2022. p. 18281-92.