ML-Based Top Taggers: Performance, Uncertainty and Impact of Tower & Tracker Data Integration

Rameswar Sahu [ORCID] [1, 2, *] and Kirtiman Ghosh [1, 2, †]

[1] *Institute of Physics, Bhubaneswar, Sachivalaya Marg, Sainik School, Bhubaneswar 751005, India*
[2] *Homi Bhabha National Institute, Training School Complex, Anushakti Nagar, Mumbai 400094, India*

Machine learning algorithms have the capacity to discern intricate features directly from raw data. We demonstrated the performance of top taggers built upon three machine learning architectures: a BDT that uses jet-level variables (high-level features, HLF) as input, while a CNN (miniature version of ResNet) trained on the jet image, and a GNN (LorentzNet) trained on the particle cloud representation of a jet utilizing the 4-momentum (low-level features, LLF) of the jet constituents as input. We found significant performance enhancement for all three classes of classifiers when trained on combined data from calorimeter towers and tracker detectors. The high resolution of the tracking data not only improved the classifier performance in the high transverse momentum region, but the information about the distribution and composition of charged and neutral constituents of the fat jets and subjets helped identify the quark/gluon origin of sub-jets and hence enhances top tagging efficiency. The LLF-based classifiers, such as CNN and GNN, exhibit significantly better performance when compared to HLF-based classifiers like BDT, especially in the high transverse momentum region. Nevertheless, the LLF-based classifiers trained on constituents' 4-momentum data exhibit substantial dependency on the jet modeling within Monte Carlo generators. The composite classifiers, formed by stacking a BDT on top of a GNN/CNN, not only enhance the performance of LLF-based classifiers but also mitigate the uncertainties stemming from the showering and hadronization model of the event generator. We have conducted a comprehensive study on the influence of the fat jet's reconstruction and labeling procedure on the efficiency of the classifiers.

I. INTRODUCTION

Since its commencement, the Large Hadron Collider (LHC) [1] at CERN has been looking for evidence of physics beyond the Standard Model (BSM). While the discovery of the Higgs [2, 3] is a remarkable success for the LHC experiment, it serves to reinforce solely the legitimacy of the Standard Model (SM). The absence of any solid evidence supporting BSM physics has motivated researchers to progressively explore higher energy scales. Such high energies facilitate the production of boosted heavy SM particles like the top quark, W/Z-boson, and the Higgs boson. The hadronic decays of the boosted SM particles lead to a collimated cluster of quarks, manifesting as large radius (large-R) single jets (fat jets) with distinctive features. At the LHC, the sub-structure features of fat jets resulting from the hadronic decay of boosted top quarks, W/Z-bosons or the Higgs boson have been widely utilized[1] to search for heavy BSM resonances within various new physics scenarios such as supersymmetry [6–8], extra-dimensional

models, leptoquark models, different gauge and field extensions of the SM [4, 5, 9–11], etc. Efficient identification of the particle identity of the fat jet becomes essential to improve the sensitivity of the LHC and future colliders. This necessitates a substantial shift in the analysis strategy and demands the development of new and innovative methodologies for tagging the particle identities of the fat jets.

Traditional methods for jet-tagging rely on constructing high-level discriminants from the jet substructure information, the so-called jet substructure observables [2] [19–30]. The usefulness of jet substructure observables in distinguishing large-R jets resulting from hadronically decaying boosted heavy SM particles over QCD quark/gluon jets has been demonstrated and widely accepted by experimental collaborations [31–36]. Incorporating machine learning-based techniques into the task of jet classification opens up new and unique directions. These methods leverage the fine granularity of the LHC detectors to construct highly specialized observables from the four-momentum of the constituents of the jet. Over the past decade, various neural network-based architectures have been developed [37–62] and have demonstrated substantial enhancement in the classifi-

---

* rameswar.s@iopb.res.in
† kirti.gh@gmail.com

[1] Considering the hadronically decaying boosted top quarks or W/Z bosons offers several advantages when designing search strategies for the heavy BSM resonances that decay into massive SM particles. On the one hand, the enhanced hadronic decay branching ratios of top quarks or W/Z bosons lead to a higher signal rate. The hadronic decay products of top quarks or W/Z bosons being visible at the LHC detectors enables the kinematic reconstruction of the decay cascade for specific BSM resonances [4, 5].

---

[2] Jet substructure observables are not only valuable for tagging boosted SM heavy particles like top quarks, Z/W-bosons, and Higgs bosons, but their significance in distinguishing between quark and gluon jets has also been demonstrated recently in the literature [12–18].

cation efficiencies compared to traditional substructure-based techniques. Apart from architectural complexity, these algorithms differ in the representation of the input dataset. While Linear classifiers [49] and BDTs [63, 64] are trained on jet-level observables constructed from the jet substructure information, Convolutional Neural Networks (CNNs) [40, 41, 44, 52, 54, 55, 65–71], Recurrent Neural Networks (RNNs) [71–74], Graph Neural Networks (GNNs) [46, 57, 75–85], Recursive Neural Networks (RvNNs) [45, 61, 86, 87], Fisher's Linear Discriminant [56], Locally Connected Networks [53], etc. are directly trained on pure or transformed four-momentum data[3] of the jet constituent. On the other hand, Multi-Layer Perceptrons (MLPs) can be trained on both jet-level observable data [48, 88–91] as well as constituents four-momentum data [39, 51, 58, 72].

In the present analysis, we focus on classifying fat jets resulting from hadronically decaying boosted top quarks from light quarks and gluon jets (here onwards, QCD jets). Theoretically, the top quark is especially interesting because of its high Yukawa coupling. The large top Yukawa coupling plays not only a crucial role in the computation of electroweak precision observables [92] and determining the vacuum stability [93] of the SM, but also significantly influences the masses and interactions of several BSM resonances, many of which have enhanced couplings with the top quark, resulting in a top quark rich final state at the LHC. As the heaviest SM particle, the top quark decays into a b quark and a $W^{\pm}$-boson. The subsequent $W^{\pm}$-boson decays can yield either a 3-quark final state or a combination of a b-quark and an SM-charged lepton, accompanied by missing transverse energy. Conventional searches at the LHC primarily rely on the leptonic decays of top quarks to suppress the huge SM QCD background. Although leptonic decays reduce the SM background contributions, the suppressed leptonic branching ratios of the top quark lead to reduced signal strength. Additionally, missing transverse energy from the elusive neutrinos in the final state complicates the reconstruction of the top quark's 4-momentum and the decay cascade of BSM resonances which lead to the top-rich final states. While the hadronic decays of the top quark simultaneously solve these two issues, the hadronic decay of the top quark into three resolved jets suffers from a huge QCD background. Effectively distinguishing these large-R jets arising from boosted top quark decays (top-tagged jets) from QCD jets is key to suppress QCD background for BSM scenarios featuring top quark-rich final

states at the LHC. To distinguish top-tagged jets from QCD jets, we have focused on three different machine-learning algorithms: a Boosted Decision Tree that uses high-level features for training, a miniature version of the ResNet [94] that uses image representation of jets as input, and LorentzNet [75], a symmetry-preserving GNN founded upon the concept of Lorenz equivariance. We have also considered composite classifiers by stacking a BDT-based classifier on top of the ResNet and LorentzNet to leverage the high-level features in BDT and low-level inputs in CNN/GNN in a single tagger.

The ubiquity of top-rich final states in the context of BSM searches has led to their status as extensively studied signatures at the LHC. Over the past decade, numerous endeavors have been undertaken in the literature to develop effective methods for efficiently distinguishing boosted top quark jets from QCD jets. Although cut-based strategies for boosted top tagging, which rely on substructure information from fat jets resulting from the hadronic decay of boosted top quarks, were introduced in the literature as early as 2008 [20, 21], recent years have witnessed a surge in the utilization of machine learning-based approaches for the classification of top-jets from QCD jets. While the CMS collaboration continues to conduct boosted top final state searches, relying on top-tagging achieved through selection cuts on jet shape observables [95–97], the ATLAS collaboration [98], in contrast, employs a diverse array of top-tagging algorithms. These span from cut-based optimization to jet-moment-based multivariate taggers, jet cluster-based deep neural network taggers, etc. For a recent review of boosted object tagging algorithms used by the LHC experimental collaborations, we refer the interested readers to Ref. [35]. In recent years, various top jet taggers leveraging modern machine learning algorithms such as BDT [63, 64], MLP [39, 51, 91], CNN [40, 41, 66, 99], RNN [73], RvNN [86], and GNN [46, 57, 75, 77, 83–85], etc. have emerged, demonstrating considerable enhancements in classification performance compared to the current classifiers used by the CMS and ATLAS collaborations. Many of these new classifiers are trained on either pure or transformed 4-momentum data of jet constituents derived from the HCAL and ECAL calorimeter towers. For instance, LorentzNet introduced in [75] was trained and tested on a dataset that uses calorimeter towers as jet constituents. The dataset was initially introduced in [100] for comparing the performance of different Neural Network-based classifiers for top-tagging. This dataset has subsequently been adopted in multiple analyses [57, 77, 101–106] to evaluate the efficiency of various top taggers. While this dataset provides a common benchmark for comparing top taggers, the absence of tracking information renders it incomplete and unsuitable for assessing a tagger's absolute performance. The calorimeters at the LHC have a fixed granularity, and as the transverse momentum of top jets increases, the energy deposition by jet constituents becomes more compact. This compactness results in reduced resolution

---

[3] While representation of a jet as a gray-scale (single layer) or color (multi-layer) image are used to train convolutional neural networks (CNNs) [40, 41, 44, 52, 54, 55, 65–71], Fisher discriminant analysis [56], locally connected networks [53], and Multi-Layer Perceptrons (MLPs) [50], the Graph Neural Networks GNNs [46, 57, 75–85] are trained on the particle cloud (graph) representation of jets. Similarly, jet-based tree-structured data [45, 61, 86, 87] can also be used in Recursive Neural Networks (RvNNs) and GNNs.

in variables constructed using calorimeter towers. To address this issue, researchers have turned to the finer spatial granularity of inner detectors, leveraging tracking information to improve their analyses. The benefits of using high-resolution tracking data for improving top-tagging in the high transverse momentum region have already been demonstrated in previous work [39]. Moreover, the principles of Quantum Chromodynamics (QCD) as well as various experimental findings, suggest that jets initiated by light quarks or gluons exhibit distinct differences in the distribution and composition of charged and neutral hadrons during their hadronization process. Exploiting the characteristics of light quark and gluon hadronization, several classifiers [107–109] based on tracking information have been developed for the classification of quark vs. gluon jets. Given this context, it is essential to investigate the impact of tracking information on top-tagging algorithms, as it can significantly enhance their performance. This work will study the critical importance of combining information from calorimeter towers and the tracker detector. This combined information provides insights into the composition and distribution of charged and neutral hadrons within a jet, ultimately playing a crucial role in determining the performance of the classifiers.

It is important to note that the efficiency (classifier's efficiency) of an ML classifier is different from the efficiency (top-tagging efficiency) of identifying a fat jet resulting from a parton-level hadronically decaying boosted top quarks as a top-tagged fatjet. The classifier's efficiency depends on how the fat jet reconstruction and labeling criteria are employed when constructing the dataset for training and testing. Using stringent criteria for truth-level top tagging during signal sample preparation results in a dataset with a high level of purity. Consequently, taggers trained on such high-purity datasets tend to show superior performance. However, such a high-performance classifier might also result in poor top-tagging efficiency if a significant number of fat jets fail to pass the truth-level tagging criteria during the preparation of signal samples for training and testing. While using the same truth level top tagging as in Ref. [100] to ensure the comparability of our results with the existing top classifiers in the literature, we tried to optimize the performance of the classifier to maximize the top-tagging efficiency by using different radius fat jets in different transverse momentum regions.

The rest of the paper is organized as follows. In section II, we discuss the dataset used to train the classifiers. Section III discusses the various model architectures used for our analysis. In Section IV and V, we discuss the effect of tracking information and truth-level identification efficiency on classifier performance. In Section VI, we discuss the variation of classifier efficiency with the transverse momentum of the fat jets. Finally, Section VII summarises our observations.

II.   DATASET

A significant portion of our analysis focuses on establishing the importance of incorporating the information from tracker and calorimeter towers into the datasets[4] used for training and testing the Machine Learning (ML) based classifiers designed to identify hadronically decaying boosted top jets over the QCD light quark and gluon jets. To fulfill this objective, we have trained our classifiers on datasets generated following two different approaches. One dataset (denoted as $DATA_{calo}$ in the rest of the manuscript) only incorporates the information stored as energy deposits in the hadronic and electromagnetic calorimeter towers. The second dataset denoted as $DATA_{trck}$ extends the previous one by incorporating the information regarding the electric charge of the charged constituents from the trackers. In the former case, we have generated the datasets following the prescription of Ref. [100]. The authors of Ref. [100] have studied various top taggers and assessed their performance based on a dataset that only contains information on the jet constituents coming from the calorimeter energy deposits. Though that dataset[5] serves its purpose of comparing the performance of various top taggers, as we will demonstrate in the subsequent sections, the same dataset is inadequate in providing a given ML algorithm's optimal performance. To make our case, we have compared the performance of LorentzNet, a symmetry-preserving Graph Neural Network (GNN) for top tagging [75], on datasets generated using these two approaches. We have also performed similar exercises for the Boosted Decision Tree (BDT) and Convolutional Neural Network (CNN) based top taggers.

Another objective of this work is to study the performance of top taggers for different transverse momentum ranges of the hadronically decaying top quark. For this section of the analysis, we have divided the p_T range between 300 to 1500 GeV into six bins of size 200 GeV each. We then generate large-R jets resulting from the hadronically decaying boosted top quarks (the signal jets) and QCD production of light quarks and gluons (the background jets) in these bins and train and test our classifiers for each p_T bin.

All large-R [6] signal and background fat jets are generated in MG5_AMC@NLO [110] with the NNPDF21LO [111] PDF. The hadronically decaying boosted top samples are generated from the SM process $pp \rightarrow t\bar{t}$. Similarly, for

————

[4] By dataset, we imply the signal (hadronically decaying boosted top quark) and background (QCD generated quark and gluon) fat jets used for our analysis.

[5] The same dataset have been used in several subsequent analyses [57, 77, 101–106] for accessing the performance of their proposed classifiers

[6] The reconstruction radius (radius of the cone used to define the fat jets in FastJet) plays a crucial role in determining the identification efficiencies of fat jets resulting from boosted top quarks. We will discuss this issue in the next section.

the background fat jets (QCD production of quark and gluon), we have used the process $pp \rightarrow jj$ (where j includes u, c, d, s, g and their anti-particles). Subsequent decay of the top quarks and showering and hadronization of the light quarks and gluons are simulated in Pythia8 [112]. To simplify the analysis, we have not included the effect of Multi Parton Interaction (MPI) and PileUp. Finally, we have used Delphes [113] with the default Atlas card to include the detector efficiencies and resolutions. The fat jets are reconstructed in Fastjet [114] using the anti-kT algorithm. To reconstruct the sub-jets inside a fat jet, we use the jet trimming [115] algorithm with default parameters $R_{trim} = 0.2$ and $p_{T,trim} = 0.05$, which gives us subjets with $R = 0.2$. For each p_T range defined in the previous paragraph, we have generated one million top quarks and one million QCD jets for our final analysis. For training purposes, we selected 600k fat jets from each category while reserving 200k from each class to validate and test the classifiers. For training and testing the composite classifiers (see section III), we have generated additional 400k fat jets from each category for training and 200k from each category for testing.

Before proceeding to the next section, we mention the convention followed in our analysis to construct the constituents of a fat jet, namely the tracks and calorimeter towers. Throughout our analysis, we adopt two different conventions; in one, we use the TrackMerger/tracks and Calorimeter/towers classes of Delphes to construct the tracks and towers. We refer to them as tracks and towers in the subsequent sections. In the second convention, we use the HCal/eflowTracks to construct the tracks while we combine the ECal/eflowPhotons and HCal/eflowNeutralHadrons classes to construct the calorimeter towers. These are referred to as Etracks and Etowers in the subsequent discussion. The only difference between the two approaches is that, in the latter case, Delphes performs a matching between the track and calorimeter energy deposits to filter out the calorimeter towers originating from the charged particles and classify them as tracks.

A. Truth-level tagging (TLT)

The quality of the training dataset has a big impact on how well a classifier performs. In our specific example, the classifier's performance is significantly influenced by the purity of the signal (hadronically decaying boosted top quarks) datasets[7]. Improved classifier efficiency results from using more pure signal data in training. Therefore, to prepare the signal samples for training

and analyzing the performance of any classifier, we need a method to select only those fat jets that are properly reconstructed. We achieve this objective by matching the fat jets and the constituent sub-jets with their partonic counterparts. This process is referred to as truth-level tagging (TLT), and we name the efficiency of a hadronically decaying boosted top quark to be associated with a properly reconstructed fat jet as the truth-level identification efficiency (ϵ_S^{truth}). To ensure a valid comparison with the performance of the available classifiers in the existing literature, we adopt the following simple truth-level tagging criteria, introduced in Ref.[100] and subsequently employed in several other references [57, 77, 101–106]: A fat jet to be tagged as a top fat jet at truth level, we require that both the partonic top and its three daughter quarks lie within the cone of that fat jet. No truth-level tagging criteria are applied to the QCD fat jets.

The truth-level identification efficiency depends on two factors: the reconstruction radius (RR) of the fat jet and the transverse momentum of the hadronically decaying top quark. If the transverse momentum of the top quark is not large enough, the decay products of the top quark will not be collimated enough, and we will require a large radius fat jet to capture all the hadrons arising from the hadronization of the three light quarks resulting from top decay. At the same time, if we have a top quark with very high transverse momentum, all the hadronic constituents resulting from the high-p_T top quark will reside inside a small cone. In this case, if we choose a very large radius of reconstruction, the fat jet will pick contributions from the background radiation, which will directly affect the resolution of various features of the fat jet and hence, the performance of the classifiers. The way out is to use a jet tagging algorithm with a variable radius of reconstruction [116, 117], which is beyond the scope of our analysis. Instead, we work with different reconstruction radii for fat jets in the six transverse momentum regimes mentioned in the previous section. In Figure 1, we present the variation of truth-level identification efficiency (ϵ_S^{truth}) with the radius of fat jets for the six p_T bins. For our final analysis, we choose the RR for the top fat jets in the different p_T bins such that we can achieve a notable ϵ_S^{truth} without being concerned about the distortion of crucial jet characteristics caused by background radiation. Ergo, we reconstruct fat jets in [300, 500] GeV and [500, 700] GeV p_T bins with a $R = 1.2$. While for the remaining four p_T bins, we fix the value of R at 0.8.

B. Extracting the data

The classifiers addressed in the remaining manuscript can be categorized into three primary groups: Boosted Decision Tree (BDT-classifiers), Convolutional Neural Network (CNN-classifiers), and Graph Neural Network (GNN-classifiers). The nature and structure of training and testing datasets for these three classifier groups differ significantly. While the BDT classifiers use high-level

---

[7] One or more constituents may reside outside the jet reconstruction cone, rendering the signal sample impure. For instance, in the case of a fat top jet, the final fat jet is fundamentally a W-jet rather than a top-jet if the b-quark sits outside the reconstruction cone.
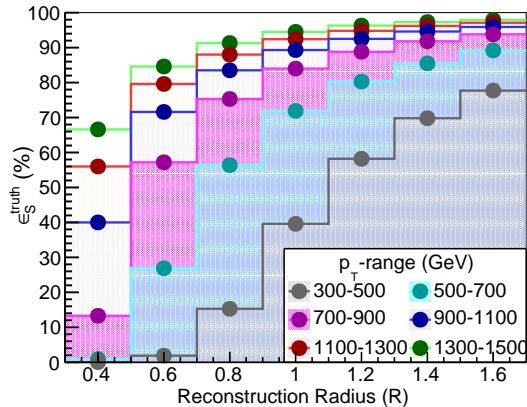
FIG. 1. The variation of truth-level identification efficiency with the jet radius in different transverse momentum ranges.

variables/features (HLF) (invariant mass of the fat jet, N-subjettiness, etc.) as input, GNN or CNN classifiers, on the other hand, use low-level features (LLF) such as the four-momentum of the constituents or jet image constructed from the transverse momentum of the jet constituents. We will discuss the nature and structure of the datasets used for training and testing these three groups of classifiers in the following.

1. BDT

The BDT uses high-level variables/features (HLFs)[8] for classification. For each signal and background fat jet, we use the information of the constituent tracks and calorimeter towers to construct the desired high-level features[9]. In this work, we studied two different BDT-based classifiers. The first BDT classifier, referred to as the tower-based BDT classifier or BDT_{calo}, utilizes the five most commonly used high-level features (HLFs) for top tagging : the invariant mass, three ratios of N-subjettiness variables, and the b-tagging information. In the case of the second BDT classifier, referred to as the track-based BDT classifier or BDT_{trck}, supplementary HLFs are designed using information from the tracker detector. The goal of the tracker detector at the LHC is to trace the paths of charged hadrons, thereby offering insights into the electrically charged constituents of the jet. The HLFs used for BDT_{calo} and BDT_{trck} classifiers are discussed in the following:
As discussed above, the BDT_{calo} uses five HLFs,

---

[8] These are functions of low-level variables like the four-momentum and position in the $\eta - \phi$ plane of the jet constituents.

[9] For this section of our analysis, we have used the track and tower class of Delphes to construct the fat jet constituents.

- **The invariant mass** of the fat jets :

$$M = \sqrt{\sum_i (E_i)^2 - \sum_i (p_i)^2}, \qquad (1)$$

where the sum runs over all constituents of the fat jet.

- **The N-subjettiness variable** τ_N :

$$\tau_N = \frac{1}{\sum_k p_{Tk} R_0^\beta} \sum_k p_{Tk} min(\Delta R_{1,k}^\beta .. \Delta R_{N,k}^\beta), \qquad (2)$$

where the sum runs over the constituents with transverse momentum $p_{T,k}$, R_0 is the radius parameter used in the clustering algorithm, $\beta = 1$ is the thrust parameter, and $\Delta R_{i,k}$ characterize the separation between the constituent k and the candidate sub-jet i. In our analyses, we use three ratios of the N-subjettiness variables τ_{43}, τ_{32}, and τ_{21}, where $\tau_{mn} = \tau_m/\tau_n$.

- **b-tag**: We consider a fat jet b-tagged when there is at least one b-tagged $\Delta R = 0.4$ sub-jets inside the cone of the fat jet, i.e., $\Delta R(J, j_b) < R_0$.

In the BDT_{trck} classifier, we extend the above list by including 21 HLFs. Most of these HLFs are discussed in [18, 109, 118–121], we summarise them here for completeness:

- N_{trk}: It characterizes the number of tracks inside a jet.

- w_{trk}: The p_T weighted width of the tracks:

$$w_{trk} = \frac{\sum_{trk \in J} p_{T,trk} \Delta R_{trk,J}}{\sum_{trk \in J} p_{T,trk}} \qquad (3)$$

- w_{calo}: the E_T weighted width, defined as:

$$w_{calo} = \frac{\sum_{i \in J} p_{T,i} \Delta R_{i,J}}{\sum_{i \in J} p_{T,i}}, \qquad (4)$$

where the sum runs over the jet constituents with transverse momentum $p_{T,i}$.

- E_{frac}: the ratio of the energy of the hardest constituent to the jet's energy:

$$E_{frac} = \frac{E_{hardest}}{E_J} \qquad (5)$$

• C_β: the two-point energy correlation function:

$$C_\beta = \frac{\sum_{i,j \in J} E_{T,i} E_{T,j} (\Delta R_{i,J})^\beta}{(\sum_{i \in J} E_{T,i})^2} \quad (6)$$

For our analysis, we use a value of 0.2 for β.

• **The Jet Charge**: the p_T weighted sum of the charge of the constituent tracks:

$$Q_k = \frac{\sum_i q_i (p_{Ti})^k}{\sum_i p_{T,i}} \quad (7)$$

Where k, the regularisation exponent, has a value of 1 for our analysis.

• ΔR_{sub}: The ΔR separation between the sub-jets inside a fat jet. They constitute a set of three variables $\Delta R_{1,2}$, $\Delta R_{2,3}$, and $\Delta R_{3,1}$. The numbers in the subscript denote the p_T-ordered sub-jets.

Note that the variables (N_{trk}, w_{trk}, w_{calo}, E_{frac}, C_β, and Q_k) are defined for each sub-jet inside a fat jet. For our analysis, we store the information of the first three highest p_T sub-jets. Therefore, they constitute a set of 18 variables for each fat jet. The absent variables are zero-padded for a fat jet with less than three sub-jets.

In summary, the BDT_{calo} classifier uses a set of five variables, and the BDT_{trck} classifier uses a list of 26 variables of which five are the ones used in BDT_{calo}, three are the ΔR separation between the three highest p_T subjets and six features (N_{trk}, w_{trk}, w_{calo}, E_{frac}, C_β, and Q_k) for each of these three subjets.

2. CNN

The Convolutional Neural Network (CNN) uses grid-shaped data or images for classification tasks. The units in an image are referred to as pixels, and each pixel is associated with the pixel intensity. For our analysis, we have used the transverse energy[10] of the tracks and towers as pixel intensities. As mentioned in Section II, we use two different datasets to demonstrate the importance of tracking information in enhancing the performance of the classifiers.

The first dataset only uses the information of the calorimeter energy deposits[11] to construct the images. Therefore these images have only one layer and are of dimension 64×64. The process of constructing these images is slightly different than the conventional methods.

We demonstrate this with a simple example. Suppose we have a fat jat with R=0.8. If we convert it into an image with dimension 64×64, we end up with pixels of dimension 0.025×0.025 — significantly smaller than the HCal resolution. To circumvent this, we first divide the jet into pixels of size 0.1×0.1, commensurate with the HCal resolution. This will result in an image with dimension 16×16. To get the final image with dimension 64×64, we further divide each pixel of intensity $E_{T,i}$ into a 4×4 grid where each final pixel caries an intensity $E_{T,i}/16$. In the subsequent discussions, we refer to the CNN trained on this dataset as CNN_{calo}.

In the Second dataset, we use the information of both tracks and calorimeter towers to construct a two layers image of dimension $2 \times 64 \times 64$. Here we make use of the Etrack and Etower classes of delphes. We adhere to the above image generation procedure for the layer constructed from the Etower class. The situation is, however, different for the layer constructed from the Etrack class. Since the tracks at LHC are recorded from particle flow information, the position of the tracks in the $\eta - \phi$ plane can be measured with high accuracy. This allows us to split the jet directly into a 64×64 image. In the subsequent discussions, we named the CNN trained using the second dataset CNN_{trck}.

To boost our taggers' performance, we process each image using a similar method as described in [40, 122]. The pre-processing steps make use of the sub-jets inside a fat jet. In Figure 2, we present the evolution of top and QCD images[12] through subsequent preprocessing stages. For better comparison, we present the top and QCD images side-by-side. First, we centralize the images such that the sub-jet with the highest E_T shifts to the origin of the $\eta - \phi$ coordinate system (see the first row of Figure 2). We see a widespread distribution of constituents in the top image. The energy in the QCD jets is distributed near the center, demonstrating its origin from a single parton. Next, we rotate the image so that the next-to-highest E_T sub-jet lies below the first sub-jet. In the absence of a second sub-jet, we rotate the image around the jet-energy centroid until the image's principal axis [56, 123] becomes vertical. We present the resulting average images in the second row of Figure 2. We see the clear appearance of a second hard structure for the top jet and a diffusive radiation pattern for the QCD jets. Then, we reflect the image such that the sum of pixel intensities on the right-hand side of the image is higher than that on the left-hand side (third row of Figure 2). Finally, we normalize the image by dividing each pixel intensity by the sum of the intensities of all pixels.

---

[10] The transverse energy is defined as $\frac{E}{cosh\eta}$.

[11] Here, we utilize the tower class of delphes to reconstruct the constituent information of the fat jets.

[12] The images demonstrated here result from averaging over 10000 individual images. This averaging makes the structures in the image more visible.
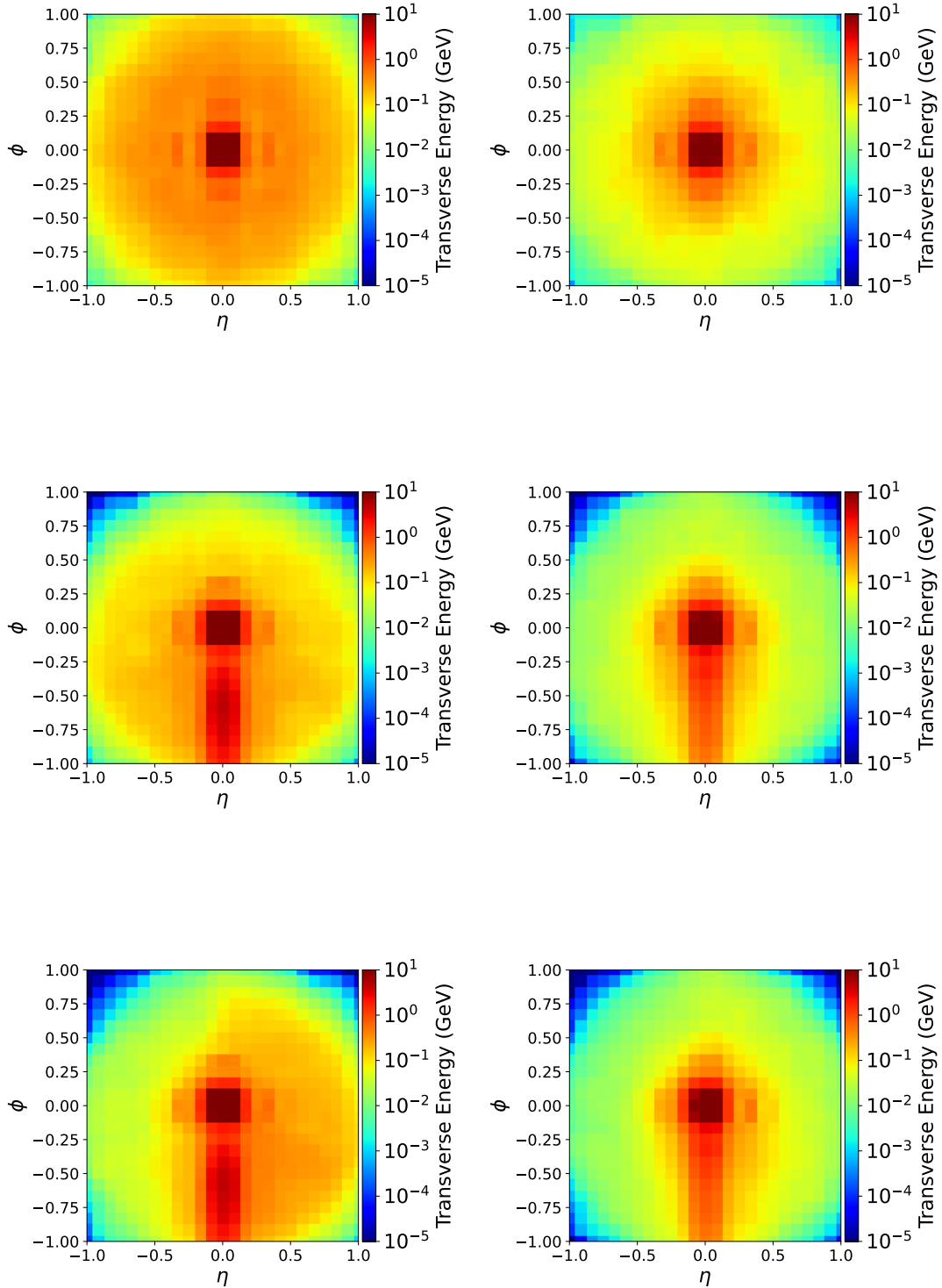
FIG. 2. Different image preprocessing stages of the top image (left) and the corresponding QCD image (right).

3. GNN

Like CNN, we construct two datasets to study the performance of GNN. The first dataset uses the tower class

of Delphes to construct the jet constituents, while the sec-

ond uses the Etrack and Etower classes. In both these cases, we store the four-momentum of the first 200 highest p_T constituents and their charge for each fat jet. If the fat jet has less than 200 constituents, we fill the remaining entries with zero. For the Etowers and towers, we set the charge to be zero, while for the Etracks, the charge can take value ± 1. There is one important point to note; at the LHC, the mass of the tracks is measured from the curvature in the magnetic field and the momentum of the tracks. Later they match this mass with the mass of a physical particle by following a matching scheme. Nevertheless, we have refrained from incorporating details regarding the particle identity of the charged track. Instead, our approach solely relies on the electric charge information of the constituents. This decision not only diminishes the classifier's sensitivity to the specifics of the hadronization model but also minimizes uncertainties stemming from the tagging or mis-tagging efficiencies of charged hadron identities. In the subsequent discussions, we follow a simplified approach and make the Etracks massless by hand to mask the identity of the charged hadron to the classifiers, i.e., we only use the information of the three momenta of the Etracks and set the energy as the magnitude of three momenta.

III. MODELS

In this section, we will discuss the architecture of the Machine Learning (ML) classifiers used in our analysis. We have organized our discussion in order of the complexity of the ML classifiers. First, we discuss a simple cut-based classifier, the Boosted Decision Tree. Next, we discuss the architecture of a CNN classifier that works with image-shaped data. Finally, we will demonstrate the architecture of a Graph Neural Network (GNN) where the input is graph-structured data.

A. BDT

The BDT_{calo} (see II B 1) classifier uses a small set of HLFs, emphasizing the importance of invariant mass, N-subjettiness variables, and b-tagging information in discriminating signal top fat jets from QCD light quark and gluon background jets. The classifier BDT_{trck} focuses more on the HLFs resulting from identifying the jet's charged constituents from the tracker detector. In addition to the above variables, BDT_{trck} includes several other track-based features that characterize the composition of charged and neutral hadrons inside the sub-jets and the fat jet. For a consistent performance comparison, both BDTs have the same hyper-parameters and are trained using the TMVA 4.3 toolkit [124] integrated into ROOT 6.24 [125] analysis framework. Table I summarizes these hyper-parameters.

BDT hyperparameter	Optimised choice
NTrees	1000
MinNodeSize	5%
MaxDepth	4
BoostType	AdaBoost
AdaBoostBeta	0.1
UseBaggedBoost	True
BaggedSampleFraction	0.5
SeparationType	GiniIndex
nCuts	40

TABLE I. Summary of optimised BDT hyperparameters.

B. CNN

The CNN model used in our analysis is a miniature version of the original ResNet model [94]. The ResNet architecture was originally designed to solve the vanishing gradient problem in very deep Neural Networks. ResNet uses the principle of residual connections that allows it to maintain a stable gradient propagation throughout the network. This residual/skipped connection passes the input of the ResNet block directly to the output along with the learned features. Mathematically,

$$x_{i+1} = x_i + F(x_i), \tag{8}$$

where x_i represents the input to the ResNet Block, x_{i+1} is the output, and $F(x_i)$ represents the residual function, a collection of non-linear operations[13]. The architecture of the ResNet block and the full CNN model is presented in Appendix A.

The ResNet model is trained with PyTorch on a single Nvidia Tesla K80 GPU. The model is trained for 35 epochs with a batch size of 32. We use the ADAMW [126] optimizer with a weight decay of 10^{-2} and an initial learning rate of 10^{-3} to minimize the Cross-Entropy loss function. We reduce the learning rate by half for the first five epochs. After that, the learning rate is reduced at a rate of 10%, and for the last five epochs, we reduce the learning rate by 90 % per epoch. We check the model's performance after every epoch on the validation dataset, and the model with the best validation accuracy is used for the final test.

C. GNN

We use the LorentzNet [75], a symmetry-preserving deep Neural Network, for the GNN part of our analysis. LorentzNet utilizes the Lorentz group equivariance principle [77] to construct the Neural Network's layers. This means under Lorentz transformation, the output of the

---

[13] The Convolution and Normalisation operations are few examples.

neural network follows the transformation of the input, i.e.,

$$x \to F(x) \ and \ \Lambda(x) \to F(\Lambda(x)) \implies F(\Lambda(x)) = \Lambda F(x) \tag{9}$$

Here, x is the input to the neural network layer, $F(x)$ is the output, and Λ represents the Lorentz transformation. The graph neural network operates on graph-structured data [127, 128]. A graph is a collection of nodes and edges, i.e., G(V, E), where $V = x \oplus h$ denotes the nodes, and E denotes the edges between the nodes. Each node is characterized by a node coordinate x, which in our case is the four-momentum of the jet constituents, and a node attribute/embedding h, which for our analysis is the charge of the constituents. LorentzNet does not assume any prior knowledge regarding the relationship between the nodes. In other words, it uses fully connected graphs. The action of the NN layers on the graph generates new or updated graphs, and this graph updating happens in three simple steps. We start with the information of the node coordinates and node attributes and use it to define the edges between the node i and j at the lth message passing step as [75],

$$m_{ij}^l = \phi_e \left(h_i^l, h_j^l, \psi(||x_i^l - x_j^l||^2), \psi(<x_i^l, x_j^l>) \right) \tag{10}$$

Here ϕ_e is a non-linear function modeled by neural networks, $\psi(.) = sgn(.)log(|.| + 1)$, $< x_i, x_j >$ is the Minkowski dot product and $||x_i - x_j||^2$ is the Minkowski norm. Note that the edges in the graph are Lorentz invariant. The next step is to update the node co-ordinate [75],

$$x_i^{l+1} = x_i^l + c \sum_{j \in [N]} \phi_x(m_{ij}^l) x_j^l \tag{11}$$

here the sum runs over the neighbourhood of the point x_i the number c is intriduced to prevent the scale of x_i^{l+1} from exploding, and $\phi_x(.)$ is a NN. The final step is to update the node attributes/scalars [75],

$$h_i^{l+1} = h_i^l + \phi_h \left(h_i^l, \sum_{j \in [N]} w_{ij} m_{ij}^l \right) \tag{12}$$

where $w_{ij} = \phi_m(m_{ij}^l) \in [0, 1]$ and $\phi_{m,h}$ are neural networks. The three steps discussed above constitute a Lorentz Group equivariant Block. Several of these blocks are stacked on top of one another to form the final model. For a detailed discussion on the model, its implementation, the optimizer used, and the learning rate scheduler, See [75].
We implemented the LorentzNet with PyTorch and trained it on a cluster with four Nvidia Tesla K80 GPUs. We pass the data in batches of size 16 on each GPU. The model is trained for a total of 35 epochs. At the end of each epoch, we test the model performance with the validation dataset, and the one with the best validation accuracy is saved for testing.

D. Composite Models

So far, we have discussed six different classifiers denoted as simple in the later part of the manuscript. A simple BDT_{calo} that only uses the features extracted from the calorimeter energy deposits of a fat jet without considering the tracking information. Next, we have an extended version of the simple BDT_{calo}, the simple BDT_{trck} classifier, which extends the previous dataset by including complementary information from the tracking detectors. Then, we discussed the one-dimensional CNN_{calo} and GNN_{calo} classifiers, which only use the information of the calorimeter towers inside a fat jet. The CNN_{trck}, on the other hand, uses 2-dimensional images where the second layer comprises the tracks that constitute the fat jet. Similarly, we have GNN_{trck}, which uses the charged hadrons and neutral hadrons four-momentum from tracks and towers to construct particle clouds/graphs.
We expect that during training, the CNN/GNN can extract important characteristics of the fat jets from these low-level features that can discriminate between the signal and background jets. However, some information about the high-level features is lost during the data pre-processing, which can be extremely valuable for the classification task. For example, as demonstrated in [55], the rotation and normalization preprocessing steps in generating the images for CNN smear the information of the invariant mass of a fat jet. Similarly, the b-tagging information of a fat jet is not included in the CNN and GNN classifiers but can be useful for the classification task. The BDTs also have one obvious disadvantage. They rely on the user-supplied HLFs rather than extracting features directly from data. This limits their ability to automatically learn the complex features present in the data.
From the above discussion, it is clear that the simultaneous use of both LLFs and HLFs can help explore complementary directions in the feature space and improve the performance of the classifiers. One naive way of incorporating both HLFs and LLFs in an analysis is to stack classifiers that use these features on top of one another. We refer to them as composite classifiers. The idea is first to use a classifier (a CNN/GNN) that uses LLFs to extract a preliminary classification score and later use this score as an HLF in a second classifier (a BDT) along with other HLFs. In this work, we have studied the performance of eight such composite Models; $CNN_{calo} + BDT_{calo}$ ($C_{calo}B_{calo}$), $CNN_{calo} + BDT_{trck}$ ($C_{calo}B_{trck}$), $CNN_{trck} + BDT_{calo}$ ($C_{trck}B_{calo}$), $CNN_{trck} + BDT_{trck}$ ($C_{trck}B_{trck}$), $GNN_{calo} + BDT_{calo}$ ($G_{calo}B_{calo}$), $GNN_{calo} + BDT_{trck}$ ($G_{calo}B_{trck}$), $GNN_{trck} + BDT_{calo}$ ($G_{trck}B_{calo}$), and $GNN_{trck} + BDT_{trck}$ ($G_{trck}B_{trck}$). In the next section, we will demonstrate the performance of all these models in discriminating top jets from QCD jets.

IV. CLASSIFIER PERFORMANCE

In this section, we will discuss the performance of the different classifiers for top tagging. For a consistent comparison with the results of [57, 75, 77, 100–106], we generate top and QCD samples in the 550 $GeV <$ $p_T <$ 650 GeV range and reconstruct the fat jets with $R = 0.8$. The generation process is the same as discussed in Section II. At the same time, to check the dependency of the classifier performance on the showering and hadronization models of the Monte Carlo event generator, we have generated a second QCD sample[14] using Herwig [129, 130]. We train the classifiers using the Pythia-generated dataset and save the model that performs best on the validation set for further analysis. We perform two final tests, one using the Pythia-generated signal and background sample and the other where the signal jets are generated using Pythia while background jets are generated using Herwig.

In Figure 3, we present the performance of the classifiers in the form of their Receiver Operator Characteristic (ROC) curves. The solid lines represent the ROC curves for the dataset where both signal and background samples are generated using Pythia. On the other hand, the dotted curves characterize the

sample where the background jets are generated in Herwig. In Table II, we present the background rejection (reciprocal of background efficiency, $1/\epsilon_B^c$) of all the classifiers corresponding to 70 and 50 % signal efficiency of the classifier denoted as classifier efficiency, ϵ_S^c [15]. The background rejection within the parentheses in Table II represents the results obtained from the dataset simulated in Herwig. The second and third columns of the Table correspond to the background rejection ($1/\epsilon_B^c$) for 50% and 70% signal efficiency (i.e., $\epsilon_S^c = 0.5$ and 0.7) of the classifier. Whereas the fourth and fifth columns show the background rejection ($1/\epsilon_B^{tag}$) for 50% and 70% top-tagging efficiency ($1/\epsilon_S^{tag}$)[16] defined as the fraction of hadronically decaying boosted top quarks resulting in top-tagged fat jets. This section will address the results in the second and third columns, i.e., the classifier efficiencies of different simple and composite classifiers. For a detailed discussion of the results in the fourth and fifth columns, i.e., the top-tagging efficiencies ($\epsilon_{S,B}^{tag}$), see Section V. In the following discussion, we will use the background rejection at 50% classifier efficiency (ϵ_S^c) as a metric to compare the performance of different simple and composite classifiers introduced here and also with existing top classifiers in the literature.

---

[14] The reason for this choice lies in the truth-level identification efficiency. The paron-level information in a Herwig-generated dataset differs from that in a Pythia-generated sample. This results in different TLIEs. Subsequently, the resulting top samples are inadequate for comparing the classifiers' performance. However, since we do not perform any truth-level identification for the QCD jets, they can be used for the task.

[15] The signal and background efficiencies (ϵ_S^c and ϵ_B^c, respectively) of the classifier are defined as the fractions of truth-level top-tagged and QCD fat jets correctly identified as top-tagged and QCD jets, respectively, by the classifier.

[16] It is important to note that the classifier efficiency ϵ_S^c differs from the efficiency of correctly identifying a fat jet, originating from the hadronically decaying boosted top quark, as a top jet, i.e., the top-tagging efficiency (ϵ_S^{tag}). The classifier efficiency depends on the truth level tagging criteria introduced to prepare the signal datasets for training and testing. Whereas for calculating the top-tagging efficiency (ϵ_S^{tag}), we have not enforced any truth-level tagging criteria. To calculate ϵ_S^{tag}, we have simulated hadronically decaying top antitop pairs using Pythia, reconstructed the fat jets in Delphes, and prepared the testing samples from fat jets falling in the given transverse momentum range without any truth level tagging.
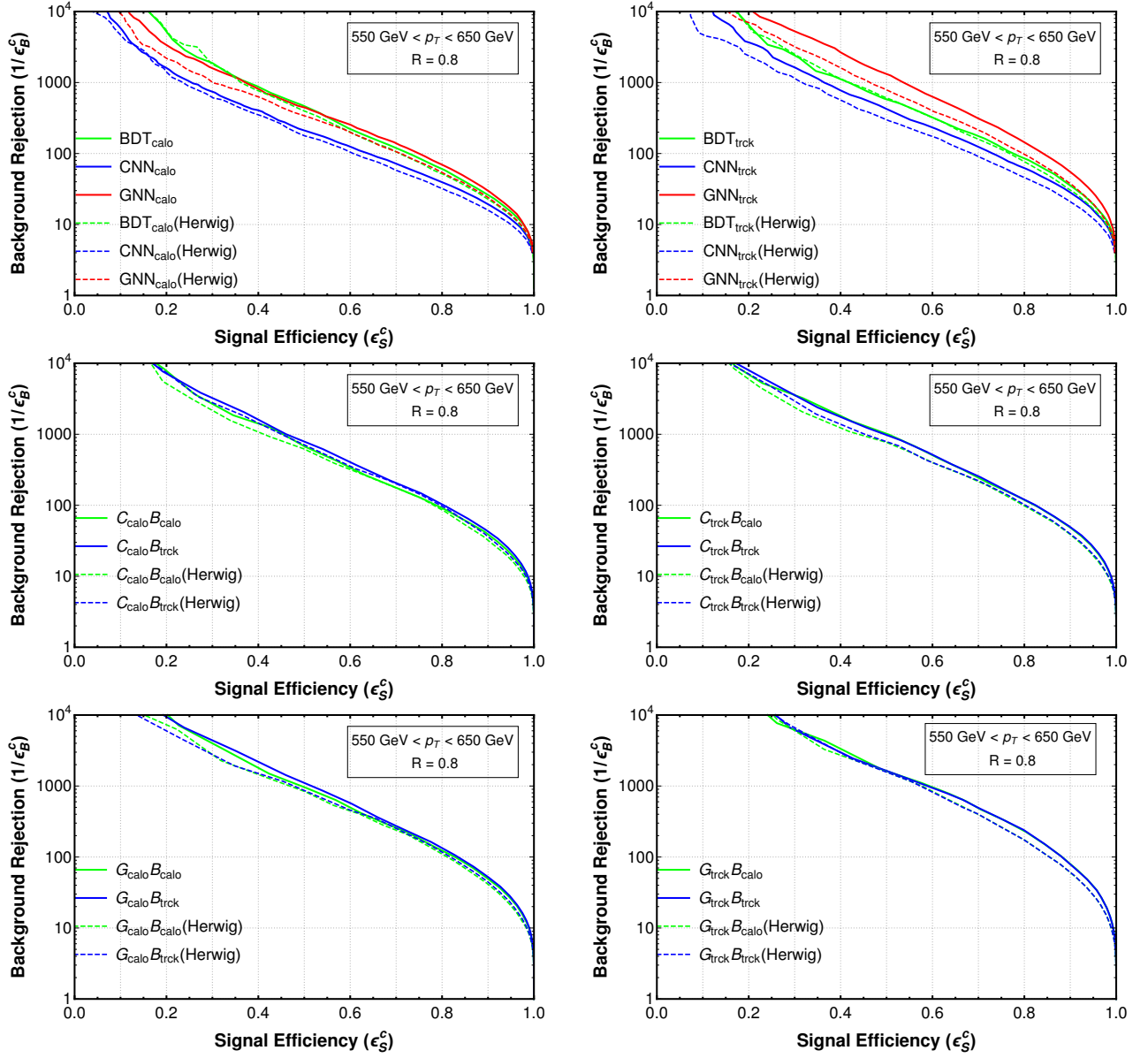
FIG. 3. Thr ROC curves for the different classifiers for top and QCD samples in the p_T range 550-650 GeV. The solid lines correspond to signal and background samples generated using Pythia, while the dotted line corresponds to the case where the background sample is generated using Herwig

Classifier	$1/\epsilon_B^c(\epsilon_S^c = 0.7)$	$1/\epsilon_B^c(\epsilon_S^c = 0.5)$	$1/\epsilon_B^{tag}(\epsilon_S^{tag} = 0.7)$	$1/\epsilon_B^{tag}(\epsilon_S^{tag} = 0.5)$
BDT_{calo}	119(105)	467(398)	22	125
CNN_{calo}	70(57)	211(178)	17	76
GNN_{calo}	139(106)	444(341)	24	139
BDT_{trck}	175(159)	579(610)	33	180
CNN_{trck}	124(90)	423(299)	25	120
GNN_{trck}	311(214)	1322(789)	42	274
$C_{calo}B_{calo}$	176(175)	682(619)	31	179
$C_{calo}B_{trck}$	208(204)	811(737)	35	222
$C_{trck}B_{calo}$	249(218)	1023(768)	43	253
$C_{trck}B_{trck}$	257(221)	995(799)	46	249
$G_{calo}B_{calo}$	260(241)	969(842)	43	261
$G_{calo}B_{trck}$	278(256)	1141(894)	52	281
$G_{trck}B_{calo}$	489(397)	1641(1604)	65	468
$G_{trck}B_{trck}$	493(399)	1736(1666)	68	500

TABLE II. Background rejection at 50 and 70 % background efficiencies. The terms in the bracket represent the results for the Herwig-generated dataset. The second and third columns correspond to a dataset where the top samples satisfy the truth-level tagging criteria. For the fourth and fifth columns, no such tagging criteria are imposed.

A. Validation and the performance of tower-based simple classifiers

Before delving extensively into the discourse of comparing the performance of various simple and composite classifiers, trained and tested on two different classes of datasets, validating our approach (event simulation, sample preparation, etc.) by comparing our results with existing literature is crucial. The top left plot of Fig. 3 shows the ROC curves for BDT_{calo}, CNN_{calo}, and GNN_{calo} classifiers in green, red, and blue, respectively. We see comparable performance between the Pythia and Herwig-generated datasets. The GNN_{calo} classifier is a slight modification of LorentzNet introduced in [75], with the difference that instead of using the mass of the constituents as node embedding, we used their charge[17], and our training process has a smaller batch size of 16. This results in a slight difference in the classifier performance. For 50% classifier efficiency (ϵ_S^c) of GNN_{calo}, we obtain a background rejection close to 444, while in [75], the corresponding background rejection was 498. In [100], the authors have demonstrated the performance of several classifiers on a similar dataset. CNN_{calo} shows comparable performance to the CNNs presented in Ref. [100]. The similar performance observed between GNN_{calo} and LorentzNet as discussed in [75], and between CNN_{calo} and the different CNN-based classifiers mentioned in Ref. [100], provides validation for our methodology.

It is important to note that all the LLF-based classifiers mentioned in Ref. [100] have a background rejection

of less than 400 at $\epsilon_S^c = 0.5$. Therefore, we can safely conclude that GNN classifiers like the LorentzNet [75] or GNN_{calo} perform better than other LLF-based classifiers. In the top left plot of Figure 3, we also present the performance of the HLF-based BDT_{calo} classifier. Interestingly, this simple cut-based classifier has a comparable performance with the GNN_{calo}. If we compare the background rejection of BDT_{calo} for $\epsilon_S^c = 0.5$, BDT_{calo} also has a better performance than not only the GNN_{calo} classifier (see Table II) but also all the CNN and GNN-based classifiers presented in Refs. [75, 100]. The poor performance of CNN_{calo} is because the jet image preprocessing steps described in Section II dilute the jet mass information, an extremely important discriminant for classifying top jets over QCD jets. Despite using complete 4-momentum information of the jet constituents for the classification task, GNN_{calo} classifier also fails to outperform HLF-based BDT_{calo} classifier. Note that while the GNN_{calo} is trained using the calorimeter tower data, the HLFs for training the BDT_{calo} are derived from the fat jets constructed with Etracks and Etowers (see section II). The superior energy/momentum resolution of tracks results in better performance for the BDT_{calo} compared to GNN_{calo}.

B. The performance of track and tower-based simple classifiers

Figure 3 (top right panel) shows the performance of the BDT, CNN, and GNN trained and tested using DATA$_{trck}$[18] namely the BDT_{trck}, CNN_{trck},

---

[17] Note that the training and testing samples used in the classifiers in the left panel of Fig. 3 are generated from the tower data and hence do not have information about the jet constituent's mass or charge.

[18] DATA$_{trck}$ was defined in section II.

and GNN_{trck} classifiers for both Pythia and Herwig-generated datasets. The systematic uncertainties of the track-based classifiers, resulting from the showering and hadronization models used in Pythia and Herwig, are large compared to the tower-based classifiers shown in the top left panel of Figure 3. While the finite resolution of the calorimeter towers reduces the classifier's dependency on the showering and hadronization models in the case of tower-based classifiers, the electric charge information of the jet constituents makes the track-based classifiers sensitive[19] to the event generator.

Figure 3 (top panel), as well as Table II shows more than 100 % improvement in the performance of the CNN and GNN-based classifiers trained and tested using DATA$_{trck}$ compared to the tower-based classifiers. The high-quality training and testing datasets (DATA$_{trck}$), made possible by the superior resolution of the LHC tracker detector, lead to improved performance in the track-based CNN and GNN classifiers. However, such a high performance enhancement can not be solely attributed to the dataset quality alone. The fat jets originating from boosted top quarks exhibit substructures that stem from the hadronization of the three quarks produced during the top decay process. Conversely, in the case of a QCD fat jet, various substructures can emerge from the branching of the initial light-flavored quarks or gluons, potentially resulting in atleast one or more substructures arising from the hadronization of gluons. It has already been known, both from theoretical principles[20] and a large collection of experimental measurements [132–135], that jets initiated by gluons exhibit differences with respect to jets from light-flavor quarks. For example, the charged particle multiplicity is higher in gluon jets than in light-quark jets; the fragmentation function of gluon jets is considerably softer than that of a quark jet; gluon jets are less collimated than quark jets, etc. These differences have been exploited [17, 58, 62, 109, 119, 121, 136–139] to construct taggers capable of discriminating jets initiated by light-quarks from those initiated by gluons. The charged hadron multiplicity and distributions derived from the tracker detector, in contrast to the neutral hadron multiplicity and distribution from calorimeter towers, play a pivotal role in the discrimination of light-quark and gluon jets. A substantial fraction of the performance improvement

achieved by the track-based CNN and GNN classifiers stems from the inclusion of information about the neutral and charged hadron composition of fat jets in the training and testing datasets ($DATA_{trck}$). This enables the classifiers to differentiate substructures originating from the hadronization of a partonic gluon or a light-quark. An evident drawback of developing a tool reliant on the hadronization of light-quarks and gluons is the inherent discrepancies in the modeling of quark and gluon jets in Monte Carlo simulations. However, event generators like Pythia and Herwig incorporate sophisticated experimentally fine-tuned models for hadronization developed through decades of experimental studies and perturbative QCD calculations. Despite minor discrepancies between these event generators (as well as between the event generators and experimental data), giving rise to the systematic uncertainty, hadronization models used in Pythia and Herwig serve as a solid foundation for building improved classifiers for top tagging.

Figure 3 (top panel) also shows around 25 % improvement in the performance of BDT_{trck} compared to BDT_{calo}. This can be ascribed to the use of subjet-based features constructed from the track and calorimeter tower constituents of the fatjet. To illustrate this point, we present in Appendix B the ranking[21] and covariance matrix[22] of a few important feature variables used in training the BDT classifiers. Appendix B shows that along with the jet mass, the features of the second sub-jet play a crucial role in the classification task. Figure 3 (top right panel) also shows a comparable performance between the BDT_{trck} and CNN_{trck} classifiers. As discussed in Section III D, the pre-processing steps in CNN smear the invariant mass distribution for the fat jets, which plays a key role in the discrimination of top from QCD jets. Therefore, performances of simple CNN classifiers (CNN_{calo} and CNN_{trck}) can be improved significantly when used in association with BDT classifiers. Potential improvements in such composite classifiers will be explored in the next section.

C. The performance of composite classifiers

Figure 3 (middle panel) shows the ROC curves for the composite classifiers involving CNN, namely $C_{calo}B_{calo}$, $C_{calo}B_{trck}$, $C_{trck}B_{calo}$, and $C_{trck}B_{trck}$ (see section III D for the definitions). Incorporating the lower-level information of CNN with the higher-level information of BDT results in a more than 100 % improvement in CNN's performance. As discussed in the previous section, the information on the HLFs, like invariant mass and b-tag,

---

[19] Simultaneous use of the LLFs and HLFs for the discrimination task might reduce the event generator dependency of the track-based classifiers. This is one of the main motivations behind constructing the composite classifiers, which we will discuss in the next section.

[20] The fundamental principle underlying the differentiation between quark and gluon jets is rooted in the observation that gluon splitting is stronger than quark splitting, as dictated by perturbative QCD. This distinction becomes evident by directly comparing the splitting probabilities for gluons, such as $g \rightarrow gg$ and $g \rightarrow u\bar{q}$, with those for quarks, like $q \rightarrow qg$ [131]. Therefore, on average, gluon jets are broader and encompass a higher particle multiplicity than quark jets with similar p_T.

[21] The variable ranking demonstrates the importance of the variables for the classification.

[22] Two variables that are least correlated represent independent directions in the feature space and, when used simultaneously, can considerably improve the performance of a classifier.

are not present in the CNN score. Therefore, when used together in a composite classifier, they significantly enhance performance. When the track-based High-Level Features (HLFs) from BDT_{trck} are combined with the score obtained from CNN_{calo} in $C_{calo}B_{trck}$, there is a notable improvement in performance compared to the $C_{calo}B_{calo}$ configuration (as depicted in the middle left panel of Fig. 3). However, using track-based HLFs with the score from CNN_{trck} ($C_{trck}B_{trck}$) does not enhance the performance (see the middle right panel of Fig. 3) of $C_{trck}B_{calo}$. Note that CNN_{trck} is trained on the datasets containing two-layer images resulting from the tracks and towers. Therefore, when used alongside BDT_{calo} and BDT_{trck}, we see a similar improvement in performance.

The ROC curves for the composite classifiers based on GNN, specifically $G_{calo}B_{calo}$ and $G_{calo}B_{trck}$, are presented in the bottom left panel of Figure 3. Likewise, the bottom right panel displays the ROC curves for $G_{trck}B_{calo}$ and $G_{trck}B_{trck}$. As shown in Figure 3 (bottom left panel) and enumerated in Table II, a substantial performance improvement of approximately 100 % (25 %) is observed for GNN_{calo} (GNN_{trck}) when the GNN score is combined with the other HLFs from BDT_{calo} within the framework of $G_{calo}B_{calo}$ ($G_{trck}B_{calo}$). The marginal enhancement in performance observed for $G_{trck}B_{calo}$ can be attributed to the presence of comprehensive details about the 4-momentum of constituent tracks and towers within the training samples of GNN_{trck}. Consequently, incorporating additional high-level features from BDT_{calo} does not yield a substantial improvement, as the training data of GNN_{trck} already contains comprehensive details for efficient classification. To illustrate this, we present the ranking of High-Level Features (HLFs) used in $G_{trck}B_{calo}$ and $G_{calo}B_{calo}$ in Table XII of Appendix C. In contrast to $G_{calo}B_{calo}$, where the highest-ranked variable is the invariant mass of the fat jets (and consequently, it is frequently employed for node splitting), the GNN score takes the top-ranking position in $G_{trck}B_{calo}$, demonstrating the importance of this variable (see Table XII). While $G_{calo}B_{trck}$ exhibits an additional performance enhancement of 20 % compared to $G_{calo}B_{calo}$, attributed to the complementary nature of the track-based HLFs in BDT_{trck} alongside the calorimeter tower-based LLFs used in GNN_{calo}, the performance of $G_{trck}B_{trck}$ is comparable to $G_{trck}B_{calo}$. Note that the training datasets for GNN_{trck} already encompass comprehensive information about the constituent tracks, making additional track-based HLFs from BDT_{trck} less impactful.

As discussed earlier, the observed improvement in the performance of $C_{trck}B_{trck}$ compared to CNN_{trck} is because of the inclusion of additional high-level features in the former that are lost during the pre-processing steps of the images used in the CNN. However, in the case of GNN_{trck}, no such pre-processing is done, and we expect GNN_{trck} with the full tracking information to be efficient enough to extract all relevant features of the fat-jet and hence to have a similar performance as $G_{trck}B_{calo}$ and

$G_{trck}B_{trck}$. But, contrary to our expectations, we still observe a significant improvement in the performance of the composite track-based GNN classifiers compared to the ordinary track-based GNN classifier. This behavior is due to masking the mass information of the fat-jet constituents arising from tracks (See Section II B 3). To validate this point, we retrain the GNN_{trck} classifier with a dataset with unmasked tracking information and present our findings in Table III. As per our expectation, the GNN_{trck} classifier with full unmasked tracking information has a comparable performance with $G_{trck}B_{trck}$. So, is it really necessary to introduce a composite classifier when we can achieve the same performance by training the original classifier with the complete information? To investigate the matter further, we performed another study where instead of testing the classifiers on a pythia-generated dataset, we tested it on data where the Background sample is generated using Herwig. The motivation for this study is to check the effect of systematic uncertainties arising from using different Monte-Carlo event generators on the performance of the classifiers. As it is evident from the secon row of Table III, the Composite track-based GNN classifier has better control over these uncertainties than the ordinary GNN classifier trained with full unmasked tracking information. We present a detailed discussion in the next section.

MC generator	GNN_{trck}	$G_{trck}B_{trck}$
Pythia8	1769	1736
Herwig7	1025	1666

TABLE III. The first row presents the background rejection at 50% classifier efficiency for GNN_{trck} classifier trained and tested using a dataset with the full unmasked tracking information and that of the $G_{trck}B_{trck}$ classifier trained and tested with the masked dataset. The second row represents similar results but the testing is performed with a background sample generated using Herwig7.

D. Systematic Uncertainties of different simple and composite classifiers

To investigate the systematic uncertainties resulting from the showering and hadronization models of the event generators, we computed the performance of the classifiers using two sets of datasets: one generated using Pythia and the other using Herwig. The solid and dotted lines in each plot of Figure 3 represent the ROC curves corresponding to Pythia and Herwig, respectively, generated testing datasets for a specific classifier. The performance of the BDT classifiers (BDT_{calo} and BDT_{trck}), which rely on HLFs[23] such as jet mass, N-subjettiness, b-

————

[23] It is worth noting that HLFs like jet mass, N-subjettiness, b-tag, p_T-weighted track width, etc. exhibit minimal sensitivity to

tag, p_T-weighted track width, etc., remains largely unaffected by the choice of Monte-Carlo generators for showering and hadronization. Due to the direct correlation between the LLFs of the fat jets and the jet modeling within the Monte-Carlo generators, it becomes evident that simple CNN/GNN classifiers trained on the LLFs of the jets display significant sensitivity to the chosen showering and hadronization model (as depicted in the top panel of Figure 3). This sensitivity contributes to large systematic uncertainties in the classification process. For classifiers based on LLFs trained using tower information (CNN_{calo} or GNN_{calo}), the associated systematic uncertainties can reach up to 30 %. In the case of track-based CNN/GNN classifiers, the potential systematic uncertainties can reach up to 40 %. The higher sensitivity of track-based CNN/GNN classifiers to the jet modeling of the Monte-Carlo generator results from the fact that these classifiers are trained on $DATA_{trck}$ that encompass the finer details of the showering and hadronization processes. In contrast, for tower-based classifiers, the limited resolution of the calorimeter leads to a smoothing effect that mitigates the influence of Monte-Carlo generator dependencies.

Remarkably, composite classifiers not only enhance top-tagging performance but also show reduced systematic uncertainties compared to simple LLF-based classifiers. Note that composite classifiers are simple BDT classifiers augmented with scores from LLF-based classifiers (track/tower-based CNN and GNN classifiers), treated as additional HLF in conjunction with other HLFs of the track and tower-based BDTs discussed in section II B 1. Interestingly, while the scores from LLF-based classifiers introduce higher systematic uncertainties, other HLFs such as jet mass, N-subjettiness, b-tag, p_T-weighted track width, etc., of the BDT classifiers remain relatively insensitive to variations in Monte-Carlo generators. The ranking of HLFs used in composite classifiers, as depicted in Tables X, XI, XII and XIII of Appendix C, illustrates that alongside the scores from LLF-based classifiers, the other HLFs also make substantial contributions to the classification task. For instance, for the composite classifiers like $C_{calo}B_{calo}$, $C_{trck}B_{calo}$, $C_{calo}B_{trck}$, $C_{trck}B_{trck}$, $G_{calo}B_{calo}$, and $G_{calo}B_{trck}$, jet mass holds the highest ranking among the HLFs. The utilization of the LLF-based score as a classifying feature is restricted to about 36 % (19 %) for $G_{trck}B_{calo}$ ($G_{trck}B_{trck}$), where score takes precedence as the highest ranking variable. This reduced reliance on the score as the main classifying feature mitigates the classifier's overall systematic uncertainties stemming from the inherent uncertainties of the score. While composite classifiers were introduced to augment the performance of simple LLF-based classifiers by incorporating high-level physical features of the fat jets, the reduction in sys-

tematic uncertainties has emerged as an additional benefit. Optimal utilization of the HLFs has the potential to boost the performance of classifiers while reducing the classifier's dependence on the Monte-Carlo generators. A comprehensive study of the performance optimization of composite classifiers while simultaneously mitigating systematic uncertainties is crucial. Nonetheless, it lies outside the boundaries of the present work, which we intend to explore in future investigations.

V. THE INTERPLAY BETWEEN TRUTH-LEVEL IDENTIFICATION, CLASSIFIER EFFICIENCY, AND TOP-TAGGING EFFICIENCY

The classifier's tagging efficiency (ϵ_S^c) and mistagging rate ($1/\epsilon_B^c$) discussed in the previous section depend on the purity of the training/testing datasets, i.e., on the truth-level tagging (TLT) criteria used for preparing the training/testing samples. A pure training/testing sample prepared with strict TLT criteria results in a superior classifier performance; however, it might lead to poor top-tagging efficiency (ϵ_S^{tag})[24]. The truth-level tagging efficiency denoted as ϵ_S^{truth}, quantifies the fraction of hadronically decaying top quark initiated fat jets with a given reconstruction radius (R) that meet the Truth-Level Tagging (TLT) criteria. The values of R and the TLT criteria together determine the truth-level identification efficiency (ϵ_S^{truth}). When we reconstruct top quark initiated fat jets with a smaller R for a given p_T range, the chances of obtaining impure jets increase. These impure jets are those where one or more top decay products fall outside the reconstruction cone. The TLT criteria remove these impure samples, resulting in a reduced ϵ_S^{truth}. For example, the $R = 0.8$ jets in the p_T range [550, 650] GeV, as discussed in Section IV, exhibit a ϵ_S^{truth} of approximately 55%. Classifiers trained on such a high-purity sample may struggle to identify top-initiated fat jets that fall outside the TLT criteria. This leads to decreased performance in top-tagging, as evidenced in the fourth and fifth columns of Table II, which display background rejection rates for top tagging efficiencies (ϵ_S^{tag}) of 70% and 50%, respectively. Comparing the background mistag rates for a given top-tagging efficiency (see fourth and fifth columns of Table II) and classifier efficiency (see second and third columns of Table II), it becomes evident that excellent classifier performance does not necessarily translate to higher top-tagging efficiency. Larger truth-level identification efficiency (ϵ_S^{truth}) can help reduce the disparities between classifier and top-tagging efficiency. To enhance TLT efficiency, one option is to relax the TLT criteria. However, doing so results in a less pure sample for training and testing, leading to poorer classifier performance. Alternatively, using appropriate fat

---

the jet modeling within Monte-Carlo generators. This leads to reduced systematic uncertainties for the BDT-classifiers.

---

[24] The efficiencies, ϵ_S^c, ϵ_B^c and ϵ_S^{tag}, are defined in section IV.

jet reconstruction radii (RR) in different p_T regions ensures that all top quark decay products remain within the reconstruction cone, thus improving ϵ_S^{truth}.

In this section, we explore the impact of varying the reconstruction radius (RR) on Truth-Level Tagging (TLT) and, consequently, on determining the classifier's tagging efficiency (ϵ_S^c) and top-tagging efficiency (ϵ_S^{tag}). We present our findings for the GNN_{trck} classifier trained and tested with two different sets of track-based samples comprising signal and background fat jets falling within the p_T range of $[550, 650]$ GeV. In one sample, fat jets were reconstructed with a radius of $R = 0.8$, while in the other sample, the fat jets were reconstructed with a radius of $R = 1.2$. The truth level tagging criteria are the same as discussed in Section II. Figure 4 shows our results, with the blue curve corresponding to $R = 0.8$ jets and the red curve to $R = 1.2$ jets.
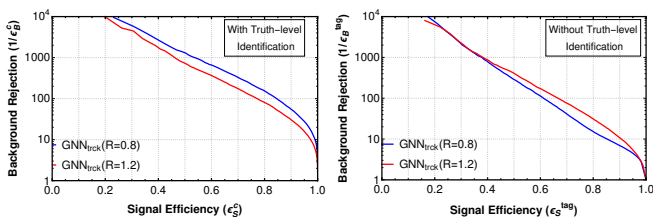
FIG. 4. The ROC curves of the GNN_{trck} classifier for two reconstruction radii of the fat jet, $R = 0.8$ (blue) and $R = 1.2$ (red). The left plot represents the performance for a dataset that satisfies truth-level identification criteria. No such criteria are imposed for the dataset used in the right plot.

Variable	$1/\epsilon_B^c$ ($\epsilon_s^c = 50\%$)	$1/\epsilon_B^{tag}$ ($\epsilon_s^{tag} = 50\%$)
$R = 0.8$	1298	274
$R = 1.2$	711	424

TABLE IV. Background rejection at 50% signal efficiency for GNN_{trck} corresponding to the datasets with $R = 0.8$ and $R = 1.2$.

Figure 4 (left panel) shows the ROC curves for the GNN_{trck} classifier trained and tested on signal and background samples with $R = 0.8$ (blue) and $R = 1.2$ (red).

In the second and third columns of Table V, we present the background rejection for the BDT_{calo} and BDT_{trck} classifiers, respectively. We see a gradual decrease in performance with increasing p_T. It is because the invariant mass of the QCD jets scales with its transverse momentum, and as we go higher in p_T, the probability of QCD jets faking as top increases. The BDT_{trck} classifier performs better than the BDT_{calo} because of the

Table IV summarises the background rejection factor corresponding to 50 % signal efficiency. The $1/\epsilon_B^c$ for the classifier trained/tested with $R = 0.8$ fat jets is close to 1300, which is 70 % higher than that of the classifier trained/tested with $R = 1.2$ fat jets. Using larger-radius jets introduces increased noise contributions from various sources. This noise can obscure the characteristic distributions of fat jets and impact the classifier's performance. While the classifier trained and tested with $R = 0.8$ jets may appear impressive in Fig. 4 (left panel) and Table IV, as discussed in the previous section, it does not guarantee optimal top-tagging performance. To illustrate top-tagging performance, we conducted tests with the same classifiers on a dataset where we did not impose any Truth-Level Tagging (TLT) criteria. The results of these tests are presented in the right panel of Figure 4. As anticipated, we observed a decrease in the performance of both classifiers when TLT criteria were not enforced. Interestingly, the classifier trained with $R = 1.2$ fat jets outperformed the one trained with $R = 0.8$ fat jets. The findings from this section have motivated us to explore the possibility of using different reconstruction radii (R) for fat jets in the six distinct p_T regions that we will discuss in the following section.

VI. FINAL RESULTS

In this section, we present the final result of our analysis. As discussed in Section II, our final result consists of the performance of six classifiers, BDT_{calo}, BDT_{trck}, CNN_{trck}, GNN_{trck}, $C_{trck}B_{calo}$, and $G_{trck}B_{calo}$ in six p_T ranges. We summarise our final result as six plots corresponding to these p_T ranges in Figure 6. Each plot presents six ROC curves, one for each classifier. For a consistent comparison of the performance of the classifiers, we present the background rejection at 50 % classifier efficiency (ϵ_S^c) in Table V. A diagrammatic representation of this result is presented in the left plot of Figure 5. Note that the fat jets in the p_T range $[300, 500]$ GeV and $[500, 700]$ GeV have different R-parameters ($R = 1.2$) and hence different truth-level identification efficiency than those in the remaining p_T bins where fat jets are constructed with a RR of $R = 0.8$ (see the discussion in II). Therefore, comparing the classifier's performances for fat jets belonging to these two groups is unsuitable.

inclusion of additional tracking information. The fall in background rejection with p_T is also smaller for BDT_{trck} than BDT_{calo}.

The fourth and fifth columns of Table V represent the background rejection for the CNN_{trck} and GNN_{trck} classifiers. In both cases, the $[300, 500]$ GeV p_T jets have a smaller $1/\epsilon_B^c$ than the $[500, 700]$ GeV p_T jets. This is because, as demonstrated in Section II A, an

p_T [GeV]	BDT_{calo}	BDT_{trck}	CNN_{trck}	GNN_{trck}	$C_{trck}B_{calo}$	$G_{trck}B_{calo}$
300-500	388	456	159	587	762	1413
500-700	136	276	184	765	455	1178
700-900	168	345	278	845	538	1409
900-1100	79	247	256	971	466	1175
1100-1300	56	167	214	882	318	872
1300-1500	39	127	217	877	273	850

TABLE V. Background rejection at 50 % classifier efficiency for the six transverse momentum range.
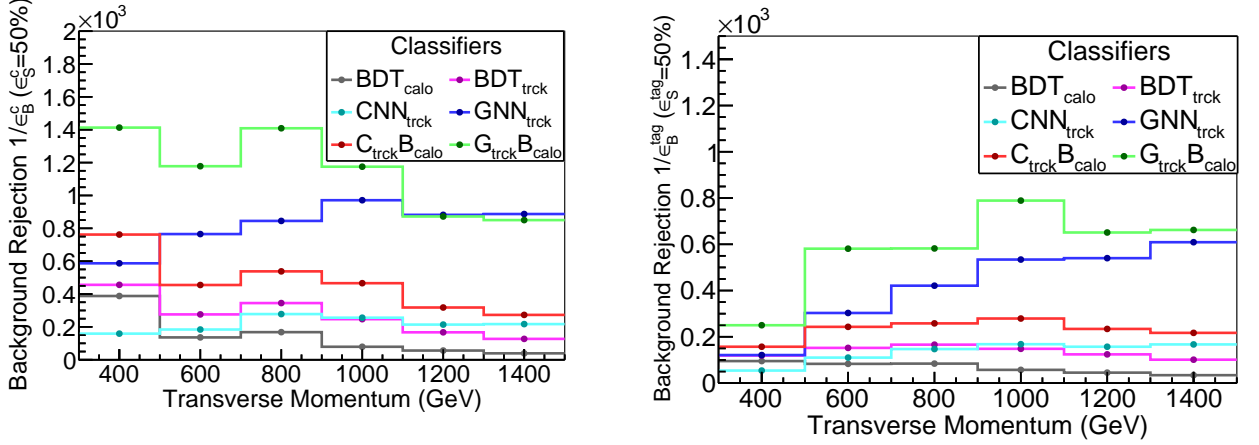
FIG. 5. ROC curves for the classifiers corresponding to the six p_T ranges considered in this paper.

p_T [GeV]	BDT_{calo}	BDT_{trck}	CNN_{trck}	GNN_{trck}	$C_{trck}B_{calo}$	$G_{trck}B_{calo}$
300-500	95	119	54	121	157	250
500-700	83	152	110	303	243	581
700-900	84	166	147	421	258	582
900-1100	57	148	168	534	279	789
1100-1300	45	124	157	540	234	651
1300-1500	34	101	167	609	217	662

TABLE VI. Background rejection at 50 % signal efficiency for the six transverse momentum range. Here, the testing is performed on a dataset without truth-level tagging.

p_T [GeV]	BDT_{calo}	BDT_{trck}	CNN_{trck}	GNN_{trck}	$C_{trck}B_{calo}$	$G_{trck}B_{calo}$
300-500	22	25	16	27	30	35
500-700	20	27	22	38	32	42
700-900	20	28	26	40	32	43
900-1100	15	26	27	43	34	46
1100-1300	12	23	26	42	29	44
1300-1500	10	21	25	42	29	43

TABLE VII. Signal efficiency corresponding to a background rejection factor ($1/\epsilon_B^c$) of 1000 for the six transverse momentum range. Here, the testing is performed on a dataset without truth-level tagging.

R-parameter 1.2 is inefficient in capturing all the constituents of the [300, 500] GeV fat jets and reduces the performance. We see almost comparable performance for the jets in the remaining p_T bins for both classifiers. The slight reduction in performance in the case of CNN_{trck} can be ascribed to the fact that with increasing p_T, the jet constituents get more collimated and look similar to that of a QCD jet.
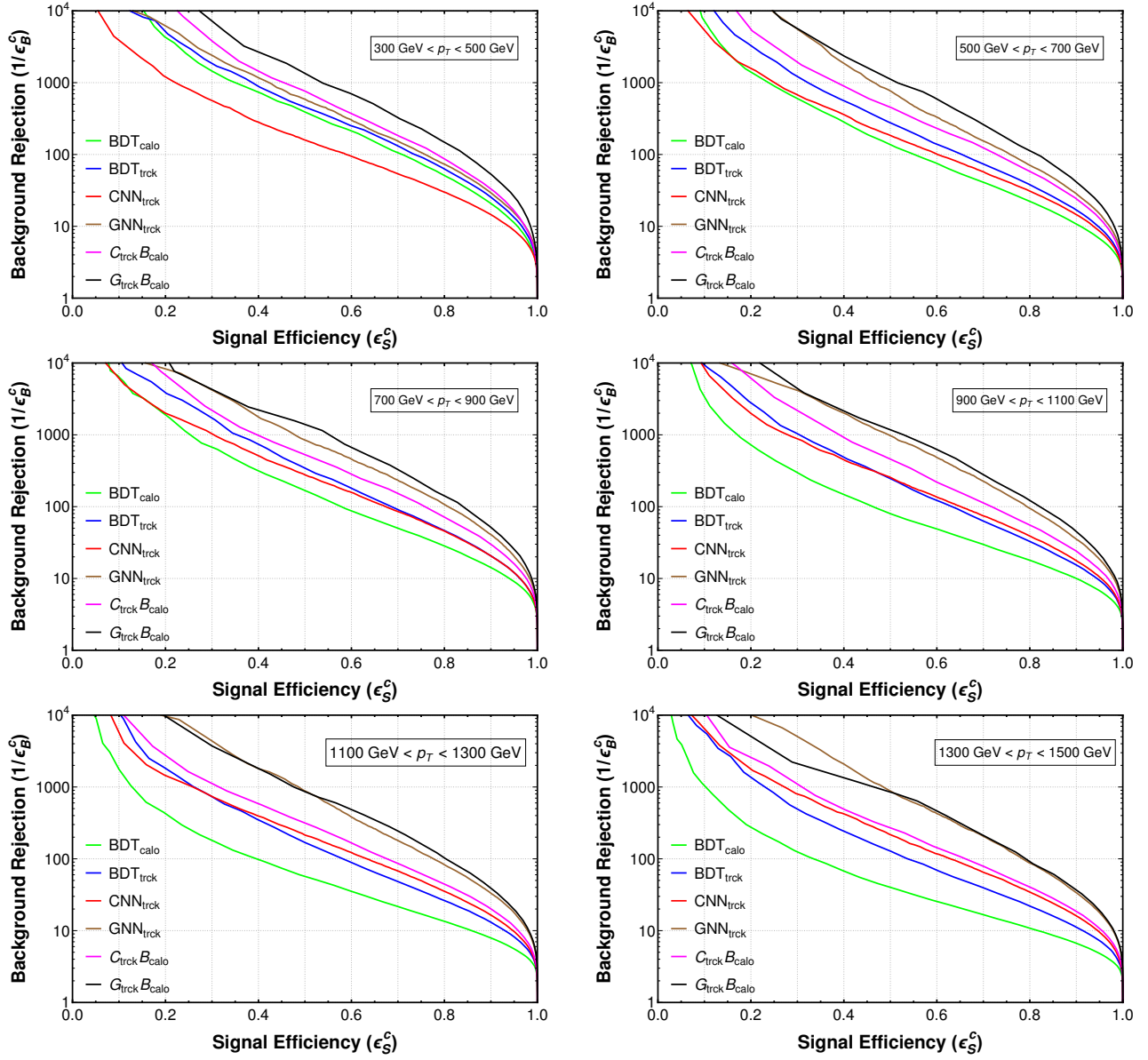
FIG. 6. ROC curves for the classifiers corresponding to the six p_T ranges considered in this paper.

Finally, in columns six and seven, we present the background rejection for the $C_{trck}B_{calo}$ and $G_{trck}B_{calo}$ classifiers. In the case of $C_{trck}B_{calo}$, we see considerable improvement compared to CNN_{trck}. This is because the preprocessing steps in CNN smear out the invariant mass of the fat jet. This information is restored when we combine CNN_{trck} with BDT_{calo} resulting in a performance improvement (for a detailed discussion see Section IV C). However, this improvement gradually decreases with increasing p_T as the performance of the BDT decreases. The GNN_{trck} classifier trained on the four-momentum data of the jet constituents can reconstruct some information about the fat jet mass. Therefore, when combined with BDT_{calo}, the perfor-

mance gain is not as high as in the case of $C_{trck}B_{calo}$. Here also, as we move higher in p_T, the performance gain gradually diminishes, and for the last two p_T bins, we see almost comparable performance between $G_{trck}B_{calo}$ and GNN_{trck}[25].

In Table VI, we present the background rejection at 50 % top-tagging efficiency (ϵ_S^{tag}) evaluated on top samples generated without any truth-level tagging criteria.

---

[25] Although $G_{trck}B_{calo}$ does not show any performance gain compared to GNN_{trck}, as discussed in Section IV D, composite classifiers come with reduced dependence on Monte-Carlo generators and hence $G_{trck}B_{calo}$ results in supressed systematic uncertainty compared to GNN_{trck}.

A diagrammatic representation of this result is presented in the right plot of Figure 5. As discussed in Section V, the motivation for this table is to demonstrate the performance of these classifiers in a collider analysis. As expected, we see an overall degradation in performance for all classifiers. This behavior can be ascribed to improperly reconstructed fat jets in the testing sample. In Table VII, we present the top-tagging efficiency (ϵ_S^{tag}) corresponding to a background rejection factor of 1000 for all the classifiers considered in our analysis. The results are evaluated using top samples generated without any truth-level tagging criteria.

VII. SUMMARY AND OUTLOOK

We have conducted an in-depth analysis of the performance of three machine learning algorithms: the high-level feature (HLF)-based BDT, and the low-level feature (LLF)-based CNN (a miniaturized version of ResNet) and GNN (Lorentznet). Our study focused on their ability to discriminate between fat jets originating from hadronically decaying top quarks and the hadronization of light quarks and gluons. The novel findings of our research are encapsulated as follows:

1. A substantial portion of our study is devoted to emphasizing the significance of leveraging combined information from the calorimeter towers and tracker detectors at the LHC. We found a significant increase in the classifier's performance due to including the jet constituents' electric charge information (tracking data for charged constituents and tower data for neutral constituents) in the training and testing of the LLF-based classifiers like the CNN and GNN. Furthermore, HLF-based classifiers like BDT also exhibit performance enhancements when incorporating track-based HLFs like the number of tracks inside a jet, the p_T weighted width of the tracks, the E_T weighted width of the jet, etc., into the classification task. We found that the high resolution of the tracking data not only improved the classifier performance in the high-p_T regions as demonstrated in Ref. [39], but the information about the distribution and composition of charged and neutral constituents of the jets coming from the tracks and towers also significantly enhance the performance of the classifiers over the whole p_T range. This performance enhancement can be attributed to the fact that, according to the QCD principles and various experimental results, jets initiated by light quarks or gluons exhibit distinct differences in the distribution and composition of charged and neutral hadrons. Consequently, information about the charged and neutral constituents of a jet in the form of tracking and tower data helps identify the quark/gluon origin of sub-jets within a fat jet and hence enhances top tagging efficiency (for an in-depth discussion, please refer to section IV B). Among the group of six simple classifiers discussed in sections IV B and IV A, we found that the track-based GNN classifier (GNN_{trck}) consistently outperformed the others. However, it is important to note that despite their high performance, LLF-based classifiers like GNN_{trck} have a significant drawback: they are heavily dependent on the jet modeling provided by the Monte Carlo simulator, such as Pythia or Herwig, which introduces substantial systematic uncertainties. We also analyzed the classifier dependence on the showering and hadronization model of the Monte-Carlo generator (see section IV D). While track-based LLF classifiers like CNN_{trck} and GNN_{trck} exhibited notable dependency on the Monte Carlo generators, composite classifiers (as discussed in section IV C) demonstrated reduced dependency.

2. To further boost the performance of our simple HLF and LLF-based classifiers, we have developed a series of composite classifiers by stacking a BDT on top of a CNN/GNN. These composite classifiers leverage the strengths of CNN/GNN in extracting specialized observables from low-level inputs and combine them with the effectiveness of BDT in handling complex features. The result is a set of classifiers that exhibit comparable or superior performance. Please refer to section IV C for a comprehensive discussion. In addition to performance enhancement, the composite classifiers demonstrate reduced dependence on the jet modeling of the Monte Carlo generators (see section IV D). The reduced Monte-Carlo generator dependency of the composite classifiers reduces the systematic uncertainties (resulting from the uncertainties in the showering and hadronisation model) to below 20%. Note that the composite classifiers do not solely rely on the event generator-sensitive scores from the LLF-based CNN/GNN classifiers. They also heavily utilize generator-insensitive HLFs such as jet mass, N-subjectness, b-tag, and others for the classification task. The combined use of CNN/GNN scores and other Monte Carlo generator-insensitive HLFs not only reduces overall generator dependency but also enhances their performance significantly.

3. We have done a comprehensive study on the interplay between truth-level identification (ϵ_S^{truth}), classifier efficiency (ϵ_S^c), and top-tagging efficiency (ϵ_S^{tag}). Strict reconstruction and identification criteria increase the purity of the sample, simultaneously decreasing ϵ_S^{truth}. A classifier trained on such pure samples is biased, and the performance cannot be efficiently generalized to new unseen data. We showed that properly selecting the reconstruction radius can improve the ϵ_S^{truth} and help mitigate this issue.

Additionally, we have demonstrated the variation in classifier performance with the transverse momentum of the fat jets. Quark and gluon jets, largely composed of QCD emissions, have an invariant mass that scales with jet p_T. This affects the performance of BDT classifiers where the invariant mass of the fat jet plays a key role in

the classification task, and we see a considerable fall in BDT performance with increasing transverse momentum. The CNN classifier also shows a slight decrease in performance with p_T. This can be ascribed to the fact that the collimation of the constituents increases with transverse momentum, resulting in a top jet that resembles more with the QCD counterpart. The LorentzNet, on the other hand, is based on a Lorentz equivariant architecture and, as claimed by [75], shows almost consistent performance with p_T.

ACKNOWLEDGMENTS

The simulations were partly supported by the SAMKHYA: High-Performance Computing Facility provided by the Institute of Physics, Bhubaneswar. The CNN model used in our analysis was discussed in the Deep Learning course 2021 at the University of Amsterdam [140]. K.G. thanks Biplob Bhattacherjee for useful discussions.

We present the full CNN architecture in Figure 8. First, we have an Input network that comprises the following sequence of operations:

$$Conv2D(c_{in}, c_{out}) \rightarrow BatchNorm2D \rightarrow Relu$$

Here, the convolution layer uses 3×3 filters with unit stride and padding, c_{in} and c_{out} represent the number of channels in the input and output images, respectively. Then, we have nine ResNet blocks stacked one after the other. The ones represented in red do not perform downsampling of the image, whereas the green ones do. then we have the Output network, which we can represent as $AdaptiveAveragePooling(1,1) \rightarrow Flatten \rightarrow Linear(d_{out}, 2)$, where d_{out} represents the number of channels in the output image. Finally, we apply the softmax activation function to get the probability as the CNN score.

Appendix B: Appendix-2 : Correlation and ranking among variables for BDT_{calo} and BDT_{trck}

Table VIII presents the ranking among the variables used in BDT_{calo}. The variables ranked higher are the ones used most frequently for node splitting. In Figure 9, we present the covariance matrix of these variables for the top jets and QCD jets.
Table IX presents the ranking among the variables used in BDT_{trck}. Note that BDT_{trck} uses 26 variables, and we present only the most important of them here. In Figure 10, we present the covariance matrix of the top

Appendix A: Appendix-1 : The CNN model

In the left panel of Figure 7, we present a diagrammatic representation of a single ResNet block. It incorporates two convolution operations with size 3×3 filters, unit stride, and padding. Therefore, these convolution blocks cannot help us reduce the size of the input image. The number of input and output channels is also the same for the convolution operations. To reduce image size, we introduce a second convolution block, as represented in the right panel of Figure 7. Here, the first convolution layer has a stride two and unit padding and hence can reduce the height and width of the input image by half. The second convolution layer in the main network is similar to the previous ResNet block. Now, for the residual connection to work, the size of the input image must match the reduced size of the output image. We achieve this using a third convolution layer with size 1×1 filters, stride two, and no padding. In the subsequent discussion, we refer to this step as downsampling.

Variable	Ranking
M	0.7832
τ_{32}	0.0878
b-tag	0.0844
$\tau_{2,1}$	0.0272
τ_{43}	0.0173

TABLE VIII. method-specific ranking of the input features of BDT_{calo}.

seven highest-ranked variables for the top jets.

Variable	Ranking
M	0.247
$C_\beta(2)$	0.0906
τ_{32}	0.068
$N_{trk}(2)$	0.0558
$\Delta R_{1,2}$	0.045
$\tau_{2,1}$	0.0415
b-tag	0.0412
$w_{trk}(2)$	0.0385
$N_{trk}(1)$	0.0339
$C_\beta(1)$	0.0316
$w_{trk}(1)$	0.0312
$\Delta R_{1,3}$	0.0305
$w_{calo}(2)$	0.0302
τ_{43}	0.0264

TABLE IX. method-specific ranking of the input features of BDT_{trck}.
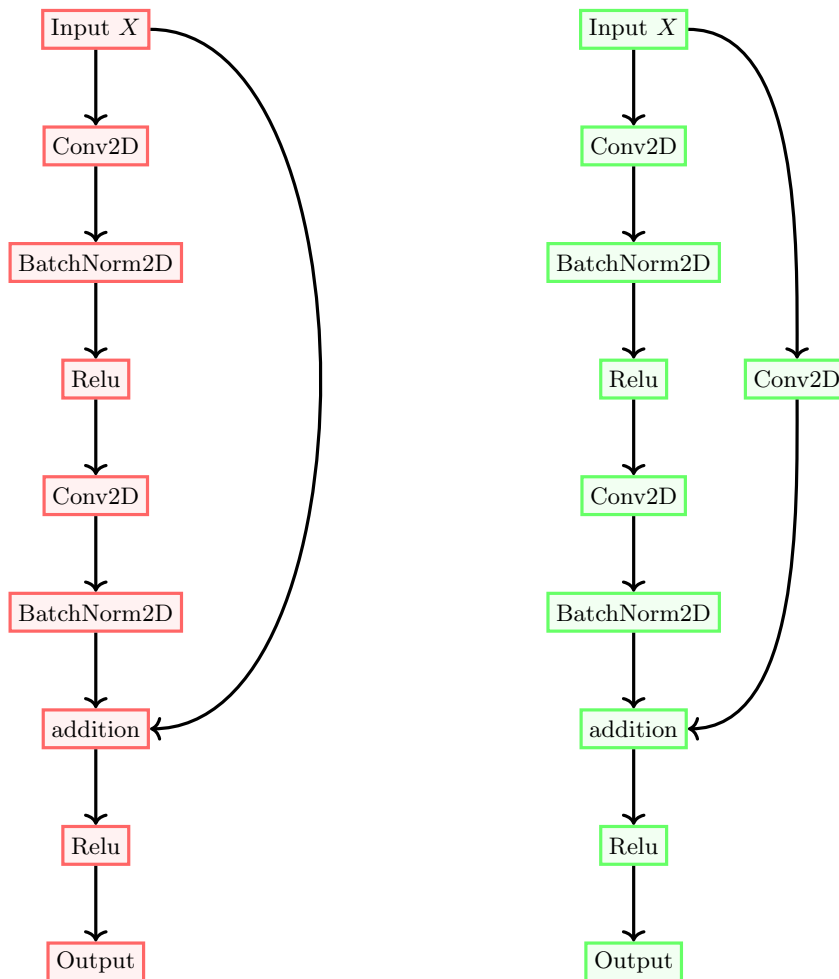
FIG. 7. Diagrammatic representation of the ResNet Block without downsampling (left) and with downsampling (right).

Appendix C: Appendix-3: Correlation and ranking among variables for composite classifiers

This section presents the model-independent ranking and correlation among the variables used in the composite classifiers. The variable ranking for $C_{calo}B_{calo}$ and $C_{trck}B_{calo}$ are presented in Table X. We have similar results for $C_{calo}B_{trck}$ and $C_{trck}B_{trck}$ in Table XI, for $G_{calo}B_{calo}$ and $G_{trck}B_{calo}$ in Table XII, and for $G_{calo}B_{trck}$ and $G_{trck}B_{trck}$ in XIII.

The covariance matrix of $C_{calo}B_{calo}$ and $C_{trck}B_{calo}$ are presented in Figure 11, for $C_{calo}B_{trck}$ and $C_{trck}B_{trck}$ in Figure 12, for $G_{calo}B_{calo}$ and $G_{trck}B_{calo}$ in Figure 13, and for $G_{calo}B_{trck}$ and $G_{trck}B_{trck}$ in Figure 14.

Variable	Ranking		Variable	Ranking
M	0.3818		M	0.3625
score	0.2685		score	0.309
$\tau_{2,1}$	0.1071		$\tau_{2,1}$	0.099
τ_{32}	0.1008		τ_{32}	0.09
b-tag	0.076		b-tag	0.0714
τ_{43}	0.0656		τ_{43}	0.0676

TABLE X. method-specific ranking of the input features of $C_{calo}B_{calo}$(left) and $C_{trck}B_{calo}$(right).

REFERENCES

[1] LHC Machine, JINST **3**, S08001.

[2] G. Aad et al. (ATLAS), Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Phys. Lett. B **716**, 1 (2012), arXiv:1207.7214 [hep-ex].

[3] S. Chatrchyan et al. (CMS), Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC, Phys. Lett. B **716**, 30 (2012), arXiv:1207.7235 [hep-ex].

[4] S. Ashanujjaman, D. Choudhury, and K. Ghosh, Search for exotic leptons in final states with two or three leptons and fat-jets at 13 TeV LHC, JHEP **04**, 150, arXiv:2201.09645 [hep-ph].
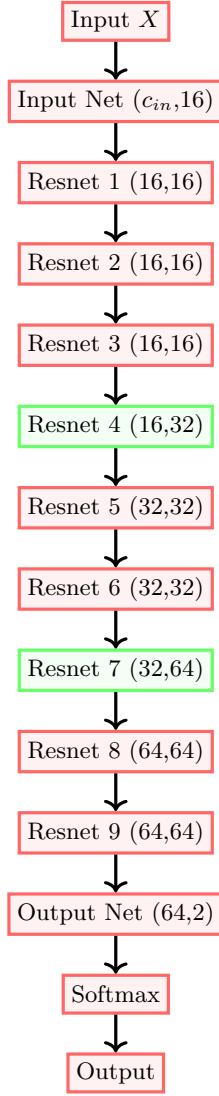
FIG. 8. Schematic Diagram of the CNN architecture used in our analysis. The bracket number represents the channels in the input and output images. For the Output Net, they represent the number of nodes in the input and output layers.
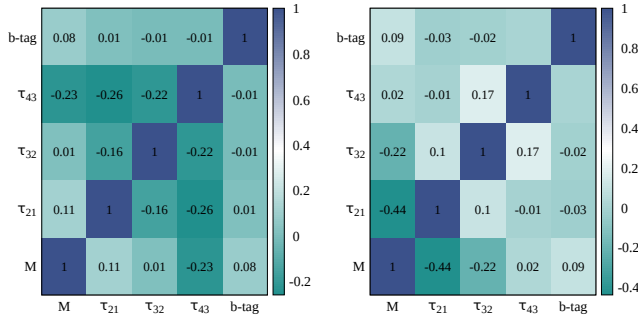
FIG. 9. Correlations among the input features of BDT_{calo} for the top jets (left) and the QCD jets (right).
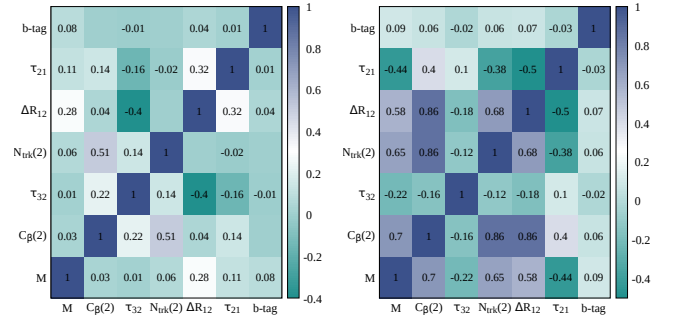
FIG. 10. Correlations among the input features of BDT_{trck} for the top jets (left) and the QCD jets (right).
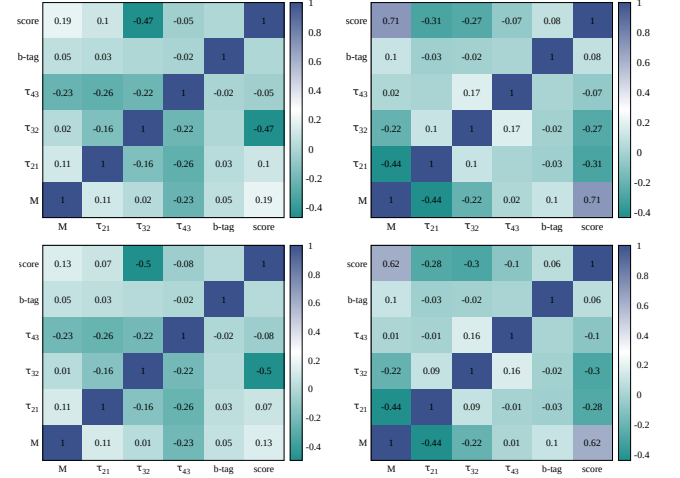
FIG. 11. Correlations among the input features of $C_{calo}B_{calo}$(top row) and $C_{trck}B_{calo}$(bottom row) for the top jets (top left) and the QCD jets (top right).

Variable	Ranking		Variable	Ranking
M	0.1811		M	0.1801
score	0.1328		score	0.1554
$w_{calo}(2)$	0.0814		$\Delta R_{1,2}$	0.0793
$\Delta R_{1,2}$	0.05		$w_{calo}(2)$	0.042
$N_{trk}(2)$	0.0493		b-tag	0.03857
$w_{trk}(2)$	0.0462		τ_{32}	0.03797
τ_{32}	0.0434		$N_{trk}(2)$	0.03731
b-tag	0.03918		$w_{trk}(2)$	0.03452
$w_{trk}(1)$	0.03639		$\tau_{2,1}$	0.03169
$\tau_{2,1}$	0.03602		$C_{\beta}(1)$	0.03011

TABLE XI. method-specific ranking of the input features of $C_{calo}B_{trck}$(left) and $C_{trck}B_{trck}$(right).

[5] S. Ashanujjaman and K. Ghosh, Type-III see-saw: Search for triplet fermions in final states with multiple leptons and fat-jets at 13 TeV LHC, Phys. Lett. B **825**, 136889 (2022), arXiv:2111.07949 [hep-ph].

[6] G. Aad et al. (ATLAS), Search for charginos and neutralinos in final states with two boosted hadronically decaying bosons and missing transverse momentum in
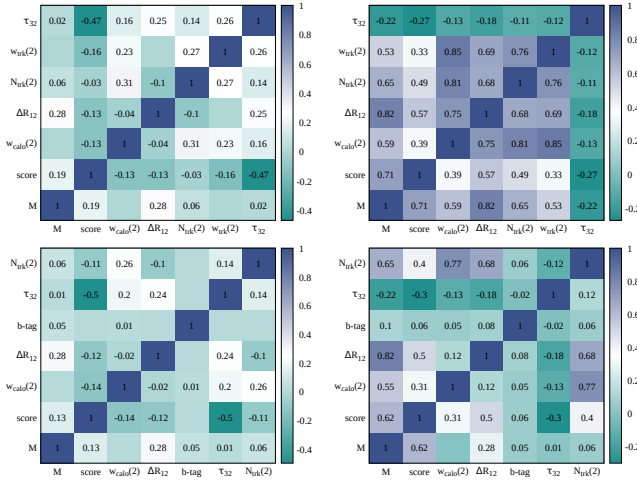
FIG. 12. Correlations among the input features of $C_{calo}B_{trck}$(top row) and $C_{trck}B_{trck}$(bottom row) for the top jets (top left) and the QCD jets (top right).

Variable	Ranking
M	0.3458
score	0.322
$\tau_{2,1}$	0.103
τ_{32}	0.094
b-tag	0.0693
τ_{43}	0.0646

Variable	Ranking
score	0.3517
M	0.3142
$\tau_{2,1}$	0.0968
τ_{32}	0.093
b-tag	0.075
τ_{43}	0.069

TABLE XII. method-specific ranking of the input features of $G_{calo}B_{calo}$(left) and $G_{trck}B_{calo}$(right).
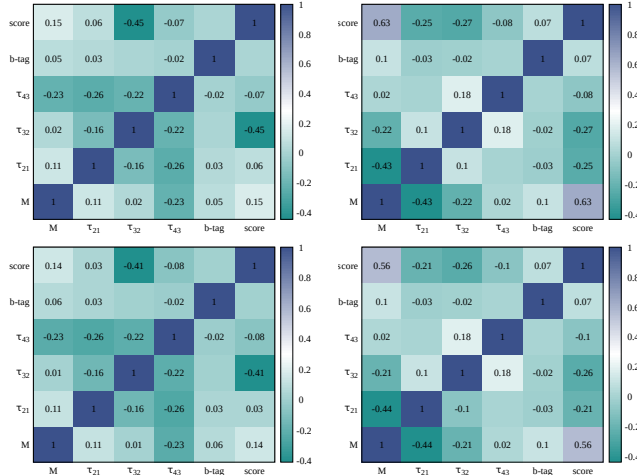
FIG. 13. Correlations among the input features of $G_{calo}B_{calo}$(top row) and $G_{trck}B_{calo}$(bottom row) for the top jets (top left) and the QCD jets (top right).

pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Rev. D **104**, 112010 (2021), arXiv:2108.07586 [hep-ex].

Variable	Ranking
M	0.1696
score	0.1652
$w_{calo}(2)$	0.067
$\Delta R_{1,2}$	0.0568
$w_{trk}(2)$	0.0438
$N_{trk}(2)$	0.0407
$w_{trk}(1)$	0.03836
b-tag	0.03775
τ_{32}	0.0369
$\tau_{2,1}$	0.0311

Variable	Ranking
score	0.1889
M	0.1624
$\Delta R_{1,2}$	0.0755
b-tag	0.0408
$w_{calo}(2)$	0.0354
τ_{32}	0.0351
$w_{trk}(2)$	0.03265
$\tau_{2,1}$	0.0321
$C_\beta(1)$	0.03
$w_{trk}(1)$	0.02931

TABLE XIII. method-specific ranking of the input features of $G_{calo}B_{trck}$(left) and $G_{trck}B_{trck}$(right).
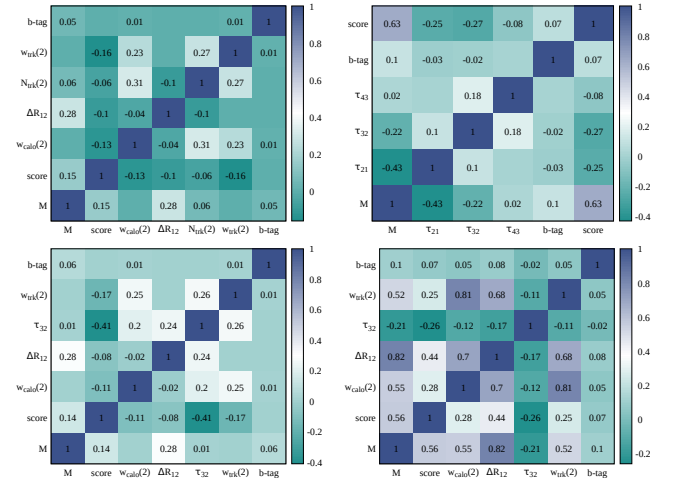
FIG. 14. Correlations among the input features of $G_{calo}B_{trck}$(top row) and $G_{trck}B_{trck}$(bottom row) for the top jets (top left) and the QCD jets (top right).

[7] G. Aad et al. (ATLAS), Search for new phenomena in final states with large jet multiplicities and missing transverse momentum using $\sqrt{s} = 13$ TeV proton-proton collisions recorded by ATLAS in Run 2 of the LHC, JHEP **10**, 062, arXiv:2008.06032 [hep-ex].

[8] K. Ghosh, K. Huitu, J. Laamanen, L. Leinonen, and J. Laamanen, Top Quark Jets as a Probe of the Constrained Minimal Supersymmetric Standard Model with a Degenerate Top Squark and Lightest Supersymmetric Particle, Phys. Rev. Lett. **110**, 141801 (2013), arXiv:1207.2429 [hep-ph].

[9] G. Aad et al. (ATLAS), Search for vector-boson resonances decaying into a top quark and a bottom quark using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, (2023), arXiv:2308.08521 [hep-ex].

[10] G. Aad et al. (ATLAS), Search for single vector-like B quark production and decay via $B \rightarrow bH(b\bar{b})$ in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, (2023), arXiv:2308.02595 [hep-ex].

[11] G. Aad et al. (ATLAS), Search for top-philic heavy resonances in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, (2023), arXiv:2304.01678 [hep-ex].

[12] A. J. Larkoski and E. M. Metodiev, A Theory of Quark vs. Gluon Discrimination, JHEP **10**, 014, arXiv:1906.01639 [hep-ph].

[13] P. T. Komiske, E. M. Metodiev, and J. Thaler, An operational definition of quark and gluon jets, JHEP **11**, 059, arXiv:1809.01140 [hep-ph].

[14] J. Davighi and P. Harris, Fractal based observables to probe jet substructure of quarks and gluons, Eur. Phys. J. C **78**, 334 (2018), arXiv:1703.00914 [hep-ph].

[15] C. Frye, A. J. Larkoski, J. Thaler, and K. Zhou, Casimir Meets Poisson: Improved Quark/Gluon Discrimination with Counting Observables, JHEP **09**, 083, arXiv:1704.06266 [hep-ph].

[16] D. Ferreira de Lima, P. Petrov, D. Soper, and M. Spannowsky, Quark-Gluon tagging with Shower Deconstruction: Unearthing dark matter and Higgs couplings, Phys. Rev. D **95**, 034001 (2017), arXiv:1607.06031 [hep-ph].

[17] J. Gallicchio and M. D. Schwartz, Quark and Gluon Jet Substructure, JHEP **04**, 090, arXiv:1211.7038 [hep-ph].

[18] J. Gallicchio and M. D. Schwartz, Quark and Gluon Tagging at the LHC, Phys. Rev. Lett. **107**, 172001 (2011), arXiv:1106.3076 [hep-ph].

[19] T. Plehn and M. Spannowsky, Top Tagging, J. Phys. G **39**, 083001 (2012), arXiv:1112.4441 [hep-ph].

[20] D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks, Phys. Rev. Lett. **101**, 142001 (2008), arXiv:0806.0848 [hep-ph].

[21] J. Thaler and L.-T. Wang, Strategies to Identify Boosted Tops, JHEP **07**, 092, arXiv:0806.0023 [hep-ph].

[22] J. Thaler and K. Van Tilburg, Identifying Boosted Objects with N-subjettiness, JHEP **03**, 015, arXiv:1011.2268 [hep-ph].

[23] J. Thaler and K. Van Tilburg, Maximizing Boosted Top Identification by Minimizing N-subjettiness, JHEP **02**, 093, arXiv:1108.2701 [hep-ph].

[24] S. Marzani, G. Soyez, and M. Spannowsky, Looking inside jets: an introduction to jet substructure and boosted-object phenomenology, Vol. 958 (Springer, 2019) arXiv:1901.10342 [hep-ph].

[25] T. Plehn, G. P. Salam, and M. Spannowsky, Fat Jets for a Light Higgs, Phys. Rev. Lett. **104**, 111801 (2010), arXiv:0910.5472 [hep-ph].

[26] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, Stop Reconstruction with Tagged Tops, JHEP **10**, 078, arXiv:1006.2833 [hep-ph].

[27] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft Drop, JHEP **05**, 146, arXiv:1402.2657 [hep-ph].

[28] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, Jet substructure as a new Higgs search channel at the LHC, Phys. Rev. Lett. **100**, 242001 (2008), arXiv:0802.2470 [hep-ph].

[29] G. P. Salam, Towards Jetography, Eur. Phys. J. C **67**, 637 (2010), arXiv:0906.1833 [hep-ph].

[30] M. Dasgupta, A. Powling, and A. Siodmok, On jet substructure methods for signal jets, JHEP **08**, 079, arXiv:1503.01088 [hep-ph].

[31] D. Adams et al., Towards an Understanding of the Correlations in Jet Substructure, Eur. Phys. J. C **75**, 409 (2015), arXiv:1504.00679 [hep-ph].

[32] A. Altheimer et al., Boosted Objects and Jet Substructure at the LHC. Report of BOOST2012, held at IFIC Valencia, 23rd-27th of July 2012, Eur. Phys. J. C **74**, 2792 (2014), arXiv:1311.2708 [hep-ex].

[33] A. Altheimer et al., Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks, J. Phys. G **39**, 063001 (2012), arXiv:1201.0008 [hep-ph].

[34] A. Abdesselam et al., Boosted Objects: A Probe of Beyond the Standard Model Physics, Eur. Phys. J. C **71**, 1661 (2011), arXiv:1012.5412 [hep-ph].

[35] R. Kogler et al., Jet Substructure at the Large Hadron Collider: Experimental Review, Rev. Mod. Phys. **91**, 045003 (2019), arXiv:1803.06991 [hep-ex].

[36] A. J. Larkoski, I. Moult, and B. Nachman, Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning, Phys. Rept. **841**, 1 (2020), arXiv:1709.04464 [hep-ph].

[37] X. Ju and B. Nachman, Supervised Jet Clustering with Graph Neural Networks for Lorentz Boosted Bosons, Phys. Rev. D **102**, 075014 (2020), arXiv:2008.06064 [hep-ph].

[38] J. A. Aguilar-Saavedra, F. R. Joaquim, and J. F. Seabra, Mass Unspecific Supervised Tagging (MUST) for boosted jets, JHEP **03**, 012, [Erratum: JHEP 04, 133 (2021)], arXiv:2008.12792 [hep-ph].

[39] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, Deep-learned Top Tagging with a Lorentz Layer, SciPost Phys. **5**, 028 (2018), arXiv:1707.08966 [hep-ph].

[40] S. Macaluso and D. Shih, Pulling Out All the Tops with Computer Vision and Deep Learning, JHEP **10**, 121, arXiv:1803.00107 [hep-ph].

[41] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, Deep-learning Top Taggers or The End of QCD?, JHEP **05**, 006, arXiv:1701.08784 [hep-ph].

[42] M. Farina, Y. Nakai, and D. Shih, Searching for New Physics with Deep Autoencoders, Phys. Rev. D **101**, 075021 (2020), arXiv:1808.08992 [hep-ph].

[43] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, QCD or What?, SciPost Phys. **6**, 030 (2019), arXiv:1808.08979 [hep-ph].

[44] F. A. Dreyer, G. P. Salam, and G. Soyez, The Lund Jet Plane, JHEP **12**, 064, arXiv:1807.04758 [hep-ph].

[45] G. Louppe, K. Cho, C. Becot, and K. Cranmer, QCD-Aware Recursive Neural Networks for Jet Physics, JHEP **01**, 057, arXiv:1702.00748 [hep-ph].

[46] E. A. Moreno, O. Cerri, J. M. Duarte, H. B. Newman, T. Q. Nguyen, A. Periwal, M. Pierini, A. Serikova, M. Spiropulu, and J.-R. Vlimant, JEDI-net: a jet identification algorithm based on interaction networks, Eur. Phys. J. C **80**, 58 (2020), arXiv:1908.05318 [hep-ex].

[47] I. Henrion, J. Brehmer, J. Bruna, K. Cho, K. Cranmer, G. Louppe, and G. Rochette, Neural message passing for jet physics (2017).

[48] A. Chakraborty, S. H. Lim, and M. M. Nojiri, Interpretable deep learning for two-prong jet classification with jet spectra, JHEP **07**, 135, arXiv:1904.02092 [hep-ph].

[49] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow polynomials: A complete linear basis for jet substructure, JHEP **04**, 013, arXiv:1712.07124 [hep-ph].

[50] L. G. Almeida, M. Backović, M. Cliche, S. J. Lee, and M. Perelstein, Playing Tag with ANN: Boosted Top Identification with Pattern Recognition, JHEP **07**, 086, arXiv:1501.05968 [hep-ph].

[51] J. Pearkes, W. Fedorko, A. Lister, and C. Gay, Jet Constituents for Deep Neural Network Based Top Quark Tagging, (2017), arXiv:1704.02124 [hep-ex].

[52] J. Barnard, E. N. Dawe, M. J. Dolan, and N. Rajcic, Parton Shower Uncertainties in Jet Substructure Analyses with Deep Neural Networks, Phys. Rev. D **95**, 014018 (2017), arXiv:1609.00607 [hep-ph].

[53] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, Jet Substructure Classification in High-Energy Physics with Deep Neural Networks, Phys. Rev. D **93**, 094034 (2016), arXiv:1603.09349 [hep-ex].

[54] P. T. Komiske, E. M. Metodiev, and M. D. Schwartz, Deep learning in color: towards automated quark/gluon jet discrimination, JHEP **01**, 110, arXiv:1612.01551 [hep-ph].

[55] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, Jet-images — deep learning edition, JHEP **07**, 069, arXiv:1511.05190 [hep-ph].

[56] J. Cogan, M. Kagan, E. Strauss, and A. Schwarztman, Jet-Images: Computer Vision Inspired Techniques for Jet Tagging, JHEP **02**, 118, arXiv:1407.5675 [hep-ph].

[57] H. Qu and L. Gouskos, ParticleNet: Jet Tagging via Particle Clouds, Phys. Rev. D **101**, 056019 (2020), arXiv:1902.08570 [hep-ph].

[58] G. Kasieczka, N. Kiefer, T. Plehn, and J. M. Thompson, Quark-Gluon Tagging: Machine Learning vs Detector, SciPost Phys. **6**, 069 (2019), arXiv:1812.09223 [hep-ph].

[59] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy Flow Networks: Deep Sets for Particle Jets, JHEP **01**, 121, arXiv:1810.05165 [hep-ph].

[60] H. Luo, M.-x. Luo, K. Wang, T. Xu, and G. Zhu, Quark jet versus gluon jet: fully-connected neural networks with high-level features, Sci. China Phys. Mech. Astron. **62**, 991011 (2019), arXiv:1712.03634 [hep-ph].

[61] T. Cheng, Recursive Neural Networks in Quark/Gluon Tagging, Comput. Softw. Big Sci. **2**, 3 (2018), arXiv:1711.02633 [hep-ph].

[62] L. Lonnblad, C. Peterson, and T. Rognvaldsson, Using neural networks to identify jets, Nucl. Phys. B **349**, 675 (1991).

[63] B. Bhattacherjee, C. Bose, A. Chakraborty, and R. Sengupta, Boosted top tagging and its interpretation using Shapley values, (2022), arXiv:2212.11606 [hep-ph].

[64] S. Bhattacharya, M. Guchait, and A. H. Vijay, Boosted top quark tagging and polarization measurement using machine learning, Phys. Rev. D **105**, 042005 (2022), arXiv:2010.11778 [hep-ph].

[65] A. M. Sirunyan et al. (CMS), Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques, JINST **15** (06), P06005, arXiv:2004.08262 [hep-ex].

[66] S. Choi, S. J. Lee, and M. Perelstein, Infrared Safety of a Neural-Net Top Tagging Algorithm, JHEP **02**, 132, arXiv:1806.01263 [hep-ph].

[67] J. Lin, M. Freytsis, I. Moult, and B. Nachman, Boosting $H \to b\bar{b}$ with Machine Learning, JHEP **10**, 101, arXiv:1807.10768 [hep-ph].

[68] Y.-L. Du, K. Zhou, J. Steinheimer, L.-G. Pang, A. Motornenko, H.-S. Zong, X.-N. Wang, and H. Stöcker, Identifying the nature of the QCD transition in relativistic collision of heavy nuclei with deep learning, Eur. Phys. J. C **80**, 516 (2020), arXiv:1910.11530 [hep-ph].

[69] J. Li, T. Li, and F.-Z. Xu, Reconstructing boosted Higgs jets from event image segmentation, JHEP **04**, 156, arXiv:2008.13529 [hep-ph].

[70] J. Filipek, S.-C. Hsu, J. Kruper, K. Mohan, and B. Nachman, Identifying the Quantum Properties of Hadronic Resonances using Machine Learning, (2021), arXiv:2105.04582 [hep-ph].

[71] K. Fraser and M. D. Schwartz, Jet Charge and Machine Learning, JHEP **10**, 093, arXiv:1803.08066 [hep-ph].

[72] D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban, and D. Whiteson, Jet Flavor Classification in High-Energy Physics with Deep Neural Networks, Phys. Rev. D **94**, 112002 (2016), arXiv:1607.08633 [hep-ex].

[73] S. Egan, W. Fedorko, A. Lister, J. Pearkes, and C. Gay, Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC, (2017), arXiv:1711.09059 [hep-ex].

[74] E. Bols, J. Kieseler, M. Verzetti, M. Stoye, and A. Stakia, Jet Flavour Classification Using DeepJet, JINST **15** (12), P12012, arXiv:2008.10519 [hep-ex].

[75] S. Gong, Q. Meng, J. Zhang, H. Qu, C. Li, S. Qian, W. Du, Z.-M. Ma, and T.-Y. Liu, An efficient Lorentz equivariant graph neural network for jet tagging, JHEP **07**, 030, arXiv:2201.08187 [hep-ph].

[76] M. Erdmann, E. Geiser, Y. Rath, and M. Rieger, Lorentz Boost Networks: Autonomous Physics-Inspired Feature Engineering, JINST **14** (06), P06006, arXiv:1812.09722 [hep-ex].

[77] A. Bogatskiy, B. Anderson, J. T. Offermann, M. Roussi, D. W. Miller, and R. Kondor, Lorentz Group Equivariant Neural Network for Particle Physics, (2020), arXiv:2006.04780 [hep-ph].

[78] E. A. Moreno, T. Q. Nguyen, J.-R. Vlimant, O. Cerri, H. B. Newman, A. Periwal, M. Spiropulu, J. M. Duarte, and M. Pierini, Interaction networks for the identification of boosted $H \to b\bar{b}$ decays, Phys. Rev. D **102**, 012010 (2020), arXiv:1909.12285 [hep-ex].

[79] V. Mikuni and F. Canelli, ABCNet: An attention-based method for particle tagging, Eur. Phys. J. Plus **135**, 463 (2020), arXiv:2001.05311 [physics.data-an].

[80] E. Bernreuther, T. Finke, F. Kahlhoefer, M. Krämer, and A. Mück, Casting a graph net to catch dark showers, SciPost Phys. **10**, 046 (2021), arXiv:2006.08639 [hep-ph].

[81] J. Guo, J. Li, T. Li, and R. Zhang, Boosted Higgs boson jet reconstruction via a graph neural network, Phys. Rev. D **103**, 116025 (2021), arXiv:2010.05464 [hep-ph].

[82] M. J. Dolan and A. Ore, Equivariant Energy Flow Networks for Jet Tagging, Phys. Rev. D **103**, 074022 (2021), arXiv:2012.00964 [hep-ph].

[83] V. Mikuni and F. Canelli, Point cloud transformers applied to collider physics, Mach. Learn. Sci. Tech. **2**, 035027 (2021), arXiv:2102.05073 [physics.data-an].

[84] P. Konar, V. S. Ngairangbam, and M. Spannowsky, Energy-weighted Message Passing: an infra-red and collinear safe graph neural network algorithm, (2021), arXiv:2109.14636 [hep-ph].

[85] C. Shimmin, Particle Convolution for High Energy Physics, (2021), arXiv:2107.02908 [hep-ph].

[86] F. A. Dreyer and H. Qu, Jet tagging in the Lund plane with graph networks, JHEP **03**, 052, arXiv:2012.08526 [hep-ph].

[87] F. Dreyer, G. Soyez, and A. Takacs, Quarks and gluons in the Lund plane, (2021), arXiv:2112.09140 [hep-ph].

[88] K. Datta and A. Larkoski, How Much Information is in a Jet?, JHEP **06**, 073, arXiv:1704.08249 [hep-ph].

[89] K. Datta and A. J. Larkoski, Novel Jet Observables from Machine Learning, JHEP **03**, 086, arXiv:1710.01305 [hep-ph].

[90] K. Datta, A. Larkoski, and B. Nachman, Automating the Construction of Jet Observables with Machine Learning, Phys. Rev. D **100**, 095016 (2019), arXiv:1902.07180 [hep-ph].

[91] A. Chakraborty, S. H. Lim, M. M. Nojiri, and M. Takeuchi, Neural Network-based Top Tagger with Two-Point Energy Correlations and Geometry of Soft Emissions, JHEP **07**, 111, arXiv:2003.11787 [hep-ph].

[92] M. Baak, J. Cúth, J. Haller, A. Hoecker, R. Kogler, K. Mönig, M. Schott, and J. Stelzer (Gfitter Group), The global electroweak fit at NNLO and prospects for the LHC and ILC, Eur. Phys. J. C **74**, 3046 (2014), arXiv:1407.3792 [hep-ph].

[93] G. Degrassi, S. Di Vita, J. Elias-Miro, J. R. Espinosa, G. F. Giudice, G. Isidori, and A. Strumia, Higgs mass and vacuum stability in the Standard Model at NNLO, JHEP **08**, 098, arXiv:1205.6497 [hep-ph].

[94] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition 10.1109/CVPR.2016.90 (2015), arXiv:1512.03385 [cs.CV].

[95] A. Tumasyan et al. (CMS), Measurement of the differential $t\bar{t}$ production cross section as a function of the jet mass and extraction of the top quark mass in hadronic decays of boosted top quarks, Eur. Phys. J. C **83**, 560 (2023), arXiv:2211.01456 [hep-ex].

[96] A. M. Sirunyan et al. (CMS), Search for a heavy resonance decaying to a top quark and a W boson at $\sqrt{s} = $ 13 TeV in the fully hadronic final state, JHEP **12**, 106, arXiv:2104.12853 [hep-ex].

[97] Boosted Top Jet Tagging at CMS, (2014).

[98] M. Aaboud et al. (ATLAS), Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC, Eur. Phys. J. C **79**, 375 (2019), arXiv:1808.07858 [hep-ex].

[99] Boosted jet identification using particle candidates and deep neural networks, (2017).

[100] A. Butter et al., The Machine Learning landscape of top taggers, SciPost Phys. **7**, 014 (2019), arXiv:1902.09914 [hep-ph].

[101] A. Bogatskiy, T. Hoffman, D. W. Miller, J. T. Offermann, and X. Liu, Explainable Equivariant Neural Networks for Particle Physics: PELICAN, (2023), arXiv:2307.16506 [hep-ph].

[102] D. Ba, A. S. Dogra, R. Gambhir, A. Tasissa, and J. Thaler, SHAPER: can you hear the shape of a jet?, JHEP **06**, 195, arXiv:2302.12266 [hep-ph].

[103] A. Bogatskiy, T. Hoffman, D. W. Miller, and J. T. Offermann, PELICAN: Permutation Equivariant and Lorentz Invariant or Covariant Aggregator Network for Particle Physics, (2022), arXiv:2211.00454 [hep-ph].

[104] D. Murnane, S. Thais, and J. Wong, Semi-Equivariant GNN Architectures for Jet Tagging, J. Phys. Conf. Ser. **2438**, 012121 (2023), arXiv:2202.06941 [hep-ph].

[105] J. Y. Araz and M. Spannowsky, Classical versus quantum: Comparing tensor-network-based quantum circuits on Large Hadron Collider data, Phys. Rev. A **106**, 062423 (2022), arXiv:2202.10471 [quant-ph].

[106] F. Mokhtar, R. Kansal, and J. Duarte, Do graph neural networks learn traditional jet substructure?, in 36th Conference on Neural Information Processing Systems (2022) arXiv:2211.09912 [hep-ex].

[107] M. Andrews, J. Alison, S. An, P. Bryant, B. Burkle, S. Gleyzer, M. Narain, M. Paulini, B. Poczos, and E. Usai, End-to-end jet classification of quarks and gluons with the CMS Open Data, Nucl. Instrum. Meth. A **977**, 164304 (2020), arXiv:1902.08276 [hep-ex].

[108] M. Andrews et al., End-to-end jet classification of boosted top quarks with the CMS open data, EPJ Web Conf. **251**, 04030 (2021), arXiv:2104.14659 [physics.data-an].

[109] Discrimination of Light Quark and Gluon Jets in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector, (2016).

[110] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, JHEP **07**, 079, arXiv:1405.0301 [hep-ph].

[111] R. D. Ball et al. (NNPDF), Parton distributions for the LHC Run II, JHEP **04**, 040, arXiv:1410.8849 [hep-ph].

[112] C. Bierlich et al., A comprehensive guide to the physics and usage of PYTHIA 8.3 10.21468/SciPost-PhysCodeb.8 (2022), arXiv:2203.11601 [hep-ph].

[113] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, JHEP **02**, 057, arXiv:1307.6346 [hep-ex].

[114] M. Cacciari, G. P. Salam, and G. Soyez, FastJet User Manual, Eur. Phys. J. C **72**, 1896 (2012), arXiv:1111.6097 [hep-ph].

[115] D. Krohn, J. Thaler, and L.-T. Wang, Jet Trimming, JHEP **02**, 084, arXiv:0912.1342 [hep-ph].

[116] D. Krohn, J. Thaler, and L.-T. Wang, Jets with Variable R, JHEP **06**, 059, arXiv:0903.0392 [hep-ph].

[117] B. Mukhopadhyaya, T. Samui, and R. K. Singh, Dynamic radius jet clustering algorithm, JHEP **04**, 019, arXiv:2301.13074 [hep-ph].

[118] D. Krohn, M. D. Schwartz, T. Lin, and W. J. Waalewijn, Jet Charge at the LHC, Phys. Rev. Lett. **110**, 212001 (2013), arXiv:1209.2421 [hep-ph].

[119] Performance of quark/gluon discrimination in 8 TeV pp data, (2013).

[120] A. J. Larkoski, G. P. Salam, and J. Thaler, Energy Correlation Functions for Jet Substructure, JHEP **06**, 108, arXiv:1305.0007 [hep-ph].

[121] J. Pumplin, How to tell quark jets from gluon jets, Phys. Rev. D **44**, 2025 (1991).

[122] Y.-C. J. Chen, C.-W. Chiang, G. Cottin, and D. Shih, Boosted W and Z tagging with jet charge and deep learning, Phys. Rev. D **101**, 053001 (2020), arXiv:1908.08256 [hep-ph].

[123] J. Sylvester, Xix. a demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **4**, 138 (1852), https://doi.org/10.1080/14786445208647087.

[124] A. Hocker et al., TMVA - Toolkit for Multivariate Data Analysis, (2007), arXiv:physics/0703039.

[125] R. Brun, F. Rademakers, and S. Panacek, ROOT, an object oriented data analysis framework, in CERN School of Computing (CSC 2000) (2000) pp. 11–42.

[126] I. Loshchilov and F. Hutter, Decoupled weight decay regularization (2019), arXiv:1711.05101 [cs.LG].

[127] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, The graph neural network model, IEEE Transactions on Neural Networks **20**, 61 (2009).

[128] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, Neural message passing for quantum chemistry (2017), arXiv:1704.01212 [cs.LG].

[129] J. Bellm et al., Herwig 7.0/Herwig++ 3.0 release note, Eur. Phys. J. C **76**, 196 (2016), arXiv:1512.01178 [hep-ph].

[130] M. Bahr et al., Herwig++ Physics and Manual, Eur. Phys. J. C **58**, 639 (2008), arXiv:0803.0883 [hep-ph].

[131] G. Altarelli and G. Parisi, Asymptotic Freedom in Parton Language, Nucl. Phys. B **126**, 298 (1977).

[132] P. D. Acton et al. (OPAL), A Study of differences between quark and gluon jets using vertex tagging of quark jets, Z. Phys. C **58**, 387 (1993).

[133] R. Akers et al. (OPAL), A Model independent measurement of quark and gluon jet properties and differences, Z. Phys. C **68**, 179 (1995).

[134] R. Barate et al. (ALEPH), Studies of quantum chromodynamics with the ALEPH detector, Phys. Rept. **294**, 1 (1998).

[135] P. Abreu et al. (DELPHI), Energy dependence of the differences between the quark and gluon jet fragmentation, Z. Phys. C **70**, 179 (1996).

[136] L. Lonnblad, C. Peterson, and T. Rognvaldsson, Finding Gluon Jets With a Neural Trigger, Phys. Rev. Lett. **65**, 1321 (1990).

[137] P. T. Komiske, S. Kryhin, and J. Thaler, Disentangling quarks and gluons in CMS open data, Phys. Rev. D **106**, 094021 (2022), arXiv:2205.04459 [hep-ph].

[138] G. Rauco (ATLAS, CMS), Distinguishing quark and gluon jets at the LHC, in Parton radiation and fragmentation from LHC to FCC-ee (2017) pp. 73–78.

[139] G. Aad et al. (ATLAS), Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector, Eur. Phys. J. C **74**, 3023 (2014), arXiv:1405.6583 [hep-ex].

[140] P. Lippe, UvA Deep Learning Tutorials (2022).