Learning to classify from impure samples with high-dimensional data

Patrick T. Komiske,[1,*] Eric M. Metodiev,[1,†] Benjamin Nachman,[2,‡] and Matthew D. Schwartz[3,§]

[1]*Center for Theoretical Physics, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, USA*
[2]*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[3]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*

A persistent challenge in practical classification tasks is that labeled training sets are not always available. In particle physics, this challenge is surmounted by the use of simulations. These simulations accurately reproduce most features of data, but cannot be trusted to capture all of the complex correlations exploitable by modern machine learning methods. Recent work in weakly supervised learning has shown that simple, low-dimensional classifiers can be trained using only the impure mixtures present in data. Here, we demonstrate that complex, high-dimensional classifiers can also be trained on impure mixtures using weak supervision techniques, with performance comparable to what could be achieved with pure samples. Using weak supervision will therefore allow us to avoid relying exclusively on simulations for high-dimensional classification. This work opens the door to a new regime whereby complex models are trained directly on data, providing direct access to probe the underlying physics.

Data analysis methods at the Large Hadron Collider (LHC) rely heavily on simulations. These simulations are generally excellent and allow us to explore the mapping between truth information (particles from collisions) and observables (reconstructed momenta and energies). In particular, simulations let us train complex algorithms to extract the truth information from the observables. Machine learning methods trained on low-level inputs have been developed for collider physics [1] to identify boosted W/Z/Higgs bosons [2–8], top quarks [9–13], b-quarks [14–16], and light quarks [17–20], to remove noise [21], and to emulate particle interactions with calorimeters [22–24]. These new methods achieve excellent performance by exploiting subtle features of the simulations, which are presumed to be similar to the features in the data. Unfortunately, the simulations are known to be imperfect. This is particularly true for subtle features in high dimensions, as illustrated clearly for boosted W bosons in Ref. [5] and by the need for non-negligible corrections ("scale factors") to be applied to multivariate classifiers used by the current LHC experiments (see e.g. Refs. [25–32]). Thus, it

is natural to question the performance of machine learning algorithms trained on simulations as we know that if a model is trained on unphysical artifacts, this is what the model will learn. This objection certainly has merit, as the power of these methods for physics applications stems precisely from their ability to find features that we do not fully understand and cannot easily interpret.

Data-driven approaches avoid the pitfalls of relying on simulations in experimental analyses. For simple observables, such as the invariant mass of a photon pair, a traditional experimental approach has been to perform sideband fits directly to the data. This avoids relying on the simulation altogether. Unfortunately, most of the sophisticated discrimination techniques developed in recent years use *full supervision*, where truth information is needed in order to train the classifier. However, real data generally consist only of mixed samples without truth information, arising from underlying statistical or quantum mixtures of two classes (henceforth referred to as "signal" and "background"). Occasionally one can find a small region of phase space where the signal or background is pure, but these regions are generally sparsely populated and may not produce representative distributions. Recent work on *weak supervision* [33] allows classifiers to be trained using only the information available from mixed samples. Two weakly supervised paradigms tailored to physics applications are learning from label proportions (LLP) [34] and classification without labels (CWoLa) [35]. Reference [34] considered the problem of discriminating the radiation pattern of quark from gluons (q/g) using three standard observables and showed how to achieve fully

*[*]pkomiske@mit.edu
[†]metodiev@mit.edu
[‡]bpnachman@lbl.gov
[§]schwartz@physics.harvard.edu

supervised discrimination power by using LLP with two samples of different but known quark fractions. In Ref. [35], it was shown that the proportions are not necessary for training since the likelihood ratio of the mixed samples is monotonically related to the signal/ background likelihood ratio, the optimal binary classifier for signal vs background.

One potential objection to the weak-learning demonstrations in Refs. [34–36] is that the dimensionality of the inputs used is small. Indeed, for a one-dimensional discriminant one can extract the exact pure distributions from mixed samples using the fractions. It is not obvious that weak supervision will succeed when trained on high-dimensional inputs where the feature space may be sparsely populated. Indeed, the most powerful modern methods are trained on high-dimensional, low-level inputs, where numerically approximating and weighting the probability distribution is completely intractable. These deep learning techniques can expose subtle correlations in many dimensions which are also much harder to model than simple low-dimensional features.

In this paper, we demonstrate that weak supervision can approach the effectiveness of full supervision on complex models with high-dimensional inputs. As a concrete illustration, we use an image representation to distinguish the radiation pattern of high-energy quarks from gluons ("jet images" [2]). Convolutional neural networks (CNNs) are applied to the quark and gluon jet images, where the dimensionality of the inputs is $\mathcal{O}(1000)$ and simulation mismodeling issues are a challenge [25,37–43]. We find that CWoLa more robustly generalizes to learning with high-dimensional inputs than LLP, with the latter requiring careful engineering choices to achieve comparable performance. Though we use a particle physics problem as an example, the lessons about learning from data using mixtures of signal and background are applicable more broadly.

We begin by establishing some notation and formulating the problem. Let $\mathbf{x}$ represent a vector of observables (*features*) useful for discriminating two classes we call *signal* (S) and *background* (B). For example, $\mathbf{x}$ might be the momenta of observed particles, calorimeter energy deposits, or a complete set of observables [7,8]. In fully supervised learning, each training sample is assigned a truth label such as 1 for signal and 0 for background. Then the fully supervised model is trained to predict the correct labels for each training example by minimizing a loss function. For a sufficiently large training set, an appropriate model parameterization, and a suitable minimization procedure, the learned model should approach the optimal classifier defined by thresholding the likelihood ratio.

Data collected from a real detector do not come with signal/background labels. Instead, one typically has two or more *mixtures* M_a of signal and background with different signal fractions f_a, such that the distribution of the features, $p_{M_a}(\mathbf{x})$, is given by

$$p_{M_a}(\mathbf{x}) = f_a p_S(\mathbf{x}) + (1 - f_a)p_B(\mathbf{x}), \qquad (1)$$

where p_S and p_B are the signal and background distributions, respectively. Weak supervision assumes *sample independence*, that Eq. (1) holds with the same distributions $p_S(\mathbf{x})$ and $p_B(\mathbf{x})$ for all mixtures. Although in most situations sample independence does not hold perfectly (see e.g. Ref. [44]), it is often a very good approximation (cf. Table Ref. II below).

LLP uses any fully supervised classification method and modifies the loss function to globally match the signal fraction predicted by the model on a batch of training samples to the known truth fractions f_a. Breaking the training set into batches, normally done to parallelize training, takes on a new significance with LLP since the loss function is evaluated globally on each batch. The batch size, which for LLP we define as the number of samples drawn from each mixture during one update of the model, is a critical hyperparameter of LLP.

The loss functions we use for LLP differ from those in Ref. [34]. Analogous to the mean squared error (MSE) loss function for fully supervised (or CWoLa) training, we introduce the weak MSE (WMSE) loss for the LLP framework:

$$\ell_{\text{WMSE}} = \sum_a \left(f_a - \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{x}_i)\right)^2, \qquad (2)$$

where N is the batch size, a indexes the mixed samples, and h is the model. Analogous to the cross entropy, we also introduce the weak cross-entropy (WCE) loss:

$$\ell_{\text{WCE}} = \sum_a \text{CE}\left(f_a, \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{x}_i)\right), \qquad (3)$$

where $\text{CE}(a, b) = -a \log b - (1 - a)\log(1 - b)$. One caveat we discovered while exploring LLP is that the range of $h(\mathbf{x})$ must be restricted to [0, 1], otherwise the model falls into trivial minima of the loss function. We also observe the effect of model outputs becoming effectively binary at 0 and 1, necessitating additional care to avoid numerical precision issues.

CWoLa classifies two mixtures, M_1 and M_2, from each other using any fully supervised classification method. The resulting classifier is then used to directly distinguish the original signal and background processes. Amazingly, the CWoLa classifier asymptotically (as the amount of training data increases) approaches an ideal classifier trained on pure samples [35,45,46]. CWoLa does not require that the fractions f_a are known for training (the fractions on smaller test sets can be used to calibrate the classifier operating points). The CWoLa framework has the nice property that

TABLE I.   The essential pros (✔), cons (✗), and open questions (?) of the CWoLa and LLP weak supervision paradigms.

Property	LLP	CWoLa
Compatible with any trainable model	✔	✔
No training modifications needed	✗	✔
Training does not need fractions	✗	✔
Smooth limit to full supervision	✗	✔
Works for >2 mixed samples	✔	?

as the samples approach complete purity ($f_1 \to 0, f_2 \to 1$) it smoothly approaches the fully supervised paradigm. CWoLa presently only works with two mixtures; if more than two are available they can be pooled at the cost of diluting their purity. The key features of CWoLa and LLP are compared in Table I. Note that no learning is possible with either method as $f_1 \to f_2$.

To explore weak supervision methods with high-dimensional inputs, we simulate $Z+q/g$ events at $\sqrt{s} = 13\,\text{TeV}$ using PYTHIA 8.226 [47] and create artificially mixed samples with various quark (signal) fractions. Jets with transverse momentum $p_T^{\text{jet}} \in [250, 275]\,\text{GeV}$ and rapidity $|y| \leq 2.0$ are obtained from final-state, non-neutrino particles clustered using the anti-k_t algorithm [48] with radius $R = 0.4$ implemented in FASTJET 3.3.0 [49]. Single-channel, 33×33 jet images [2,3,17] are constructed from a patch of the pseudorapidity-azimuth plane of size 0.8×0.8 centered on the jet, treating the particle p_T values as pixel intensities. The images are normalized so the sum of the pixels is 1 and standardized by subtracting the mean and dividing by the standard deviation of each pixel as calculated from the training set.

All instantiations and trainings of neural networks were performed with the python deep learning library Keras [50] with the TensorFlow [51] backend. A CNN architecture similar to that employed in Ref. [17] was used: three 32-filter convolutional layers with filter sizes of 8×8, 4×4, and 4×4 followed by a 128-unit dense layer. Maxpooling of size 2×2 was performed after each convolutional layer with a stride length of 2. The dropout rate was taken to be 0.1 for all layers. Keras VarianceScaling initialization was used to initialize the weights of the convolutional layers. Due to numerical precision issues caused by the tendency of LLP to push outputs to 0 or 1, a softmax activation function was included as part of the loss function rather than the model output layer. Validation and test sets were used consisting each of 50k 50:50 mixtures of quark and gluon jet images. Training was performed with the Adam algorithm [52] with a learning rate of 0.001 and a validation performance patience of 10 epochs. Each network was trained 10 times and the variation of the performance was used as a measure of the uncertainty. Unless otherwise specified, the following are used by default: exponential linear unit (ELU) [53] activation functions for all nonoutput layers, the CE loss function for CWoLa, and the WCE loss function for LLP.

The performance of a binary classifier can be captured by its receiver operating characteristic (ROC) curve. To condense the classifier performance into a single number, we use the area under the ROC curve (AUC). The AUC is also the probability that the classifier output is higher for signal than for background. Random classifiers have AUC = 0.5 and perfect classifiers have AUC = 1.0. We also confirmed that our conclusions are unchanged when using the background mistag rate at 50% signal efficiency as a performance metric instead.

As previously noted, the LLP paradigm works by matching the predicted fraction of signal events to the known fraction for multiple mixed samples. In Ref. [34], the averaging took place over the entire mixed sample. Averaging over the entire training set at once is effectively impossible for high-dimensional inputs such as jet images because the graphics processing units (GPUs) that are needed to train the CNNs in a reasonable amount of time typically do not have enough memory to hold the entire training set at one time. Hence, the ability to train with batches is highly desirable for using LLP with high-dimensional inputs.

There are many tradeoffs inherent with choosing the LLP batch size. Smaller batch sizes are susceptible to shot noise in the sense that the actual signal fraction on that batch may differ significantly from the fraction for the entire mixed sample, an effect which decreases as the batch size increases. Smaller batch sizes result in longer training times per epoch (because the full parallelization capabilities of the GPU cannot be used) but often require fewer epochs to train. Larger batch sizes have shorter training times per epoch but typically require more epochs to train. For CWoLa, the batch size plays the same role as in full supervision, with the
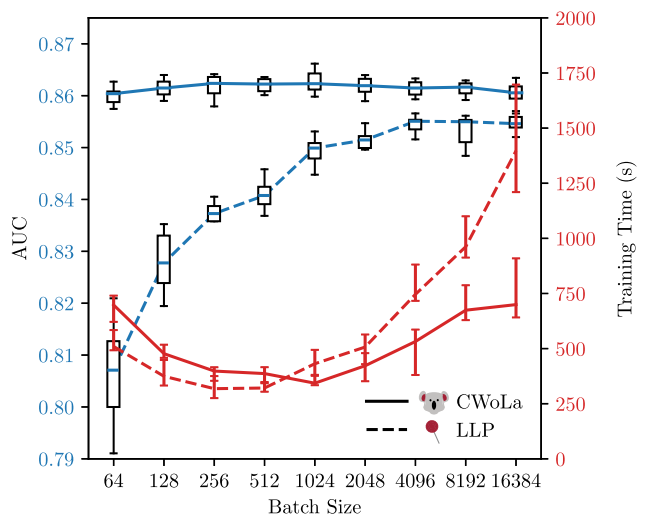

FIG. 1.   The AUC and training time of CWoLa (solid) and LLP (dashed) as the batch size is varied. Training times are measured on an NVIDIA Tesla K80 GPU using CUDA 8.0, TensorFlow 1.4.1, and Keras 2.1.2. AUC is a measure of classifier performance and is 1 for a perfect classifier and 0.5 for a completely random one.

TABLE II.   AUCs for training with CWoLa and LLP on Z + jet and dijet samples as well as on artificial mixtures of $Z + g$ and $Z + q$ samples. The error given is the interquartile range. There is no significant difference in classifier performance between the naturally mixed (Z + jet vs dijets) samples and the artificially mixed ($Z + q/g$) samples with the same signal fractions.

Learning	Sample	AUC
CWoLa	Z + jet vs dijets	0.8626 ± 0.0020
	Artificial $Z + q/g$	0.8621 ± 0.0019
LLP	Z + jet vs dijets	0.8544 ± 0.0019
	Artificial $Z + q/g$	0.8549 ± 0.0018

performance being largely insensitive to it but the total training time varying slightly. These tradeoffs are captured in Fig. 1, which shows both the performance and training time for CWoLa and LLP models as the batch size is swept in powers of two from 64 to 16384, trained on two mixtures with $f_1 = 0.2$ and $f_2 = 0.8$. The expected independence of CWoLa performance and the degradation of LLP performance for low batch sizes can clearly be seen. The training time curves are concave with optimum batch sizes toward the middle of the swept region. Based on this figure, we choose default batch sizes of 4000 for LLP and 400 for CWoLa.

In order to explore a slightly more realistic scenario than artificially mixing samples from the same distribution of quarks and gluons, we generate Z + jet and dijet events with the same generation parameters and cuts as described previously. These "naturally" mixed samples have quark fractions $f_{Z+jet} = 0.88$ and $f_{dijets} = 0.37$. The signal and background fractions have been systematically explored for these and many other processes in Ref. [54]. As indicated by Table Ref. II, there is no significant difference in performance on the naturally mixed or artificially mixed samples. Hence, artificially mixed samples are used in the rest of this study in order to evaluate weak supervision performance at different quark purities.

Figure 2 compares CWoLa and LLP performance for various quark/gluon purities as a function of the number of training samples. Each network is trained using two samples, one with quark fraction f_1 and the other with quark fraction $f_2 = 1 - f_1$. Each point in the figure is the median of 10 independent network trainings and the error bars show the 25th and 75th percentiles. Full supervision performance corresponds to CWoLa with $f_1 = 0$. The most important takeaway from Fig. 2 is that we have achieved good performance with both weak supervision methods over a large variety of sample purities and training sample sizes. We also see that CWoLa consistently outperforms LLP and continues to get better as additional training samples are used, likely a result of the increasingly populated feature space, whereas LLP performance tends to level off. It should be noted that given the binary output nature of LLP models, classifiers trained in this way
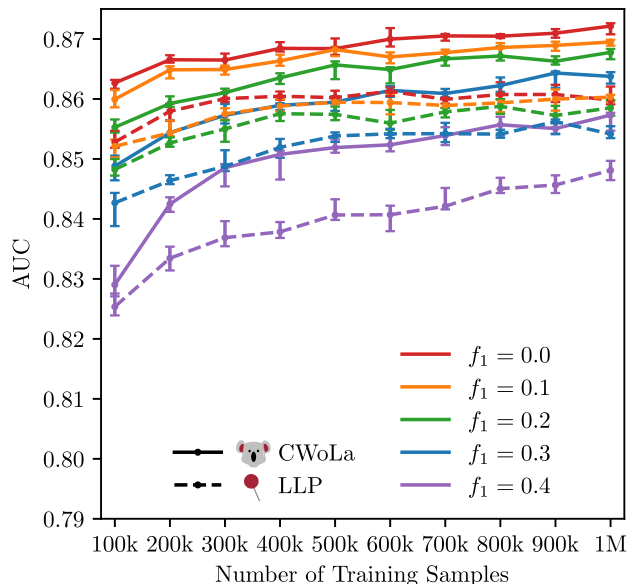

FIG. 2.   Classifier performance (AUC) shown for both CWoLa (solid) and LLP (dashed) trained on two mixed samples with various signal fractions f_1, $1 - f_1$ as the number of training data is varied between 100k and 1M. Each training is repeated 10 times and the 25th, 50th, and 75th percentiles are shown. The $f_1 = 0.0$ CWoLa curve corresponds to full supervision. CWoLa outperforms LLP by this metric, though both methods work quite well.

effectively come with a working point and sweeping the threshold to produce a ROC curve may not be ideal. The purity/data tradeoff analysis of Fig. 2 can provide valuable information for practical applications of weak supervision methods in physics, particularly in cases where more data can be acquired at the expense of worsening sample purity.

The sensitivity of LLP to different choices of loss function and activation function was examined. We studied the choices of the symmetric squared loss of Eq. (2) and the weak cross-entropy loss of Eq. (3) with rectified linear unit (ReLU) [55] and ELU activation functions. We found a significant improvement in LLP classification performance in using ELU activations instead of ReLU activations, particularly at high signal efficiencies. The choice of loss function was found to be less important than the choice of activation function, but minor improvements in AUC were observed with the WCE loss function over WMSE. We also studied the dependence of CWoLa on the choice of activation function and found consistent performance between ELU and ReLU activations. These results justify our default choices of ELU activation and WCE loss functions. With the choice of ELU activation, LLP achieves almost the same performance to our CWoLa-trained network near the operating point with equal signal and background efficiencies. We suspect this is a result of the tendency of LLP to output binary predictions (near 0 or 1) rather than a continuous output that can be easily thresholded.

Lastly, LLP has the potential advantage over the present implementation of CWoLa that it can naturally encompass multiple mixed samples with different purities. While in principle adding more samples should help, it is not obvious whether the network will effectively take advantage of them. Indeed, we did not find significant improvement to LLP when adding additional samples with intermediate purities, even after significant, dedicated architecture engineering.

In conclusion, we have shown that machine learning approaches using very high-dimensional inputs can be trained directly on mixtures of signal and background, and therefore on data. This addresses one of the main objections to the use of modern machine learning in jet tagging: sensitivity to untrustworthy simulations. We have implemented and tested weakly supervised learning with both LLP and CWoLa, finding that for the quark/gluon discrimination problem considered here CWoLa outperforms LLP and is less sensitive to particular hyperparameter choices. We have developed a method for training LLP with high-dimensional inputs in batches and demonstrated that the batch size is a critical hyperparameter for both performance and training time. Given any fully supervised classifier, CWoLa works "out-of-the-box," whereas LLP requires additional engineering to achieve good performance and is generally harder to train. Nonetheless, the success in using both of these weak supervision approaches on high-dimensional data is encouraging for the future of modern machine learning techniques in particle physics and beyond.

[1] 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, arXiv:1709.04464.

[2] J. Cogan, M. Kagan, E. Strauss, and A. Schwarztman, Jet-Images: Computer vision inspired techniques for jet tagging, J. High Energy Phys. 02 (2015) 118.

[3] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, Jet-images deep learning edition, J. High Energy Phys. 07 (2016) 069.

[4] 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).

[5] 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).

[6] G. Louppe, K. Cho, C. Becot, and K. Cranmer, QCD-aware recursive neural networks for jet physics, arXiv:1702.00748.

[7] K. Datta and A. Larkoski, How much information is in a jet?, J. High Energy Phys. 06 (2017) 073.

[8] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow polynomials: A complete linear basis for jet substructure, J. High Energy Phys. 04 (2018) 013.

[9] L. G. Almeida, M. Backović, M. Cliche, S. J. Lee, and M. Perelstein, Playing tag with ANN: Boosted top identification with pattern recognition, J. High Energy Phys. 07 (2015) 086.

[10] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, Deep-learning Top Taggers or The End of QCD?, J. High Energy Phys. 05 (2017) 006.

[11] J. Pearkes, W. Fedorko, A. Lister, and C. Gay, Jet constituents for deep neural network based top quark tagging, arXiv:1704.02124.

[12] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, Deep-learned Top Tagging with a Lorentz Layer, arXiv:1707.08966.

[13] 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, arXiv:1711.09059.

[14] CMS Collaboration, Heavy flavor identification at CMS with deep neural networks, Report No. CMS-DP-2017-005, 2017.

[15] ATLAS Collaboration, Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment, Report No. ATL-PHYS-PUB-2017-003, 2017.

[16] A. M Sirunyan et al. (CMS Collaboration), Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV, J. Instrum. 13, P05011 (2018).

[17] P. T. Komiske, E. M. Metodiev, and M. D. Schwartz, Deep learning in color: towards automated quark/gluon jet discrimination, J. High Energy Phys. 01 (2017) 110.

[18] CMS Collaboration, New Developments for Jet Substructure Reconstruction in CMS, Report No. CMS-DP-2017-027, 2017.

[19] ATLAS Collaboration, Quark versus gluon jet tagging using jet images with the ATLAS detector, Report No. ATL-PHYS-PUB-2017-017, 2017.

[20] W. Bhimji, S. A. Farrell, T. Kurth, M. Paganini, Prabhat, and E. Racah, Deep Neural Networks for Physics Analysis on low-level whole-detector data at the LHC, in *Proceedings of 18th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2017) Seattle, Washington, USA, 2017.*

[21] P. T. Komiske, E. M. Metodiev, B. Nachman, and M. D. Schwartz, Pileup mitigation with machine learning (PUMML), J. High Energy Phys. 12 (2017) 051.

[22] L. de Oliveira, M. Paganini, and B. Nachman, Learning particle physics by example: Location-aware generative adversarial networks for physics synthesis, Comput. Softw. Big Sci. **1,** 4 (2017).

[23] M. Paganini, L. de Oliveira, and B. Nachman, Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multi-Layer Calorimeters, Phys. Rev. Lett. **120,** 042003 (2018).

[24] M. Paganini, L. de Oliveira, and B. Nachman, CaloGAN: Simulating 3D high energy particle showers in multi-layer electromagnetic calorimeters with generative adversarial networks, Phys. Rev. D **97,** 014021 (2018).

[25] G. Aad *et al.* (ATLAS Collaboration), 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).

[26] CMS Collaboration, Identification of b-quark jets with the CMS experiment, J. Instrum. **8,** P04013 (2013).

[27] CMS Collaboration, Performance of quark/gluon discrimination in 8 TeV pp data, Report No. CMS-PAS-JME-13-002, 2013.

[28] V. Khachatryan *et al.* (CMS Collaboration), Identification techniques for highly boosted W bosons that decay into hadrons, J. High Energy Phys. 12 (2014) 017.

[29] CMS Collaboration, Boosted top jet tagging at CMS, Report No. CMS-PAS-JME-13-007, 2014.

[30] G. Aad *et al.* (ATLAS Collaboration), Performance of b-Jet identification in the ATLAS experiment, JINST **11,** P04008 (2016).

[31] G. Aad *et al.* (ATLAS Collaboration), Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8$ TeV, Eur. Phys. J. C **76,** 154 (2016).

[32] G. Aad *et al.* (ATLAS Collaboration), Identification of high transverse momentum top quarks in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector, J. High Energy Phys. 06 (2016) 093.

[33] J. Hernández-González, I. Inza, and J. A. Lozano, Weak supervision and other non-standard classification problems: a taxonomy, Pattern Recogn. Lett. **69,** 49 (2016).

[34] L. M. Dery, B. Nachman, F. Rubbo, and A. Schwartzman, Weakly supervised classification in high energy physics, J. High Energy Phys. 05 (2017) 145.

[35] E. M. Metodiev, B. Nachman, and J. Thaler, Classification without labels: Learning from mixed samples in high energy physics, J. High Energy Phys. 10 (2017) 174.

[36] T. Cohen, M. Freytsis, and B. Ostdiek, (Machine) Learning to do more with less, J. High Energy Phys. 02 (2018) 034.

[37] ATLAS Collaboration, Discrimination of light quark and gluon jets in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector, Report No. ATLAS-CONF-2016-034, 2016.

[38] CMS Collaboration, Performance of quark/gluon discrimination in 8 TeV pp data, Report No. CMS-PAS-JME-13-002, 2013.

[39] CMS Collaboration, Performance of quark/gluon discrimination in 13 TeV data, Report No. CMS-DP-2016-070, 2016.

[40] CMS Collaboration, Jet algorithms performance in 13 TeV data, Report No. CMS-PAS-JME-16-003, 2017.

[41] ATLAS Collaboration, Quark versus gluon jet tagging using charged particle multiplicity with the ATLAS detector, Report No. ATL-PHYS-PUB-2017-009, 2017, https://cds.cern.ch/record/2263679.

[42] J. Gallicchio and M. D. Schwartz, Quark and gluon jet substructure, J. High Energy Phys. 04 (2013) 090.

[43] G. Aad *et al.* (ATLAS Collaboration), Measurement of the charged-particle multiplicity inside jets from $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector, Eur. Phys. J. C **76,** 322 (2016).

[44] P. Gras, S. Hche, D. Kar, A. Larkoski, L. Lnnblad, S. Pltzer, A. Sidmok, P. Skands, G. Soyez, and J. Thaler, Systematics of quark/gluon tagging, J. High Energy Phys. 07 (2017) 091.

[45] G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott, Classification with asymmetric label noise: Consistency and maximal denoising, Electron. J. Statist. **10,** 2780 (2016).

[46] K. Cranmer, J. Pavez, and G. Louppe, Approximating likelihood ratios with calibrated discriminative classifiers, arXiv:1506.02169.

[47] T. Sjöstrand, S. Mrenna, and P. Z. Skands, A brief introduction to PYTHIA 8.1, Comput. Phys. Commun. **178,** 852 (2008).

[48] M. Cacciari, G. P. Salam, and G. Soyez, The Anti-k(t) jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.

[49] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, Eur. Phys. J. C **72,** 1896 (2012).

[50] F. Chollet *et al.*, Keras, https://github.com/fchollet/keras.

[51] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, Tensorflow: A system for large-scale machine learning, in OSDI **16,** 265 (2016), https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

[52] D. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[53] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, Fast and accurate deep network learning by exponential linear units (elus), arXiv:1511.07289.

[54] J. Gallicchio and M. D. Schwartz, Pure samples of quark and gluon jets at the LHC, J. High Energy Phys. 10 (2011) 103.

[55] V. Nair and G. E Hinton, Rectified linear units improve restricted boltzmann machines, in *Proceedings of the 27th international conference on machine learning (ICML-10)* (Omnipress, Israel, 2010), p. 807.