Distilling Normalizing Flows for Real-Time Anomaly Detection at the LHC

Tara P. A. Tahseen[1,3*] **Noah Clarke Hall**[2*]
Nikolaos Konstantinidis[1] **Paula Martínez Suárez**[2,3]
[1]Department of Physics and Astronomy, University College London, WC1E 6BT, UK
[2]Experimental Physics, CERN, Espl. des Particules 1, Meyrin, 1217, Switzerland
[3]CERN openlab, CERN, Espl. des Particules 1, Meyrin, 1217, Switzerland
tara.tahseen.22@ucl.ac.uk
noah.clarke.hall@cern.ch

Abstract

Normalizing flows are principled anomaly detectors, selecting anomalies using a probabilistic per-event likelihood. Extreme latency and resource constraints have prevented the deployment of flow likelihoods within the hardware triggers at the Large Hadron Collider. We bypass these limitations by distilling the likelihood from a large normalizing flow into lightweight student estimators suitable for deployment on a field-programmable gate array. Our use of a conditional normalizing flow enables precise likelihood estimation in the presence of missing input features. Both decision tree and neural network students are considered, the latter optimized under advanced quantization techniques. By simultaneously improving likelihood quality and lowering inference cost, we demonstrate both state-of-the-art physics performance and latency compared with existing flow-based approaches.

1 Introduction

The Large Hadron Collider (LHC) produces collisions at 40 MHz, far more than can be stored. A hardware Level-1 Trigger (L1T) performs the first stage of data reduction, implemented on field-programmable gate arrays (FPGAs) with a latency budget of a few microseconds. Unsupervised anomaly detection offers an unbiased approach to event selection, selecting outliers in the absence of labelled examples. Deploying such a model at the L1T requires inference in $\mathcal{O}(100\text{ ns})$ using a small fraction of the FPGA resources [1].

The ATLAS [2] and CMS [3] experiments have deployed several unsupervised anomaly detection algorithms at the L1T level [4, 5, 6, 7, 8]. All are based on autoencoder neural networks, scoring events using an input-output reconstruction error or latent coordinates. Autoencoders based on decision trees [9] and spiking networks [10] have also been proposed for L1T deployment. Normalizing flows provide an attractive alternative, scoring events using the exact likelihood $p(x)$ [11] of the event occurring within the training set. However, exact computation of flow likelihoods on FPGAs is impractical, requiring the numerical integration of a neural vector field (continuous flows, [12]) or the evaluation of complex sequential transformations and associated Jacobians (discrete flows, here). A prior study by [12] addressed this by replacing the exact likelihood with a cheaper proxy such as the vector field itself. While L1T-compatible thanks to advanced quantization techniques, this approach was found to be less anomaly-efficient than the full likelihood.

In the work presented here, we seek to bypass the efficiency-cost tradeoff using knowledge distillation, regressing the outputs of a large, FPGA-incompatible "teacher" using a lightweight, FPGA-compatible

*These authors contributed equally to this work.

Preprint.

"student". This approach was originally proposed for anomaly detection at the LHC by Pol et al. [6], and implemented for autoencoders by [5, 13]. In the absence of FPGA-compatibility constraints, the teacher can be large and heterogeneous and leverage feature engineering in order to model the likelihood with maximal precision. Our use of a conditional normalizing flow enables precise likelihood estimation in the presence of missing input features, a challenge not addressed by [12]. Likelihood precision is closely aligned with anomaly detection, since incorrectly assigning high likelihoods to normal events will increase the rate of false positives, and vice versa. We regress the flow likelihood directly from detector inputs using both a boosted decision tree (BDT) and a dense neural network (DNN). The resulting FPGA implementations improve anomaly discrimination and lower latency compared with previous flow-based approaches.

2 Datasets

We use the datasets employed by Govorkova et al. [1], published on Zenodo [14, 15, 16, 17, 18], and discussed in detail in [19]. The dataset contains simulated and pre-filtered proton–proton collisions containing an electron or a muon with a transverse momentum $p_{\mathrm{T}} > 23$ GeV and pseudorapidity $|\eta| < 3$ (electron) and $|\eta| < 2.1$ (muon). This selection is representative of the L1T of a multipurpose LHC experiment. The Standard Model (SM) data sample represents a typical proton–proton collision dataset [14] and contains 13.5 million events. The SM dataset is split 40-10-50 into training, validation, and testing datasets. An additional four new-physics scenarios are benchmarked:

- A leptoquark with a mass of 80 GeV, decaying to a b quark and a τ lepton ($LQ \to b\tau$) [15],
- A neutral scalar boson with a mass of 50 GeV, decaying to two off-shell Z bosons, each forced to decay to two leptons ($A \to 4l$) [16],
- A scalar boson with a mass of 60 GeV, decaying to two τ leptons ($h^0 \to \tau\tau$) [18],
- A charged scalar boson with a mass of 60 GeV, decaying to a τ lepton and a neutrino ($h^+ \to \tau\nu$) [17].

Each event is represented by the $(p_{\mathrm{T}}, \eta, \phi)$ for 18 reconstructed particles (4 electrons, 4 muons, and 10 jets) alongside the magnitude and ϕ-coordinate of the missing transverse energy ($E_{\mathrm{T}}^{\mathrm{miss}}$). The $E_{\mathrm{T}}^{\mathrm{miss}}$ ϕ coordinate is subtracted from each ϕ, and the now-redundant $E_{\mathrm{T}}^{\mathrm{miss}}$ ϕ is removed. The momenta of missing particles are represented by imputing zeros. The resulting vector, x, of unnormalized particle-level features has 55 components.

3 Normalizing flow

Feature engineering is applied on x in order to make the density of SM events easier to learn. The decomposition $\phi \to (\sin\phi, \cos\phi)$ is taken in order to capture ϕ periodicity. The transverse momenta transform under $p_{\mathrm{T}} \to \log(1 + p_{\mathrm{T}})$ to suppress the variance introduced by high-p_{T} tails. The integer multiplicities of muons, n_μ, electrons, n_e and jets, n_j, are appended to the input vector. The engineered features consist of 73 normalized particle-level features (x') plus 3 multiplicities (m). All features are scaled to unit variance using scales measured on the SM training dataset, not including imputed zeros. Imputed zeros in x' are replaced by a small dequantization noise $\epsilon \sim \mathcal{N}(0, 10^{-2})$.

A key challenge for improving the learned likelihood is the treatment of missing data. Normalizing flows are not naturally suited to learning sparse representations, since they cannot directly model data on a lower-dimensional manifold of unknown dimension [20]. As a result, using a normalizing flow to model $p(x)$ or $p(x', m)$ directly is unlikely to produce precise likelihood estimates, especially if model capacity is limited by FPGA compatibility as in [12]. We resolve this by using a conditional normalizing flow [21] to model the conditional likelihood $p(x'|m)$. Conditioning on m guarantees the tractability of the modeling task by breaking down the target distribution into a mixture of low-dimensional continuous distributions. Since every possible combination of (n_μ, n_e, n_j) can be assigned an integer index, the prior $p(m)$ can be modeled using a categorical distribution. Combining the conditional flow and prior, we recover the joint log-likelihood

$$-\log p(x', m) = -\underbrace{\log p(x'|m)}_{\text{flow}} - \underbrace{\log p(m)}_{\text{prior}} \tag{1}$$

The flow uses the rational-quadratic spline bijectors proposed by Durkan et al. [22] and implemented in the `nflows` package [23]. The flow consists of 10 spline transformations. Each spline has 3 knots, with the knot locations and gradients controlled by an m-conditioned ResNet [24] with 2 residual blocks and 128 nodes per hidden layer. Each hidden layer uses the ReLU activation, with batch normalization and a dropout probability of 0.1. The flow is trained for 500 epochs with a batch size of 4096. The Adam optimizer is used with a learning rate of 10^{-4}. In total, the flow has 1.3 M trainable parameters. Classifying between flow samples and SM data using a BDT, we measure the cross-entropy to be 0.68, indicating that the learned density is almost indistinguishable from the SM distribution.

4 Knowledge distillation

Since $-\log p(x', m)$ is the result of both a normalizing flow and a prior, it is more challenging to define an FPGA-compatible score such as the vector field v_t used in Vaselli et al. [12]. Instead, we seek to transfer $-\log p(x', m)$ into an FPGA-compatible student $s(x)$ using knowledge distillation. Regressing $-\log p(x', m)$ directly from x incorporates the feature engineering into the student, avoiding the need to preprocess features on the FPGA. The students are trained by minimizing the weighted square error

$$\mathcal{L}_{\text{student}} = w \times ||s(x) + \log p(x', m)||^2, \tag{2}$$

where w is a normalized importance weight such that $w \propto 1/\sqrt{p(x', m)}$. Importance weighting improves the precision of the student at extreme anomaly scores, removing the need to supervise the student on signals, in contrast to “outlier exposure" distillation methods [5, 6].

Two student model architectures are evaluated: a boosted decision tree (BDT) and a dense neural network (DNN). The BDT is implemented using `xgboost` [25] with a learning rate of 0.3, a maximum ensemble size of 100 and a maximum tree depth of 3. Post-training quantization (PTQ) is performed using `conifer` [26] to compress the BDT in preparation for implementation on an FPGA. The minimum precision needed was found to be 18 bits for the inputs and tree nodes, and 14 bits for the leaves. The DNN consists of 2 hidden layers of 32 nodes with a ReLU activation. Input normalization is incorporated into the DNN by applying batch normalization before the first hidden layer. Quantization-aware training is performed using the HGQ method [27] to compress the DNN in preparation for FPGA implementation. The DNN is trained for 500 epochs with a batch size of 4096. The Adam optimizer is used with a learning rate of 10^{-4} and the gradient norm clipped to 1. The targets for the DNN are scaled to unit variance, which stabilizes training while preserving ranking in $-\log p(x', m)$. The BDT produces an (un)weighted R^2 score on the validation set of 0.86 (0.90), compared with 0.84 (0.95) for the DNN.

5 Physics performance

Model physics performance is evaluated using the four new-physics scenarios. We quantify anomaly detection performance by plotting the receiver operating characteristic (ROC) curve for each model, shown in Figure 1. We quote the area under the curve (AUC) as well as the true-positive rate (TPR) at a fixed false-positive rate (FPR) of 10^{-5}. The AUC measures the overall discrimination power of a given model, whereas the TPR measures discrimination power in the extreme tails of the SM score distribution. The AUC and TPR are compared to relevant results obtained by [12] in Table 1.

The $-\log p(x', m)$ from our conditional normalizing flow achieves the highest AUC for all new physics benchmarks. Both the BDT and DNN produce AUCs within $\leq 1\%$ of the teacher model. Crucially, the teacher and both students produce AUCs that exceed those obtained by [12]. Our flow has the highest TPR for $LQ \to b\tau$, $h^+ \to \tau\nu$ and $h^0 \to \tau\tau$ but the lowest TPR for $A \to 4l$. The TPRs from the BDT and DNN fluctuate about that of the teacher model, with the BDT producing consistently higher TPRs than the DNN. We note that measuring the TPR in the extreme tails of the score distribution makes the TPR highly sensitive to effects from finite statistics, feature engineering, random initialization, and hyperparameter choice.

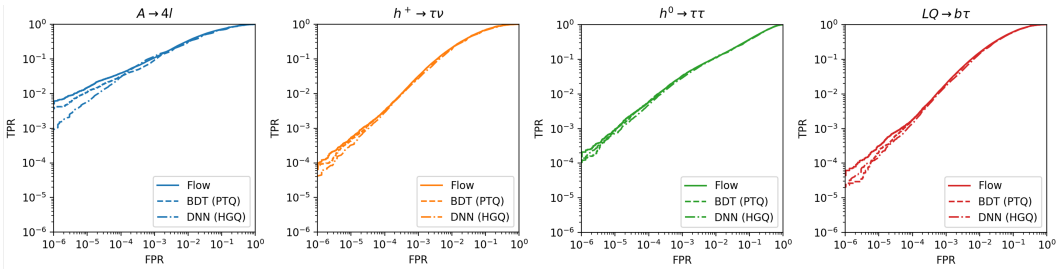

Figure 1: ROC curves for the four new-physics benchmarks, comparing the flow teacher score $-\log p(x', m)$ with its BDT and DNN student surrogates.

Model	AD score	TPR @ FPR 10^{-5} [%]				AUC [%]			
		$LQ \to b\tau$	$A \to 4l$	$h^+ \to \tau\nu$	$h^0 \to \tau\tau$	$LQ \to b\tau$	$A \to 4l$	$h^+ \to \tau\nu$	$h^0 \to \tau\tau$
Flow from [12]	ODE	0.04	3.8	0.04	0.05	80	88	86	69
	v_t	0.04	2.8	0.04	0.06	80	82	84	68
	v_t (HGQ)	0.04	3.4	0.05	0.06	77	86	82	66
Flow (ours)	$-\log p(x', m)$	0.05	1.5	0.05	0.10	88	90	90	76
	BDT (PTQ)	0.04	1.1	0.05	0.10	88	89	89	76
	DNN (HGQ)	0.04	0.58	0.03	0.07	88	89	89	76

Table 1: Comparison of AUC and TPR for each model, with comparison to prior work in [12].

6 FPGA implementation

The BDT and DNN are converted to firmware using the `conifer` and `da4ml` bridges respectively. We target a Xilinx Virtex UltraScale+ FPGA with a 200 MHz clock, using the same part number as Vaselli et al. [12] to ensure consistency. Resource and timing estimates, displayed in Table 2, are obtained via RTL synthesis in Vitis HLS. The DNN student performs similarly to the HGQ implementation of v_t, demonstrating that the large size of the teacher does not compromise the FPGA compatibility of the student. The BDT student can be evaluated in just a single clock cycle thanks to the unrolling of the decision tree structure [26]. The latency of both the BDT and DNN students is less than that demonstrated previously by both autoencoder [1] and flow-based [12] approaches. All the resource usages quoted represent $< 1\%$ of the resources of the whole FPGA.

Model	DSP	LUT	FF	Latency [ns]	II [clk]
v_t (HGQ) from [12]	28	5,798	1,683	35	1
BDT (PTQ)	0	12,869	260	5	1
DNN (HGQ)	0	6,633	1,798	25	1

Table 2: Resource and timing estimates for a Xilinx Virtex UltraScale+ FPGA with a 200 MHz clock.

7 Conclusion

We demonstrate that knowledge distillation enables large conditional normalizing flows to serve as high-performance anomaly detectors while remaining deployable under strict hardware constraints. By training lightweight surrogate models on the flow likelihood, we transfer its discrimination power into FPGA-compatible boosted decision tree and neural network students. The conditional flow improves likelihood modeling in the presence of missing features and achieves strong physics performance, reaching AUCs of up to 90% on benchmark signals. The students reproduce the teacher AUC to within 1% while achieving nanosecond-scale inference latency (5–25 ns) on FPGA hardware. This approach delivers state-of-the-art likelihood-based anomaly detection with a practical path to real-time deployment within LHC hardware triggers.

Acknowledgements

We gratefully acknowledge the support of the UK's Science and Technology Facilities Council (STFC). TT is supported by the STFC UCL Centre for Doctoral Training in Data Intensive Science (ST/W00674X/1) and funding from CERN openlab during a 6-month work placement at CERN. The support and collaboration of CERN openlab are gratefully acknowledged. TT thanks R. P. Nathan for helpful and interesting discussions around anomaly detection. This work has been partially funded by the Eric & Wendy Schmidt Fund for Strategic Innovation through the CERN Next Generation Triggers project under grant agreement number SIF-2023-004.

References

[1] Ekaterina Govorkova, Ema Puljak, Thea Aarrestad, Thomas James, Vladimir Loncar, Maurizio Pierini, Adrian Alan Pol, Nicolò Ghielmetti, Maksymilian Graczyk, Sioni Summers, Jennifer Ngadiuba, Thong Q. Nguyen, Javier Duarte, and Zhenbin Wu. Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider. *Nature Machine Intelligence*, 4(2):154–161, February 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00441-3. URL `http://dx.doi.org/10.1038/s42256-022-00441-3`.

[2] The ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3(08):S08003, Aug 2008. doi: 10.1088/1748-0221/3/08/S08003. URL `https://doi.org/10.1088/1748-0221/3/08/S08003`.

[3] CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08004, 2008. doi: 10.1088/1748-0221/3/08/S08004. URL `https://doi.org/10.1088/1748-0221/3/08/S08004`.

[4] Abhijith Gandrakota. Realtime Anomaly Detection at the L1 Trigger of CMS Experiment. In *Proceedings of 42nd International Conference on High Energy Physics — PoS(ICHEP2024)*, volume 476, page 1025, 2025. doi: 10.22323/1.476.1025.

[5] CMS Collaboration. Model-Independent Real-Time Anomaly Detection at the CMS Level-1 Calorimeter Trigger with CICADA. Technical report, CMS Collaboration, 2024. URL `https://cds.cern.ch/record/2917884`.

[6] Adrian Alan Pol, Ekaterina Govorkova, Sonja Gronroos, Nadezda Chernyavskaya, Philip Harris, Maurizio Pierini, Isobel Ojalvo, and Peter Elmer. Knowledge Distillation for Anomaly Detection. In *ESANN 2023 proceedings*, 2023.

[7] Kaito Sugizaki and ATLAS Collaboration. GELATO: A Generic Event-Level Anomalous Trigger Option for ATLAS in LHC Run 3. Poster presented at Lepton Photon 2025, 2025. URL `https://cds.cern.ch/record/2941530`.

[8] Paula Martinez Suarez. ARTEMIS: exploring anomaly detection in the ATLAS Level-1 Trigger. Poster presented at LHCP2026, 2026. URL `https://cds.cern.ch/record/2963506`.

[9] S. T. Roche, Q. Bayer, B. T. Carlson, W. C. Ouligian, P. Serhiayenka, J. Stelzer, and T. M. Hong. Nanosecond anomaly detection with decision trees and real-time application to exotic Higgs decays. *Nature Communications*, 15(1):3527, 2024. doi: 10.1038/s41467-024-47704-8. URL `https://doi.org/10.1038/s41467-024-47704-8`.

[10] Barry M Dillon, Jim Harkin, and Aqib Javed. Anomaly detection with spiking neural networks for LHC physics. *Machine Learning: Science and Technology*, 7(2):025019, Mar 2026. doi: 10.1088/2632-2153/ae4a85. URL `https://doi.org/10.1088/2632-2153/ae4a85`.

[11] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. URL `http://jmlr.org/papers/v22/19-1028.html`.

[12] Francesco Vaselli, Chang Sun, Thea Aarrestad, Dimitrios Danopoulos, Roope Oskari Niemi, Maciej Mikolaj Glowacki, Katya Govorkova, Vladimir Loncar, Felice Pantaleo, and Maurizio Pierini. It's not a FAD: first demonstration of flows for unsupervised anomaly detection at 40 MHz for use at the Large Hadron Collider. *Machine Learning: Science and Technology*, 7(2):025052, 4 2026. doi: 10.1088/2632-2153/ae51dd. URL `https://doi.org/10.1088/2632-2153/ae51dd`.

[13] Rajat Gupta and ATLAS Collaboration. NomAD: Low-Latency Unsupervised Anomaly Detection for the ATLAS Trigger. Poster presented at Fast Machine Learning for Science 2025, 2025. URL https://cds.cern.ch/record/2942542.

[14] Thea Aarrestad, Ekaterina Govorkova, Jennifer Ngadiuba, Ema Puljak, Maurizio Pierini, and Kinga Anna Wozniak. Unsupervised New Physics detection at 40 MHz: Training Dataset , June 2021. URL https://doi.org/10.5281/zenodo.5046389.

[15] Thea Aarrestad, Ekaterina Govorkova, Jennifer Ngadiuba, Ema Puljak, Maurizio Pierini, and Kinga Anna Wozniak. Unsupervised New Physics detection at 40 MHz: $LQ \to b\tau$ Signal Benchmark Dataset , July 2021. URL https://doi.org/10.5281/zenodo.5055454.

[16] Thea Aarrestad, Ekaterina Govorkova, Jennifer Ngadiuba, Ema Puljak, Maurizio Pierini, and Kinga Anna Wozniak. Unsupervised New Physics detection at 40 MHz: $A \to 4$ leptons Signal Benchmark Dataset , June 2021. URL https://doi.org/10.5281/zenodo.5046446.

[17] Thea Aarrestad, Ekaterina Govorkova, Jennifer Ngadiuba, Ema Puljak, Maurizio Pierini, and Kinga Anna Wozniak. Unsupervised New Physics detection at 40 MHz: $h^+ \to \tau\nu$ Signal Benchmark Dataset , July 2021. URL https://doi.org/10.5281/zenodo.5061688.

[18] Thea Aarrestad, Ekaterina Govorkova, Jennifer Ngadiuba, Ema Puljak, Maurizio Pierini, and Kinga Anna Wozniak. Unsupervised New Physics detection at 40 MHz: $h^0 \to \tau\tau$ Signal Benchmark Dataset , July 2021. URL https://doi.org/10.5281/zenodo.5061633.

[19] Ekaterina Govorkova, Ema Puljak, Thea Aarrestad, Maurizio Pierini, Kinga Anna Woźniak, and Jennifer Ngadiuba. LHC physics dataset for unsupervised New Physics detection at 40 MHz. *Scientific Data*, 9(1):118, March 2022. doi: 10.1038/s41597-022-01187-8. URL https://doi.org/10.1038/s41597-022-01187-8.

[20] Brendan Ross and Jesse Cresswell. Tractable density estimation on learned manifolds with conformal embedding flows. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26635–26648. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/dfd786998e082758be12670d856df755-Paper.pdf.

[21] Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019. doi: 10.1561/2200000056.

[22] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019. URL https://arxiv.org/abs/1906.04032.

[23] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. nflows: normalizing flows in PyTorch, November 2020. URL https://doi.org/10.5281/zenodo.4296287.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2015. URL https://arxiv.org/abs/1512.03385.

[25] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, Aug 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

[26] S. Summers, G. Di Guglielmo, J. Duarte, P. Harris, D. Hoang, S. Jindariani, E. Kreinar, V. Loncar, J. Ngadiuba, M. Pierini, D. Rankin, N. Tran, and Z. Wu. Fast inference of Boosted Decision Trees in FPGAs for particle physics. *JINST*, 15(05):P05026, May 2020. doi: 10.1088/1748-0221/15/05/P05026. URL https://dx.doi.org/10.1088/1748-0221/15/05/P05026.

[27] Chang Sun, Zhiqiang Que, Thea Årrestad, Vladimir Loncar, Jennifer Ngadiuba, Wayne Luk, and Maria Spiropulu. HGQ. *ACM DL*, FPGA '26:79–91, 2026. doi: 10.1145/3748173.3779200. URL https://cds.cern.ch/record/2951900.