Towards FPGA-Deployed Anomaly Detection for Real-Time Neutrino Data Selection in LArTPCs

Seokju Chung*
Department of Physics
Columbia University
New York, NY 10027

Georgia Karagiorgi
Department of Physics
Columbia University
New York, NY 10027

Abstract

Raw wire readout data from Liquid Argon Time Projection Chambers (LArTPCs) in future experiments such as the Deep Underground Neutrino Experiment will vastly exceed available storage bandwidth, necessitating intelligent online data selection. We present a compressed autoencoder architecture for model-independent anomaly detection operating directly on LArTPC wire waveforms, benchmarked for deployment on resource-efficient hardware including field-programmable gate arrays (FPGAs). Drawing on anomaly trigger strategies developed at the Large Hadron Collider, we adapt these methods to LArTPC ionization readouts and assess physics performance across diverse signal categories. Model compression and quantization strategies are evaluated for compatibility with FPGA resource constraints. We report end-to-end comparisons of latency, throughput, and logic utilization across FPGA, GPU, and CPU backends, providing a concrete pathway toward online anomaly-driven triggering within LArTPC data acquisition (DAQ) systems.

1 Introduction

Next-generation neutrino detectors generate data volumes that far exceed the storage and processing capacities available for offline analysis. This challenge is acute for the Deep Underground Neutrino Experiment (DUNE) Abi et al. [2020], which aims to operate continuously for over a decade to search for rare phenomena including supernova neutrino bursts and proton decay. Such experiments require intelligent, high-efficiency data acquisition (DAQ) systems capable of filtering relevant activity in real time and with minimal latency Bartoldus et al. [2022]. The task is made harder when the signals of interest are rare and not precisely modeled, which makes conventional, model-specific triggers brittle.

Model-independent anomaly detection offers an attractive alternative. An autoencoder trained only on background data learns to reconstruct nominal activity well; inputs that reconstruct poorly are flagged as anomalous, regardless of whether they match any particular signal hypothesis. This strategy has been pioneered for real-time triggering at the Large Hadron Collider (LHC), most notably by the CMS CICADA algorithm Gerlach et al. [2025], The CMS Collaboration [2024], which performs anomaly detection on calorimeter images within the Level-1 trigger latency budget on field-programmable gate arrays (FPGAs). We adapt this paradigm to Liquid Argon Time Projection Chamber (LArTPC) ionization readouts, where the input images are considerably larger and the detector physics is markedly different.

A central obstacle to real-time deployment is that high-capacity anomaly-detection networks are too large for the resource-limited hardware found in DAQ systems. We address this with knowledge

*Corresponding author: seokju.chung@columbia.edu.

Submitted to 7th Fast Machine Learning for Science Conference (FastML 2026). Do not distribute.

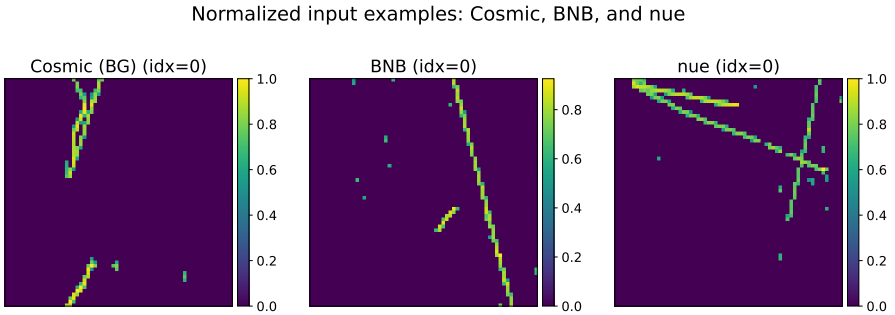


Figure 1: Example normalized 64×64 inputs after preprocessing for cosmic, BNB, and ν_e CC interactions.

distillation (KD) Hinton et al. [2015], Pol et al. [2023], compressing a large unsupervised "Teacher" autoencoder into a lightweight "Student" that regresses the Teacher's reconstruction error. The Student is quantized and synthesized for an FPGA with `hls4ml` FastML Team [2025], Duarte et al. [2018], Aarrestad et al. [2021]. This extended abstract is based on our broader study of this end-to-end framework, and builds on our earlier demonstration of unsupervised anomaly detection in LArTPCs Chung et al. [2026], whose distilled network was too large for practical FPGA deployment. Here we focus on the FPGA-deployable regime and on quantifying end-to-end cost across FPGA, GPU, and CPU backends.

2 Framework

Input data and preprocessing. We use the publicly available MicroBooNE open dataset MicroBooNE Collaboration [2023], Acciarri et al. [2017], consisting of real cosmic-ray interactions recorded in the detector with simulated Booster Neutrino Beam (BNB) interactions overlaid. We use only the collection plane, where a single event spans 6400 time ticks and 3456 wire channels, with one ADC value per pixel. Following the SparsePixels approach Tsoi et al. [2025], cosmic interactions serve as background and charged-current electron-neutrino (ν_e CC) interactions as the target signal. For each event a 256-tick $\times$ 512-channel window is selected around the interaction, along with five randomly chosen cosmic windows of the same size; each window is resized to a 64×64 image to reduce network size. An ADC-sum threshold removes low-energy depositions, and min-max normalization is applied. Representative inputs are shown in Figure 1.

Autoencoder and knowledge distillation. The Teacher is a convolutional autoencoder (four encoder layers, a convolutional bottleneck, and a mirrored decoder) trained on cosmic interactions to minimize the mean squared error (MSE) between input and output. The Student is a compact convolutional network (three convolutional layers followed by pooling and dense layers) that outputs a single anomaly score and is trained to regress the Teacher MSE. Following Pol et al. [2023] we combine co-learning with outlier exposure: a small fraction of ν_e CC interactions is injected during training and penalized when reconstructed too well, sharpening the decision boundary without committing the network to a single signal model. The Teacher (59,375 parameters) is compressed to a Student of only 897 parameters, with the Student output tracking the Teacher MSE closely (Figure 2).

FPGA deployment. The Student is built with QKeras for quantization-aware training and converted to high-level synthesis (HLS) with `hls4ml` using the `VitisAccelerator` backend, targeting a Xilinx Alveo U250. Per-layer fixed-point precision is tuned so that the synthesized and FPGA-deployed Student reproduce the Python Student output with negligible loss (Figure 2).

3 Physics performance

We evaluate discrimination with the efficiency–mis-identification (mis-id) curve, which is more natural than the receiver operating characteristic in high-energy-physics analyses where the signal-to-background ratio carries direct physical meaning. Efficiency is the fraction of ν_e CC interactions

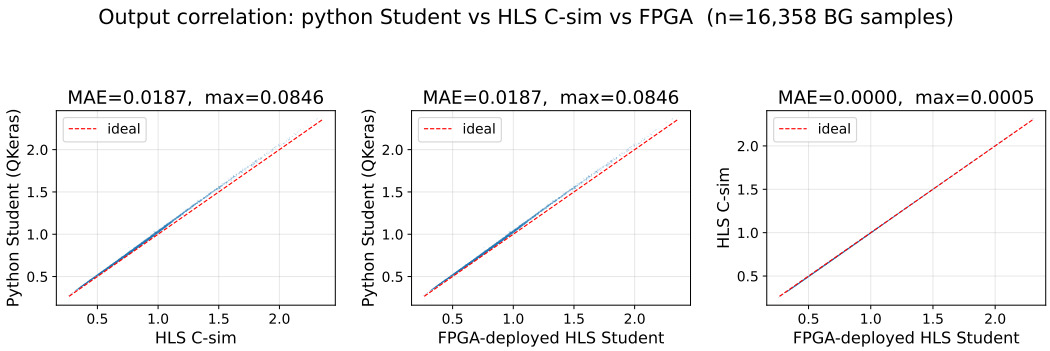
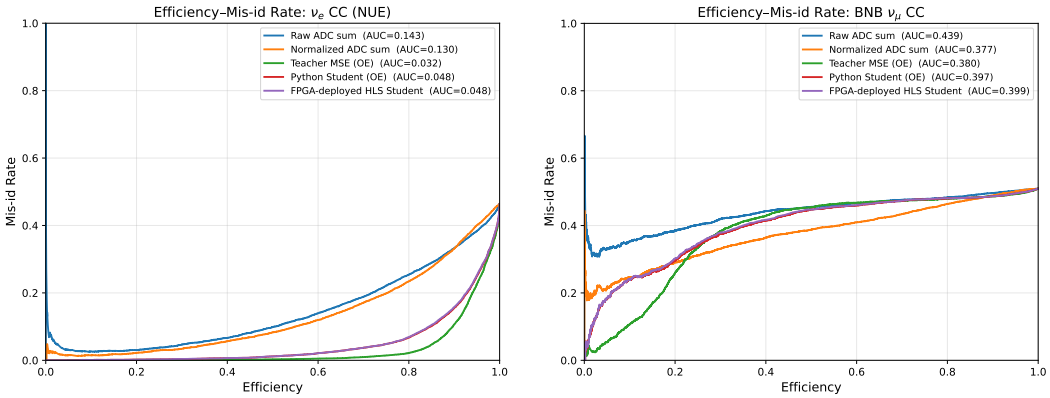


Figure 2: Anomaly score of the Python Student versus the synthesized and FPGA-deployed HLS Student. The synthesized Student reproduces the Python Student output, confirming that pre-quantization preserves discrimination power.



(a) ν_e CC interactions as the target signal.

(b) BNB interactions as the target signal.

Figure 3: Efficiency–mis-id rate for (a) ν_e CC and (b) BNB selection, with cosmic interactions as background. For ν_e CC the quantized FPGA-deployed Student (897 parameters) tracks the full Teacher (59,375 parameters) and far exceeds the ADC-sum baselines; performance degrades for the muon-like BNB sample, which is topologically similar to the cosmic training data.

selected; mis-id rate is the fraction of selected interactions that are background (one minus purity). As a single summary metric we report the area under this curve (AUC), where 0 indicates perfect separation and a value near the background fraction (≈ 0.5 for our near-balanced samples) corresponds to random selection. The evaluation uses 16,358 cosmic interactions as background and 18,903 ν_e CC interactions as signal.

Figure 3a compares the Teacher, the Python Student, and the FPGA-deployed HLS Student against raw and normalized ADC-sum baselines. The learned models substantially outperform the simple ADC-sum baselines, and crucially the quantized FPGA Student retains essentially all of the Teacher's discrimination power despite a $66\times$ reduction in parameters. Performance is considerably weaker when BNB interactions are taken as the target signal (Figure 3b), as expected: both the cosmic training sample and the generic BNB sample are dominated by muon-like topologies, so they are not anomalous relative to one another. This confirms that the network responds to genuine topological novelty rather than to beam timing or sample provenance.

4 Resource utilization and latency

Table 1 summarizes end-to-end cost across backends. GPU measurements use an NVIDIA A100; CPU measurements use the `hls4ml` C-simulation on a single Intel Xeon Gold 6226 core; FPGA measurements are end-to-end wall-clock values (including PCIe transfer) for a single compute

Table 1: End-to-end cost across backends for the anomaly-detection network. The quantized Student is small enough that a single FPGA CU achieves microsecond-scale compute latency at negligible incremental power, with ample room for parallel replication. GPU latencies are per-input at batch size 1 (sequential) and 512 (bulk); FPGA/CPU values are single-unit. Parameter counts: Teacher 59,375, Student 897.

Model	Platform	Latency [µs]		Power Δ [W]
		batch $= 1$	batch $= 512$	
Teacher	GPU (A100)	$10{,}274 \pm 1{,}471$	28.8	$+3$
Python Student	GPU (A100)	$27{,}020 \pm 6{,}398$	50.6	$+3$
HLS Student	CPU (Xeon 6226, C-sim)	$1{,}874 \pm 231$	–	$+34$
HLS Student	FPGA (Alveo U250)	53.49	–	< 1
HLS Student FPGA utilization		DSP 4%, LUT 2%, FF/BRAM $< 3\%$ (total device)		

unit (CU) on the Alveo U250 at 500 MHz. These numbers span different machines and include per-inference overhead, so they are indicative rather than strictly comparable. The pure-compute HLS estimate for the Student is 24.73 µs (12,365 cycles at 500 MHz); the remaining ≈ 29 µs of the measured 53.49 µs end-to-end FPGA latency is host-to-device data-movement overhead, dominated by the PCIe round trip that transfers the input image to the accelerator and reads the score back. This overhead is an artifact of the host-driven benchmark rather than the algorithm, and a streaming DAQ integration that feeds the kernel directly would largely eliminate it.

The synthesized Student is extremely lightweight, using 4% of total DSPs (16% per super logic region, SLR) and $< 3\%$ of other resources on the U250. This small footprint implies substantial headroom for parallelism: up to 6 CUs fit per SLR and up to 24 across the four SLRs, which would bring the pure-compute throughput to ≈ 1 µs per sample. The FPGA board power is essentially unchanged between idle and inference ($\Delta < 1\,W$), reflecting how little of the device the Student occupies.

5 Conclusion and outlook

We have demonstrated an end-to-end, model-independent anomaly-detection framework for LArTPCs that distills a large autoencoder into a quantized network deployable on an FPGA. Trained only on cosmic backgrounds, it cleanly separates ν_e CC interactions, and the FPGA-deployed Student preserves the Teacher's physics performance at microsecond-scale compute latency, a tiny logic footprint, and negligible incremental power. The end-to-end comparison across FPGA, GPU, and CPU backends quantifies the cost of online anomaly-driven selection and shows substantial headroom for parallel deployment within DAQ resource budgets. Future work will integrate the framework into a streaming readout path to remove PCIe overhead and apply it to data from running detectors such as the ICEBERG demonstrator and the Short-Baseline Near Detector, as a step toward online anomaly-driven triggering for DUNE.

Acknowledgments

This work was supported by the National Science Foundation under OAC-2209917. The authors thank Giuseppe Cerati for valuable help in utilizing the MicroBooNE Open Dataset. Large language models were used solely to proofread and improve the wording of this manuscript; all technical content, analysis, and results are the authors' own.

References

Thea Aarrestad et al. Fast convolutional neural networks on FPGAs with hls4ml. *Mach. Learn. Sci. Tech.*, 2(4): 045015, 2021. doi: 10.1088/2632-2153/ac0ea1.

B. Abi et al. Volume I. Introduction to DUNE. *Journal of Instrumentation*, 15(08):T08008, aug 2020. doi: 10.1088/1748-0221/15/08/T08008. URL `https://dx.doi.org/10.1088/1748-0221/15/08/T08008`.

R. Acciarri et al. Design and construction of the MicroBooNE detector. *Journal of Instrumentation*, 12(02): P02017–P02017, February 2017. ISSN 1748-0221. doi: 10.1088/1748-0221/12/02/p02017. URL https://iopscience.iop.org/article/10.1088/1748-0221/12/02/P02017. Publisher: IOP Publishing.

Rainer Bartoldus, Catrin Bernius, and David W. Miller. Innovations in trigger and data acquisition systems for next-generation physics facilities. 2022. URL https://arxiv.org/abs/2203.07620.

S. Chung, J. Cleeve, A. Malige, G. Karagiorgi, L. Gerlach, A.A. Pol, and I. Ojalvo. Real-time anomaly detection for liquid argon time projection chambers. *Journal of Instrumentation*, 21(06):P06030, jun 2026. doi: 10.1088/1748-0221/21/06/P06030. URL https://doi.org/10.1088/1748-0221/21/06/P06030.

Javier Duarte et al. Fast inference of deep neural networks in FPGAs for particle physics. *JINST*, 13(07):P07027, 2018. doi: 10.1088/1748-0221/13/07/P07027.

FastML Team. fastmachinelearning/hls4ml, 2025. URL https://github.com/fastmachinelearning/hls4ml.

Lino Gerlach, Elliott Kauffman, and Abhishikth Mallampalli. Evaluation of novel fast machine learning algorithms for knowledge-distillation-based anomaly detection at cms, 2025. URL https://arxiv.org/abs/2510.15672.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2015. URL https://arxiv.org/abs/1503.02531.

MicroBooNE Collaboration. Microboone open samples. https://github.com/uboone/OpenSamples, 2023.

Adrian Alan Pol, Ekaterina Govorkova, Sonja Gronroos, Nadezda Chernyavskaya, Philip Harris, Maurizio Pierini, Isobel Ojalvo, and Peter Elmer. Knowledge distillation for anomaly detection. In *31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 141–146, 2023. doi: 10.14428/esann/2023.ES2023-159. URL https://doi.org/10.14428/esann/2023.ES2023-159.

The CMS Collaboration. Model-Independent Real-Time Anomaly Detection at the CMS Level-1 Calorimeter Trigger with CICADA. CMS Detector Performance Summary CMS-DP-2024-121, 2024. URL https://cds.cern.ch/record/2917884.

Ho Fung Tsoi, Dylan Rankin, Vladimir Loncar, and Philip Harris. Sparsepixels: Efficient convolution for sparse data on fpgas, 2025. URL https://arxiv.org/abs/2512.06208.