
Edge Machine Learning for Cluster Counting in Next-Generation Drift Chambers

Deniz Yilmaz

Department of Physics
Stanford University
450 Jane Stanford Way, Stanford, CA 94305, United States of America
dyilmaz@stanford.edu

Liangyu Wu

Department of Physics
Stanford University
450 Jane Stanford Way, Stanford, CA 94305, United States of America
liangyu.wu@stanford.edu

Julia Gonski

SLAC National Accelerator Laboratory
2575 Sand Hill Road, Menlo Park, CA 94025, United States of America
jgonski@slac.stanford.edu

Dylan Rankin

University of Pennsylvania
209 S 33rd St, Philadelphia, PA 19104, United States of America
dsrankin@sas.upenn.edu

Christian Herwig

University of Michigan
450 Church St, Ann Arbor, MI 48109, United States of America
herwig@umich.edu

Abstract

Drift chambers have long been central to collider tracking, but future machines like a Higgs factory motivate higher granularity and cluster counting for particle ID, posing new data processing challenges. Machine learning (ML) at the "edge", or in cell-level readout, can dramatically reduce the off-detector data rate for high-granularity drift chambers by performing cluster counting at-source. We present machine learning algorithms for cluster counting in real-time readout of future drift chambers. These algorithms outperform traditional derivative-based techniques based on achievable pion-kaon separation. When synthesized to FPGA resources, they can achieve latencies consistent with real-time operation in a future Higgs factory scenario, thus advancing both R&D for future collider detectors as well as hardware-based ML for edge applications in high energy physics.

1 Introduction

Drift chambers are gaseous detectors that track charged particles with high precision and aid in particle identification (PID). As particles pass through the gas, primary ionization electrons drift under an electric field toward sense wires, where arrival times reveal the particle’s position; secondary electrons can form through further ionization. With fine spatial resolution, wide acceptance, and low material budget, drift chambers have long been central to collider experiments, including Belle at KEK [1] and BaBar at SLAC [2], among others.

Recent strategic planning processes conducted by the high energy physics community have pointed to the importance of a future e^+e^- Higgs factory in achieving prioritized future physics goals of the field [3]. The low mass and high precision of drift chambers make them attractive for future e^+e^- trackers, but adapting them to unprecedented luminosities poses novel challenges. High occupancies complicate signal separation, while long drift times limit performance at rapid bunch crossings. In addition, stringent demands on tracking and PID performance call for high-granularity designs that generate unprecedented off-detector data rates (towards TB/s) [4].

The introduction of machine learning (ML) methods to drift chamber data analysis can help address these challenges and enable performant and realizable next-generation drift chambers. Specifically, this work focuses on ML-assisted *cluster counting* (dN/dx) for PID [5]. Traditionally, drift chambers have done PID using the average energy loss from ionization over length (dE/dx) [6], which suffers from large Landau fluctuations due to secondary electrons, thereby limiting the ultimate resolution for particle separation. dN/dx alternatively quantifies the number of primary ionization clusters per unit length, which follows a Poisson-like distribution, thus improving statistical resolution for more accurate PID measurements. Implementing dN/dx in practice is difficult (especially at high track densities), as it requires fine-grained time and spatial resolution to distinguish closely spaced individual clusters, readout of single-ionization electron signal, and a gas mixture with favorable ionization and drift properties. Previous works have studied both the feasibility of dN/dx at future e^+e^- colliders [7] and the ability of ML to improve dN/dx using drift chamber simulation [8, 9].

This work is the first to demonstrate ML methods that aim to simultaneously address the data rate and dN/dx challenges of future drift chamber designs, by performing cluster counting at the “edge”. This feature extraction enables a dramatic reduction of the off-detector data rate by transmitting only the cluster count per waveform, while enhancing PID with respect to traditional methods. While a valuable proof-of-concept, this approach is limited by its use of a toy detector design, lack of precise data rate and power estimates needed for a full readout design, and full reliance on simulation, all of which would need to be addressed in subsequent works to fully realize and deploy edge ML in a future experiment.

2 Methodology

This study utilizes the dataset from Ref. [9], generated through Garfield++ [10] simulations based on the fourth conceptual detector design for the Circular Electron Positron Collider (CEPC) [11]. The simulation employs a 1.5 GHz sampling rate with square drift cells of 18 mm \times 18 mm. Each drift cell features a central sense wire surrounded by eight field wires in a square geometry. The detector operates with a gas mixture of 90% He and 10% iC_4H_{10} to achieve optimal primary ionization density. The training dataset comprises 5×10^5 π^\pm events with momenta ranging from 0.2 to 20.0 GeV/c, along with pion and kaon samples with 10^5 events for each of seven momenta ranging from [5, 20] GeV at 2.5 GeV intervals for evaluating pion-kaon separation performance.

Each waveform in this study is truncated to the first 500 samples (corresponding to a maximum drift time of approximately 350 ns) to emulate the reduction of the CEPC cell size expected for future drift chamber designs such as the IDEA detector [12] at the Future Circular Collider (FCC) [13]. Figure 1 provides an example waveform used in the model development, with background, primary ionization, and secondary ionization truth labels provided.

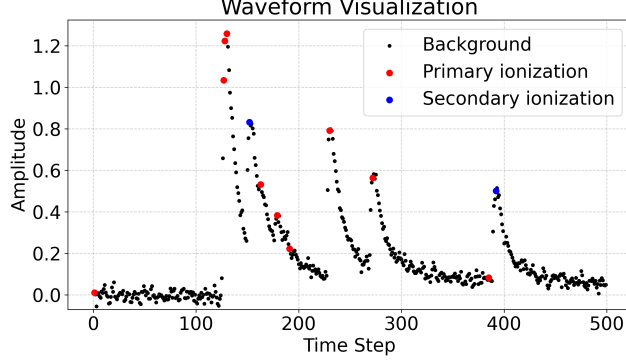


Figure 1: Example waveform from Ref. [9], including primary and secondary electron truth labels, truncated to 500 samples.

2.1 Traditional cluster counting

Traditional cluster counting approaches follow a two-step process: first applying peak finding algorithms to identify all candidate electron peaks, then merging adjacent peaks into single primary ionization clusters based on gas mixture properties and electron diffusion characteristics.

For peak-finding, the second-derivative-based (D2) algorithm is used in this work, which applies a second-order derivative to the raw waveform to identify signal peaks while maintaining a balance between detection efficiency and noise sensitivity. A baseline threshold T_1 (approximately three times the RMS noise amplitude) excludes background noise, while a second threshold T_2 is applied to the integrated second derivative for peak detection. Only positive derivatives corresponding to the rising edge are retained, as steep rising edges are characteristic features of ionization signals.

Following peak identification, two clustering algorithms are considered to merge electron peaks into primary ionization clusters. The Fixed-Clusterization Algorithm (FCA) uses fixed parameters across the entire waveform, including a time window size for forward-looking peak search, a normalization factor based on gas mixture cluster size characteristics, and a maximum number of peaks per window to limit noise and secondary ionization effects. The Adapted-Clusterization Algorithm (ACA) extends FCA by pre-dividing the waveform into time segments based on drift electron collection statistics, with each segment employing optimized parameter sets to adapt to time-dependent ionization physics such as cluster size variations with diffusion distance.

2.2 ML-based direct cluster counting

State-of-the-art ML models outperform classical methods for PID with cluster counting but often use computationally costly complex architectures, which are impractical under strict power and latency constraints. This work therefore focuses on fully connected (dense) neural networks, which are sufficiently small that no specific computing resources (eg. GPUs) are needed for model development. The ML-based approach removes the need to do peak finding and cluster counting as separate steps, and instead does a direct regression of the number of clusters (eg. number of primary ionizations) for each track in a supervised way using the true number of clusters from Garfield++ as a label. The baseline model is a deep neural network (DNN) consisting of four layers following the 500 dimensional input: 8 units in the first, 32 in the second, 8 in the third, and a single output for cluster count regression. All dense layers use ReLU activation, with the output layer providing direct regression values.

The baseline model was trained for 75 epochs using the Adam optimizer with a learning rate of 0.0006. 400,000 events were used for training, 100,000 events were used for validation, and 100,000 events were used for testing. To calculate pion-kaon separation, the model trained over exclusively pion events was used to predict the number of clusters for both kaon and pion events, using 50,000 events for each momentum value. For each momentum point, we obtained distribution histograms for both particle types, extracted their respective means and standard deviations, and calculated the separation power S scaled to a two meter track length.

To study real-time implementation, two compressed model variants were developed: pruned, which gradually removed 60% of low-weight nodes over 25 epochs with minimal performance loss, and additionally a model with quantization-aware training via QKeras [14], which retrained the pruned model using 10-bit fixed-point precision (with 5 fractional bits).

3 Results

Figure 2 shows the capability of the baseline ML model to determine the predicted number of clusters for the test set, along with the correlation between predicted and truth cluster counts for each event, indicating good capability of the model to learn the cluster count for a given track.

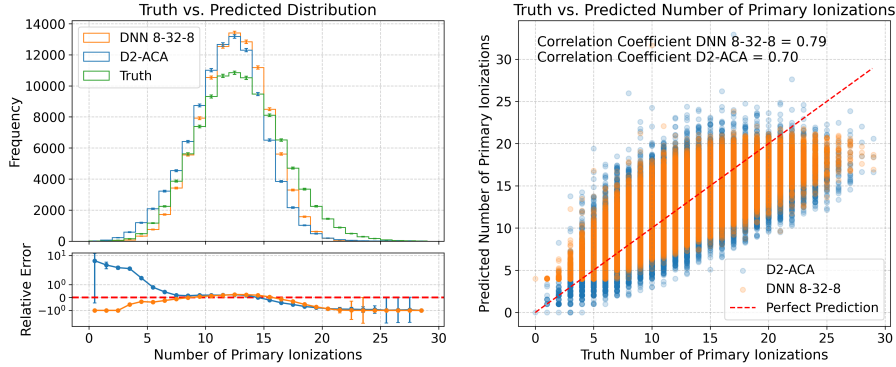


Figure 2: Performance of the baseline ML model for the task of predicting the truth number of clusters, comparing the predicted vs. truth cluster number distribution across the test set (left) and the correlation between the predicted and truth cluster number per test set event (right).

PID performance is evaluated using the pion-kaon separation power defined as: $S = \frac{|\frac{dN}{dx}\pi - \frac{dN}{dx}K|}{\frac{\sigma_\pi + \sigma_K}{2}}$ (taken from Ref. [9]), where $\frac{dN}{dx}\pi(K)$ and $\sigma_\pi(K)$ are the mean and standard deviation of primary ionizations per unit length for pions (kaons). The truncated dataset has an average of 12.5 clusters per waveform (in good agreement with the expected cluster density for the gas [15] and the Garfield++ simulation), corresponding to a mean track length per cell of approximately 9.3 mm. To project the separation power for realistic track lengths, results are scaled from a single cell assuming a 2 m average particle trajectory. Figure 3 shows the obtained pion-kaon separation power over the test sets with track momentum range of [5, 20] GeV. A separation of $> 3\sigma$ is achieved across the full tested track momentum range for an assumed 2 m track length.

Trained models were synthesized using hls4ml [16, 17] for FPGA implementation. This approach imagines a future target of embedded FPGAs for reconfigurable digital logic in a chip, which have been previously studied for front-end applications in high energy physics detectors [18]. To optimize for the latency constraint, the synthesis was configured with a reuse factor of 1 and io_parallel option. The FPGA resource utilization for each algorithm is expressed in terms of Look-Up Tables (LUTs), Flip-Flops (FFs), and Digital Signal Processors (DSPs).

Table 1 shows the achieved latency and FPGA resources for the baseline, pruned, and pruned+quantized models. The baseline direct regressor model achieves a total latency of 55 ns, putting these in the realm of real-time/trigger applications for future colliders with $\mathcal{O}(10)$ ns bunch crossing rates. Pruning and quantization significantly reduce resource usage and thus on-chip power consumption, improving feasibility for front-end implementation.

Future studies into chip design and power studies are required to understand any further optimizations needed for realizable design, as well as codesign of analog and ML readout components. While this work presents the use of ML to perform front-end data compression by extracting the cluster count from the ionization waveform, a variety of other strategies exist that may prove superior given drift chamber readout specifications. For example, data reduction could also be achieved by reducing the waveform to transmit the times of all primary ionizations, determining and transmitting PID directly

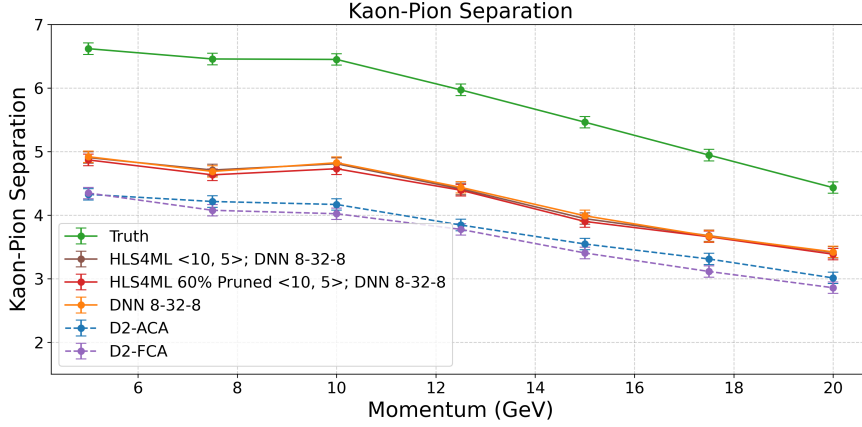


Figure 3: Performance of cluster counting methods for the task of separating pions and kaons. The ML-based methods outperform both traditional derivative-based methods across the full track momentum range, and the application of compression techniques can achieve a model that has $\mathcal{O}(10)$ ns latency.

from the front-end, or reducing the waveform data itself through autoencoder-based compression [19]. Additional work is needed to study all such options, which offer varying levels of compression, power, and data integrity, to determine which is best suited to future physics performance.

Model	Latency [ns]	LUTs	FFs	DSPs
DNN 8-32-8	55	183,726	44,946	3,399
HLS4ML Quantized <10,5> DNN 8-32-8	55	83,417	22,029	39
HLS4ML 60% Pruned; Quantized to <10,5> DNN	45	127,818	26,002	19

Table 1: FPGA resources (expressed in LUTs, DSPs, and FFs) and latency for the 3 ML regressor models considered.

4 Conclusions

A demonstration of ML-based cluster counting for drift chambers is presented. This work demonstrates the first ML approaches that have both improved pion-kaon separation with respect to traditional methods, as well as a clear path toward edge implementation given FPGA synthesis studies. Future work will focus on evaluating the power consumption of such designs and ensuring compatibility with the specifications of future drift chambers, and look into different on-chip implementations that afford a wider variety of data compression options. Code to reproduce the studies shown here can be found at <https://anonymous.4open.science/r/drift-chamber-ml-2DE0>.

Acknowledgments and Disclosure of Funding

The authors thank Charlie Young for helpful discussions. This work is supported by the U.S. Department of Energy under contract number DE-AC02-76SF00515 and the Office of the Vice Provost for Undergraduate Education at Stanford University.

References

- [1] A. Abashian et al. The Belle detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 479(1):117–232, 2002. Detectors for Asymmetric B-factories.
- [2] D. Boutigny et al. BaBar technical design report. 3 1995.

- [3] H. Murayama, K. Heeger, et al. Exploring the Quantum Universe: Pathways to Innovation and Discovery in Particle Physics, Dec 2023. <https://www.usparticlephysics.org/2023-p5-report/>.
- [4] Giovanni F. Tassielli. A proposal of a drift chamber for the IDEA experiment for a future e^+e^- collider. *Proceedings of Science, 40th International Conference on High Energy Physics (ICHEP2020)*, 390:877, 2021.
- [5] L. Rolandi, W. Riegler, and W. Blum. *Particle Detection with Drift Chambers*. Particle Acceleration and Detection. Springer, 2008.
- [6] W. W. M. Allison and J. H. Cobb. Relativistic Charged Particle Identification by Energy Loss. *Ann. Rev. Nucl. Part. Sci.*, 30:253–298, 1980.
- [7] Dian Yu, Houqian Ding, Yunyun Fan, Yongfeng Zhu, and Ming Qi. Evaluation of PID Performance at CEPC and Optimization with Combined dN/dx and Time-of-Flight Data, 2025.
- [8] Guang Zhao, Linghui Wu, Francesco Grancagnolo, Nicola De Filippis, Mingyi Dong, and Shengsen Sun. Peak finding algorithm for cluster counting with domain adaptation. *Computer Physics Communications*, 300:109208, 2024.
- [9] Zhe-Fei Tian, Guang Zhao, Ling-Hui Wu, Zhen-Yu Zhang, Xiang Zhou, Shui-Ting Xin, Shuai-Yi Liu, Gang Li, Ming-Yi Dong, and Sheng-Sen Sun. Cluster counting algorithm for the CEPC drift chamber using LSTM and DGCNN. *Nuclear Science and Techniques*, 36(7), May 2025.
- [10] Ibrahim Alsamak, Kim Baraka, Stefano Caiazza, Asmund Folkestad, Egor Frolov, Jakob Haimberger, Kevin Heijhoff, Michał Jagielski, Djunes Janssens, Irina Kempf, François Lagarde, Grigory Latyshev, Corentin Lemoine, Pere Mato Vila, Joseph McKenna, James Mott, Tom Neep, Lorenzo Neri, Shinsuke Ota, Dorothea Pfeiffer, Joshua Renner, Heinrich Schindler, Ali Sheharyar, Nicholi Shiell, Dario Stocco, Rob Veenhof, Ann Wang, and Klaus Zenker. Garfield++. <https://garfieldpp.docs.cern.ch/>.
- [11] The CEPC Study Group. CEPC Conceptual Design Report: Volume 2 - Physics & Detector, 2018.
- [12] Mogens Dam. Detector requirements, design, and technologies for the FCC-ee Higgs, electroweak, and top factory. *Nucl. Instrum. Meth. A*, 1080:170648, 2025.
- [13] Michael Benedikt, Alain Blondel, Olivier Brunner, Mar Capeans Garrido, Francesco Cerutti, Johannes Gutleber, Patrick Janot, Jose Miguel Jimenez, Volker Mertens, Attilio Milanese, Katsunobu Oide, John Andrew Osborne, Thomas Otto, Yannis Papaphilippou, John Poole, Laurent Jean Tavian, and Frank Zimmermann. FCC-ee: The Lepton Collider. Technical Report 2, CERN, Geneva, 2019.
- [14] Qkeras: a quantization deep learning library for tensorflow keras, 2017.
- [15] Yasser Assran and Archana Sharma. Transport Properties of operational gas mixtures used at LHC. 10 2011.
- [16] FastML Team. fastmachinelearning/hls4ml, 2025.
- [17] Javier Duarte et al. Fast inference of deep neural networks in FPGAs for particle physics. *JINST*, 13(07):P07027, 2018.
- [18] J. Gonski, A. Gupta, H. Jia, H. Kim, L. Rota, L. Ruckman, A. Dragone, and R. Herbst. Embedded FPGA developments in 130 nm and 28 nm CMOS for machine learning in particle detector readout. *Journal of Instrumentation*, 19(08):P08023, August 2024.
- [19] Alexander Yue, Haoyi Jia, and Julia Gonski. Variational autoencoders for at-source data reduction and anomaly detection in high energy particle detectors. *Machine Learning: Science and Technology*, 6(3):035017, jul 2025.