Machine Learning for Event Reconstruction at the EIC

MITACS Globalink Research Internship — Summer 2025

Tomas Sosa

Supervisor: Prof. Dr. Wouter Deconinck

University of Manitoba

June 27, 2025

Outline

Introduction

- The Electron-Ion Collider (EIC) is a next-generation facility for nuclear physics.
- ePIC is the first detector to be built at the EIC.
- Canada plays a leading role in software and computing via the EIC Canada collaboration.
- This project is hosted by the University of Manitoba as part of the MITACS Globalink program.

Goal of the Internship

Integrate and validate a Machine Learning solution for particle identification in the ePIC detector.

Why Machine Learning?

- Traditional reconstruction methods are not optimal for high-granularity data.
- BIC high-granularity data produces complex energy-deposition patterns beyond a simple ratio.
- Machine Learning can identify patterns in complex energy deposition profiles.
- CNNs can learn spatial correlations across layers and hits, capturing subtle shower shape differences.
- Improves accuracy in particle identification (PID), which is crucial for many physics analyses.

Application Area

Particle identification using calorimeter shower profiles in the ePIC Barrel Imaging Calorimeter.

Barrel Imaging Calorimeter (BIC)

- The BIC is part of the central calorimetry system of ePIC.
- Measures energy deposited by particles passing through.
- Electrons, pions, and photons leave distinct energy showers.
- Electrons generate compact and well-defined showers; pions show wider and less regular showers.

Physics Motivation

Electron/pion separation is critical in measurements like $\pi^0 \to \gamma\gamma$.

Machine Learning Pipeline

- We have divided our methodology into two steps: a classical cutoff using the E/p ratio and an ML cutoff using a CNN.
- Configuration:

  Beam energy: $E_{\text{beam}} = 1.0\,\text{GeV}$
  Polar angles $\theta = 45^\circ$–135°
- The objective is to have a total electron efficiency of 0.95 ($\varepsilon_e = 0.95$) and at the same time maximize pion rejection (R_π).

Total Efficiency and Rejection Definitions

$$\varepsilon_e = \frac{N_e^{\text{pass}}}{N_e^{\text{total}}}, \qquad R_\pi = \frac{1}{\frac{N_\pi^{\text{pass}}}{N_\pi^{\text{total}}}} = \frac{N_\pi^{\text{total}}}{N_\pi^{\text{pass}}}$$

First cut: E/p Preselection

- We first exploit the classic calorimeter-to-track ratio $E/p = \frac{\sum_{i=1}^{L} E_{\mathrm{SciFi}}(i)}{p_{\mathrm{track}}}$.
- Physically:

  Electrons shower electromagnetically $\rightarrow$ deposit $E \approx p$.
  Pions leave minimum-ionizing signal $\rightarrow$ $E/p \ll 1$.
- We scan $L = 1 \ldots 12$ SciFi layers, for each finding the E/p threshold that keeps 97
- We select the best separation between all layers based on the maximum pion rejection and use the E/p ratio to obtain the initial cutoff

Second cut: CNN Classifier

- At this point all events have already passed the E/p cut (keeps $\approx 97\%$ of electrons, rejects pions by $R_\pi^{E/p} \approx 23$).
- CNN's job consists of learning residual differences in shower shape to further separate electrons from pions.
- We must choose a CNN output threshold $P_{e^-}^{\mathrm{cut}}$ such that

$$\varepsilon_{e^-}^{\mathrm{tot}} = \varepsilon_{e^-}^{E/p} \times \varepsilon_{e^-}^{\mathrm{ML}} \approx 0.95$$

  (i.e. overall 95% electron efficiency).
- Our goal is to maximize the combined pion rejection $R_\pi^{\mathrm{tot}} = R_\pi^{E/p} \times R_\pi^{\mathrm{ML}}$ at this 95% efficiency.
- raw events $\rightarrow$ [E/p pre-cut] $\rightarrow$ [CNN classifier]

Total Efficiency and Rejection

$$\varepsilon_e^{\mathrm{tot}} = \varepsilon_e^{E/p} \times \varepsilon_e^{\mathrm{ML}}, \quad R_\pi^{\mathrm{tot}} = R_\pi^{E/p} \times R_\pi^{\mathrm{ML}}$$

Data & Features

- Data Loading:
  `hits.snappy.parquet` $\rightarrow$ tensor $(N_{\mathrm{evt}}, N_{\mathrm{layers}}, N_{\mathrm{hits}}, N_{\mathrm{feat}} = 5)$
  `labels.snappy.parquet` $\rightarrow$ PDG codes $\rightarrow \{e^-, \pi^-\}$
- Preprocessing:
  - Reshape to [event, layer, hit, feature]
  - Map PDG codes to binary labels ($1 = e^-$, $0 = \pi^-$)
  - Pion weight: $w_\pi = \min\left(\frac{N_e}{N_\pi} \times t_{\mathrm{imb}}, w_\pi^{\mathrm{max}}\right)$ with $t_{\mathrm{imb}} = 0.1$, $w_\pi^{\mathrm{max}} = 1.0$
  - Split train/val/test: 70 / 10 / 20

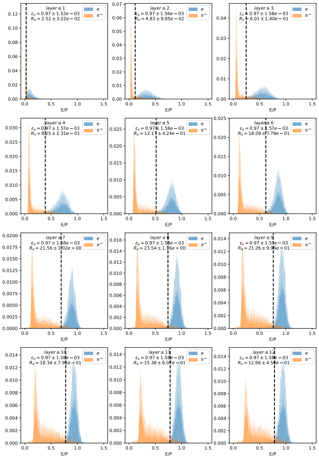
Data Features (per hit)

- e_{norm}: hit energy fraction
- r_{norm}: radial coordinate
- $\Delta\eta$ and $\Delta\phi$ from shower centroid
- layer-type flag (Astropix or SciFi)

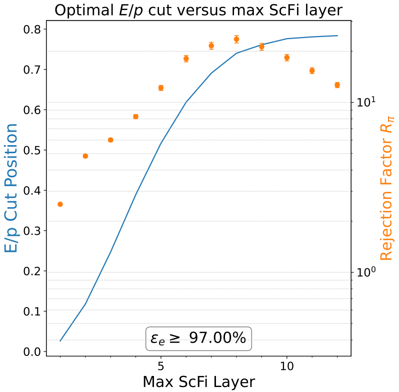
Model Architecture & Training

- Model (VGG-v2):
  $\left[\text{Conv2D(64,3)}\times 2 \rightarrow \text{MaxPool}\right] \rightarrow$
  $\left[\text{Conv2D(128,3)}\times 3 \rightarrow \text{MaxPool}\right] \rightarrow \text{Flatten} \rightarrow \text{Dense(1024)}\times 2 \rightarrow \text{Softmax(2)}$
- Training:
  Adam(lr=1e-3), weighted sparse CCE; 30 epochs; batch 2000 (train) / 1000 (val)
- Evaluation:
  Loss/Acc curves ($\rightarrow$ `ML_learning.pdf`); test inference $\rightarrow$ ϵ_{ML}, $R_{\pi,ML}$; P(e^-) histogram ($\rightarrow$ `ML_rejection.pdf`)

E/p Layer Scan

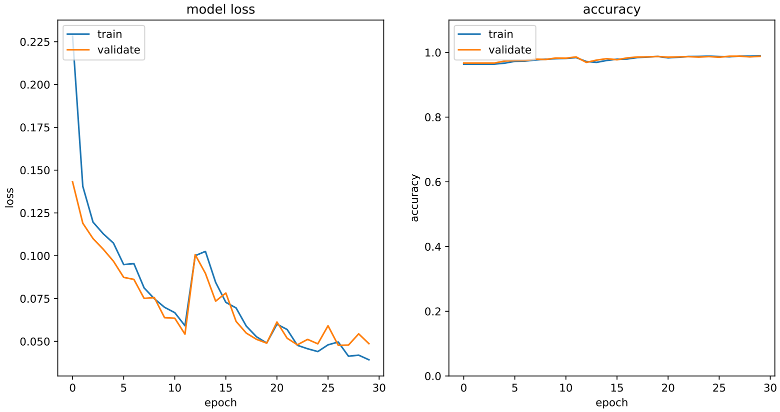


E/p Results

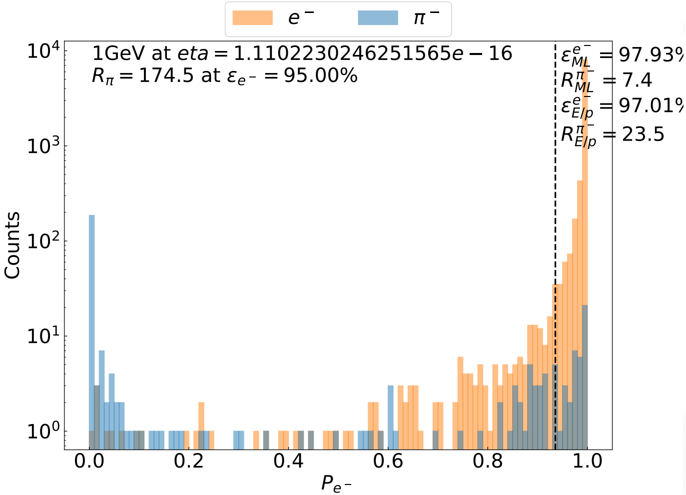


- Blue curve: chosen E/p threshold vs. max SciFi layer.
- Orange points (log-scale): pion-rejection factor R_π.
- Peak at layer 8 $E/p > 0.74$ maximizes R_π while keeping 97

Training Validation Curves



ML Rejection Histogram



EICrecon Integration

- Converting the keras model to an onnx model.
- Create `C++` inference to integrate the E/p and ML algorithms properly into the `EICrecon` framework.
- Validation of the `EICrecon` inference algorithm using simulated data.

Output

A working and reproducible ML-based PID module for ePIC.

Expected Impact

- Improved particle identification in BIC.
- Application in analyses.
- Reusable training pipeline and inference module for future upgrades.

Conclusions

- We demonstrated a two-step PID workflow in the ePIC Barrel Calorimeter:
  - An optimized E/p cut (8 SciFi layers, $E/p > 0.7403$) $\rightarrow$ 97
  - A CNN-based secondary cut on shower “images” $\rightarrow$ net 95
- Our 5-channel per-hit feature representation (e_{norm}, r_0, $\Delta\eta$, $\Delta\phi$, layer-flag) successfully captures subtle EM vs. hadronic shower shapes.
- The VGG-v2 CNN learns layer–hit spatial correlations and boosts pion suppression by nearly an order of magnitude beyond E/p alone.

References

- ⟨https://doi.org/10.1016/j.nuclphysa.2022.122447⟩
- ⟨https://eicweb.phy.anl.gov/Argonne_EIC/becal/ai-reconstruction⟩