

Proposal of CALORICH AI (Draft)

An UBC MDS Capstone Project with TRIUMF, 2023

Crystal Geng, Daniel Merigo, Kelvin Wong, Peng Zhang

May 13, 2023

Contents

1	Executive summary	1
2	Introduction	1
2.1	NA62 Experiment	1
2.2	RICH detector	2
2.3	Project deliverable	2
3	Data science techniques	4
3.1	Data	4
3.2	Machine learning	4
4	Timeline	6
5	Appendix A: Data Structure	6
6	Appendix B: Meeting Schedule	7
	References	7

1 Executive summary

The capstone project titled “*CALORICH AI*” is proposed by TRIUMF, Canada’s national particle accelerator center, aimed at improving particle identification performance of the Ring-imaging Cherenkov detector (RICH detector). Our work this year is an extension of the “*RICH AI*” project, attempted by an MDS capstone team of similar goals last year.

The project dataset comes from the NA62 experiment at CERN, the European Organization for Nuclear Research, which involves particle physics experiment designed to study rare decays of charged kaons. The goal of the capstone project is to build a model that can accurately classify a particle as pion produced in a decay event. The formal deliverable for our project is a trained model that could be used for classification. We aim to train the model using the data from a subset of the 2021A NA62 experiments, which contains around 2.4 million decay events, labeled by calorimeter (Gil et al. 2017).

2 Introduction

2.1 NA62 Experiment

The NA62 experiment at CERN is a particle physics experiment designed to study rare decays of charged kaons (K^+ mesons), which are subatomic particles composed of one up quark and one strange antiquark.

The primary goal of the NA62 experiment is to probe the Standard Model of particle physics, the current theoretical framework that describes the fundamental particles and their interactions, by measuring the ultra-rare decay of a charged kaon into a charged pion and a neutrino-antineutrino pair:

$$K^+ \rightarrow \pi^+ \nu \bar{\nu}$$

By observing and measuring this rare decay, physicists hope to probe the accuracy and completeness of the Standard Model. If the observed rate of the decay differs significantly from the predicted value, it could indicate the existence of new, undiscovered particles or interactions, providing clues to new physics beyond the Standard Model.

The NA62 experiment utilizes a high-intensity beam of charged kaons provided by the Super Proton Synchrotron (SPS) at CERN. The experiment employs various detectors and techniques to identify and measure the decay products of charged kaons, allowing researchers to reconstruct the decay processes and perform precision measurements.

2.2 RICH detector

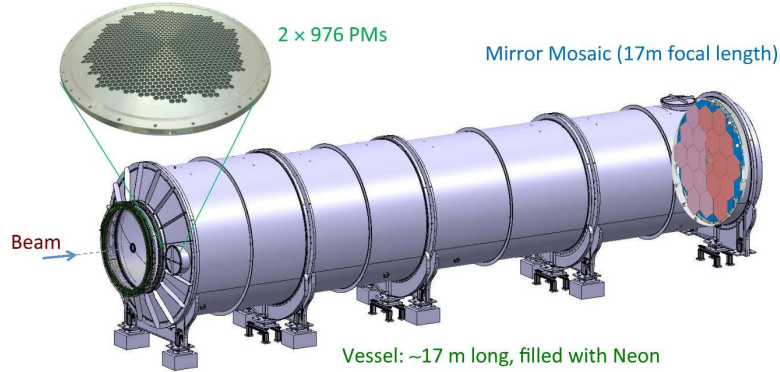


Figure 1: RICH detector (Gil et al. 2017)

Our data comes from the Ring-imaging Cherenkov detector, or a RICH detector, shown in Figure 1 (Anzivino et al. 2020). The RICH detector collects data on photon signals that is emitted during each particle decay (an event). Other detectors, shown in Figure 2 (Anzivino et al. 2020), measure supplementary data such as particle momentum, velocity, and time of passage for additional data.

The data collected by the RICH detector forms an image of a ring, as it is the projection of the conical light emitted due to the Cherenkov radiation (Gil et al. 2017), and we can therefore represent the signal as a 2D scatterplot ((Lazzeroni and Collaboration 2019)).

2.3 Project deliverable

The project deliverable is a trained model for classification of the decay events, as well as the full code with adequate comments to reproduce the results and, potentially, train on new data. The goal in building the

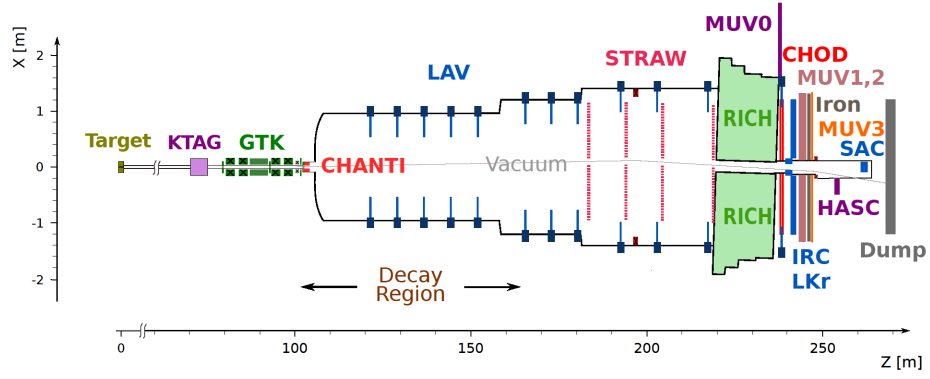


Figure 2: NA62 experiment beam and detector set up (Gil et al. 2017)

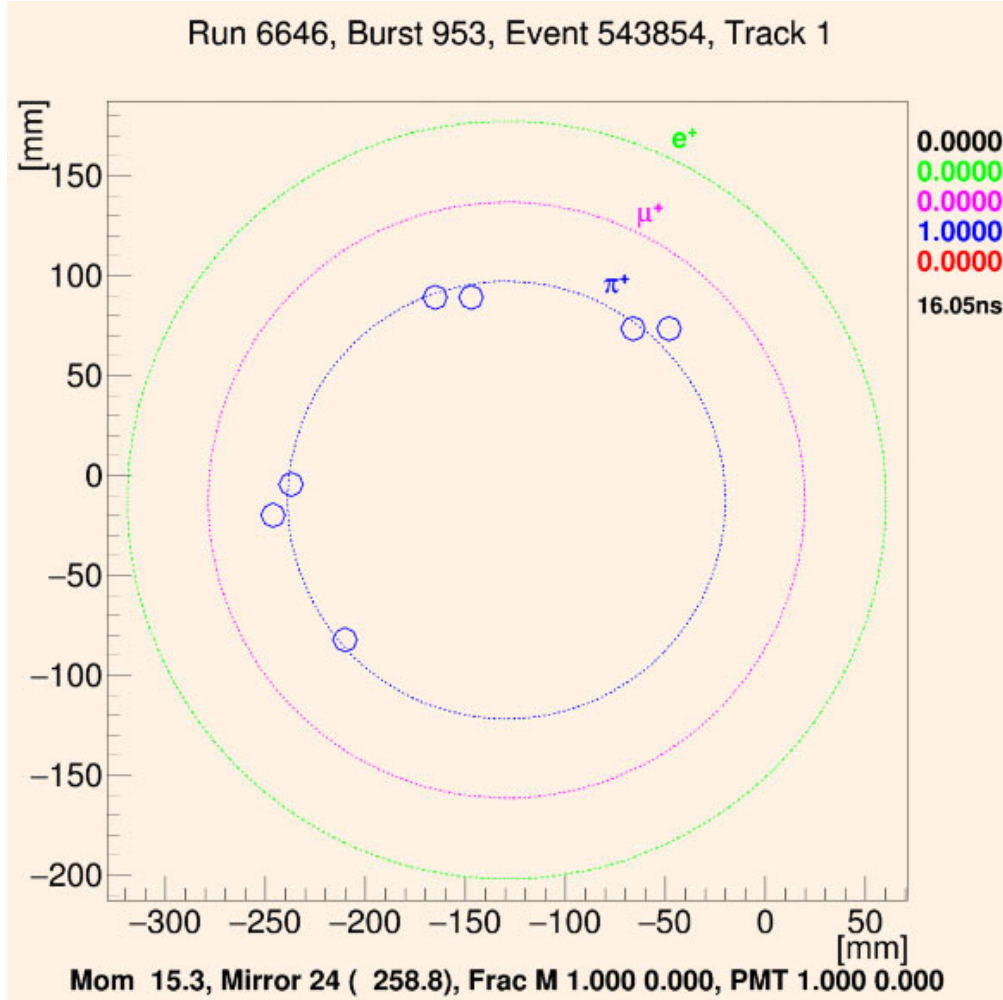


Figure 3: Ring data sample, with fitted probabilities (Lazzeroni et al. 2019)

model would be to accurately classify each particle decay, with an emphasis on pion decays. A knowledge transfer session is also proposed, so that it can be utilized and further developed by the NA62 project team at TRIUMF if needed.

3 Data science techniques

3.1 Data

3.1.1 Data collection

The data we are working with are collected from the NA62 experiment. In each run, after calibrating the instruments, the following steps are performed ((Gil et al. 2017)):

1. A kaon-rich beam is sent into the beam and detector setup in “bursts” every four to five seconds (shown in Figure 2).
2. During each burst, multiple particle decays take place, with each decay assigned a unique “event” ID.
3. As the decay product accelerates through a chamber filled with neon gas in the RICH detector (shown in Figure 1), it generates a cone of light.
4. The light cone is reflected by a mirror mosaic onto a photomultiplier tube (PMT) array, ideally forming a “ring” pattern
5. Each individual PMT in the array records whether it was struck by light and the time of arrival for each light hit.
6. The hodoscope counters (CHOD) detector (shown in Figure 2), registers the time at which the particle decay occurs.

3.1.2 Dataset

We are working on a small slice of data collected from the 2021A NA62 experiment. which is around 2.7GiB in HDF5 format. It contains around 2.4 million labeled events. Each event has a variable number of hits. In total, there are around 101 million records of hits across all events.

The raw data structure can be found in Appendix A.

3.2 Machine learning

3.2.1 Task

Our project’s objective is to develop a machine learning model capable of accurately fitting the radius of the ring, enabling the identification of individual particle decay events as either pions or muons. The experiment primarily concentrates on the rate of rare kaon decay into pions, while muon and positron decays are of lesser interest. The challenge of this task stems from the class imbalance in the data, with pions being considerably rarer compared to muons.

3.2.2 Current approach

The state-of-the-art algorithm currently employed by TRIUMF utilizes an analytical approach without incorporating machine learning for particle classification. Specifically, they implement a statistical algorithm based on the maximum likelihood estimation method to fit a ring to the PMT hit data, using the resulting radius for particle decay classification.

With this methodology, TRIUMF achieves the following results:

1. Pion efficiency (or true positive rate for pion classification) of 95%
2. Misclassification of a pion as a muon (or false negative) of 1%

3.2.3 Goals

In building our machine learning model, we have the following goals:

1. Achieve a pion efficiency of above 95%
2. Reduce pion/muon misclassification to 0.01%

3.2.4 Metrics

In building our machine learning model, we will attempt to maximize pion efficiency, which is defined as:

$$\text{pion efficiency} = \frac{\text{total number of pions detected}}{\text{total number of pions in dataset}}$$

In order to compare different machine learning models, we propose using the metrics root mean square error (RMSE) and coefficient of determination (R^2).

3.2.5 Baseline model

We will use a gradient boosted tree regressor as our baseline model. In our preliminary experiment, we will test both XGBoost and LightGBM, training them on our raw dataset. The model exhibiting superior performance will be selected as the baseline.

3.2.6 Deep learning models

The major concern with XGBoost and LightGBM models is their inability to handle hit data of variable lengths. To effectively incorporate hit positions as features, we need to employ a more sophisticated approach using deep learning.

The first step towards deep learning is that we need to convert the position data corresponding to in-time hits as an input, and use that as a feature instead. The reason for that is that the in-time hits themselves are sparse, compared with the number of the detectors (typically no more than 50 hits in 2×976 detectors), leaving a large portion of the array with no information. Consequently, a traditional CNN method would be disadvantageous due to the sparse input data. To address this issue, we have identified two potential model architectures:

1. PointNet (Qi et al. 2017)
2. Graph Convolutional Neural Networks (“Graph CNN”) (Zhou et al. 2018)

PointNet is designed to directly consume point cloud data without the need to convert positional data to grids or voxels. This can resolve the sparsity problem associated with the traditional CNN approach. Furthermore, PointNet uses a symmetric max pooling function, making it robust to changes in the order of the coordinates comprising the point cloud data.

Dynamic Graph CNN is an architecture capable of handling dynamic data to capture both temporal and spatial dependencies. As a result, Dynamic Graph CNN is better suited for capturing local information within the point cloud data. Additionally, it has a shorter training time requirement, improving computational efficiency and reducing time consumption.

3.2.7 Regression methods

There are three different regression methods that we will attempt in our neural networks:

1. Simple regression
2. Quantile regression
3. Mixture density networks

3.2.8 Implementation

The baseline model will be implemented with the Python libraries for XGBoost (Chen and Guestrin 2016) or LightGBM (Ke et al. 2017), respectively. The deep learning models will be built with PyTorch (Paszke et al. 2019) and PyTorch Geometric (Fey and Lenssen 2019).

4 Timeline

The total time allocated for our capstone project is nine weeks. We propose the following high-level project timeline:

Week 1 & 2 (May 1 to May 12)

- MDS hackathon
- Gain understanding of the scientific problems in the specific NA62 experiment at CERN
- Brainstorm data science-driven solutions
- Conduct exploratory data analysis (EDA) to identify patterns and potential biases in raw data
- Carry out preliminary machine learning (ML) experiments
- Clarify and define the TRIUMF team’s needs and expectations for this year’s MDS capstone project
- Prepare and complete proposal presentation and report for MDS mentor and the TRIUMF team

Week 3 & 4 (May 15 to May 26)

- Finalize preliminary ML experiments
- Select appropriate models and input features
- Implement model training on the full dataset provided
- Conduct hyperparameter tuning

Week 5 (May 28 to June 2)

- Test models against test data and evaluate their performance
- Compare results with current TRIUMF NA62 algorithm and previous MDS capstone modeling efforts
- Prepare and submit a draft model product

Week 6 (June 5 to June 9)

- Gather feedback from the TRIUMF team and MDS mentor
- Further iterate and improve models
- Modularize code and prepare necessary documentation

Week 7 (June 12 to June 16)

- Finalize project deliverables
- Prepare final presentation
- Present the project to the MDS faculty members and class

Week 8 (June 19 to June 23)

- Submit final report and product to MDS mentor for review and feedback
- Revise final report and product based on mentor’s comments

Week 9 (June 26 to June 28)

- Submit the final report and product to TRIUMF
- Present the project to TRIUMF

5 Appendix A: Data Structure

There are **Events**, **Hits**, and **Hitmapping** datasets within the HDF5 file.

The **Events** dataset contains the recorded decays. It contains the following attributes:

- ID (composite of Run ID, Burst ID, Event ID, and Track ID)
- Track momentum
- CHOD time
- Track position
- Features from TRIUMF’s maximum likelihood fit:
 - Ring radius
 - Ring centre (x, y)
 - Ring likelihood
- Decay label, labeled by calorimeter (pion, muon, positron)

The **Hits** dataset corresponds to each of excited sensor. It contains the following attributes:

- Assigned flag, output from the current NA62 algorithm
- Hit position (calculated by the composite IDs)
- Hit time

There is also a **Hitsmapping** dataset, which maintains a one-to-many relationship between the **Events** and **Hits**. An event has a variable number of hits, and each hit must belong to one and only one event.

In addition to the dataset, we also have a file that converts the sensor ID into its relative (x, y) coordinates, so that it could be used in training.

The total number of decays (“events”) for each class are as follows:

- 2,160,219 muon decays
- 215,955 pion decays
- 28,515 positron decays

6 Appendix B: Meeting Schedule

To ensure effective communication throughout the project, we propose the following meeting schedule:

- Meetings with TRIUMF team Every Tuesday, 11:00 am at TRIUMF onsite boardroom
- Meetings with MDS mentor Every Tuesday, 1:00 pm at UBC ICCS 204
- Group meetings Stand-up meeting on Mondays, Wednesdays, and Thursdays on 12:00 pm via Slack huddle Weekly in-person meetings for every Friday, 3:00 pm at UBC after weekly seminars

References

- Anzivino, G, M Barbanera, A Bizzeti, F Brizioli, F Bucci, A Cassese, P Cenci, et al. 2020. “Light Detection System and Time Resolution of the Na62 RICH.” *Journal of Instrumentation* 15 (10): P10025.
- Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. <https://doi.org/10.1145/2939672.2939785>.
- Fey, Matthias, and Jan Eric Lenssen. 2019. “Fast Graph Representation Learning with PyTorch Geometric.” arXiv. <https://doi.org/10.48550/ARXIV.1903.02428>.
- Gil, E. Cortina, E. Martin Albarran, E. Minucci, G. Nüssle, S. Padolski, P. Petrov, N. Szilasi, et al. 2017. “The Beam and Detector of the NA62 Experiment at CERN.” *Journal of Instrumentation* 12 (05): P05025–25. <https://doi.org/10.1088/1748-0221/12/05/p05025>.
- Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. “Lightgbm: A Highly Efficient Gradient Boosting Decision Tree.” *Advances in Neural Information Processing Systems* 30: 3146–54.
- Lazzeroni, Cristina, and on behalf of the NA62 Collaboration. 2019. “First Results from the Na62 Experiment at CERN.” *Journal of Physics: Conference Series* 1137 (1): 012001. <https://doi.org/10.1088/1742-6596/1137/1/012001>.
- Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In *Advances*

- in *Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, 8024–35. Curran Associates, Inc. <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- Qi, Charles R, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. “Pointnet: Deep Learning on Point Sets for 3d Classification and Segmentation.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 652–60.
- Zhou, Jie, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. “Graph Neural Networks: A Review of Methods and Applications.” *CoRR* abs/1812.08434. <http://arxiv.org/abs/1812.08434>.