RICH AI - MDS Team Proposal

Nico Van den Hooff, Mukund Iyer, Rakesh Pandey, Shiva Jena

May 13, 2022

Contents

1 Executive Summary

The NA62 experiment at CERN aims to verify the Standard Model which explains the rate of decay of the Kaon into a pion using the RICH detector and other instruments. However, the pion is only one of the three particles that may be produced in the decay event due to its stochastic nature. Hence the main challenge is to increase the pion efficiency, that is, the fraction of matched detected pions out of the all the pion particles. The ultimate goal of this project is to develop a Python package that applies a combination of machine learning and deep learning techniques to accurately classify the particle produced in a decay event. The data for this project is from the NA62 experiments run in 2018 and contains examples of 11 million decay events. For each event, the subatomic particle produced in the decay is tagged, and contains features specific to the decay such as the particle momentum, CHOD time and the PMT data. This data is the hit times and X and Y positions of the photons produced as the subatomic particle passes through the RICH detector. The machine learning and deep learning algorithms will be developed with the objective of achieving a 95% pion efficiency whilst reducing the misclassification rate of muons and positrons to 0.001%. The baseline classification model will be a boosted tree approach (LightGBM or XGBoost) that will only utilise the base features. The more sophisticated classification models will involve deep learning. Due to the sparse and scattered nature of the images from the experiment, the neural network structures that will be explored are PointNet and Graph Convolution Neural networks. These algorithms are specifically designed to handle the kind of image data described above, and are preferred over a traditional CNN. In the implementation of these neural networks, we expect to use the hit times of the photons as the third dimension. The receiver operating characteristic (ROC) curves will be used to assess the performance of each of these classifiers. Additionally, the pion efficiency will be maximised. The project will be executed in 6 phases: proposal, implementation, evaluation, code refactoring, documenting and reporting, and the final handover.

2 Introduction

Particle physics involves the study of subatomic particles that are the most elementary constituents of matter. However these entities are extremely small and exist at very high energies, and therefore complex instruments are needed to study and characterise them. The NA62 experiment at CERN is one such experiment which is focused on the study of the decay of Kaon particles into a pion particle. The objective of this experiment is to verify the theoretical Standard Model which explains this rate of decay by precisely comparing the relationships between the quarks.

The main detector in the NA62 experiment is the RICH detector, which captures the ring of light produced by every product after a decay event. Other detectors in the instrument measure the particle momentum, velocity and time of passage. The images formed are in the shape of the ring, as it is the 2D projection of the conical light emitted due to the Cherenkov radiation. The light is emitted in packets of energy called photons, and therefore the final image is a scatter of X and Y coordinates.

The challenges of this study arise due to the rarity of the decay event, and the stochastic and uncontrolled nature of the particle decay. In the experimental setup of NA62, the Kaon can randomly decay into 3 particles: the pion, muon or positron. The probability of each of these decays is non-uniform. Hence the efficiency in identifying the pion after the decay process is of key interest.

The structure of this problem sets up perfectly for an image classification to identify the three types of particles that are imaged using a combination of machine learning and deep learning techniques. The metric of focus for these algorithms will be the efficiency of detection with regards to the pion. The ML models can be built based on the features associated with each decay such as the momentum and ring radius of each decay. On the other hand, the relevant deep learning methods to this problem can involve PointNet or graph neural networks as they will be effective in modelling the sparse and scattered structure of the image data. These deep learning architectures can also be used as feature extractors.The technique applied must achieve the desired pion efficiency whilst maintaining the overall accuracy by limiting the misclassification of muons and positrons as pions.

Based on the targets for this project, the final product will be a python package that has functions to convert the raw data of each decay event to image data and other relevant features that can be fed into the main algorithm, a classification algorithm based on deep learning and machine learning techniques and another function that can be used to visualise the results and run an evaluation. These will be developed under the poetry framework so that the package dependencies are effectively managed. The product will be scalable and generalised so that it can be applied to the NA62 data from all the years. This is based on the assumption that the structure of the raw data such as the indexing will be consistent throughout the periods of the experiment. The scripts will be modularized so that the package can easily be incorporated into the existing architecture at TRIUMPH/CERN.

3 Data Science Techniques

- *TODO: Change "Figure X" with dynamic figure captions in markdown*

3.1 Data

3.1.1 Data Generation process

The RICH AI dataset was created as part of the NA62 experiment at CERN in Switzerland in 2016, 2017, 2018, and 2021. In order to generate data, several experiment "runs" are performed. For each run, the experiment configuration is fixed and then the following steps are performed:

1. A beam rich in kaon particles is delivered in "bursts" every four or five seconds into the beam and detector set up as shown in Figure X (Anzivino et al. 2020).
2. During a burst, several particle decays occur. Each particle decay has an individual "event" ID associated with it.

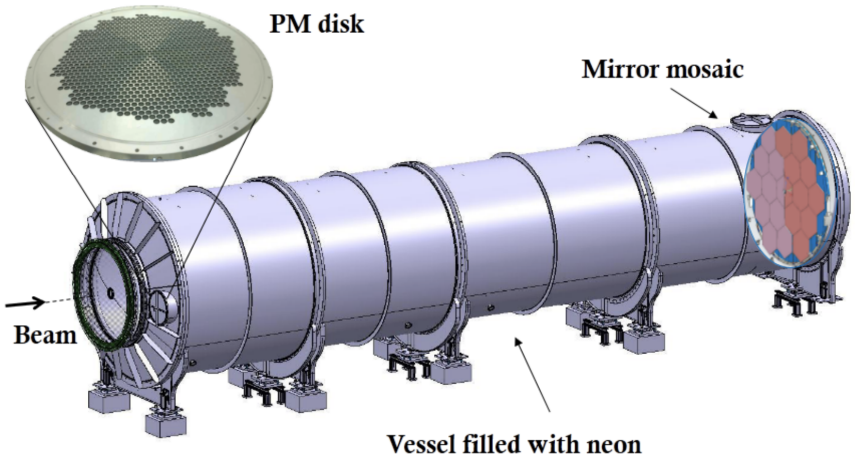

Figure 1: RICH AI beam and detector set up.

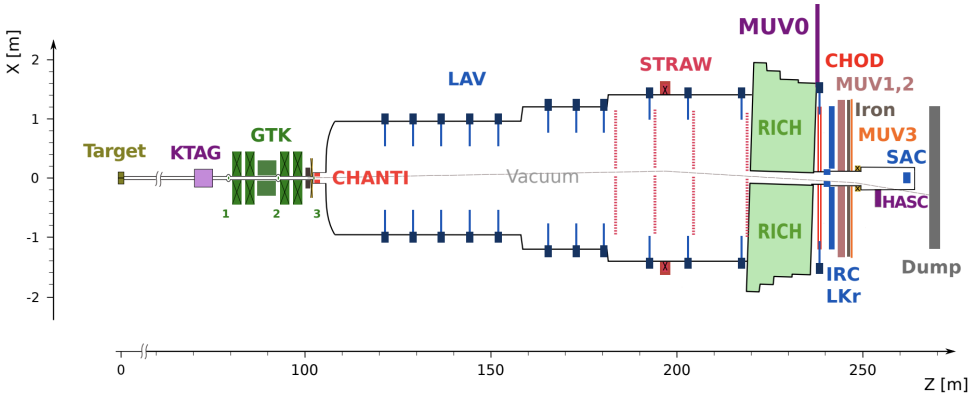

Figure 2: RICH AI beam and detector set up.

3. The product of the decay is accelerated through a chamber of neon gas in the RICH detector and produces/emits a cone of light. The RICH detector is shown in Figure X (Anzivino et al. 2020).
4. The cone of light is reflected by a mosaic of mirrors onto an array of photomultiplier tubes ("PM"). In an ideal situation, the cone of light forms a "ring" on the PM array.
5. Each individual PM in the array records whether or not it was hit by light, as well as the time of arrival for each hit of light.
6. The CHOD detector shown in Figure X (Anzivino et al. 2020) records the time that the particle decay occurs.

3.1.2 RICH AI dataset

Our project will be utilising the 2018 data generated as part of the NA62 experiment. The 2018 dataset contains information on approximately 11 million particle decays and is stored in HDF5 format. Each particle decay contains the following features:

- Decay label (pion, muon, positron)
- PMT data
  - Hit locations (x, y)
  - Hit times
- Particle momentum
- CHOD time
- Features from TRIUMFs maximum likelihood fit:
  - Ring radius
  - Ring centre (x, y)
  - Ring likelihood
- Metadata (Run ID, Burst ID, Event ID etc.)

The total number of decays for each class is as follows:

- 10,281,301 muon decays
- 1,169,556 pion decays
- 113,112 positron decays

3.2 Machine Learning

3.2.1 Task

The task of our project is to build a machine learning model that can accurately classify an individual particle decay event as a pion, muon, or positron. The rarest and most interesting decay is a pion, whereas muon and positron decays are more common and less interesting. We note that the rarity of pion decays is demonstrated by the imbalanced dataset.

3.2.2 TRIUMF's Current Approach

TRIUMF currently employs an analytical approach without the use of machine learning to classify particle decays. Specifically, maximum likelihood estimation is used to fit a ring to the PM hit data, and the ring radius is used to classify the particle decay. With this methodology, TRIUMF achieves the following results:

1. pion efficiency of 95% (true positive rate for pion classification)
2. Misclassification of a pion as a muon 1% of the time.

3.2.3 Goals

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

1. Achieve a pion efficiency of at least 95% or greater.
2. Reduce pion/muon misclassification by 10x to 0.01% of the time.

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 comparing multiple machine learning models, we will utilize receiver operating characteristic ("ROC") curves.

3.2.5 Baseline model

We have decided to use a gradient boosted tree as our baseline model. Specifically, we will try both XGBoost and LightGBM. These models will be trained on raw feature data, and the one that performs better will be selected as the baseline mode to use in assessing the performance of our more complex models.

3.2.6 Deep learning models

In building a more complicated machine learning model we will employ deep learning. We have identified two potential models that may work well in achieving our task and goals:

1. PointNet or PointNet++
2. Graph Convolutional Neural Networks ("Graph CNN")

The first model architecture is called PointNet, which was originally designed to work with point cloud coordinate data (Qi et al. 2017). We will first try out the "vanilla" version of PointNet, and if needed we will also implement the more complex updated architecture called PointNet++.

The second model architecture is called a Graph CNN. Depending on the performance of our PointNet model, we may or may not implement this architecture for our task.

These two architectures are relevant for the type of images produced using the RICH detector as the scatter of photons are both sparse and in point form. Running a traditional CNN would be disadvantageous as a significant portion of the image cintains no information.

3.2.7 Implementation

The baseline model will be implemented with the python libraries for XGBoost or LightGBM, respectively. The deep learning models will be built with PyTorch and PyTorch Geometric.

4 Communication and timeline

4.1 Communication between TRIUMF and MDS

The following modes of communication exist for communication between the two teams:

- TRIUMF/MDS slack channel
- Thursday Zoom meetings between TRIUMF and MDS
- Email

4.2 Overall Project Timeline

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

Week 1 (May 2 to May 6)

- MDS hackathon

- Initial meetings with TRIUMF to learn about data generation process, machine learning options and helper scripts provided
- Brainstorming problem solutions
- Prepare and complete proposal presentation to MDS faculty

Week 2 (May 9 to May 13)

- Perform exploratory data analysis
- Create input data pipeline for machine learning models
- Begin implementation of machine learning models
- Prepare and submit formal written proposal to TRIUMF

Week 3, 4, 5 (May 16 to June 3)

- Build machine learning models

Week 6 (June 13 to June 17)

- Evaluate models on test set against target metrics
- If needed, further iterate and improve models
- Modularize code
- Create final deliverable python package

Week 7 (June 20 to 24)

- Document code base and steps required to run in production
- Write final project report

Week 8 (June 27 to June 30)

- Submit final product to TRIUMF and perform handover session

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.

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.