Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Machine Learning Techniques for Space Calorimeter Experiments

Maria Bossa, Federica Cuna, Fabio Gargano

Referee: Gianpaolo Carlino (INFN Sezione di Napoli), Pasquale Lubrano (INFN Sezione di Perugia)

Spoke 3 General Meeting, Elba 5–9 / 05, 2024

Scientific Rationale

Space-based experiments for direct detection of high-energy cosmic rays often employ optimized calorimeters, designed to achieve high energy resolution and broad acceptance capabilities. The significant volume of data collected demands innovative approaches for analysis and interpretation.

Technical Objectives, Methodologies and Solutions

In this study, we introduce our efforts to develop an AI algorithm dedicated to **classify electromagnetic and hadronic showers**.

- MonteCarlo simulation
- The AI used techniques
- First preliminary results

The montecarlo simulations have been accomplished by using HTCondor on Recas infrastructure in Bari.
The ML software development and the model training has been accomplished by using the JupyterHub instance in Recas-Bari.
For now, the training has been done by using CPU, for the next tests with a large amount volume of data, we will use GPUs.

MonteCarlo Simulations

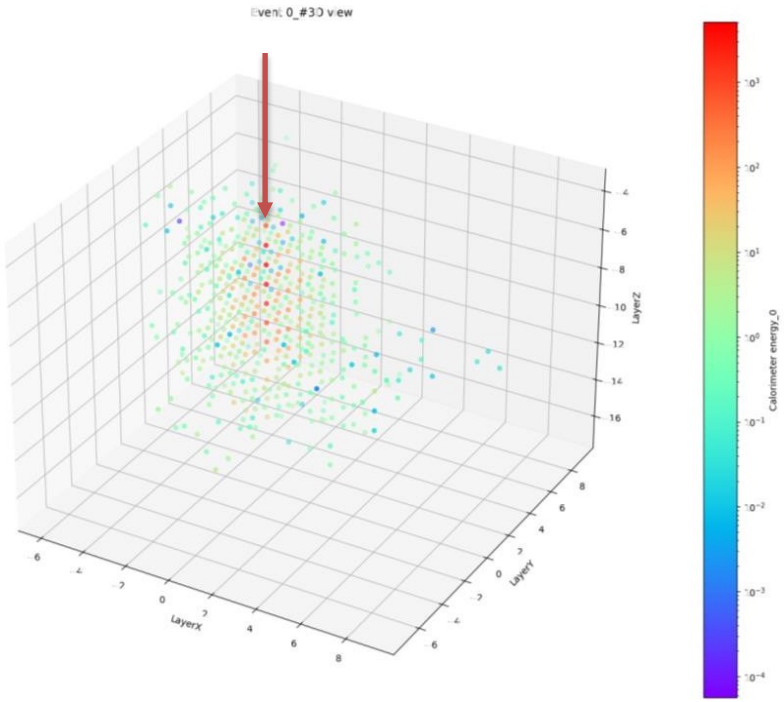
First of all, it was necessary to have a Toy MonteCarlo model of a spatial calorimeter in order to fine-tune a machine learning algorithm, or a neural network model

For this purpose, we simulated, with Geant4 toolkit, a cubic layered calorimeter, composed of 25 layers of LYSO, each measuring 3x3x3 cm, resulting in a total side length of 75 cm.

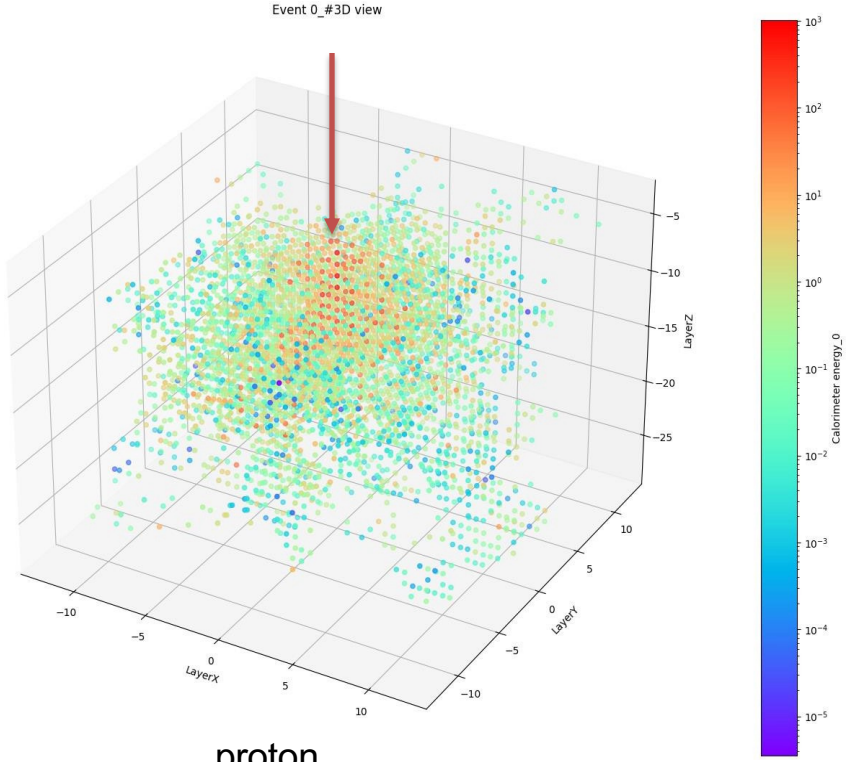
In this preliminary study, the simulation was conducted using monoenergetic primary particles propagated in a linear fashion.

MonteCarlo Simulations – Results

Electrons vs protons of 20 GeV



e-



proton

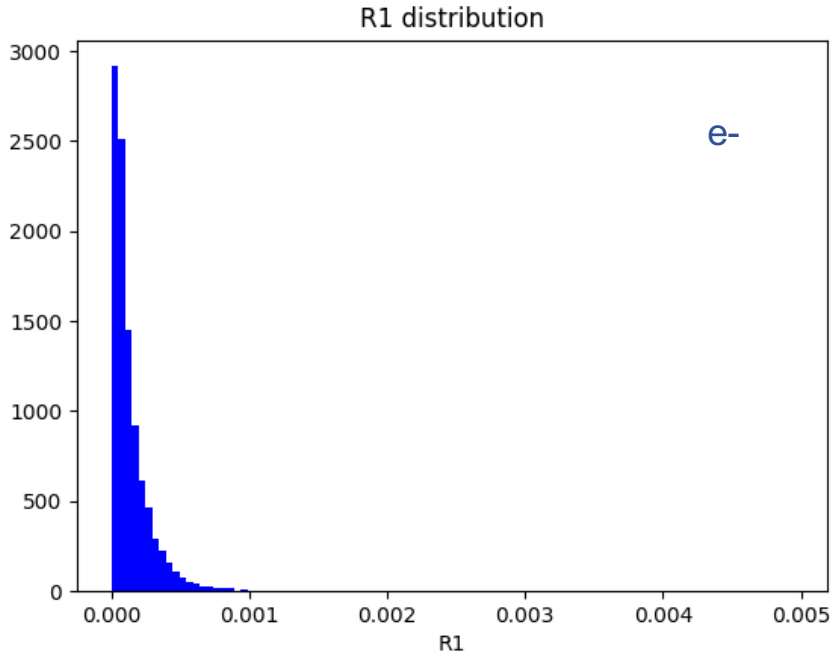
The AI used techniques

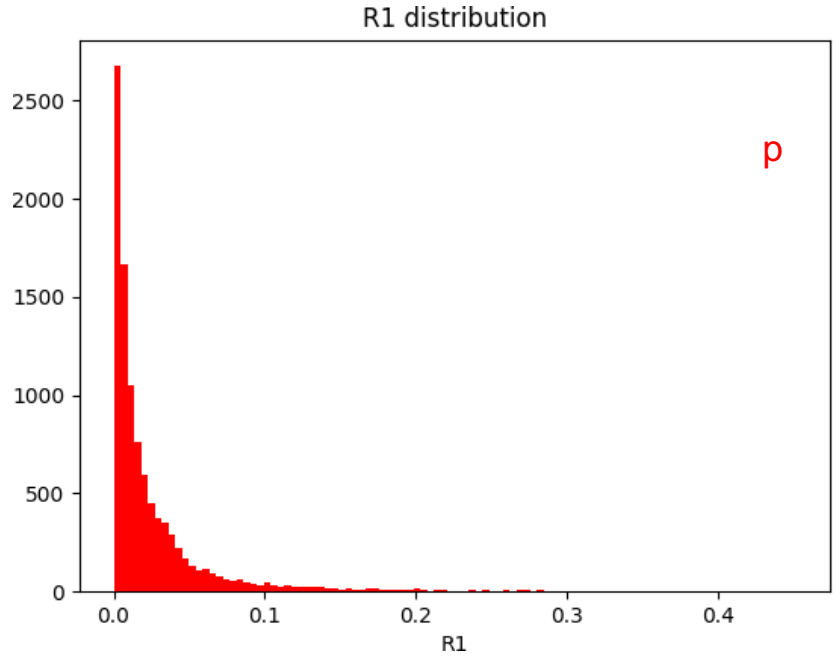
For the machine learning model, we identified the following parameters:

- R1: Ratio of energy deposited in the last layer to the total energy deposited in the calorimeter.
- R2: Ratio of the maximum energy deposited, in a layer, to the total energy deposited in the calorimeter.
- R3: Ratio of energy released in each layer to the total energy deposited in the calorimeter (25 parameters in total).
- R4: Containment radius, the radius within which 90% of the deposited energy is contained in each layer.
- R5: Z-coordinate of the last hit layer.
- R6: Z-coordinate of the maximum energy deposited.

Parameters distributions

Electrons vs protons of 20 GeV

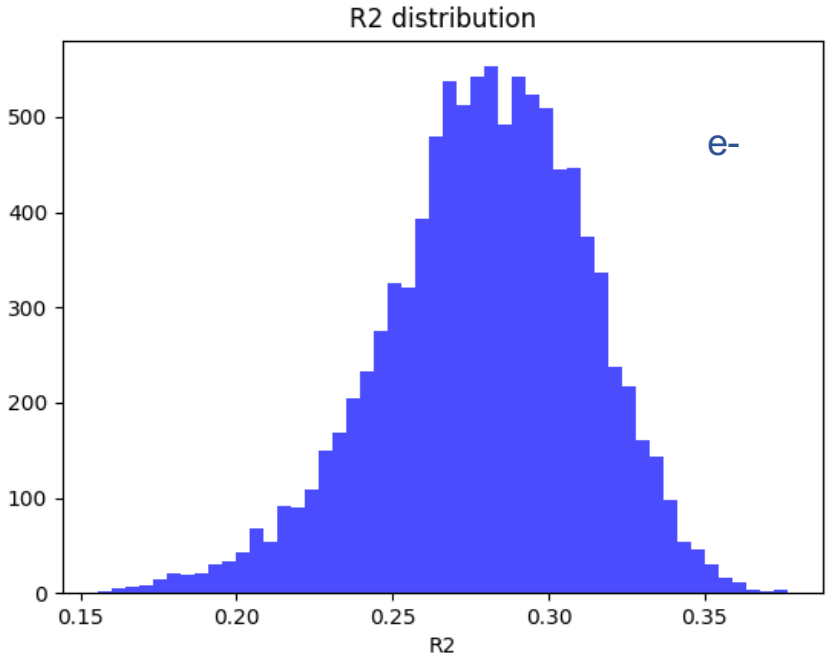


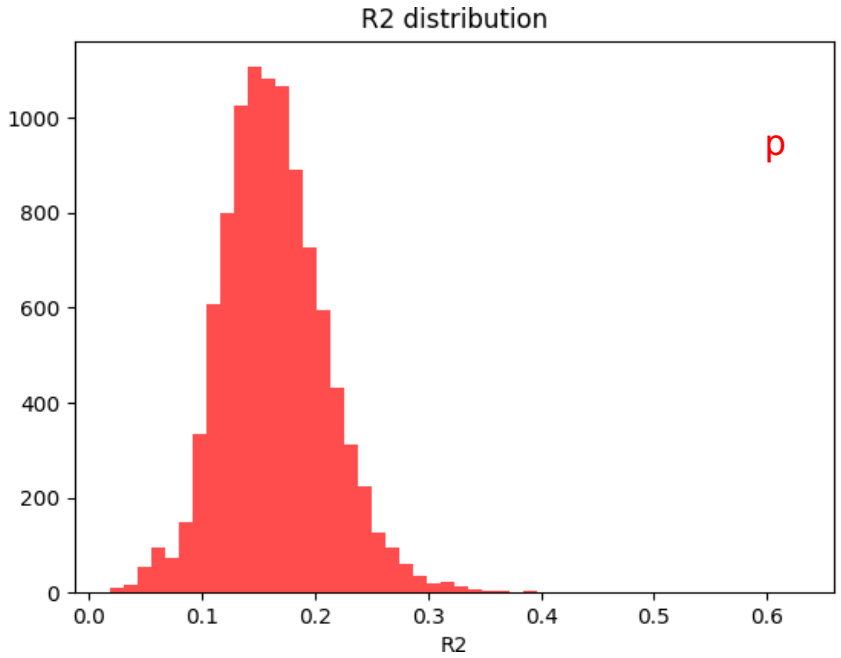


$$R1 = \frac{E_{LastLayer}}{E_{dep}^{tot}}$$

Parameters distributions

Electrons vs protons of 20 GeV

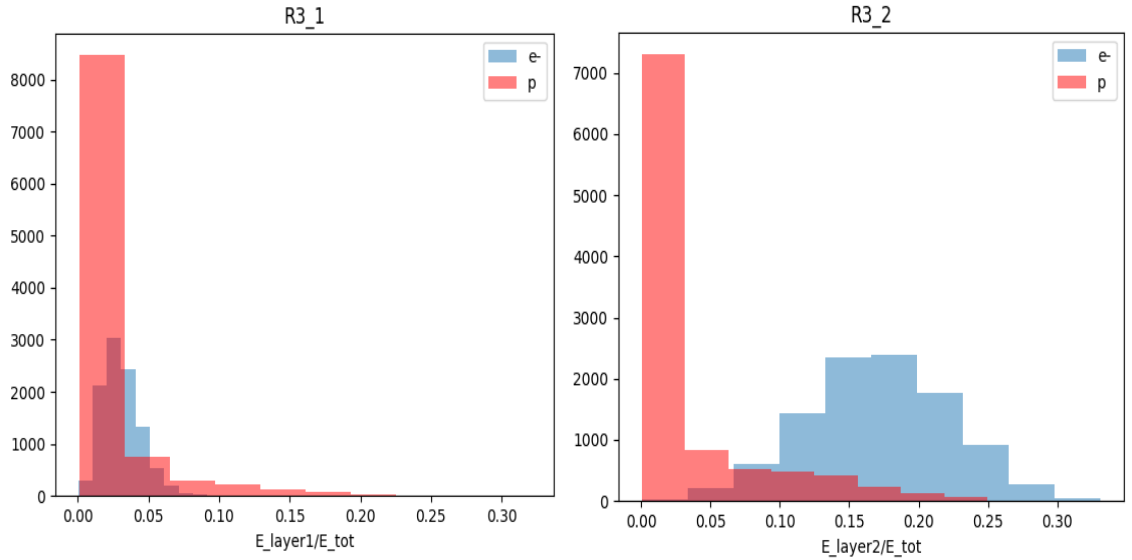




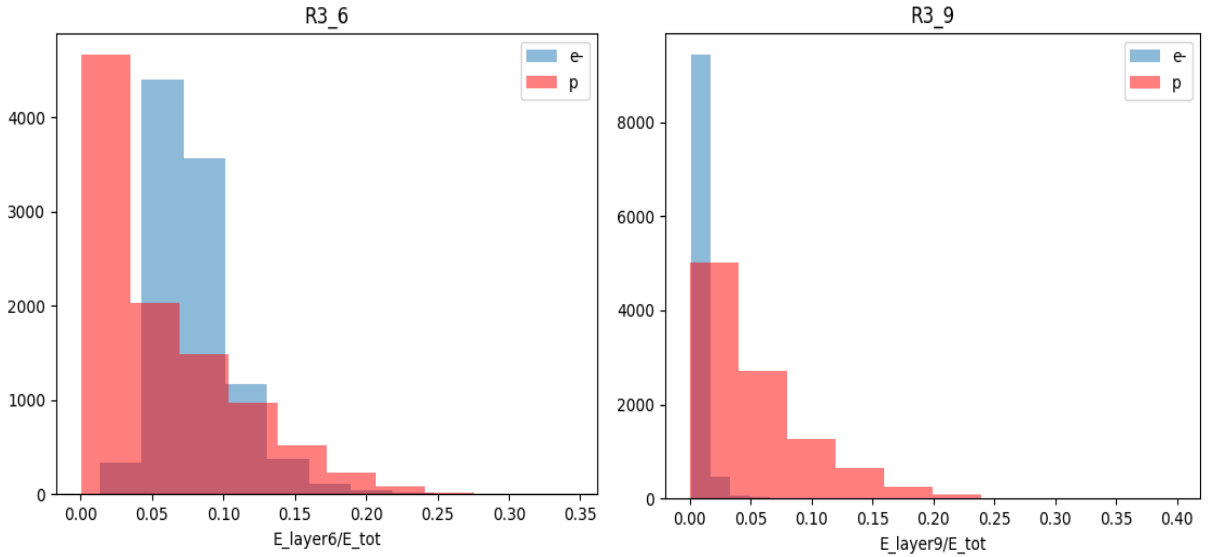
$$R2 = \frac{E_{dep}^{max}}{E_{dep}^{tot}}$$

Parameters distributions

Electrons vs protons of 20 GeV

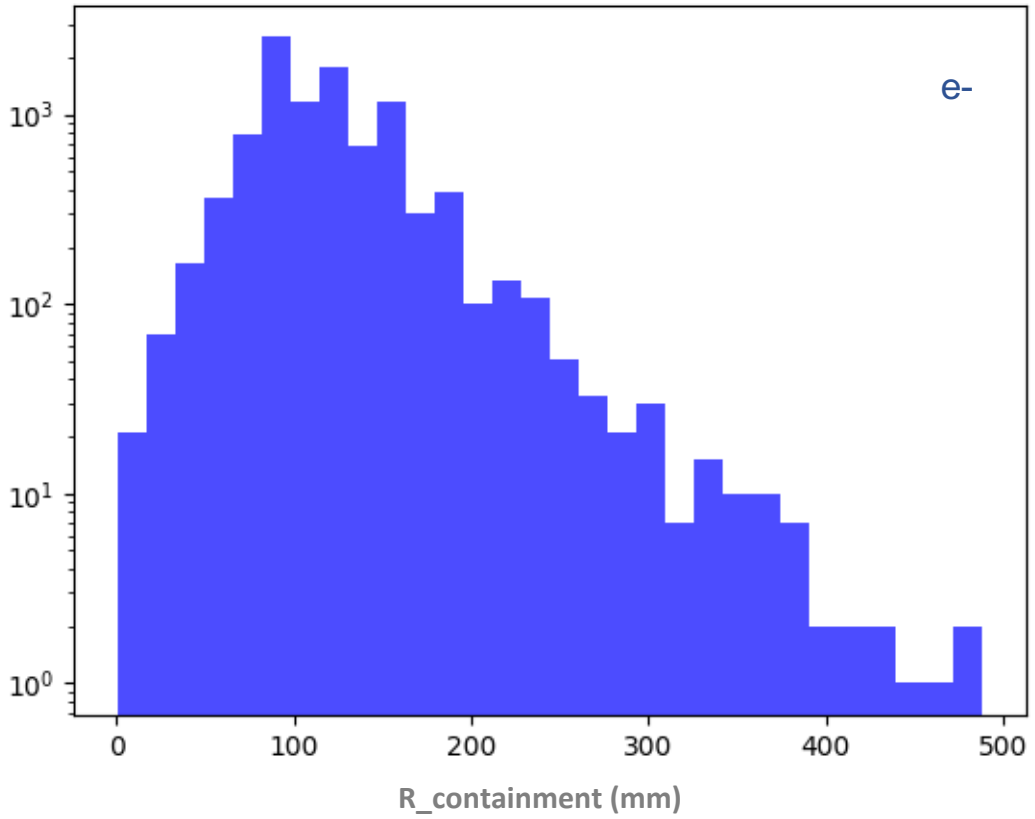


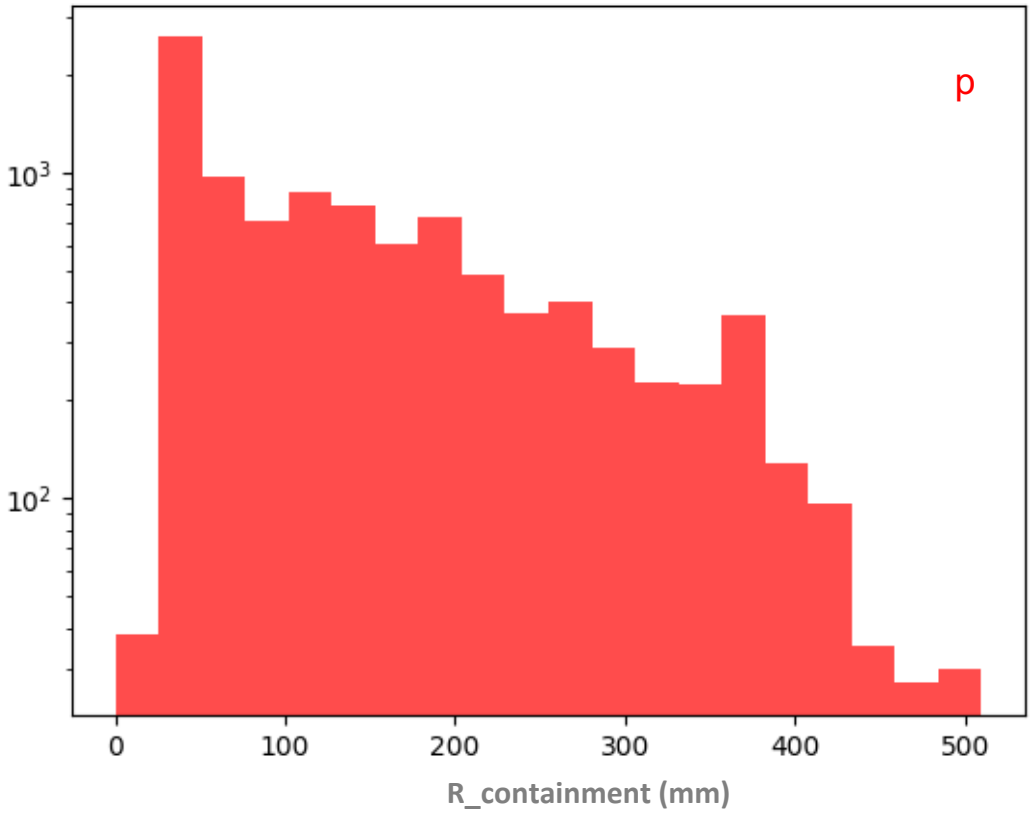
As we move deeper into the different layers, the energy deposition of electrons decreases, while that of protons continues to be significant.



Parameters distributions

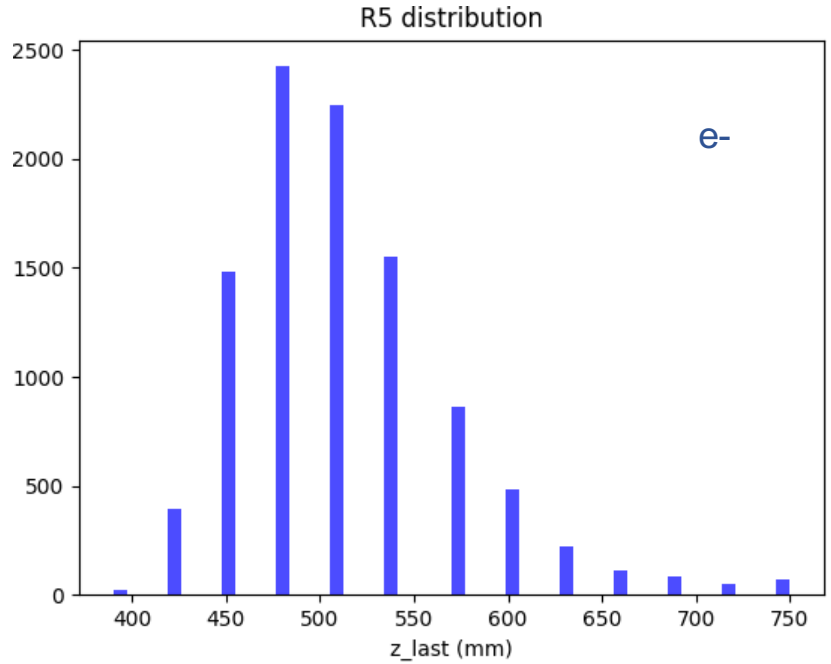
Electrons vs protons of 20 GeV



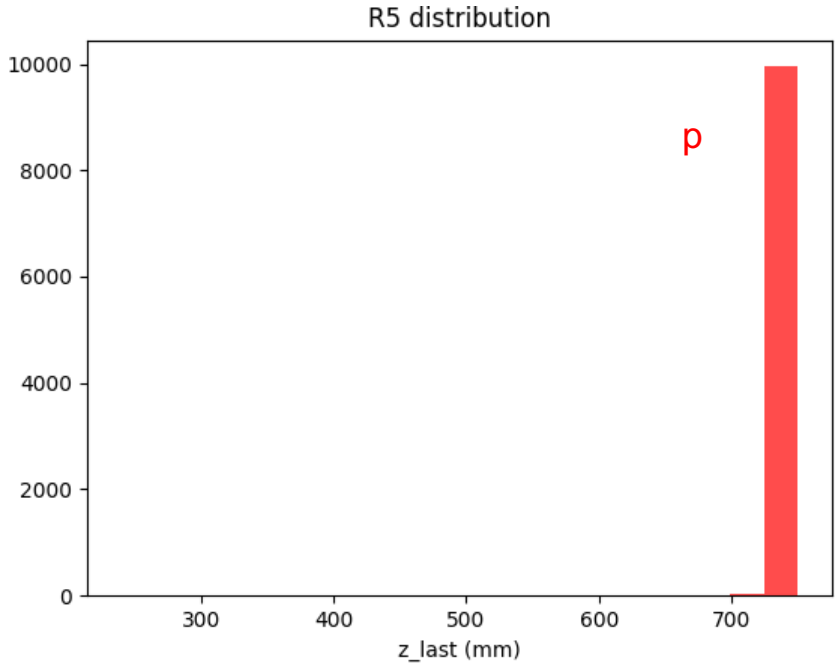


Parameters distributions

Electrons vs protons of 20 GeV: **last hit layer**



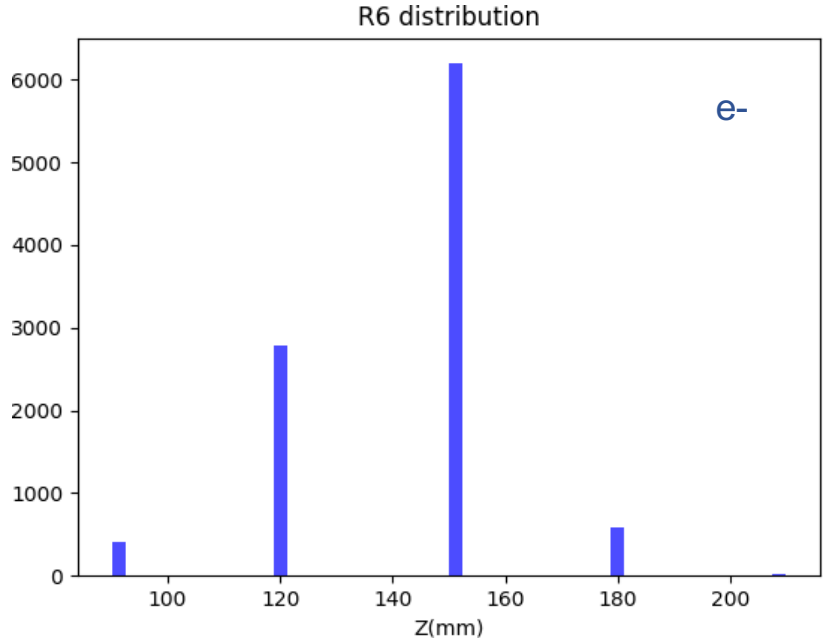
layer ~15th–16th



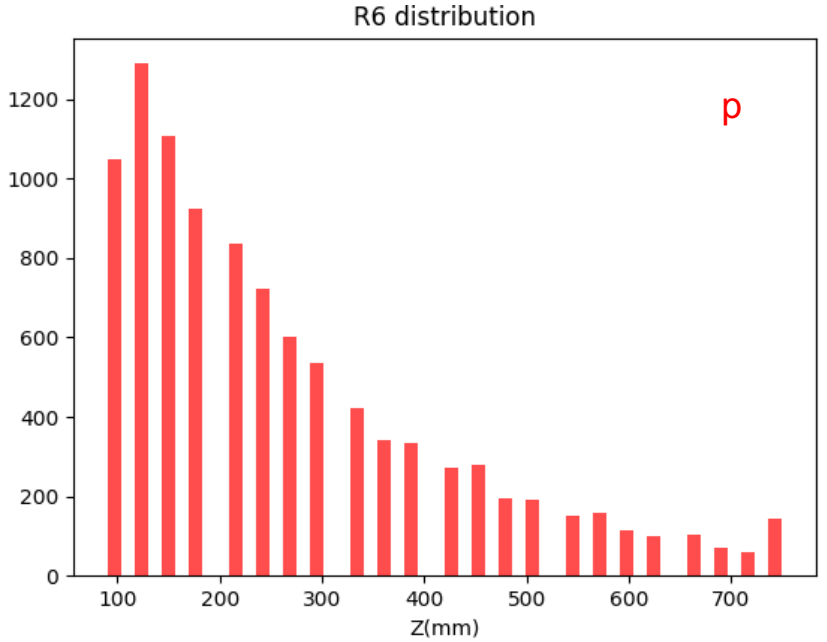
layer ~25th

Parameters distributions

Electrons vs protons of 20 GeV: **z-coord of maximum energy deposited**



layer ~ 5th



layer ~ 4th

XGBoost algorithm

Extreme Gradient Boosting

- The main goal of XGBoost is to find the best balance between the complexity of the trees (how deep and complex they are) and the accuracy of the prediction

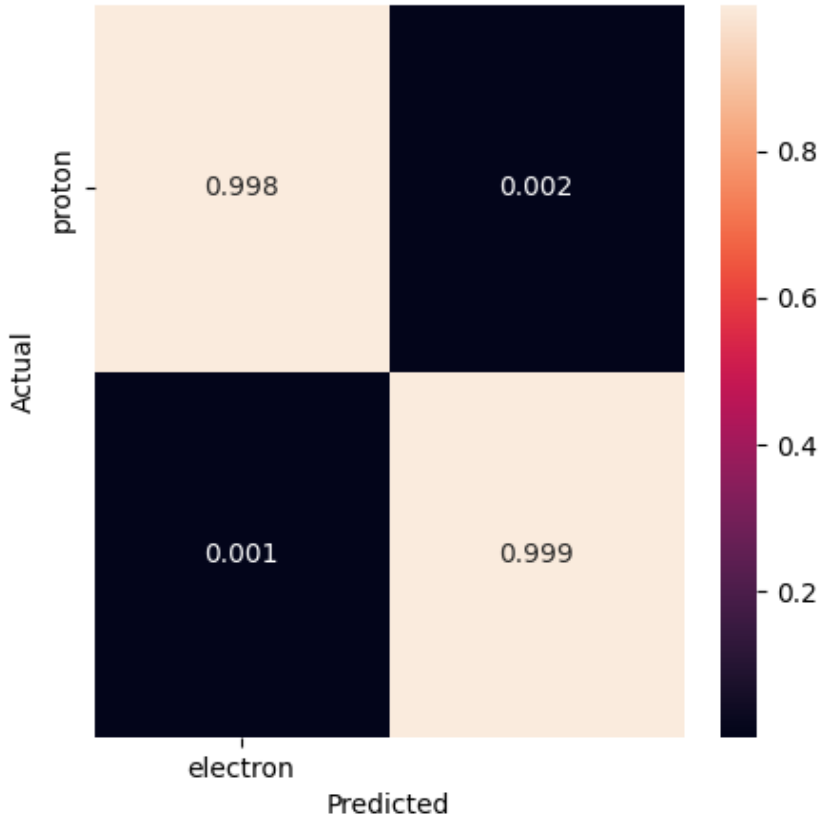
- XGBoost is based on decision trees, similar to random forest. The difference lies in the fact that XGB trains these trees one at a time. It starts with one tree and then adds more incrementally. Each new tree tries to correct the errors made by the previous ones.

- Weak trees have associated weights - these weights represent how skilled each tree is at solving the problem. XGBoost assigns a higher weight to trees that contribute more to the overall error reduction.

XGBoost algorithm Results

Training an algorithm of machine learning with XGBoost, on a sample of 20k events, the results are:

- Accuracy XGB Classifier: 99.85%
- Recall XGB Classifier: 99,90%
- Precision XGB Classifier: 99,80%



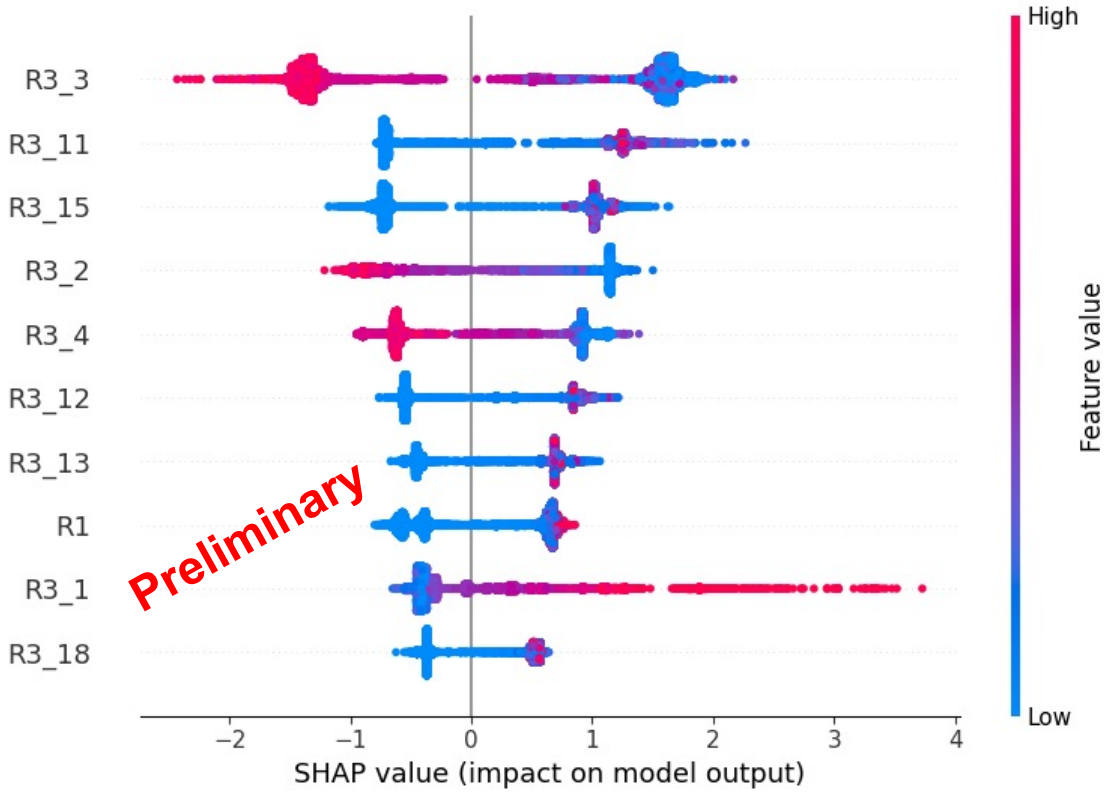
EXplanable Artificial Intelligence (XAI)
SHAP Analysis

- SHAP stands for **SHapley Additive exPlanations**, is the most powerful method for explaining how machine learning models make predictions.

- In particular Beeswarm plots are a more complex and information-rich display of SHAP values that reveal not just the relative importance of features, but their actual relationships with the predicted outcome.

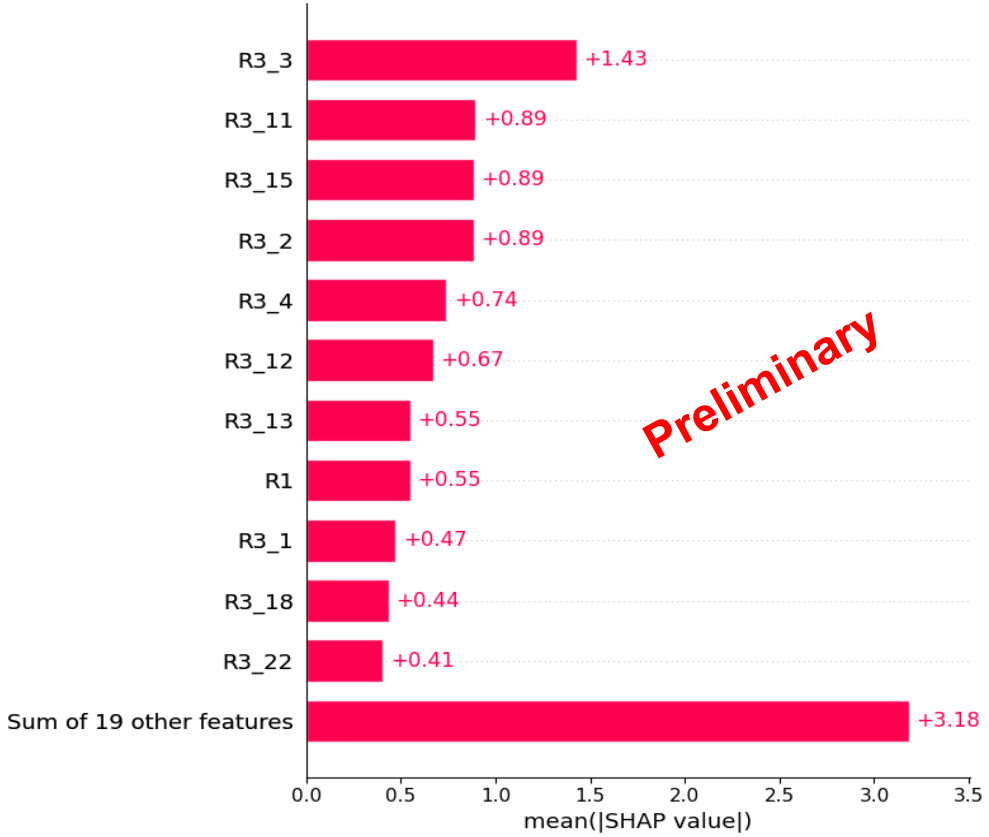
SHAP Analysis

Beeswarm plot



- In a beeswarm plot, for each features, every instance of the dataset appears as it's own point. The points are distributed horizontally along *x*-axis according to their SHAP value.
- Examining how the SHAP values are distributed reveals how a variable may influence the model's predictions.
- Color is used to display the original value of a feature.

SHAP Analysis
Bar Plot

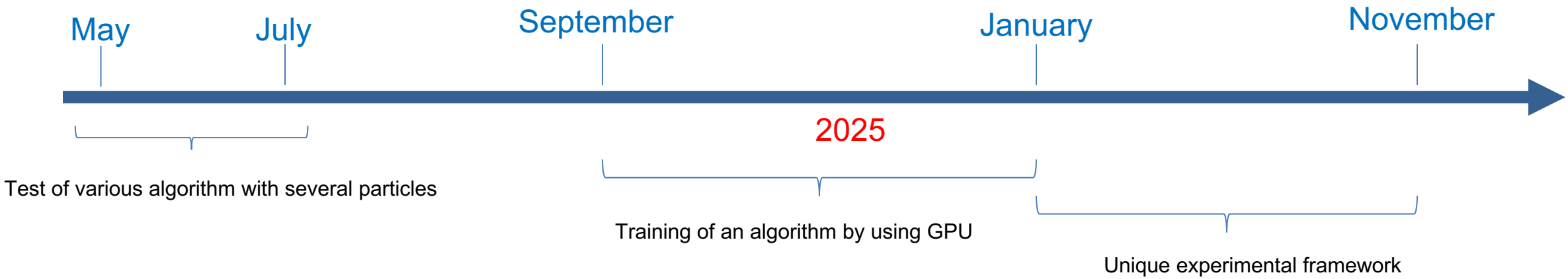


- The simplest starting point for global interpretation with SHAP is to examine the *mean absolute SHAP value* for each feature across all of the data that quantifies, on average, the magnitude of each feature's contribution
- Features with higher mean absolute SHAP values are more influential.

Next Steps and Expected Results

- Conduct tests with various algorithms, exploring their performance across different particle types and energies.
- Explore the possibility of training a Convolutional Neural Network (CNN) using GPU acceleration.
- Perform classification tests using more sophisticated and robust simulations.
- Develop an integrated framework to facilitate spatial experiments, incorporating both tracker (see Federica Cuna's talk) and calorimeter functionalities.

Timescale, Milestones and KPIs and Expected Results



Thank you for your attention

Backup slides

R3 parameter for each layer