

Machine Learning Techniques for Space Calorimeter Experiments

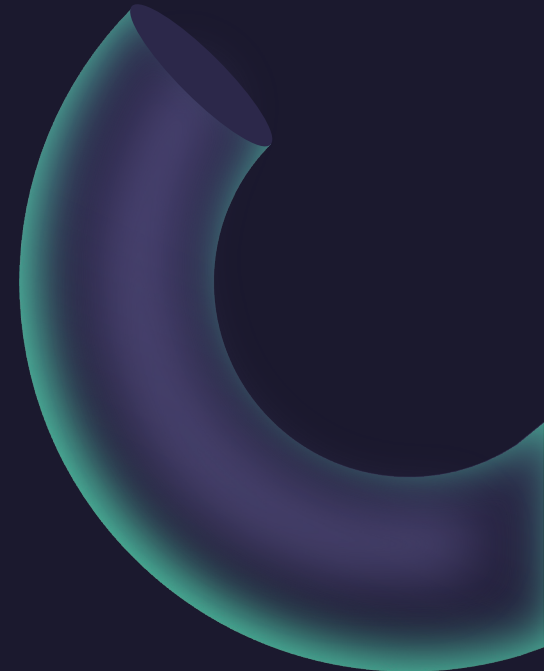
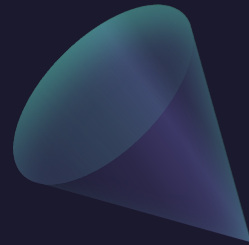
Maria Bossa, Federica Cuna, Fabio Gargano
Referee: Gianpaolo Carlino (INFN Sezione di Napoli),
Pasquale Lubrano (INFN Sezione di Perugia)



Overview

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



MonteCarlo Simulations



First, it was necessary to develop a toy Monte Carlo model of a spatial calorimeter to fine-tune a machine learning algorithm or neural network model.

The initial simulation provided only a single layer of pixels along the z-axis, where the pixel's z-dimension corresponded to the z-dimension of the calorimeter. At that stage, adding additional layers of pixels was not feasible.

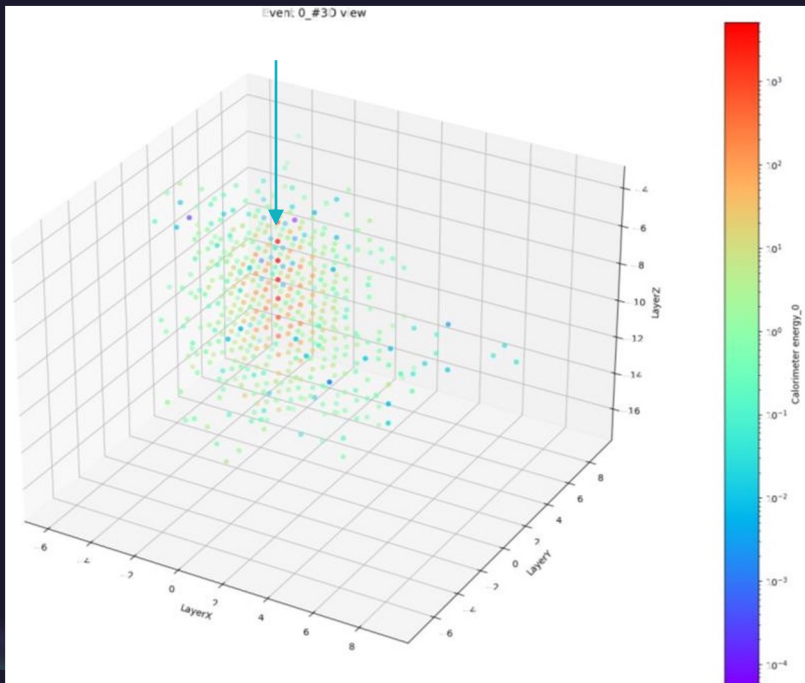
However, we have since overcome this limitation and can now modify the number of layers in each dimension and reconstruct events accordingly.

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

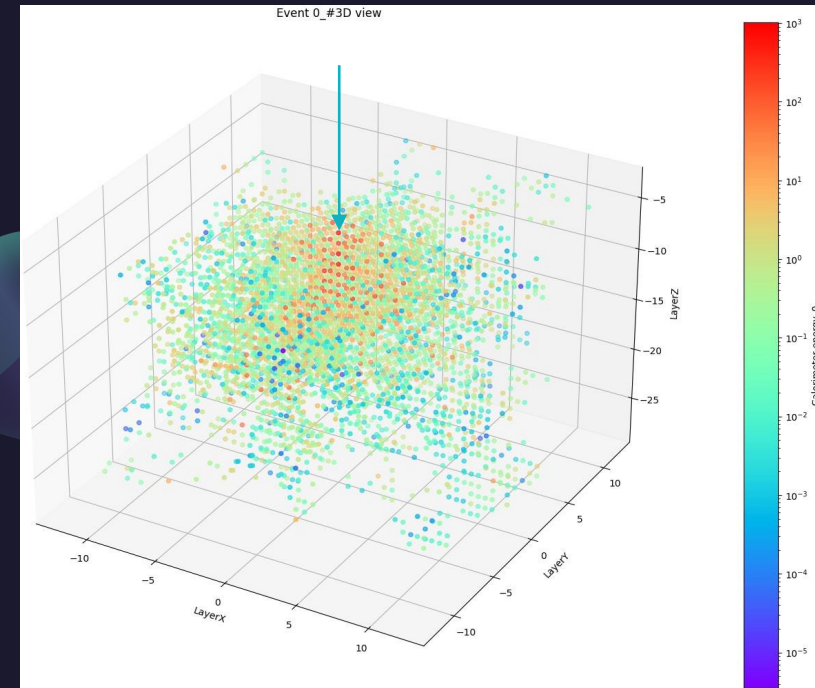
MonteCarlo Simulations

Electrons vs protons of 20 GeV

The current configuration consists of a cubic layered calorimeter composed of 25 layers of LYSO, each measuring 3 cm on each side, resulting in a total length of 75 cm per side.



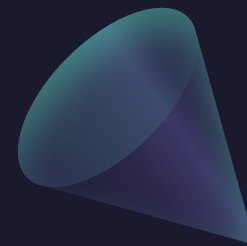
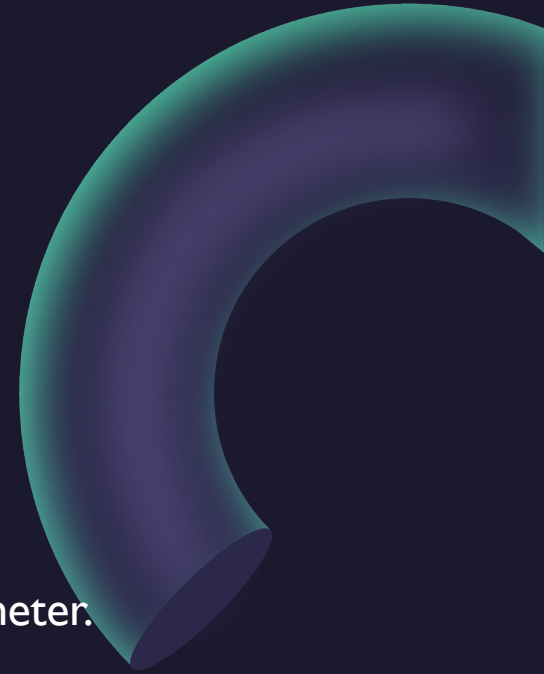
Electron event



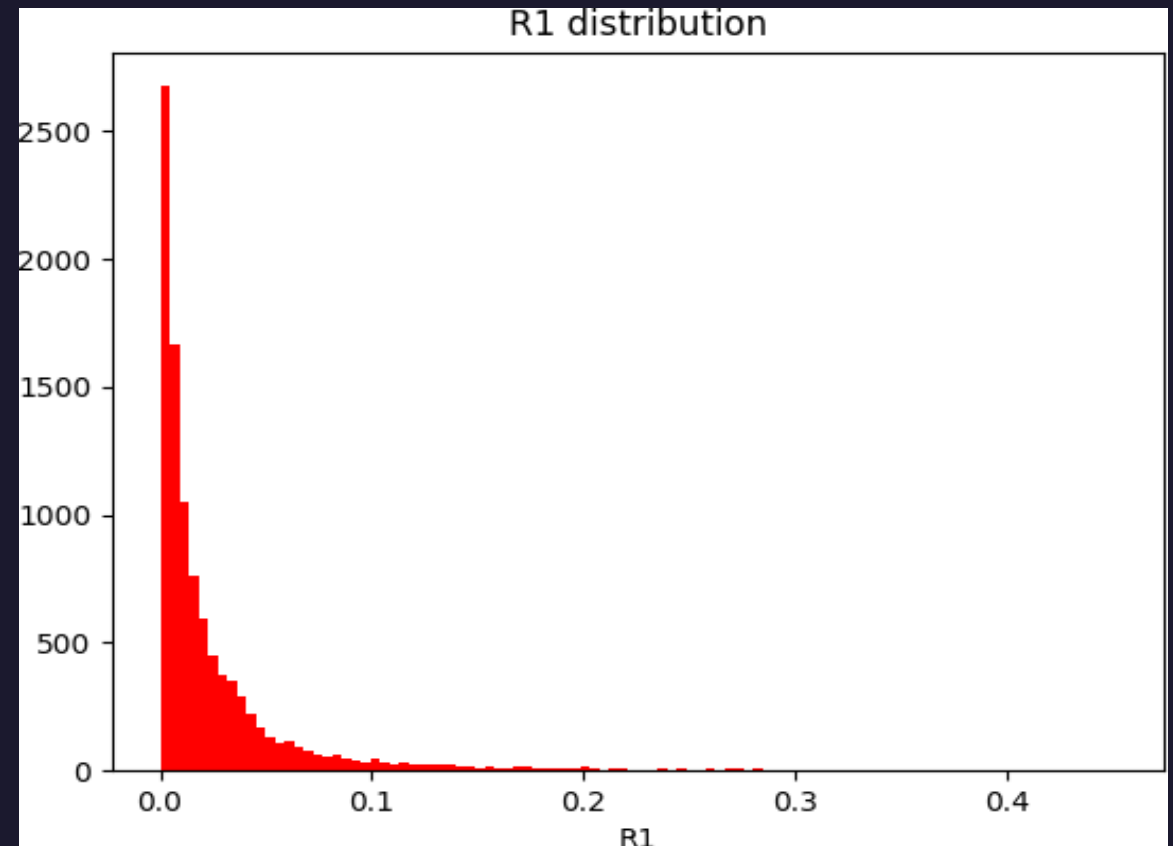
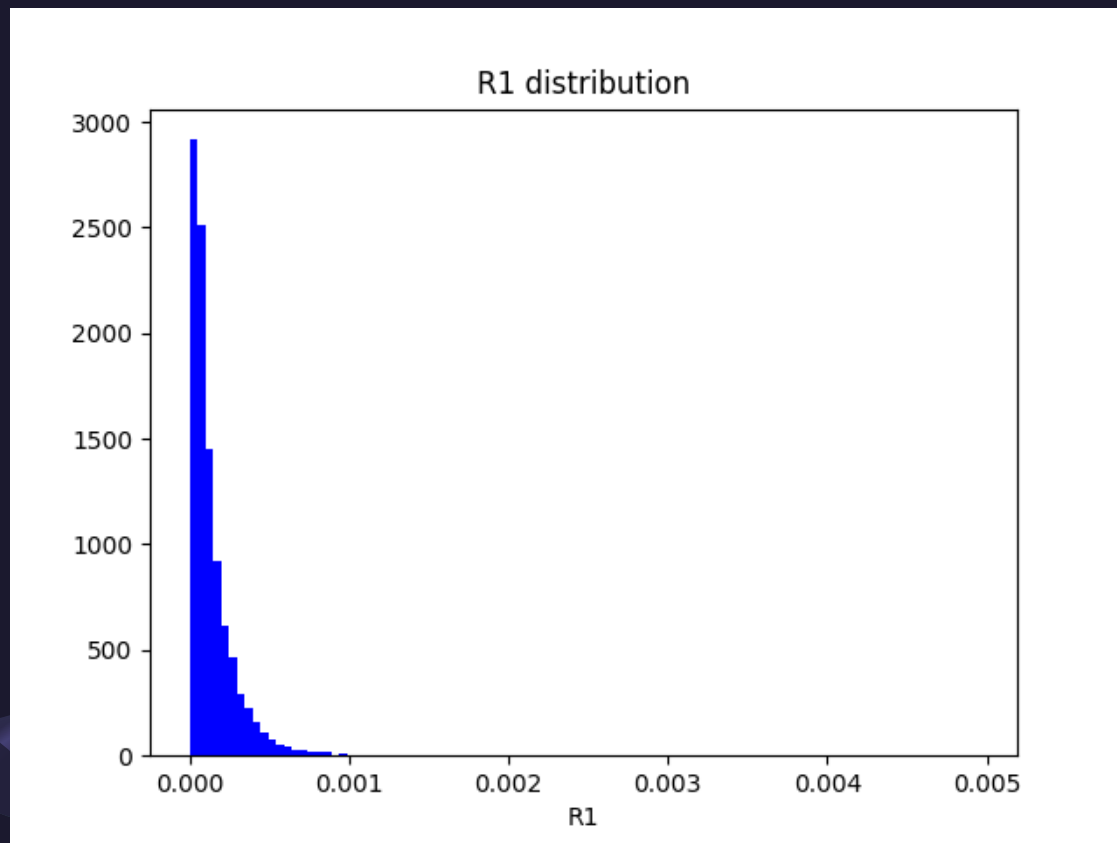
Proton event

Chosen parameters for the AI model

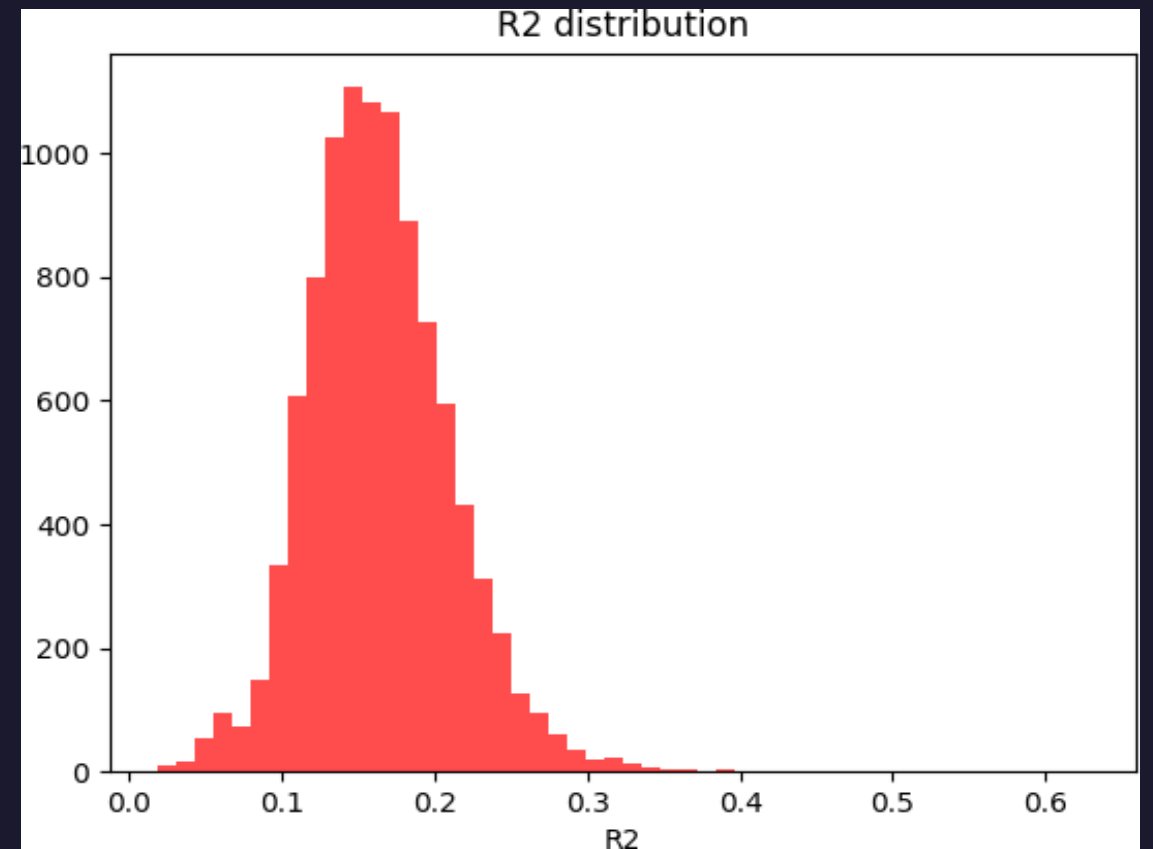
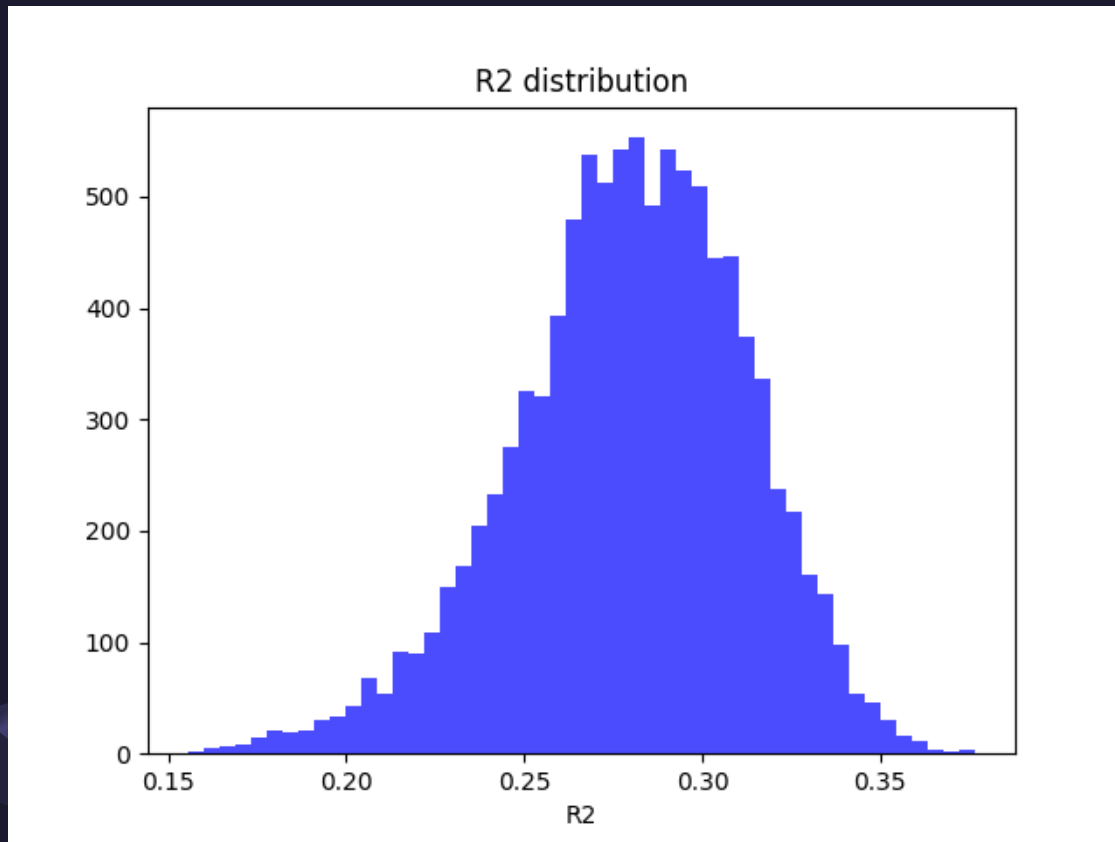
- For the machine learning model, we identified the following parameters:
- R1: Ratio of energy deposited in the last layer to the total energy deposited in the calorimeter.
- R2: Ratio of the maximum energy deposited, in a layer, to the total energy deposited in the calorimeter.
- R3: Ratio of energy 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.



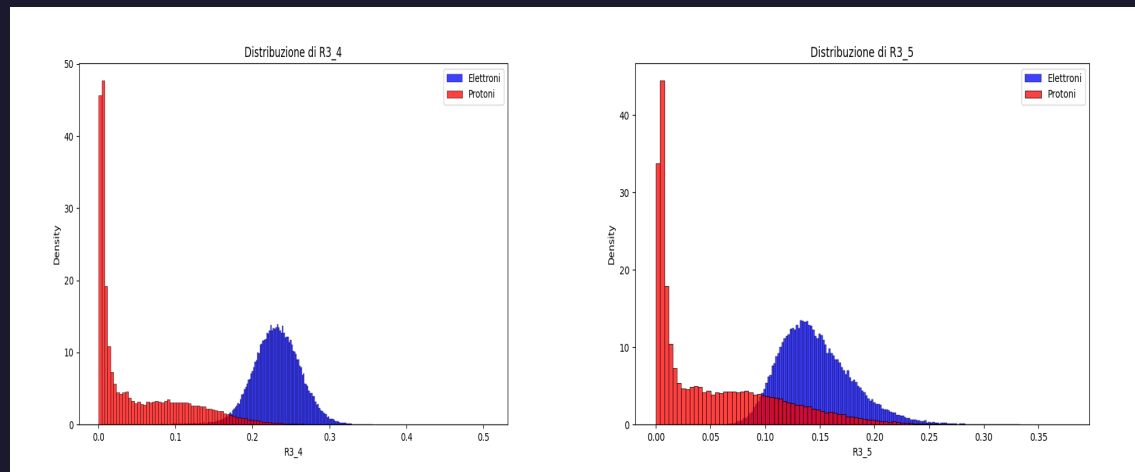
Parameters distributions – R1 (Ratio of energy deposited in the last layer to the total energy deposited) Electrons vs protons at 20 GeV



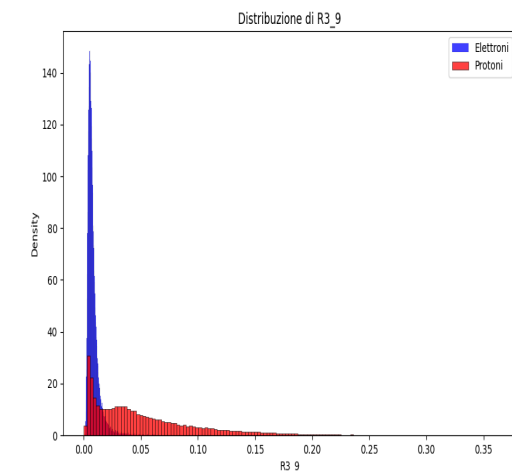
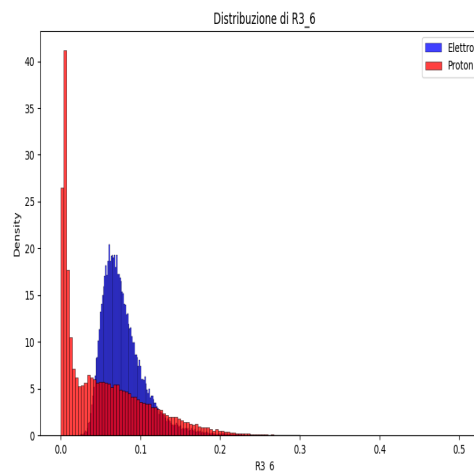
Parameters distributions – R_2 (maximum energy deposited, in a layer, to the total energy deposited) Electrons vs protons at 20 GeV



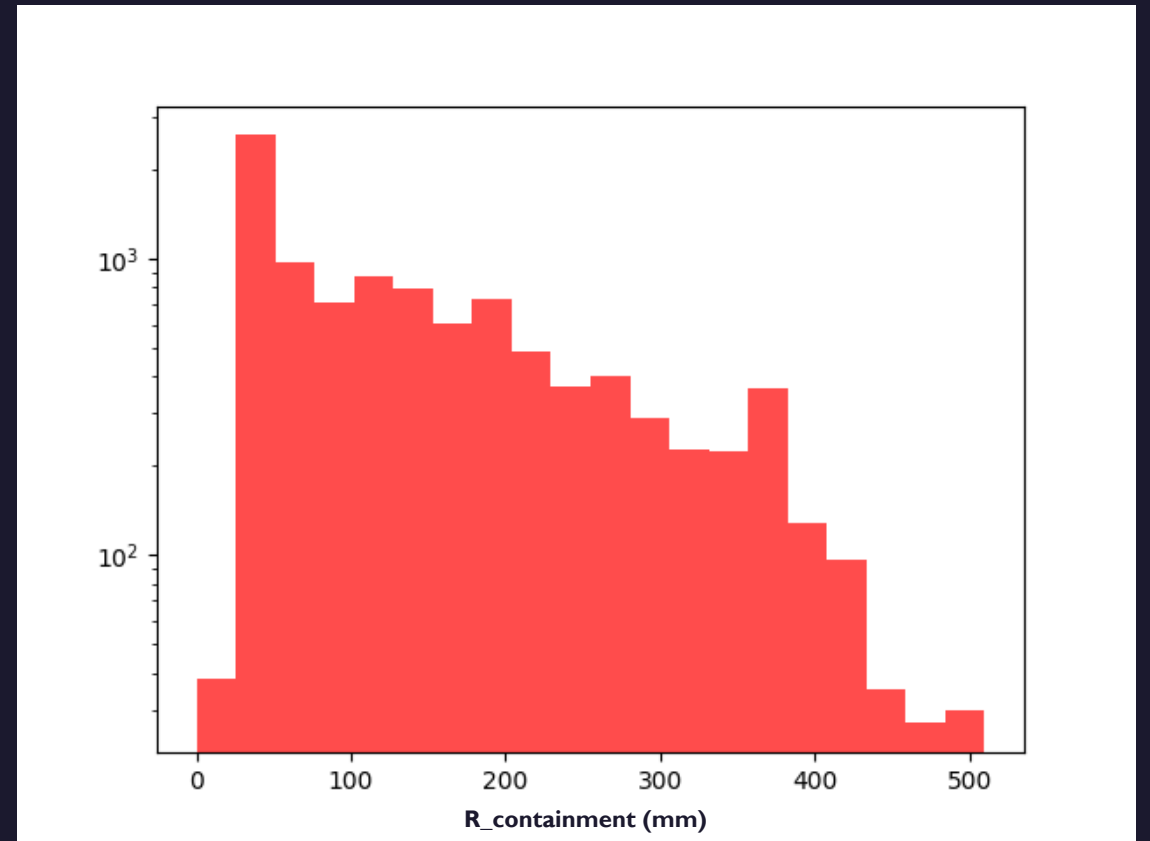
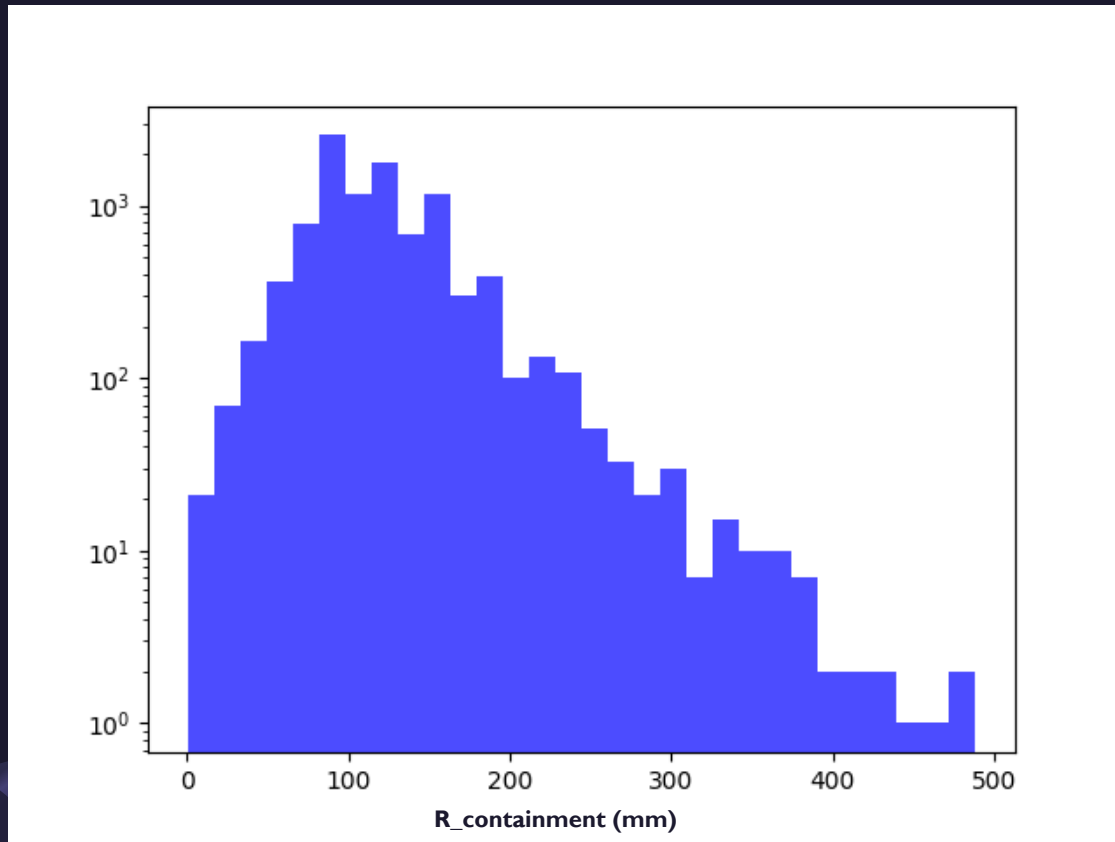
Parameters distributions – R_3 (Ratio of energy released in each layer to the total energy deposited) Electrons vs protons at 20 GeV



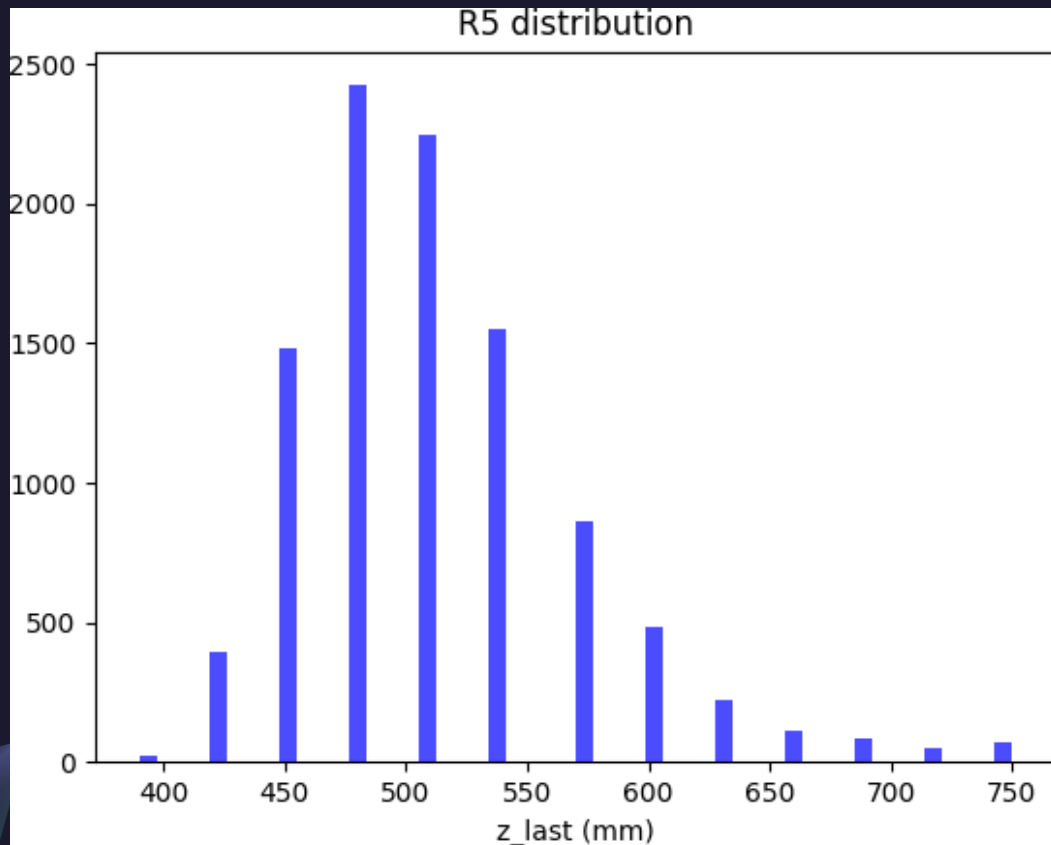
As we move deeper into the different layers, the energy deposition of electrons decreases, while that of protons continues to be significant



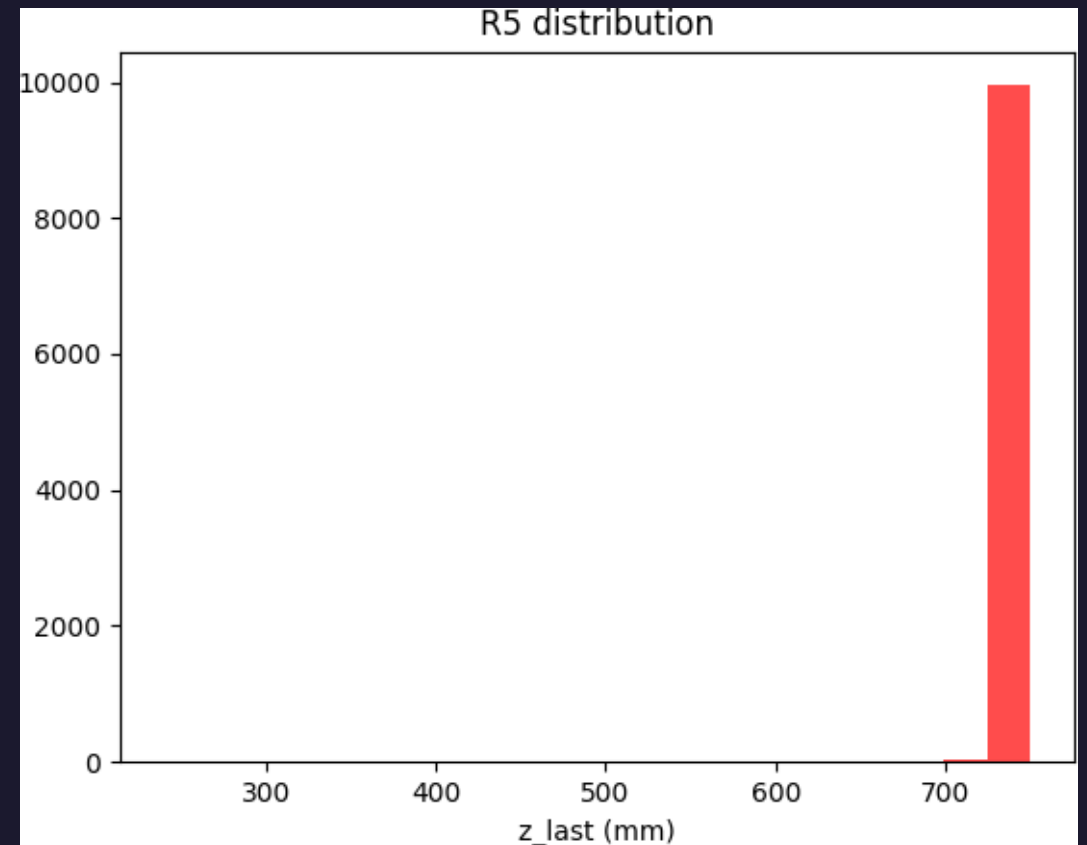
Parameters distributions – R4(containment radius) Electrons vs protons at 20 GeV



Parameters distributions – R5(Z of the last hit layer) Electrons vs protons at 20 GeV



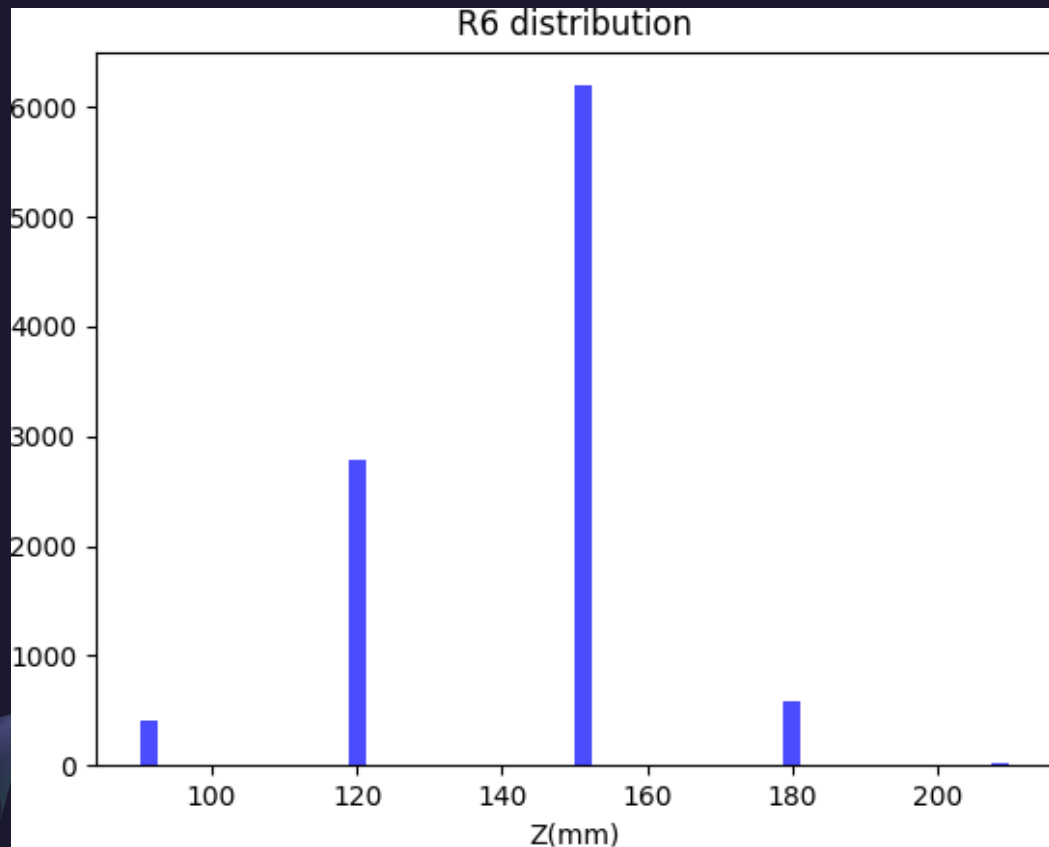
layer ~15th-16th



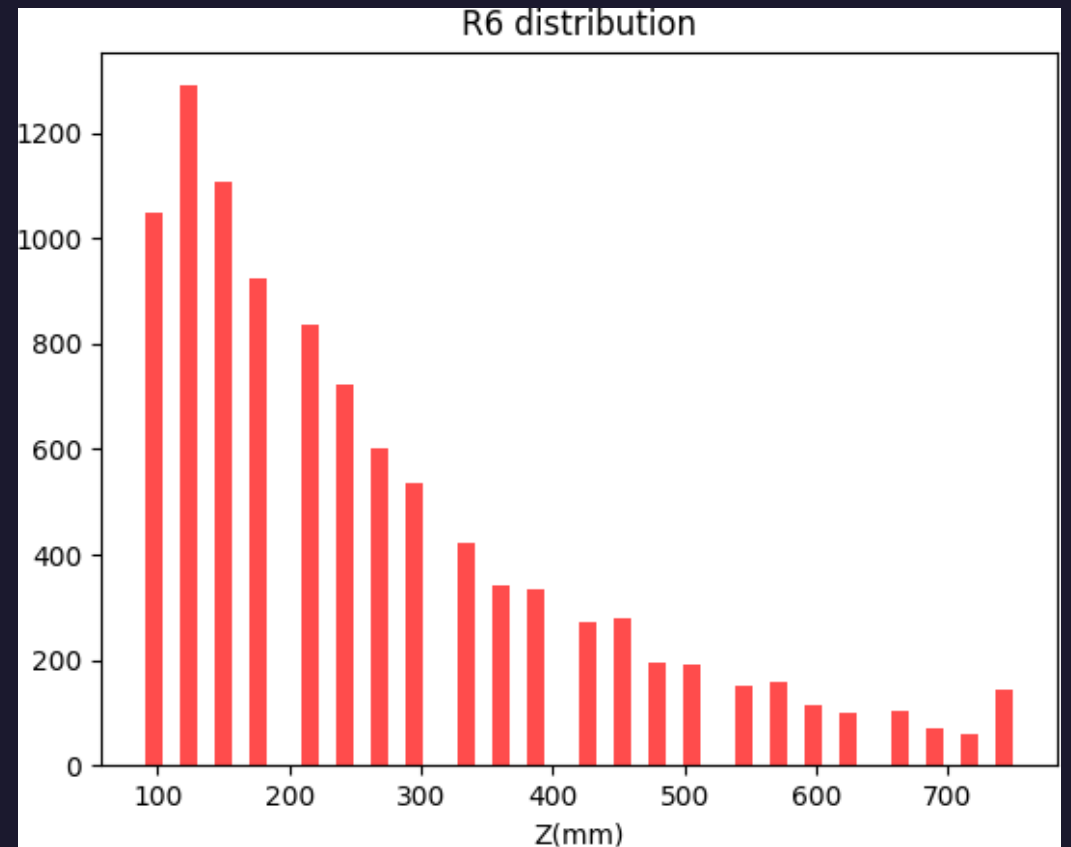
layer ~25th

Parameters distributions – R6(Z of the maximum energy deposited)

Electrons vs protons at 20 GeV



layer ~ 5th

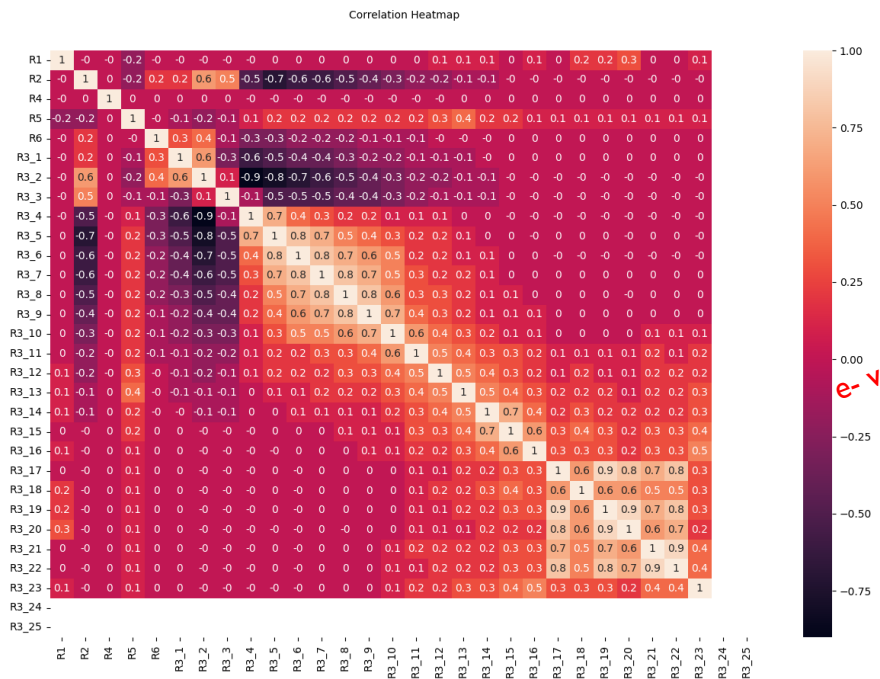


layer ~ 4th

The background of the slide is a dark blue gradient with a complex digital pattern. It features numerous vertical lines of varying heights and colors, including cyan, magenta, and orange. Scattered throughout are small circles and dots in white, red, and blue, some connected by thin lines, suggesting a network or data flow. The overall aesthetic is futuristic and technological.

AI Techniques Applied

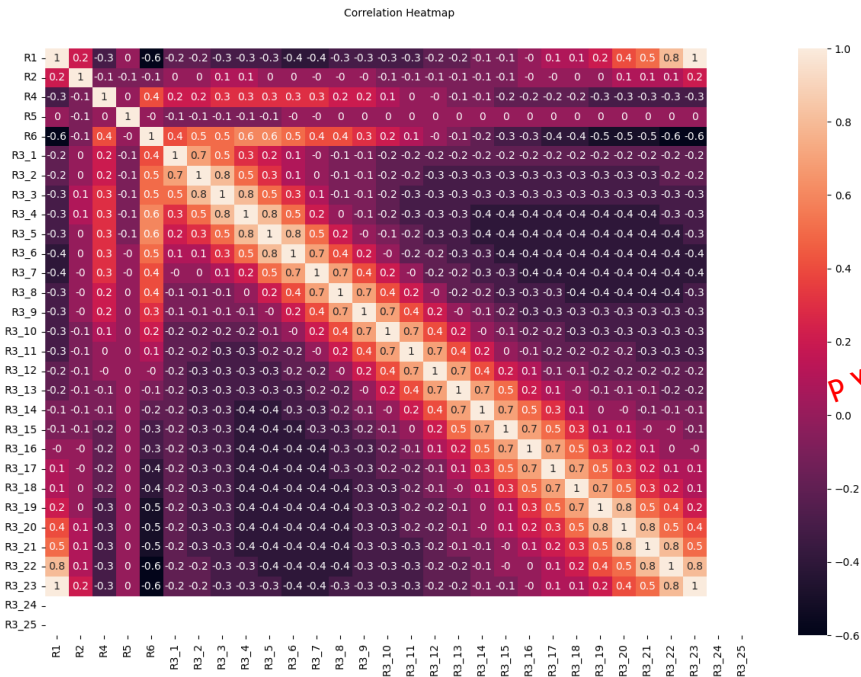
Correlation matrix



The correlation matrix shows that the selected variables have minimal to no interdependence, suggesting that they contribute independently to the analysis.

e- vertical tracks

Correlation matrix

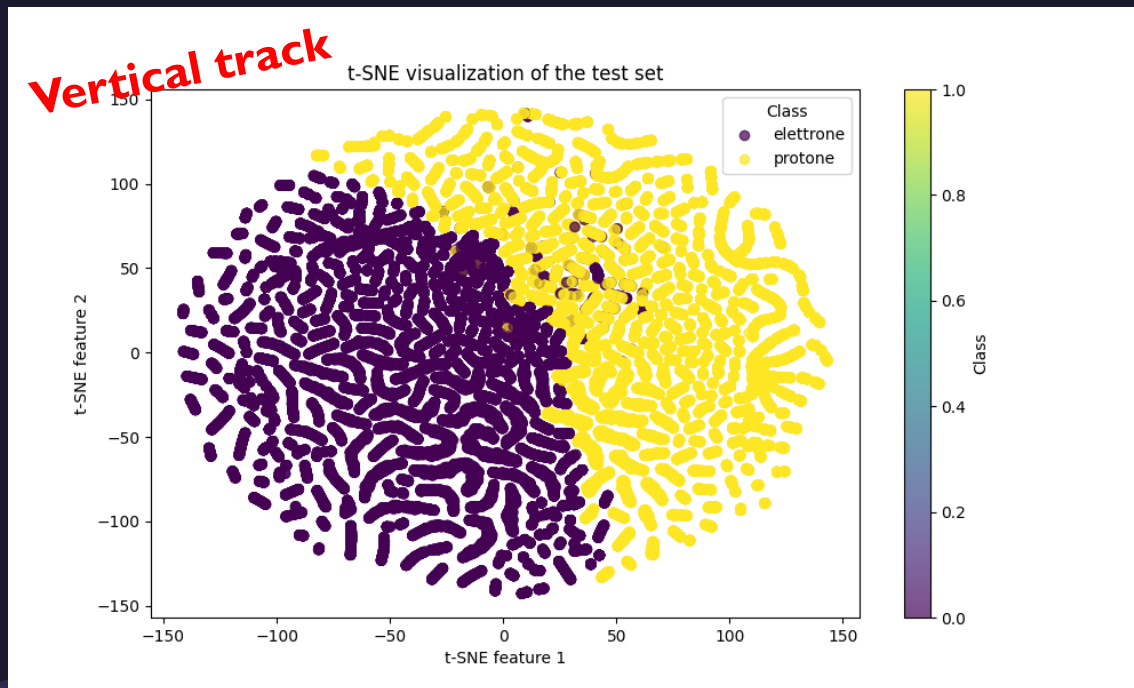


The correlation matrix shows that the selected variables have minimal to no interdependence, suggesting that they contribute independently to the analysis.

p vertical tracks

t-SNE

(t-distributed stochastic neighbour embedding)

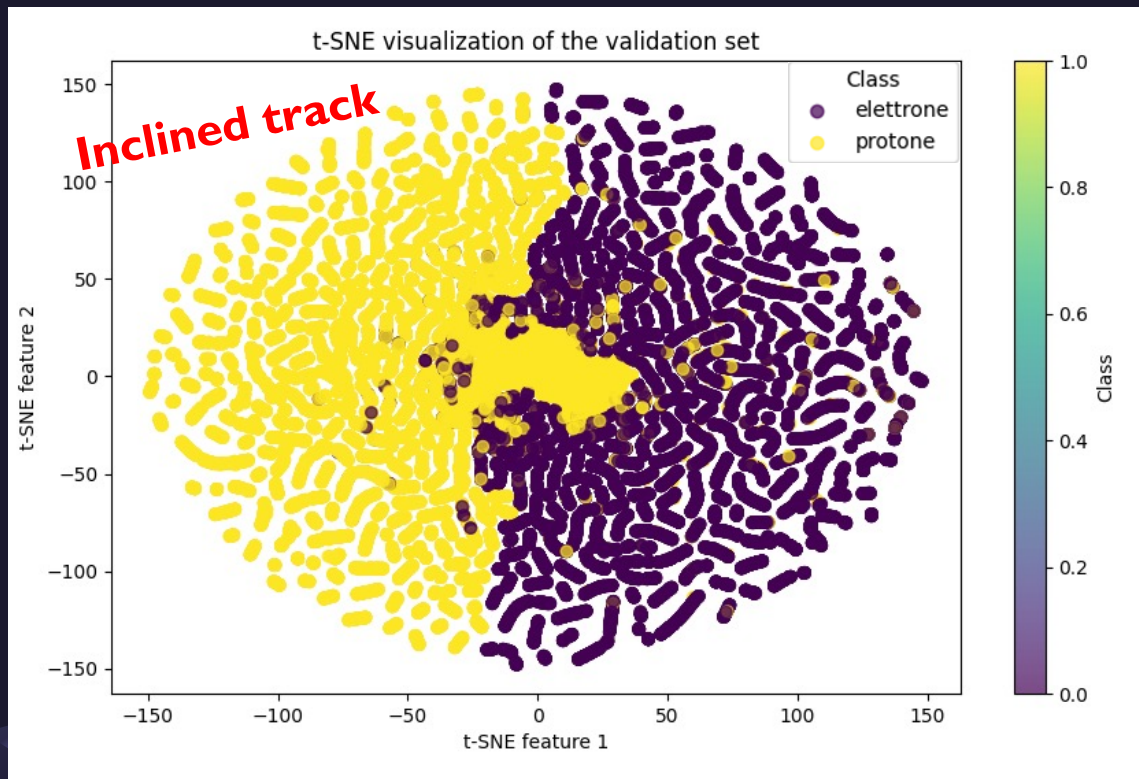


t-distributed stochastic neighbor embedding (t-SNE) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map

With a larger dataset of simulated particles, we can determine if the feature data can be divided into clusters before applying any machine learning algorithm.

t-SNE

(t-distributed stochastic neighbour embedding)

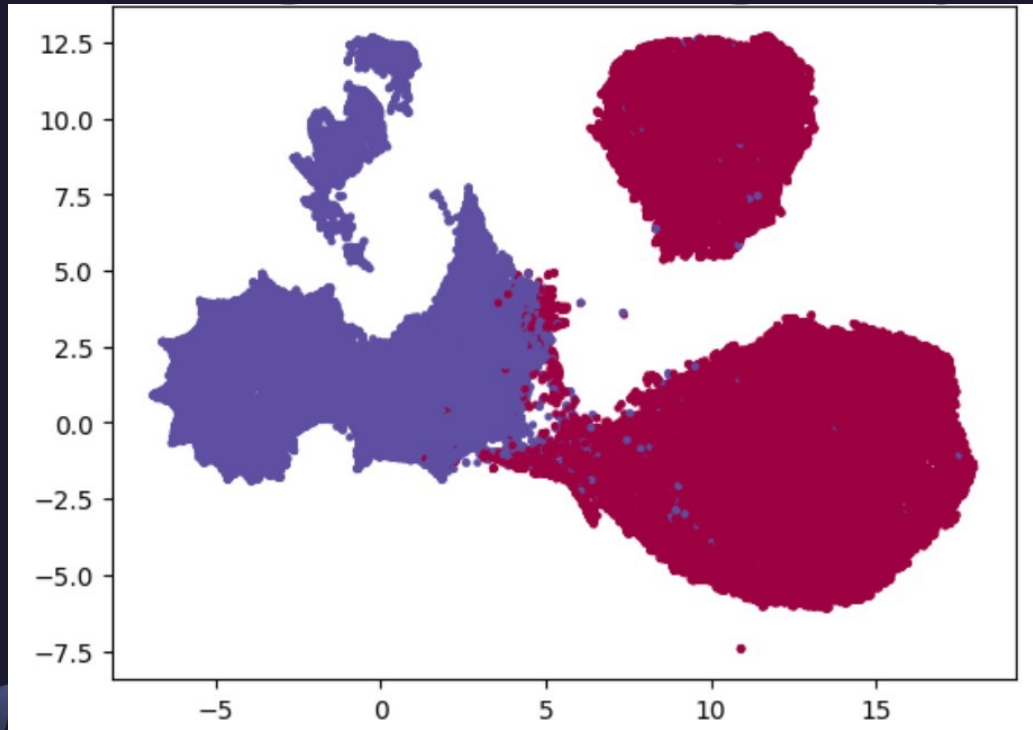


t-distributed stochastic neighbor embedding (t-SNE) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map

With a larger dataset of simulated particles, we can determine if the feature data can be divided into clusters before applying any machine learning algorithm.

UMAP

(Uniform Manifold Approximation and Projection for Dimension reduction)



UMAP is a dimension reduction technique that can be used for visualisation similarly to t-SNE, but also for general non-linear dimension reduction.

XGBoost algorithm



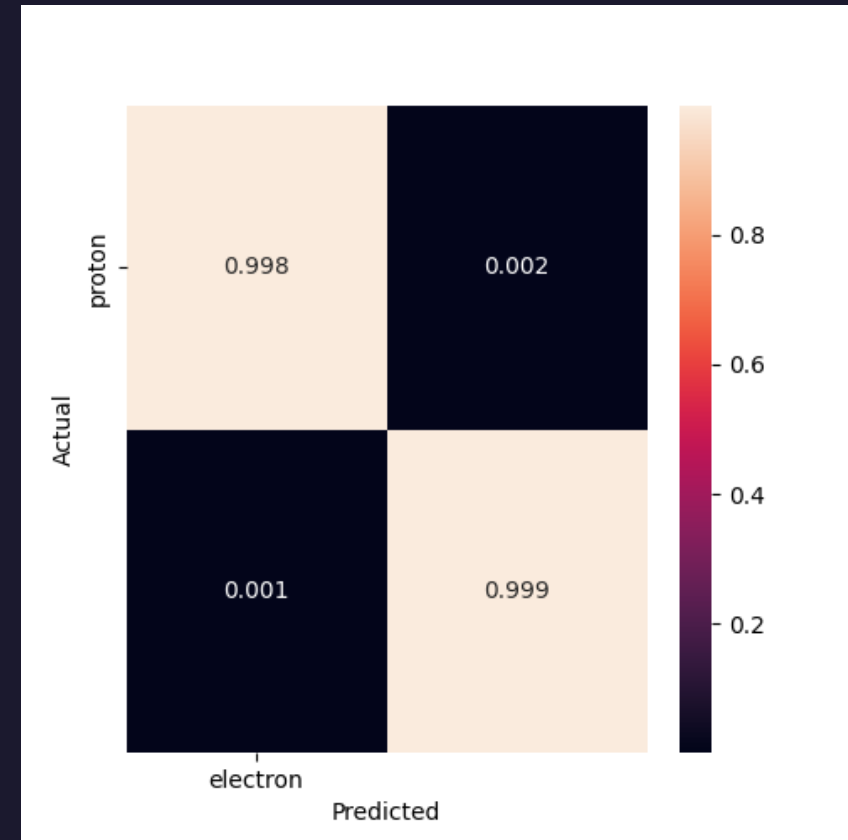
- The main goal of XGBoost is to find the best balance between the complexity of the trees (how deep and complex they are) and the accuracy of the prediction
- XGBoost is based on decision trees, similar to random forest. The difference lies in the fact that XGB trains these trees one at a time. It starts 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 and performance on vertical tracks

Training a machine learning algorithm using XGBoost yielded the following results:

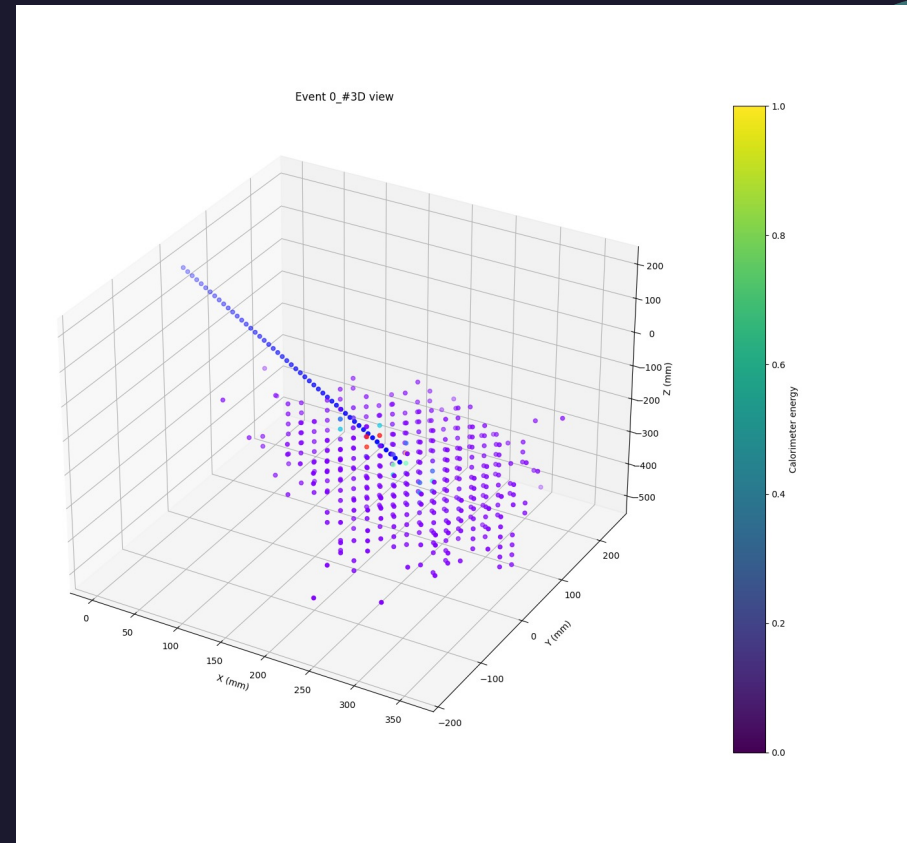
- Accuracy XGB Classifier: 99.85%
- Recall XGB Classifier: 99.90%
- Precision XGB Classifier: 99.80%



XGBoost algorithm

Results and performance on inclined tracks

Electron event of 20 GeV, with momentum generated within a cone having an angular opening between -30 and 30 degrees

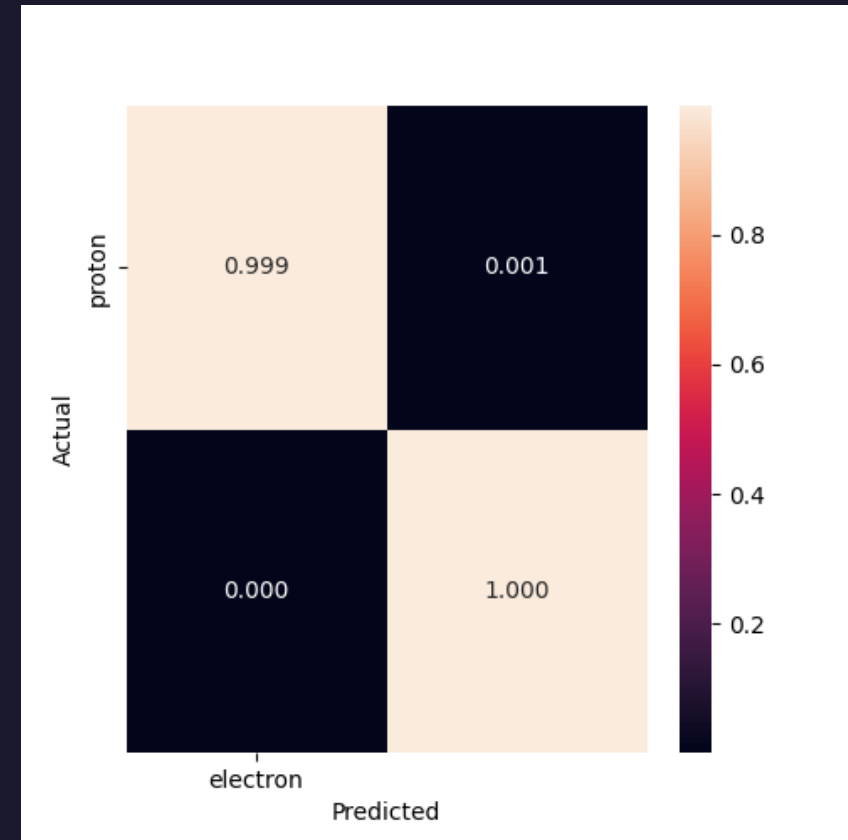


XGBoost algorithm

Results and performance on inclined tracks

Train of XGBoost algorithm on a sample of 100k events

- Accuracy XGB Classifier: 99.95%
- Recall XGB Classifier: 99.95%
- Precision XGB Classifier: 99.94%

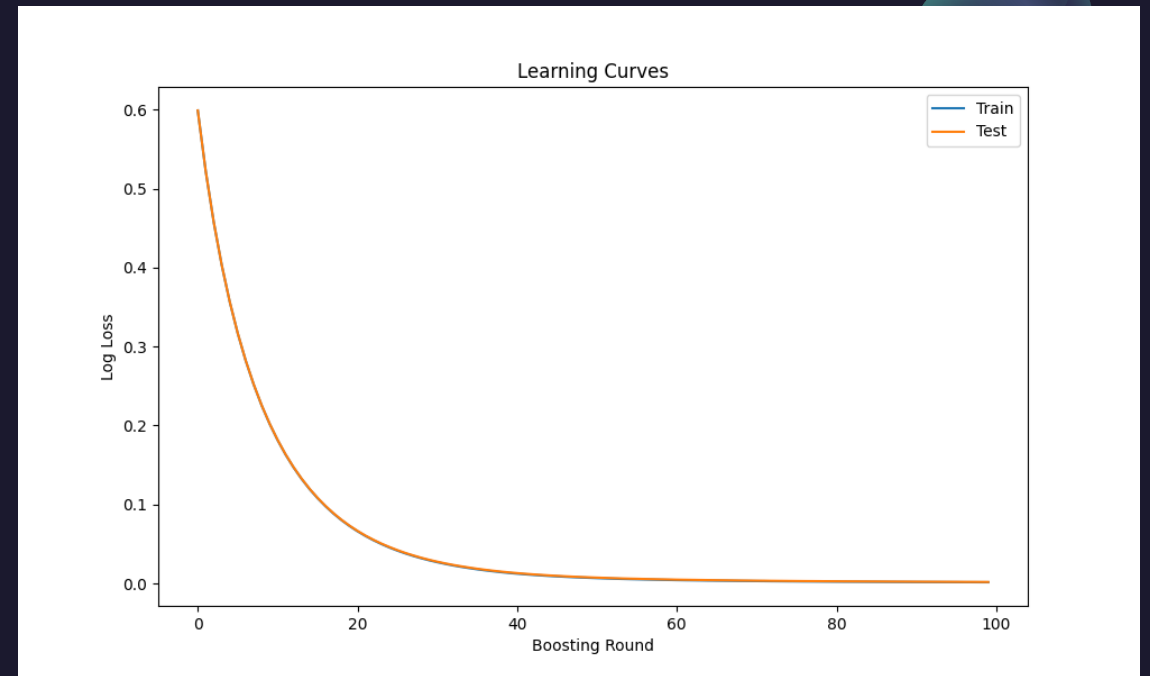



XGBoost algorithm

Results and performance on inclined tracks

Train of XGBoost algorithm on a sample of 100k events

- Accuracy XGB Classifier: 99.95%
- Recall XGB Classifier: 99.95%
- Precision XGB Classifier: 99.94%





EXplanable
Artificial
Intelligence (XAI)
SHAP Analysis

EXplanable Artificial Intelligence (XAI)

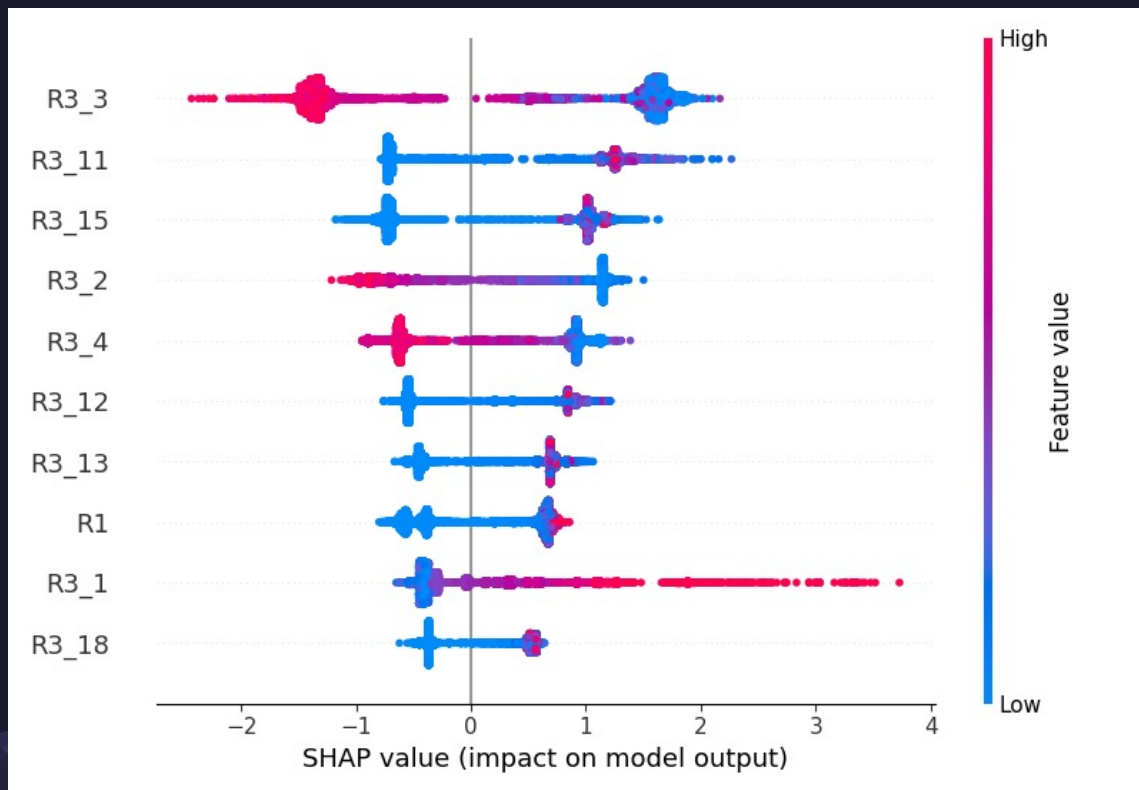
SHAP Analysis



- 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

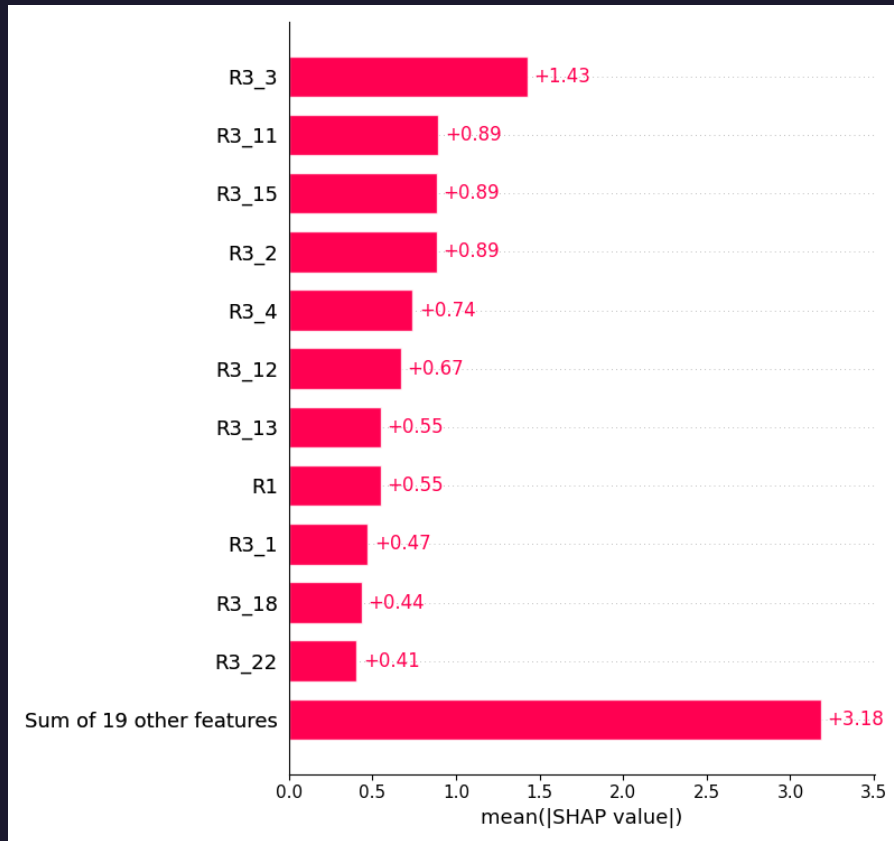
Beeswarm plot



- 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.

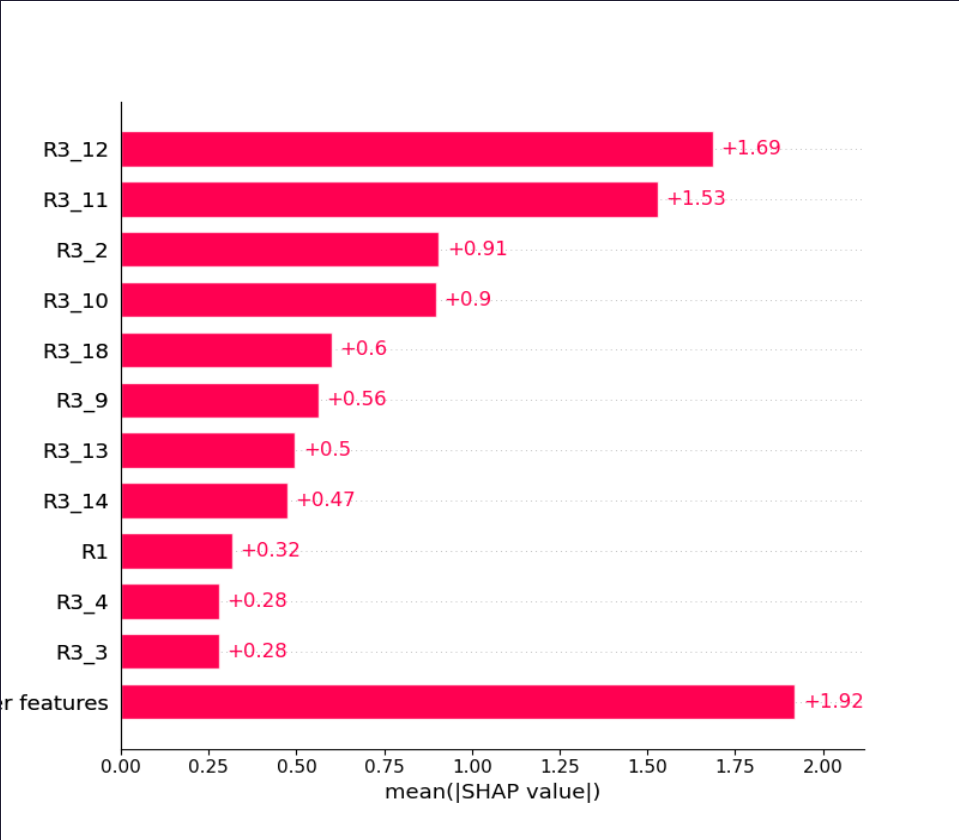
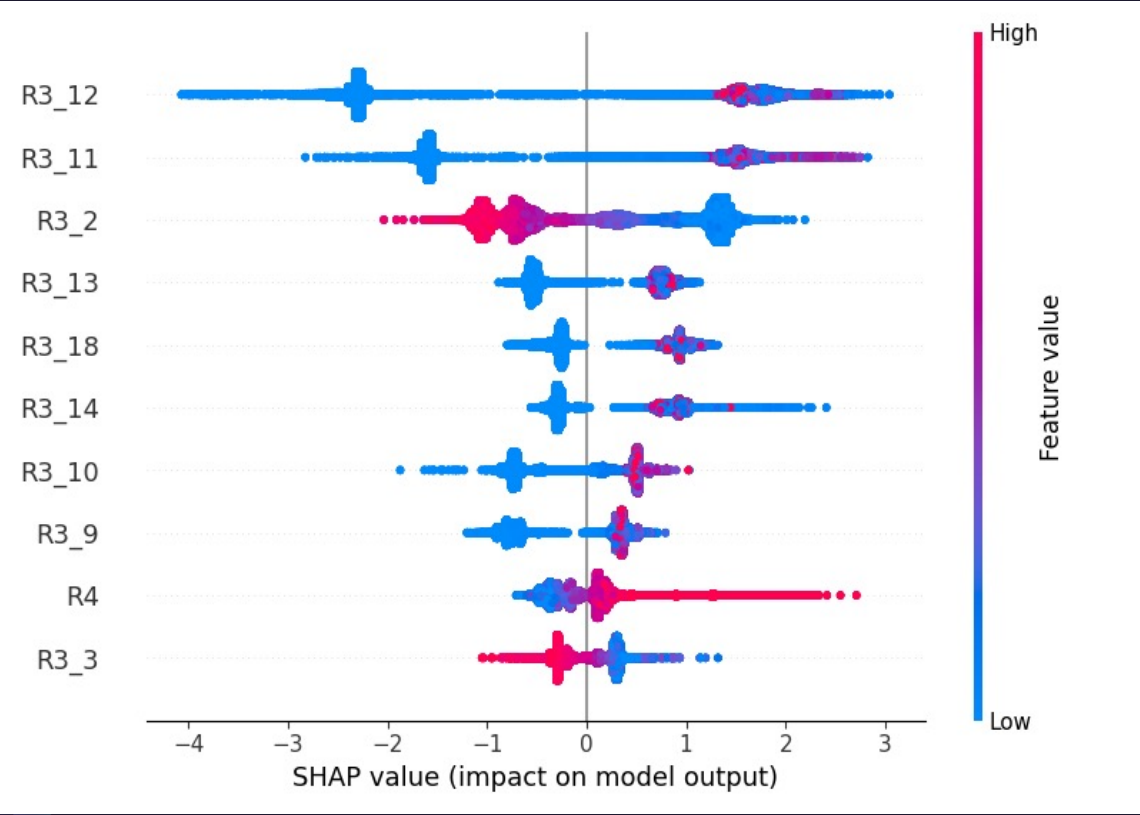
SHAP Analysis

Bar plot



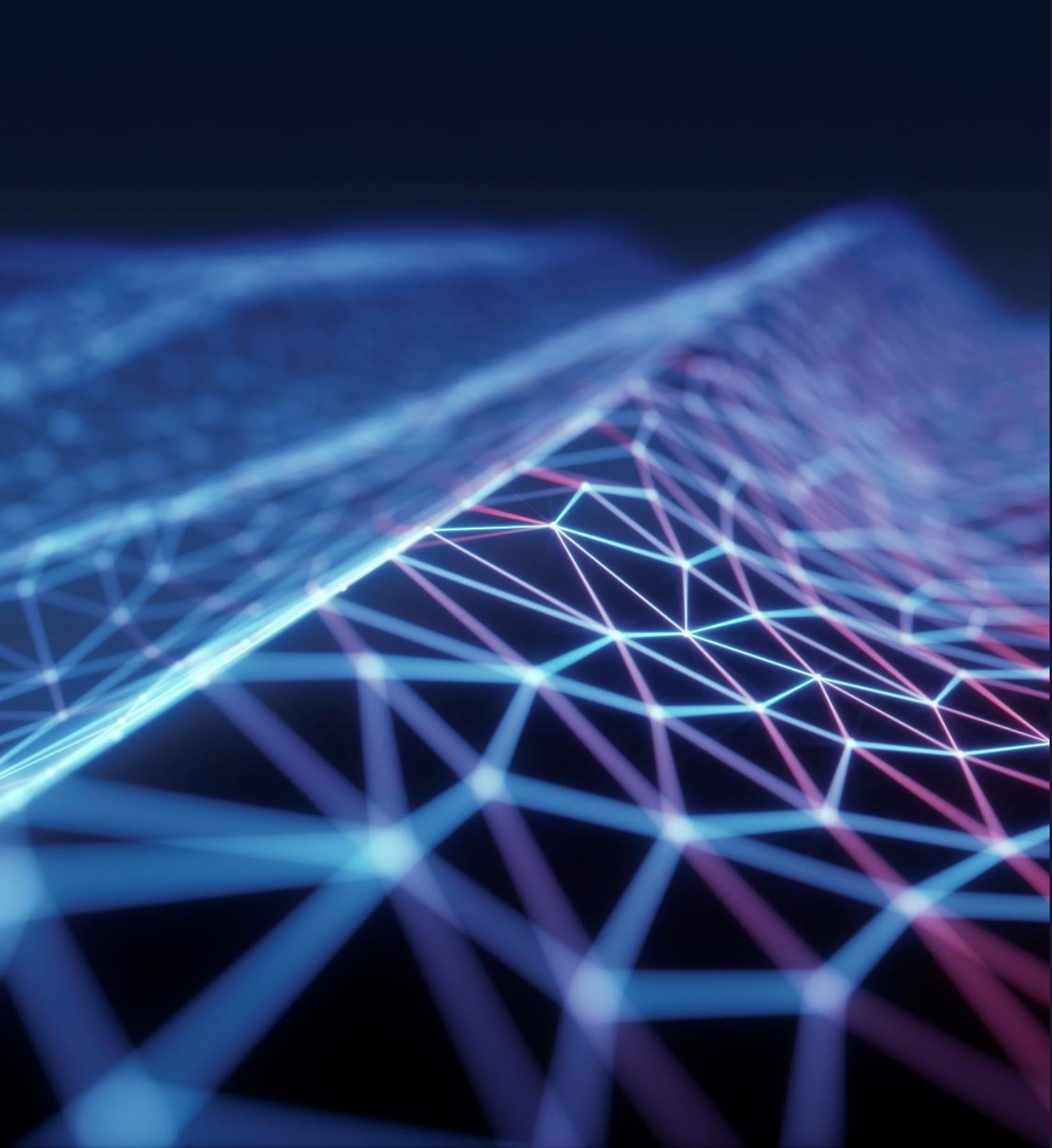
- The simplest starting point for global interpretation with SHAP is to examine the mean absolute SHAP value for each feature across all of the data that quantifies, on average, the magnitude of each feature's contribution
- Features with higher mean absolute SHAP values are more influential.

SHAP Analysis on inclined tracks



Next steps and possible new scenarios

- Evaluation of the XGBoost classifier on particle datasets with energy distributions that follow a power spectrum.
- Now we will work on a new dataset from the HERD detector simulation. In this initial phase, we are carefully studying the simulation's output before applying any AI algorithm
- Explore the potential of training Convolutional Neural Networks (CNNs) with GPU acceleration to enhance computational efficiency.
- Investigate alternative machine learning paradigms, including reinforcement learning and unsupervised learning, to broaden the scope of model development.
- Test and validate deep learning models based on transformer architectures for their applicability in the context of particle energy distribution analysis.

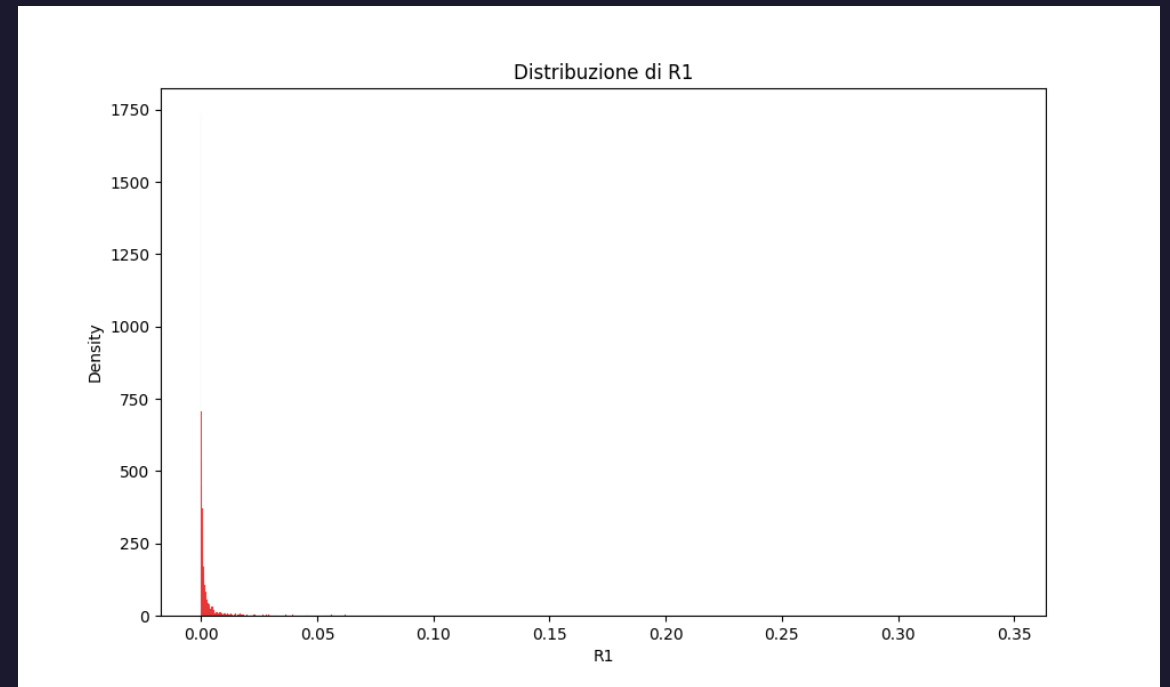
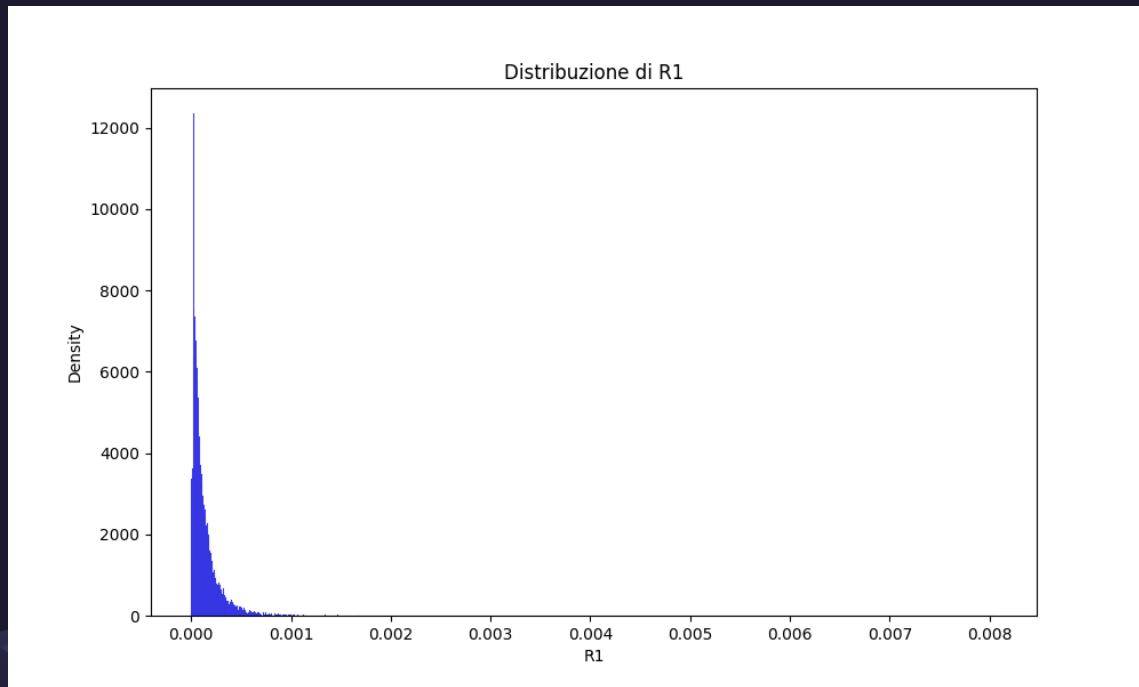


Thank you

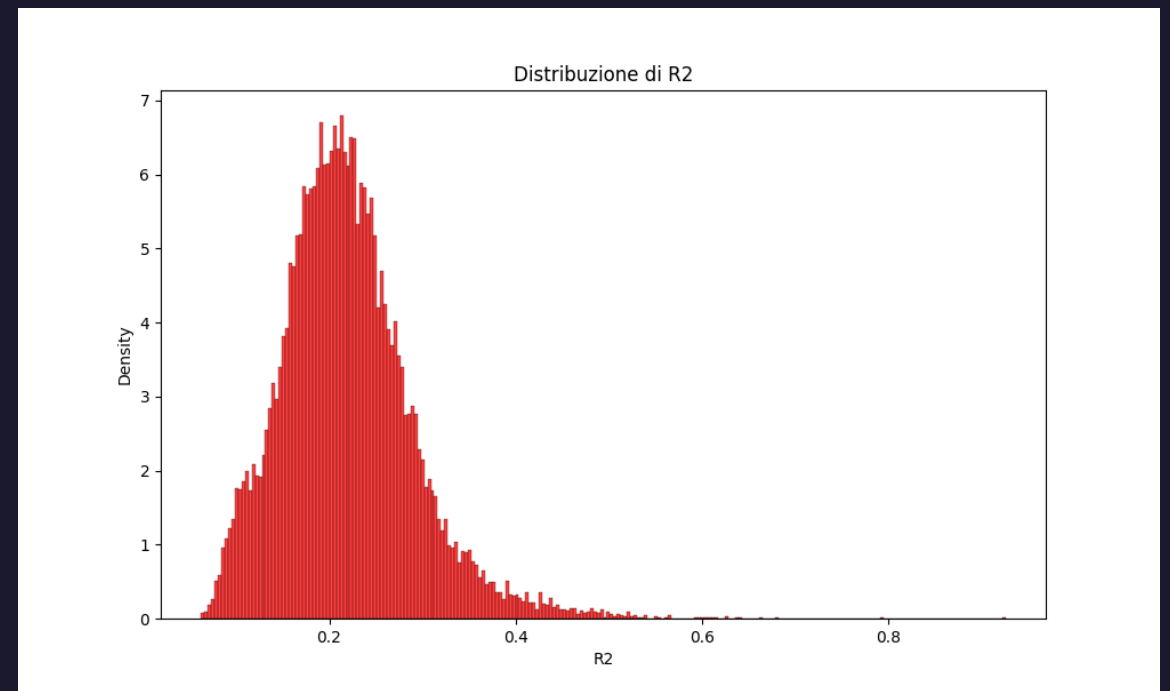
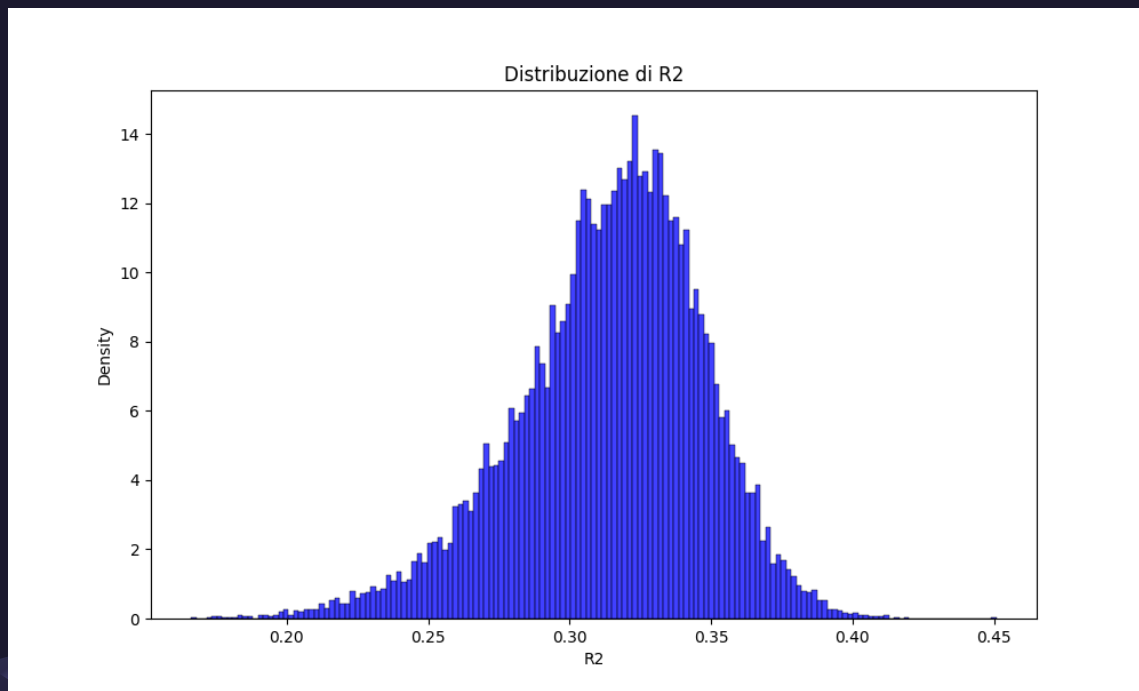


Backup

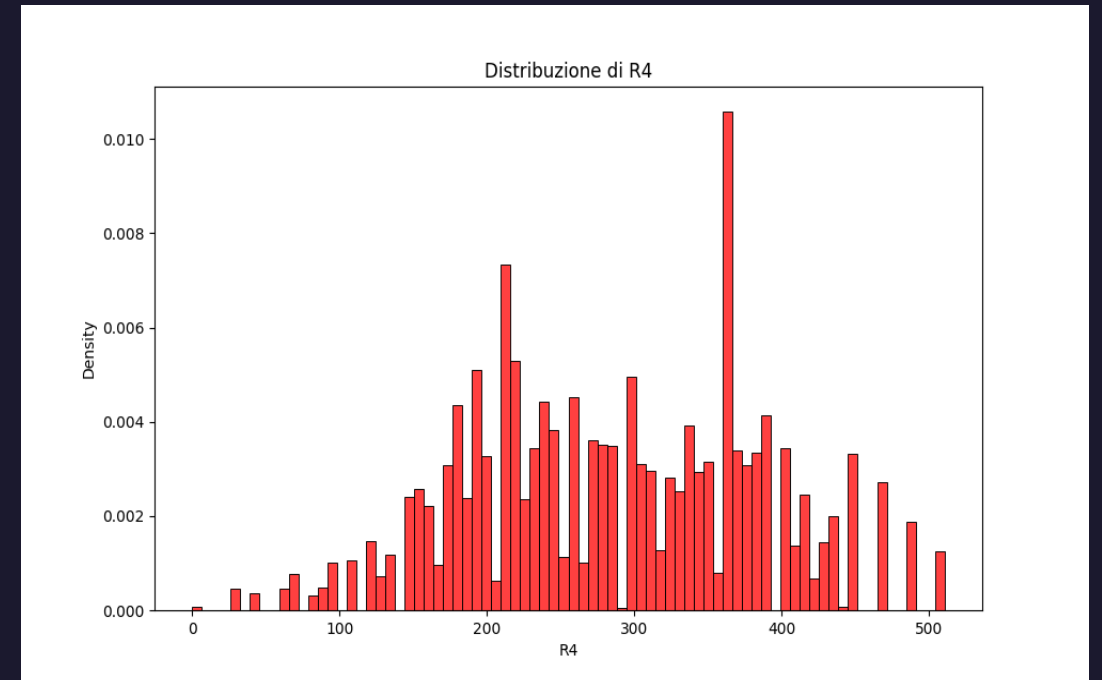
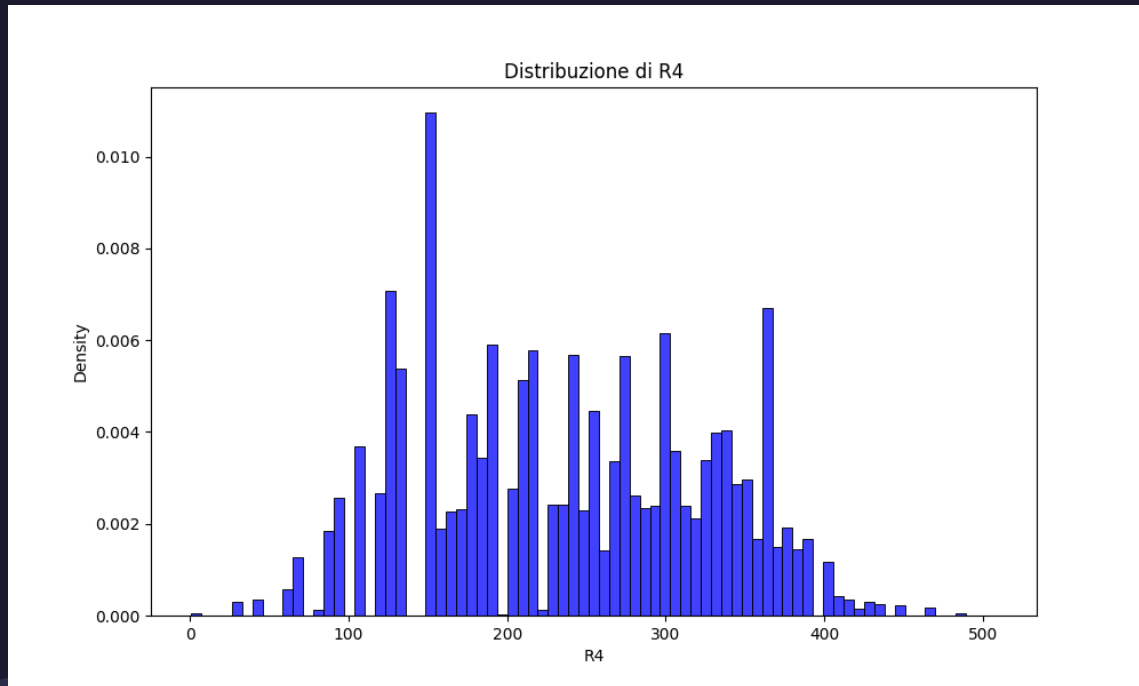
Parameters distributions – R1 (Ratio of energy deposited in the last layer to the total energy deposited) Electrons vs protons at 20 GeV inclined



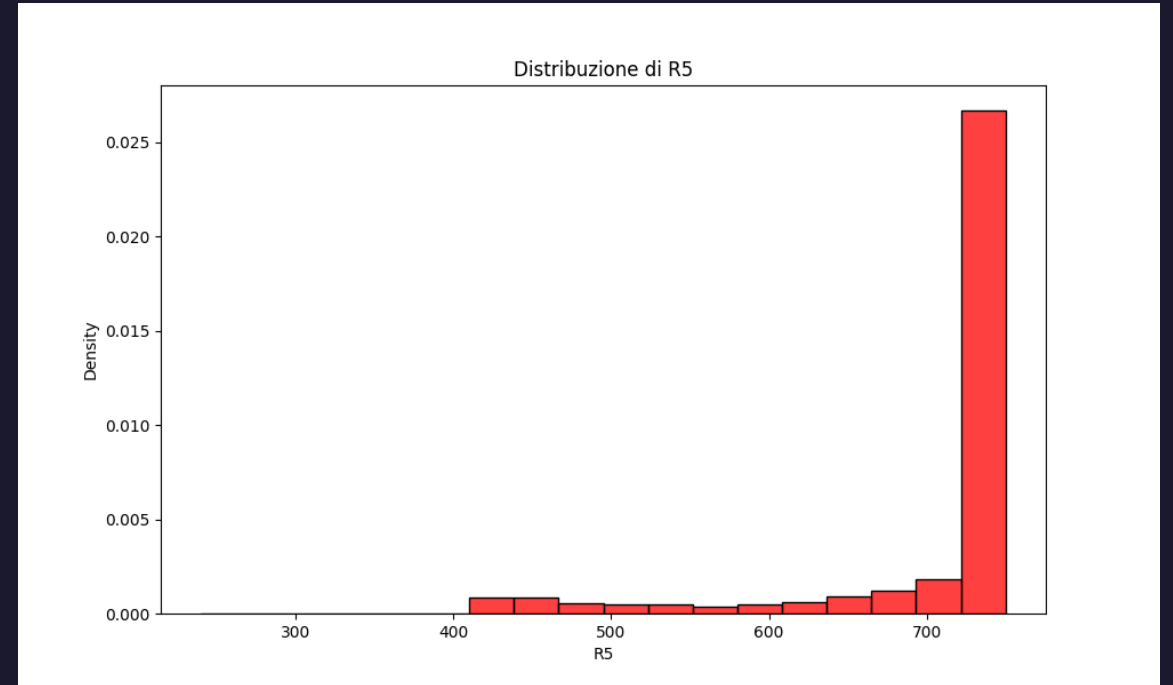
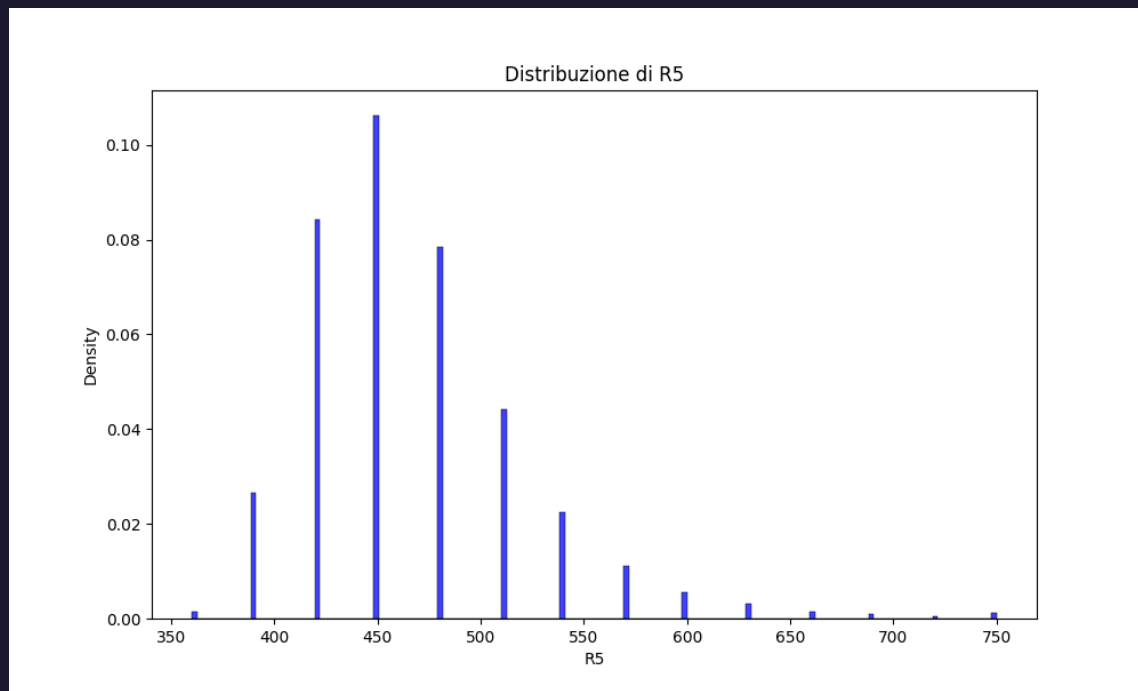
Parameters distributions – R2(maximum energy deposited, in a layer, to the total energy deposited) Electrons vs protons at 20 GeV inclined



Parameters distributions – R4(containment radius) Electrons vs protons at 20 GeV inclined



Parameters distributions – R5(Z of the last hit layer) Electrons vs protons at 20 GeV inclined



Parameters distributions – R6(Z of the maximum energy deposited)

Electrons vs protons at 20 GeV inclined

