Convolutional Neural Network Processing of Radio Emission for Nuclear Composition Classification of Ultra-High-Energy Cosmic Rays

Cosmina Mihoreanu[2], Tudor Calafeteanu[2], Gina Isar[1], Emil-Ioan Slușanschi[2]

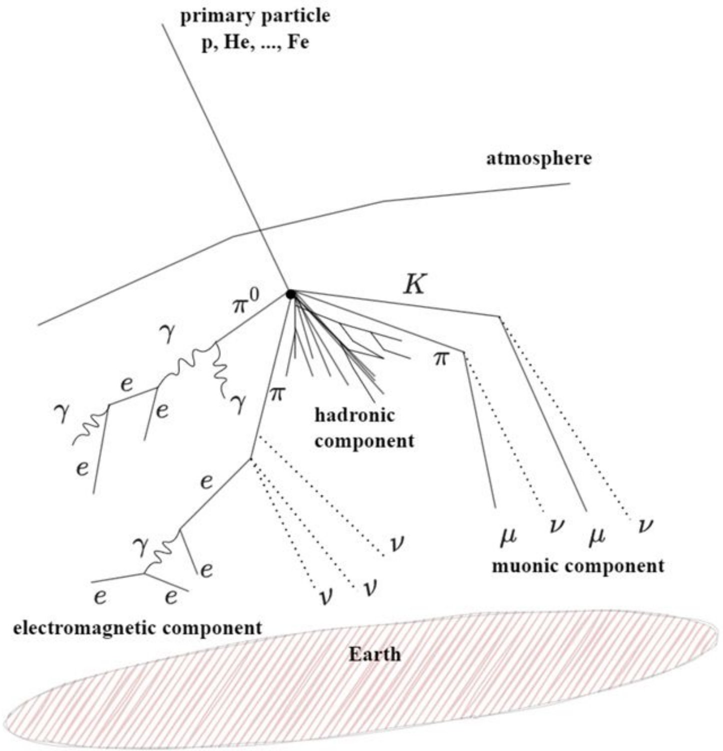
[1]Institute of Space Science (ISS), Bucharest - Magurele, Romania
[2]Politehnica University of Bucharest, Romania

Workshop on Machine Learning for Analysis of High-Energy Cosmic Particles, 27-31 January 2025, University of Delaware

Outline

- Motivation:
  - Bachelor’s Thesis in Computer Science (with application in Astroparticle Physics)
- Case study
  - Previous work: Nuclear Composition Classification of 2 Primary Particles
  - Current work and further steps: Nuclear Composition Classification of UHECRs for 4 Primaries
- Conclusion
- Acknowledgements
- References

Introduction: Extensive air showers

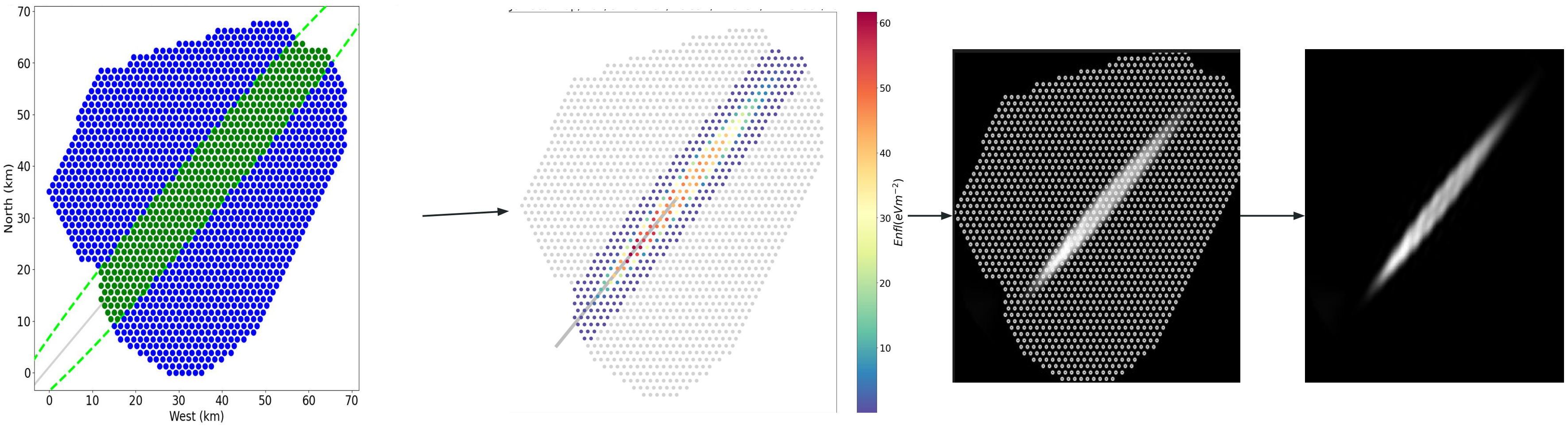


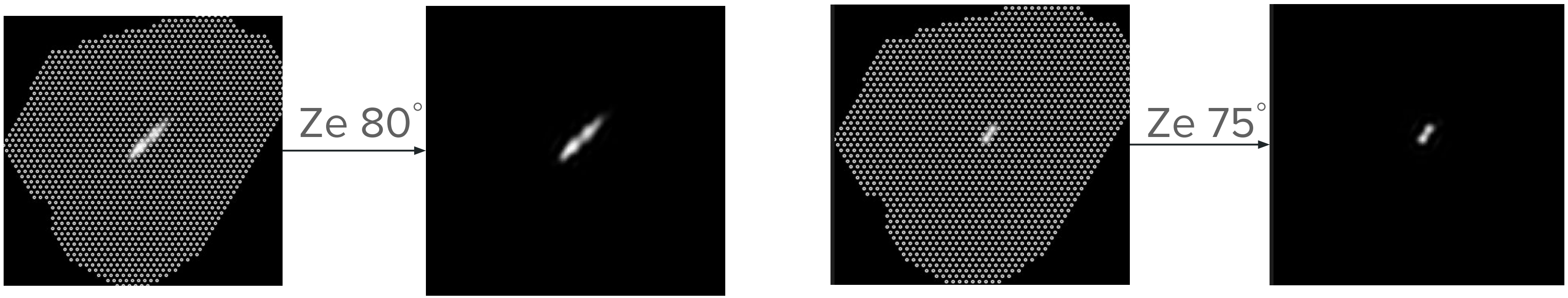
Ref. 6

- Primary cosmic ray particles produce a cascade of subatomic particles when entering the atmosphere.
- The radiation emitted by such air showers can be recorded with radio antennas at the Pierre Auger experiment.

Previous work: Radio imaging

- Radio imaging technique (grayscale coloring on the energy fluence)





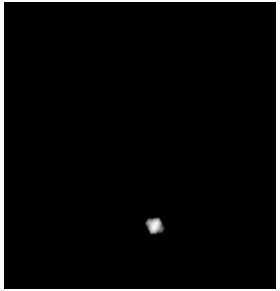
Ref. 6

Current work: Radio imaging techniques

4 radio imaging techniques:

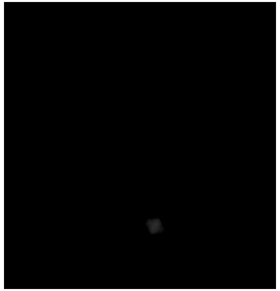
- **Max local method:** Each RD's energy fluence is MinMax scaled, where the maximum value is determined per simulation.

- **Max global method:** Each RD's energy fluence is MinMax scaled, where the maximum value is determined across all simulations.

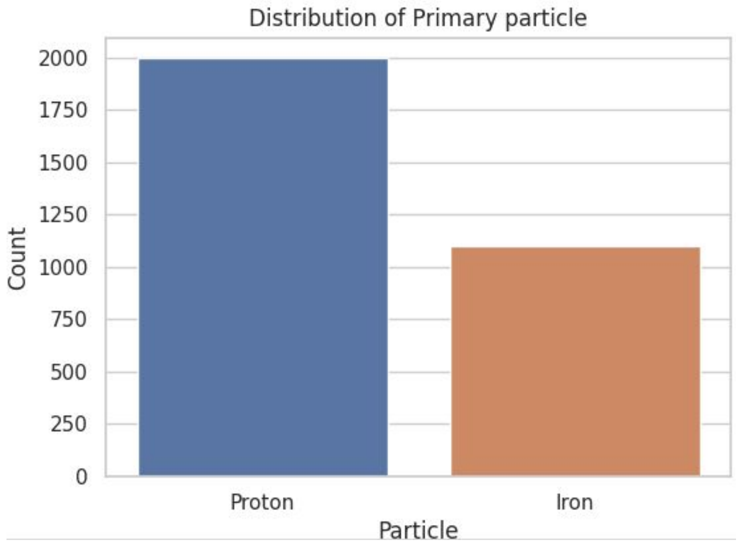


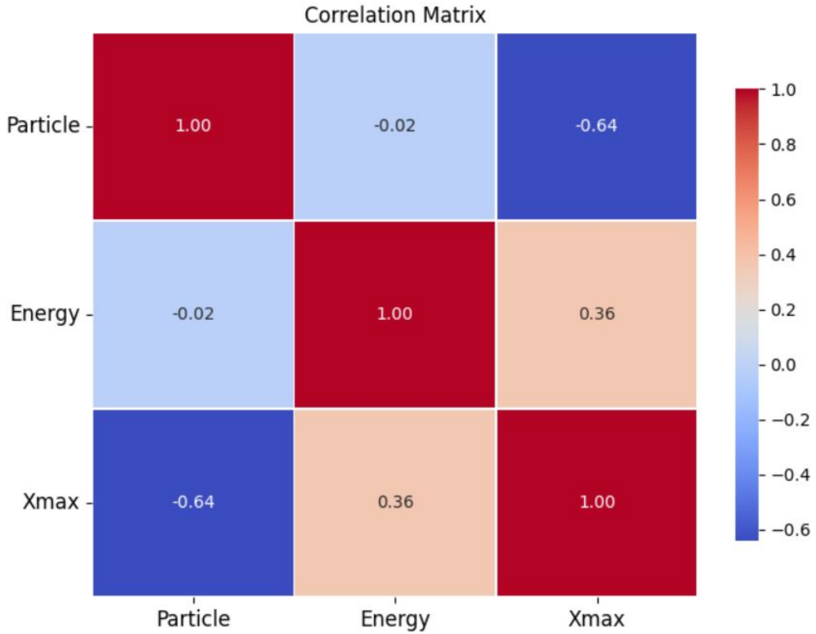
- **Log max local method:** Similar to the Max local method, but with a log10 transformation applied to the energy fluence.

- **Log max global method:** Each RD's energy fluence is MinMax scaled; maximum value is determined across all simulations - $\approx 4.24^{*}10^{5}eVm^{-2}$ (iron, $10^{20}eV$, vertical, South-East). log10 transformation is then applied to the energy fluence.



Previous work: Data exploration and preprocessing





- Work described in the article referenced at (6)
- Proton - 0, Iron - 1
- Chemical composition of the UHECR has a greater effect on the depth of the shower maximum, than the energy with which it arrives in our atmosphere

Ref. 6

Previous work: Dataset description

Next steps:

1. **Data exploration and preprocessing:** This includes tasks such as applying log10 transformations and feature scaling.
2. **Splitting the data:** Dividing the data into training and test sets (70% - 30%).
3. **Create the dataset:** Dataset used for training and testing
4. **Training of the convolutional neural network (CNN):** Using a modified architecture of a ResNet-18 CNN to train on the labeled training data for image classification between primary particles; ResNet-18 was chosen due to its simplicity and wide applicability in image recognition tasks; it also gave the best preliminary results
5. **Evaluation of the CNN**

Features:

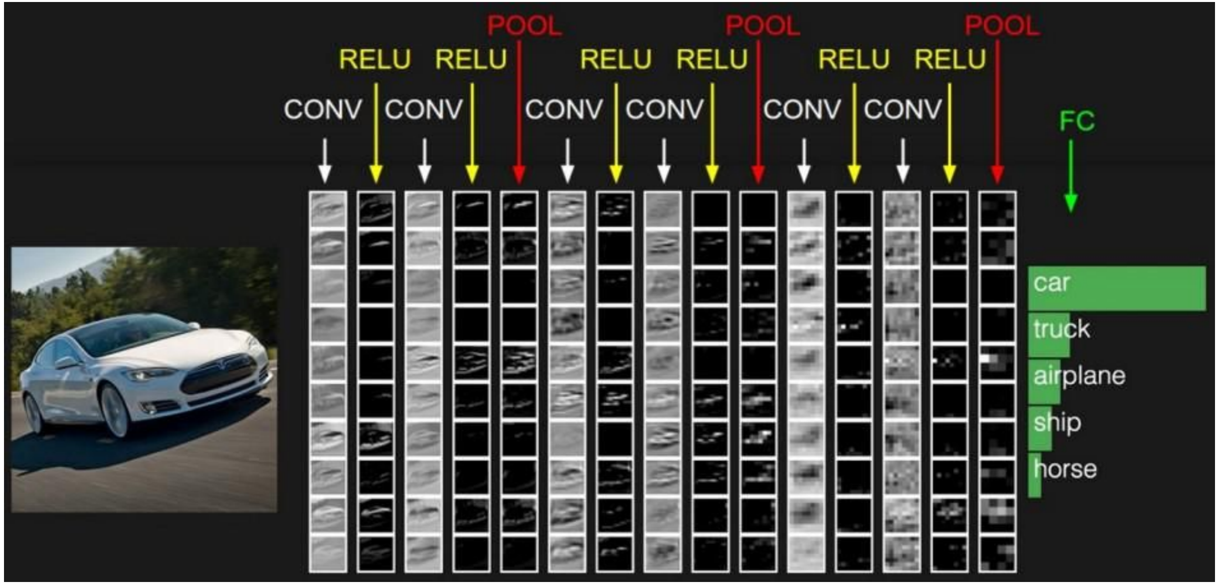
- **4 numerical features:** Zenith (MinMax scaling), Azimuth (MinMax scaling), Energy (MinMax scaling), Xmax (Standard scaling)
- **4 images:** Max local method, Max global method, Log max local method, Log max global method

Labels: Particle type (Proton - 0, Iron - 1)

Previous work: Convolutional Neural Networks

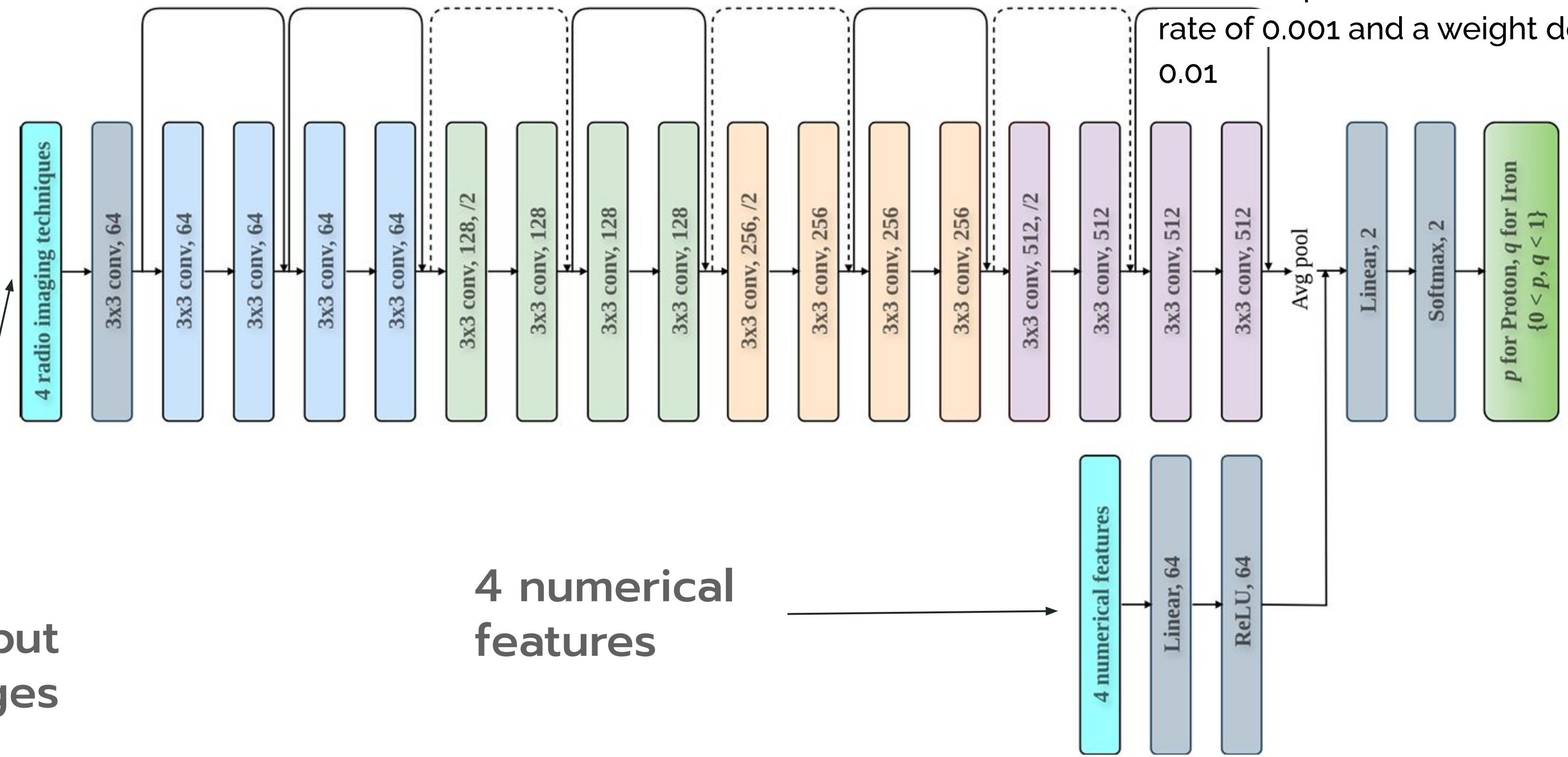
- A convolutional neural network (CNN) is an algorithm used in image recognition and processing that is inspired by the biological processes in the visual cortex of animals. They are made up of neurons that have learnable weights and biases.

- The model we use, ResNet-18, is a public CNN, 18 layers deep, with the first convolutional layer switched for one with 4 input channels (for each imaging method)



Previous work: CNN architecture

CrossEntropyLoss loss function
AdamW optimizer with a learning rate of 0.001 and a weight decay of 0.01



4 input images

4 numerical features

Ref. 6

Previous work: Training and evaluation

- An epoch in machine learning is one complete pass through the entire training dataset. For example, if we are training a model on a 1000 samples dataset, one epoch would involve training on all 1000 samples at one time.

- The model's weights are updated based on the training data during each epoch, and the model's performance is evaluated on the training and validation sets.

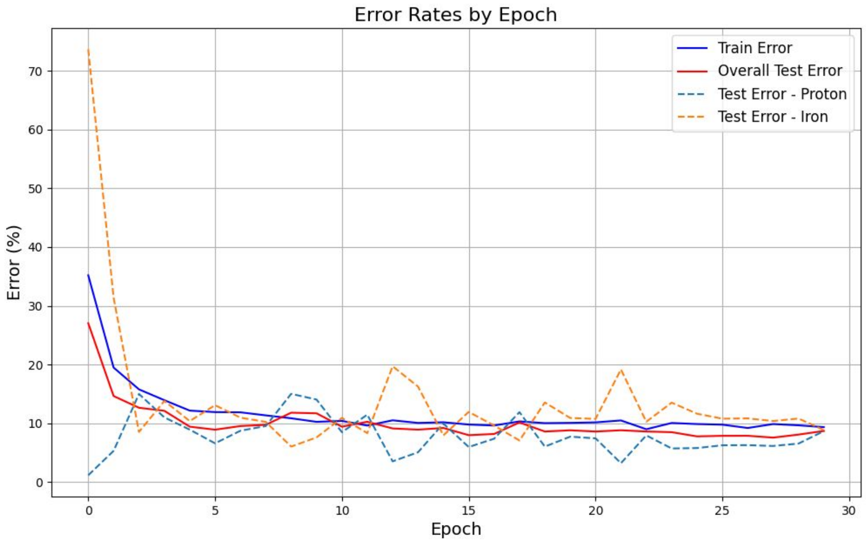
True Positive (TP): - predicted label = actual label	**False Positive (FP):** - samples incorrectly labeled as a given label
False Negative (FN): - samples with a given label that have been incorrectly labeled	**True Negative (TN):** - predicted label != actual label

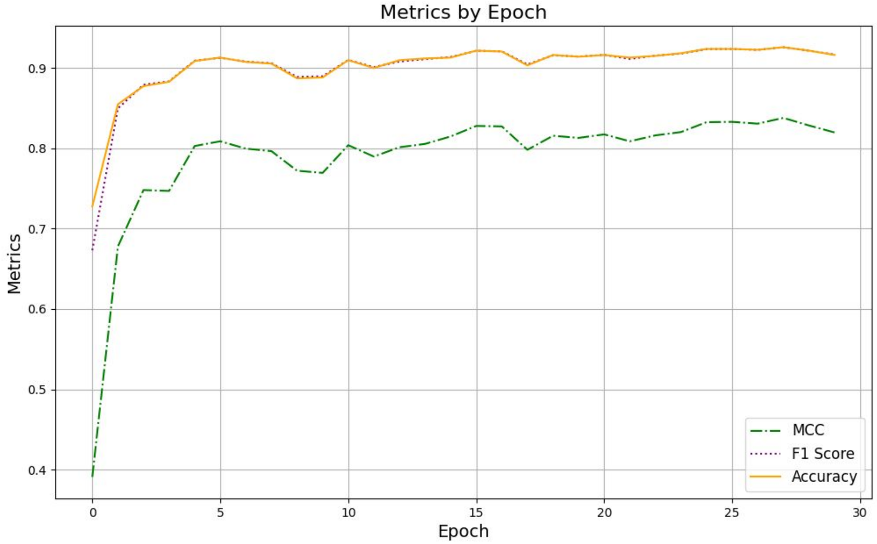
$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Previous work: Nuclear composition classification - Results





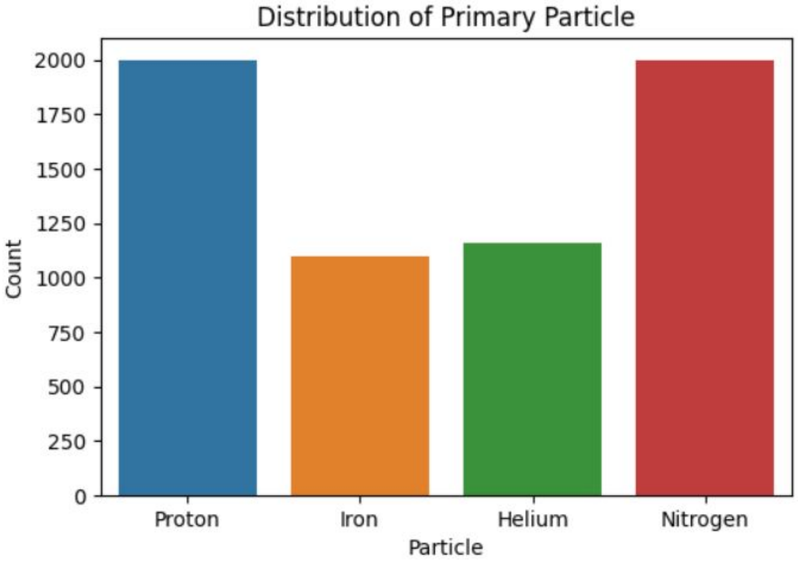
- The **test errors** for both proton and iron follow a similar decreasing trend, stabilizing around 10%

- The **MCC** increases rapidly and stabilizes around 0.8, indicating a strong correlation between predicted and actual values.

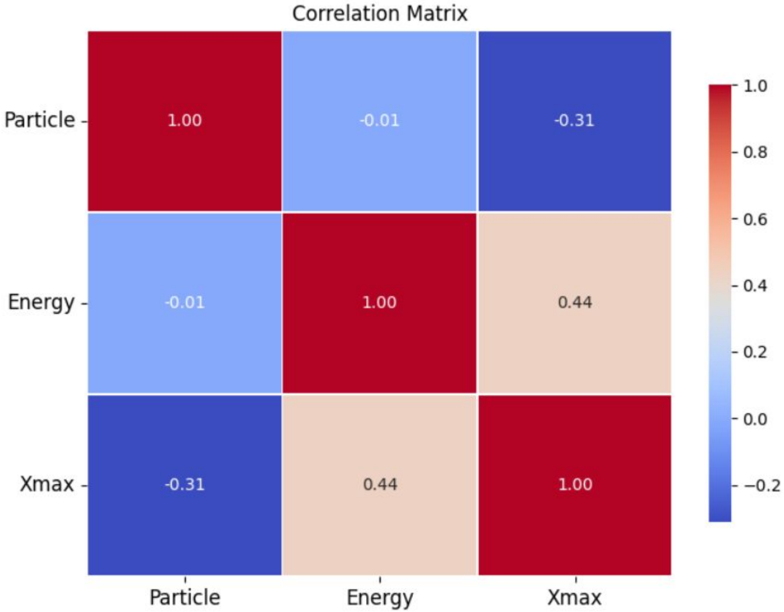
- **Accuracy** and **F1 Scores** increase sharply at the beginning and stabilize around 0.9, indicating a good balance between precision and recall, and that the model performs well on the classification task.

Ref. 6

Current work: Data exploration and preprocessing

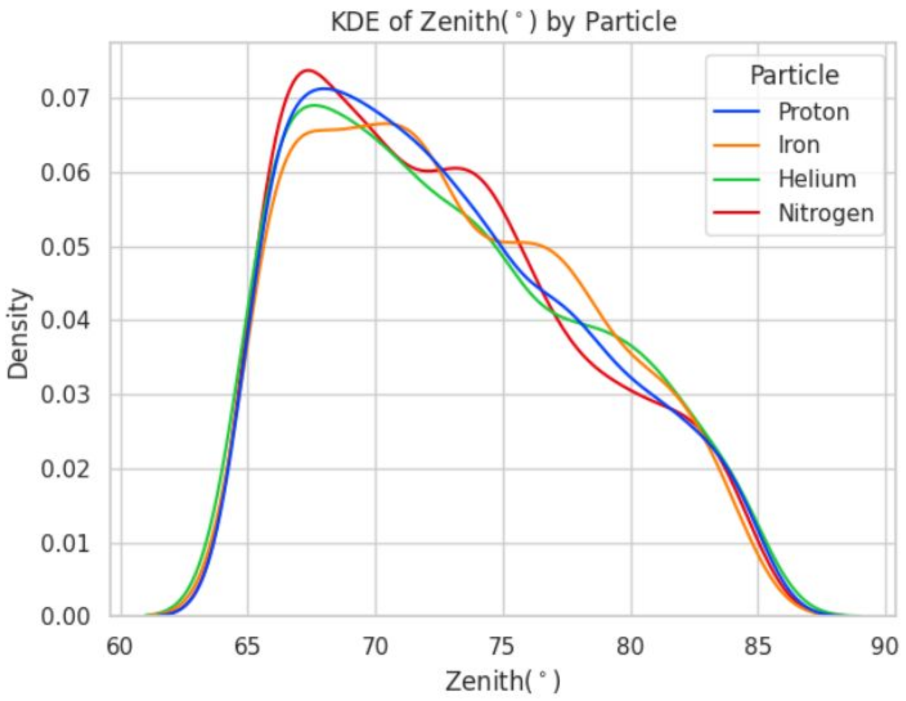


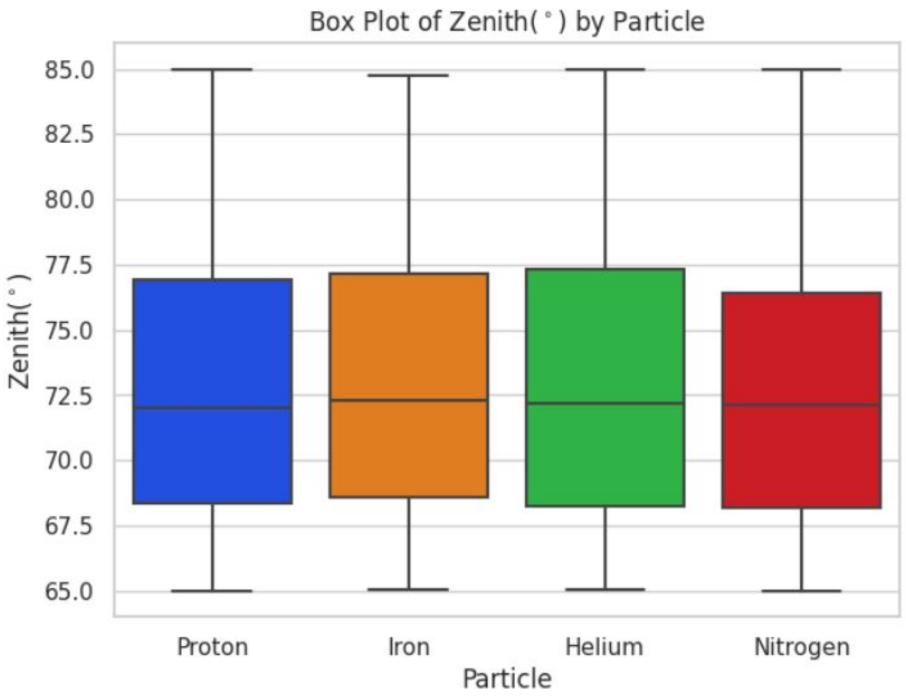
Proton: 1996
Iron: 1099
Helium: 1161
Nitrogen: 2000



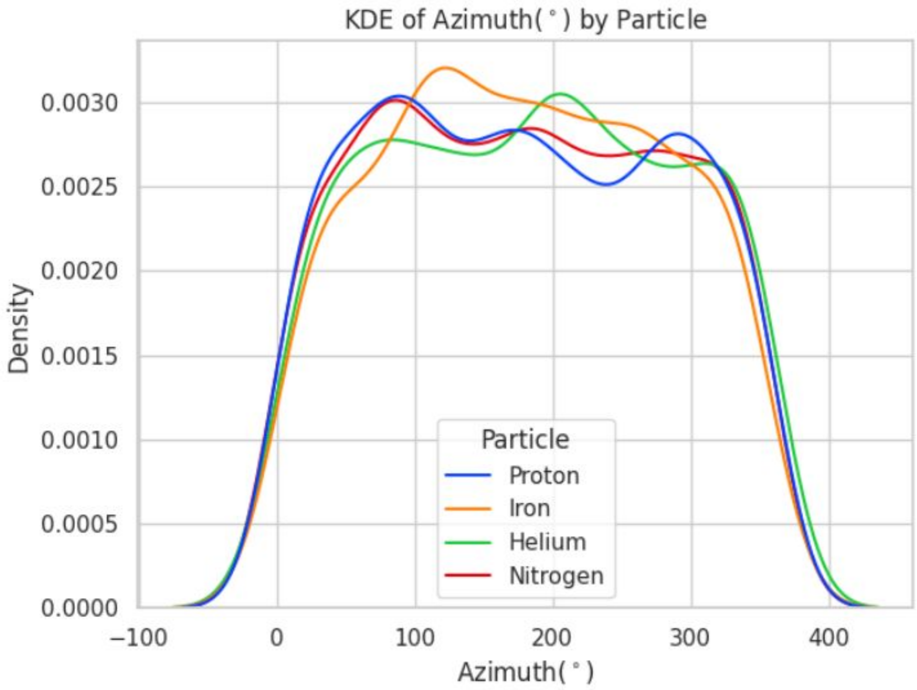
- Proton - 0, Iron - 1, Helium - 2, Nitrogen - 3
- Weaker correlation between nuclear composition and shower maximum depth and energy, indicating similarities between the previous primaries and the newly included ones.

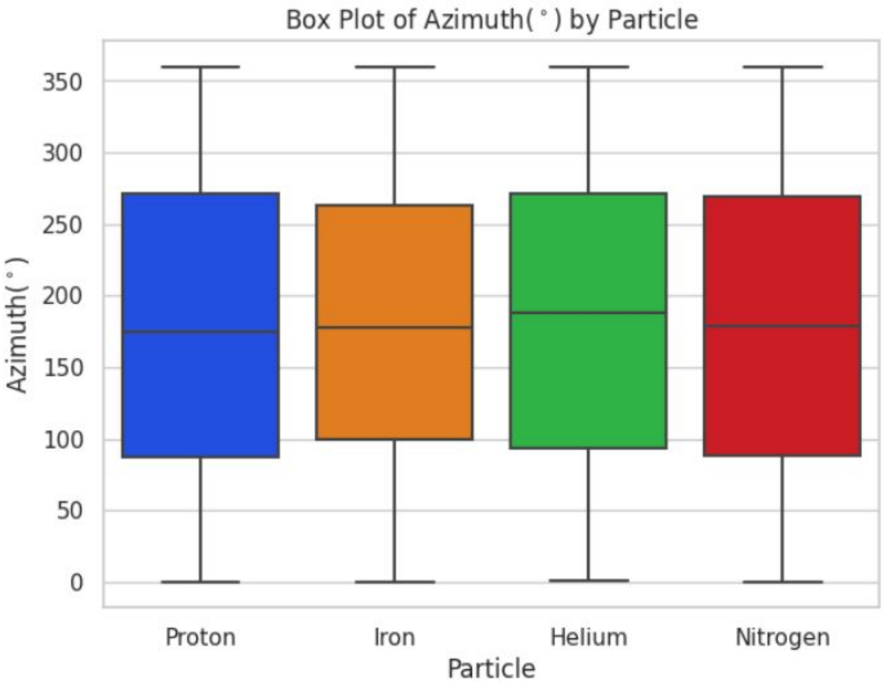
Current work: Data exploration and preprocessing



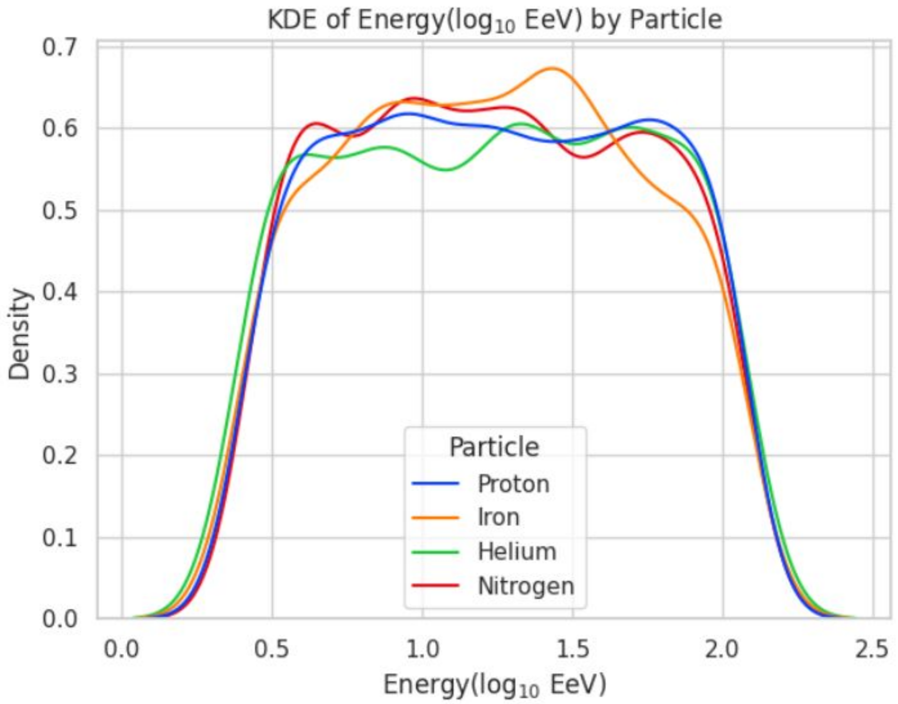


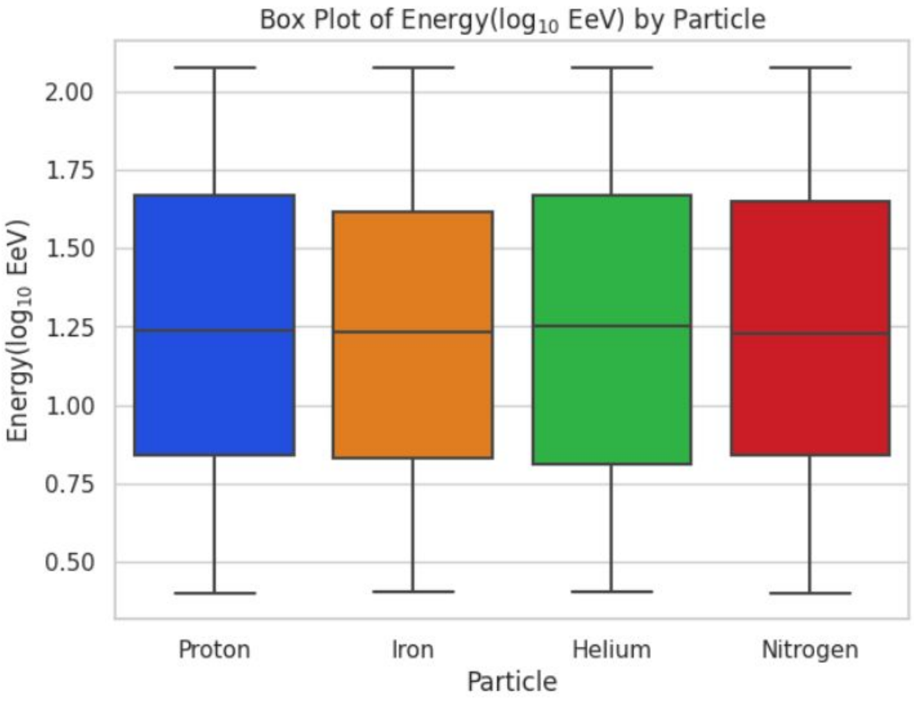
Current work: Data exploration and preprocessing



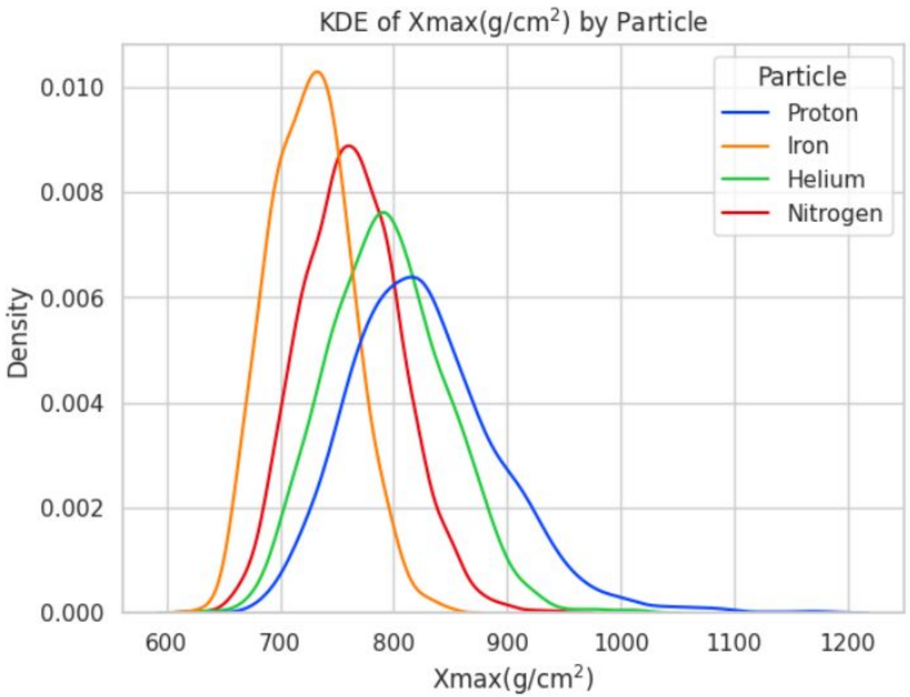


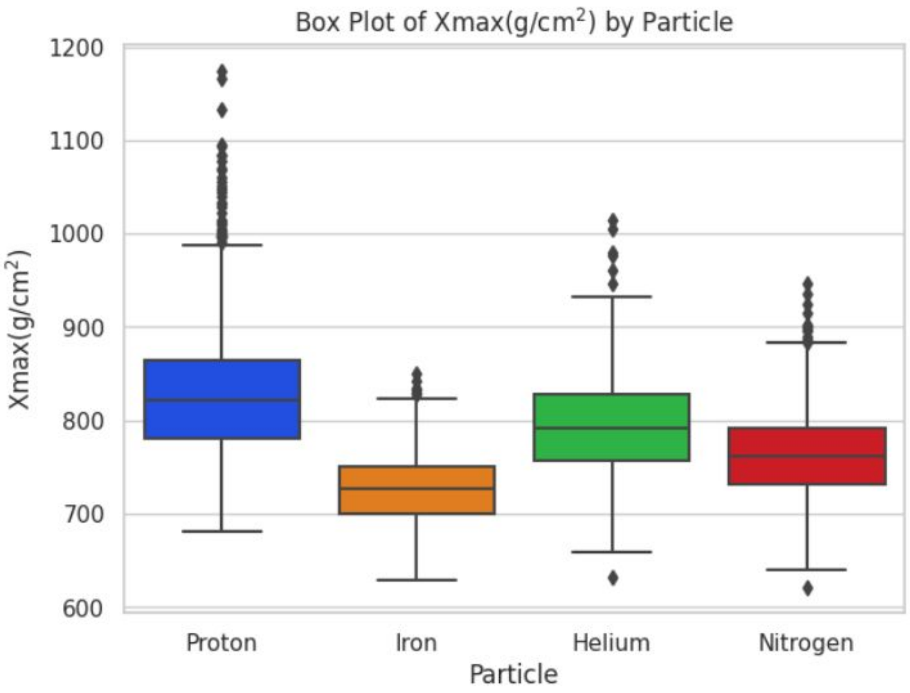
Current work: Data exploration and preprocessing





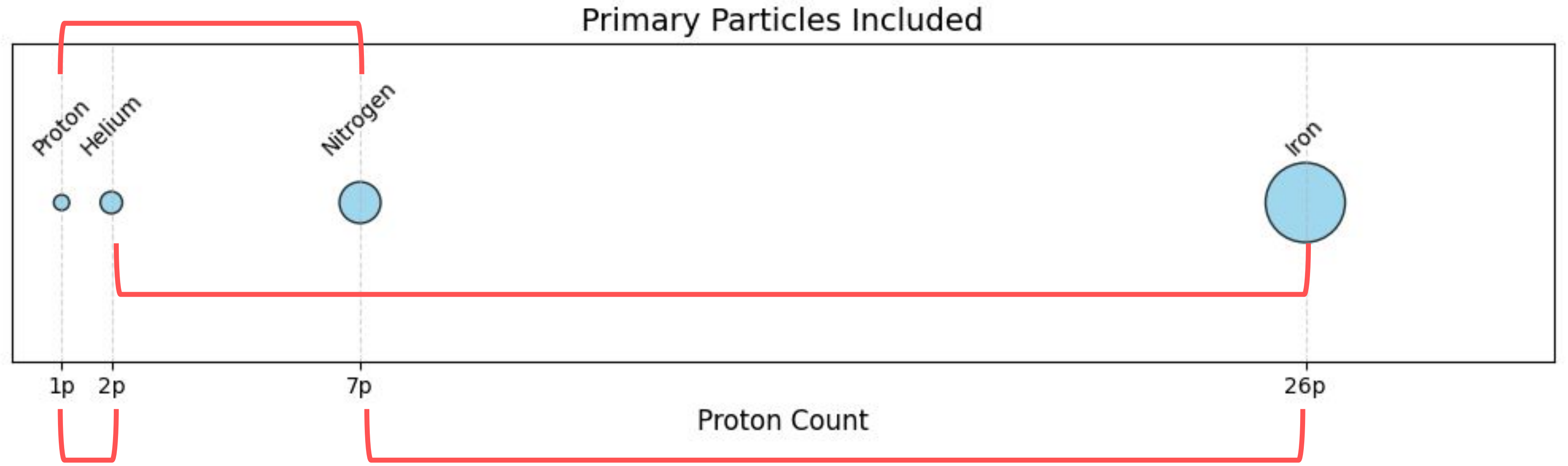
Current work: Data exploration and preprocessing





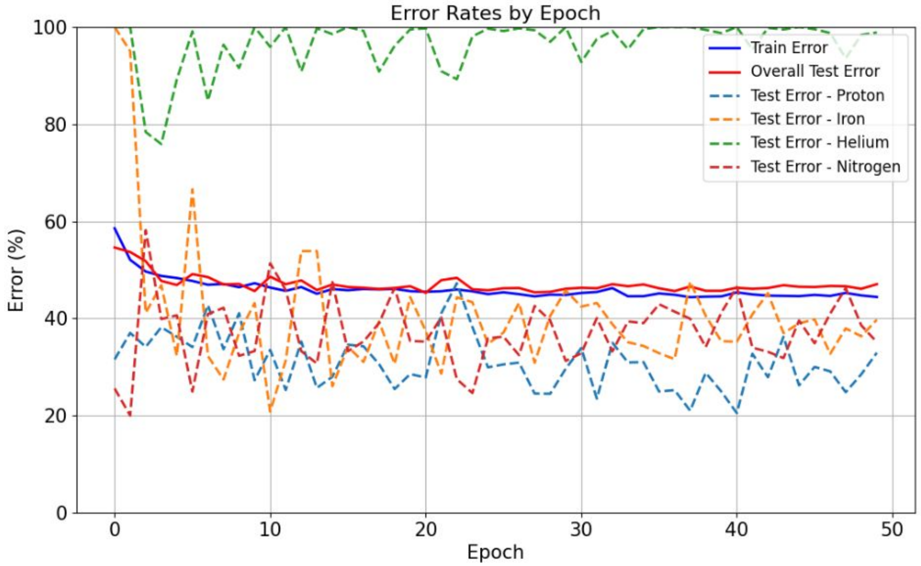
Current work: Data exploration and preprocessing

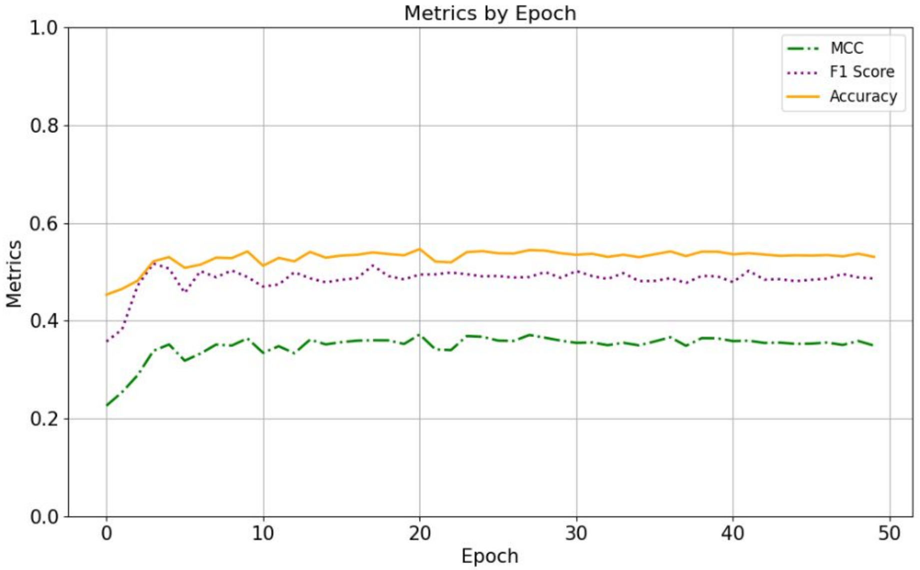
- The mass of UHECR particles directly relates to its shower-maximum depth in the atmosphere.
- The number of protons heavily influences the accuracy with which different particles are being recognized by the Machine Learning algorithm.



Current work: Nuclear composition classification of UHECRs

- Proton - Iron - Helium - Nitrogen



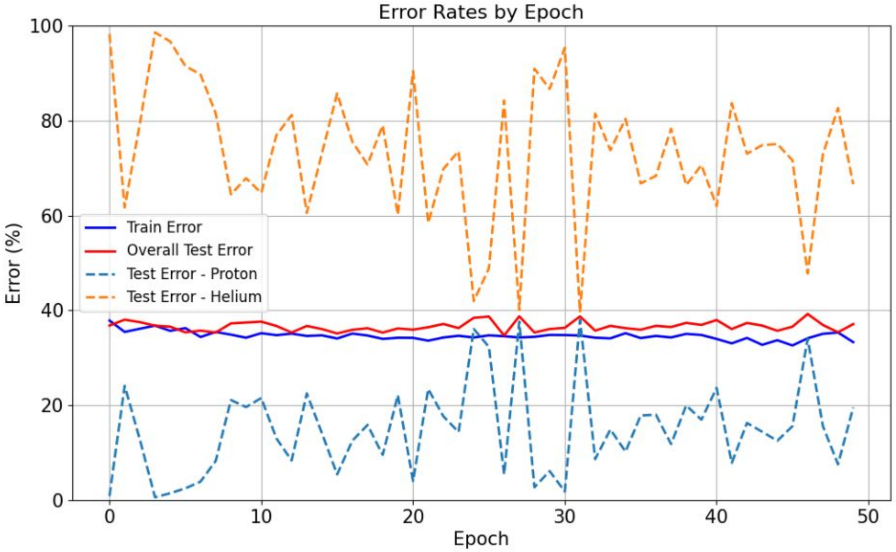


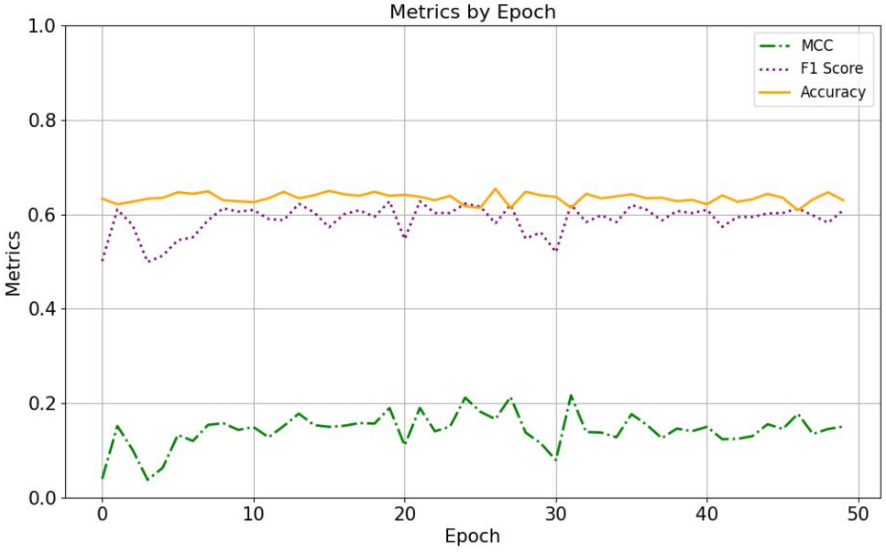
- Test errors decrease rapidly, then fluctuate in the 20-50% range for Proton, Nitrogen and Iron, and keep above 90% for Helium.

- This could be because of the similarity between Proton and Helium.

Current work: Nuclear composition classification of UHECRs

- Proton - Helium

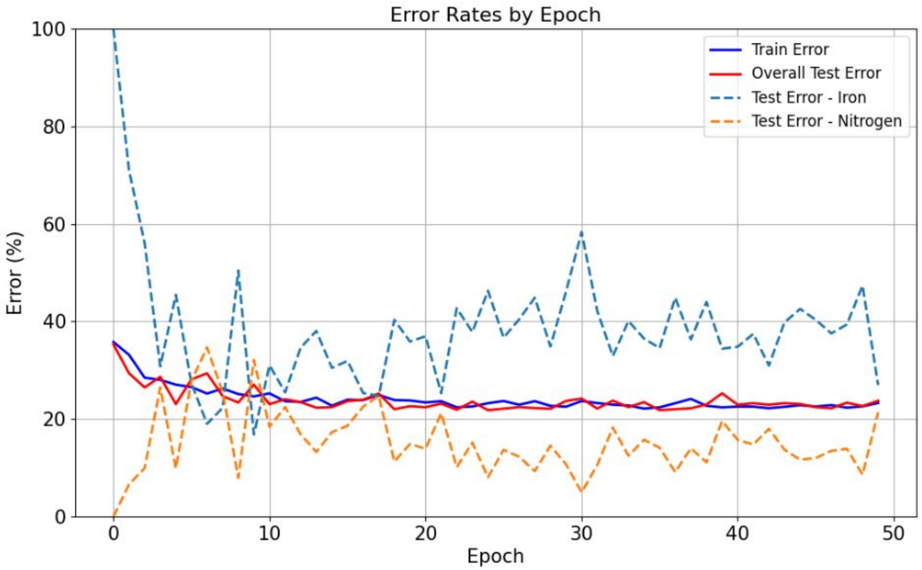


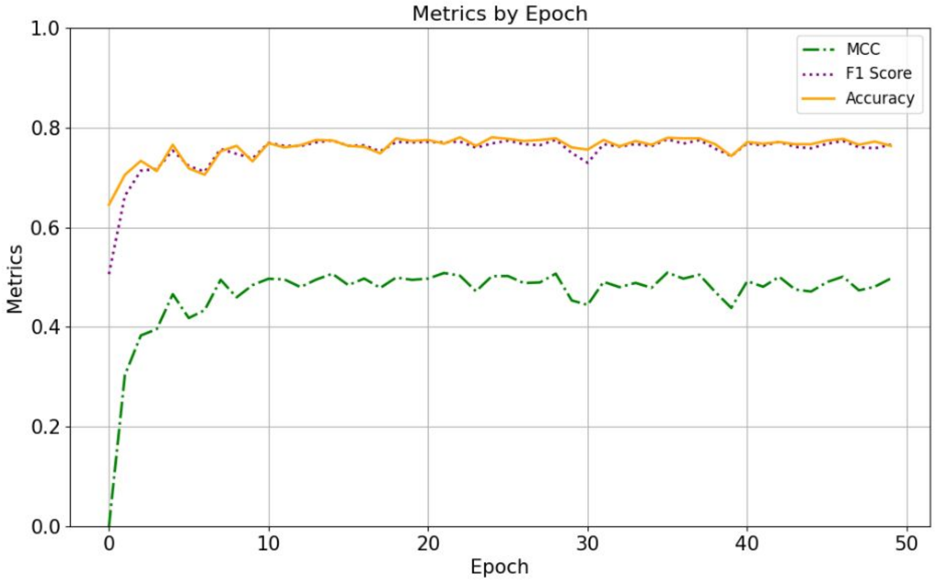


- Similar characteristics, different numbers of samples (Proton - 1996, Helium - 1161)

Current work: Nuclear composition classification of UHECRs

- Iron - Nitrogen

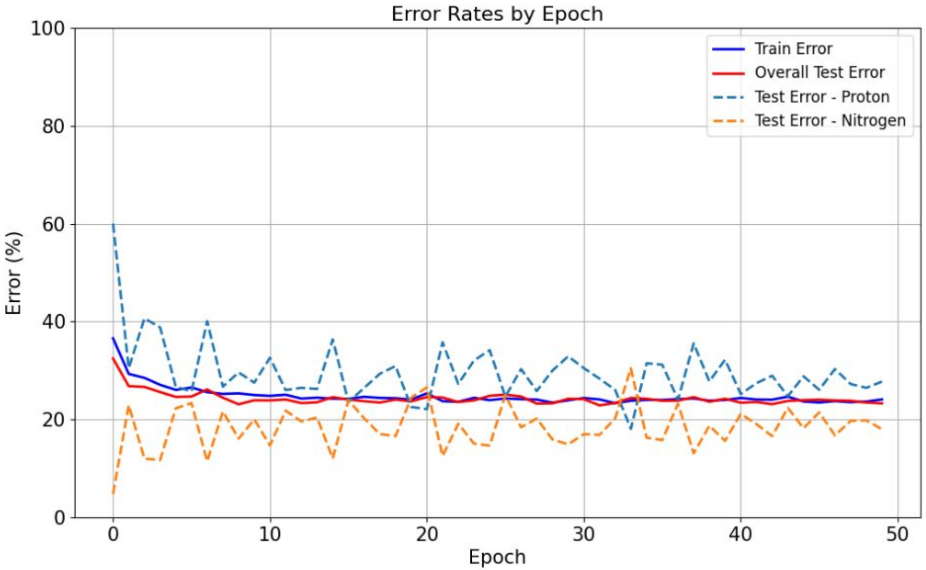


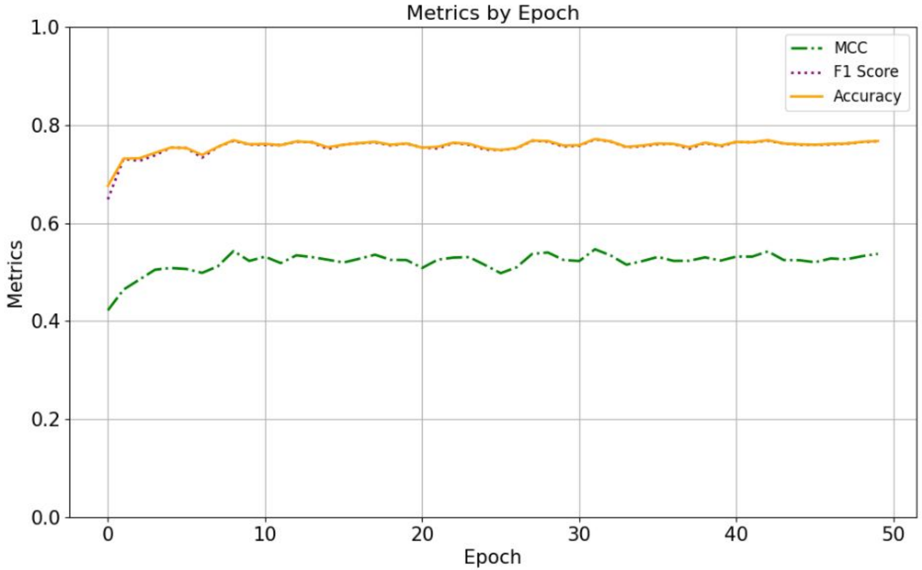


- Larger difference in nuclear mass, different numbers of samples (Iron - 1099, Nitrogen - 2000)

Current work: Nuclear composition classification of UHECRs

- Proton - Nitrogen

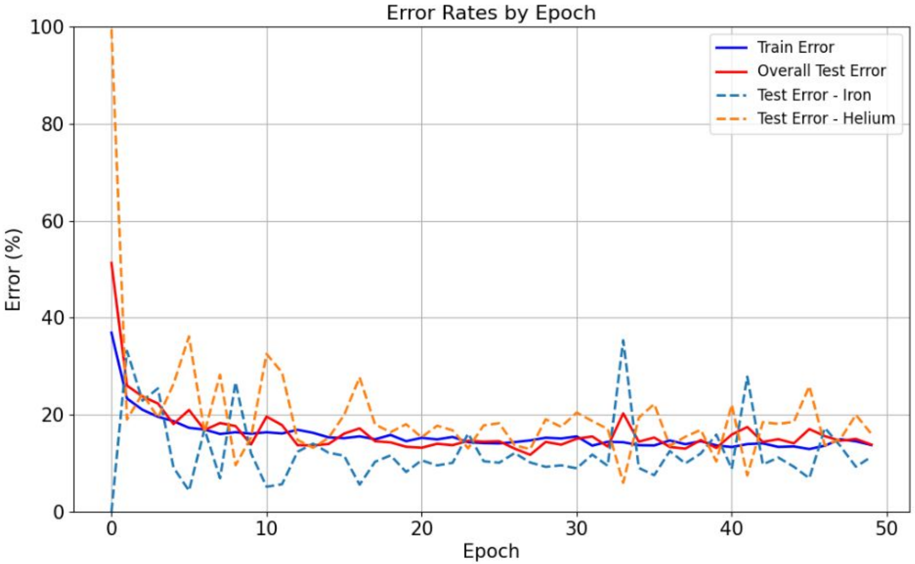


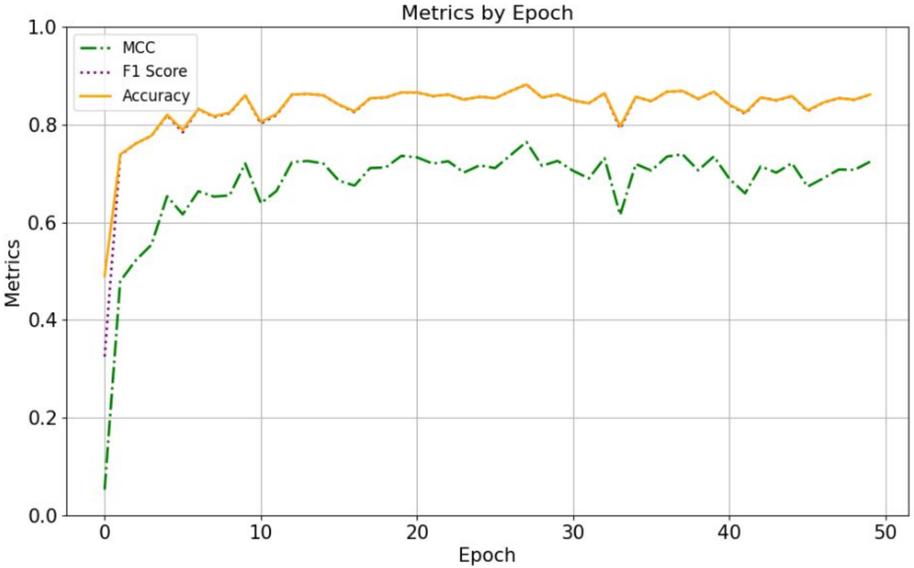


- Larger difference in nuclear mass, but not as large as for proton-iron or helium-iron pairings, similar numbers of samples (Proton - 1996, Nitrogen - 2000)

Current work: Nuclear composition classification of UHECRs

- Iron - Helium





- Large difference in nuclear mass, comparable with the previous proton-iron pairing, similar numbers of samples (Iron - 1099, Helium - 1161)

Conclusions

- The scope is to develop a framework for cosmic ray classification that works on multiple primary particles.
- Next steps:
  - Move training on CUDA/GPU instead of CPU
  - Test with a larger and more balanced simulation dataset.
  - Improve CNN model accuracy on the dataset (Grid search, data augmentation)
  - Add the number of muons from the air shower as a feature for training.
- This work serves as a Computer Science Bachelor's Thesis to be finished by summer 2025.

Particles in dataset	Error [%]	Metric MCC | F1 | Acc
P-Ir (initial)	10	0.82 | 0.91 | 0.91
P-Ir-He-N	44	0.28 | 0.45 | 0.47
P-He	39	0.18 | 0.60 | 0.61
Ir-N	22	0.50 | 0.79 | 0.79
P-N	22	0.57 | 0.79 | 0.79
Ir-He	18	0.76 | 0.83 | 0.83

Acknowledgements

- Thanks to the people providing the public repositories we used:

  https://gitlab.com/harmscho/AtmosphereCal/-/tree/master

  https://gitlab.com/harmscho/earsim

  https://github.com/nu-radio/radiotools

  https://github.com/psampathkumar/RadioPlotter

- We would like to thank the Pierre Auger Collaboration for the simulation library used in the machine learning training.

- This work was supported by the Romanian Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, grant number PN19150201/16N/2019 within the National Nucleus Program, and project number PN-III-P1-1.1-TE-2021-0924/TE57/2022, within PNCDI III.

References

(1) An analytic description of the radio emission of air showers based on its emission mechanisms, Christian Glaser, arXiv:1806.03620, 2018

(2) A proposed method for measurement of cosmic-ray chemical composition based on geomagnetic spectroscopy, Rajat K Dey, arXiv:1603.07835, 2016

(3) Cosmic-Ray Composition Measurements Using Radio Signals, Fabrizia Canfora, PhD Thesis, 2021

(4) Absolute Energy Calibration of the Pierre Auger Observatory using Radio Emission of Extensive Air Showers, Jens Christian Glaser, PhD Thesis, 2017

(5) A high-precision interpolation method for pulsed radio signals from cosmic-ray air showers, A. Corstanje et al., JINST 18, P09005, 2023

(6) Convolutional Neural Network Processing of Radio Emission for Nuclear Composition Classification of Ultra-High-Energy Cosmic Rays, Tudor Alexandru Calafeteanu, Paula Gina Isar, Emil Ioan Slusanchi, Universe 10(8), 327, 2024

Thank you! Questions?