Search for ultra-high energy neutrinos using Pierre Auger Observatory

Abha Khakurdikar

Co-authors: Prof. Jörg R. Hörandel and Dr. Washington R. Carvalho Jr.

Radboud Universiteit

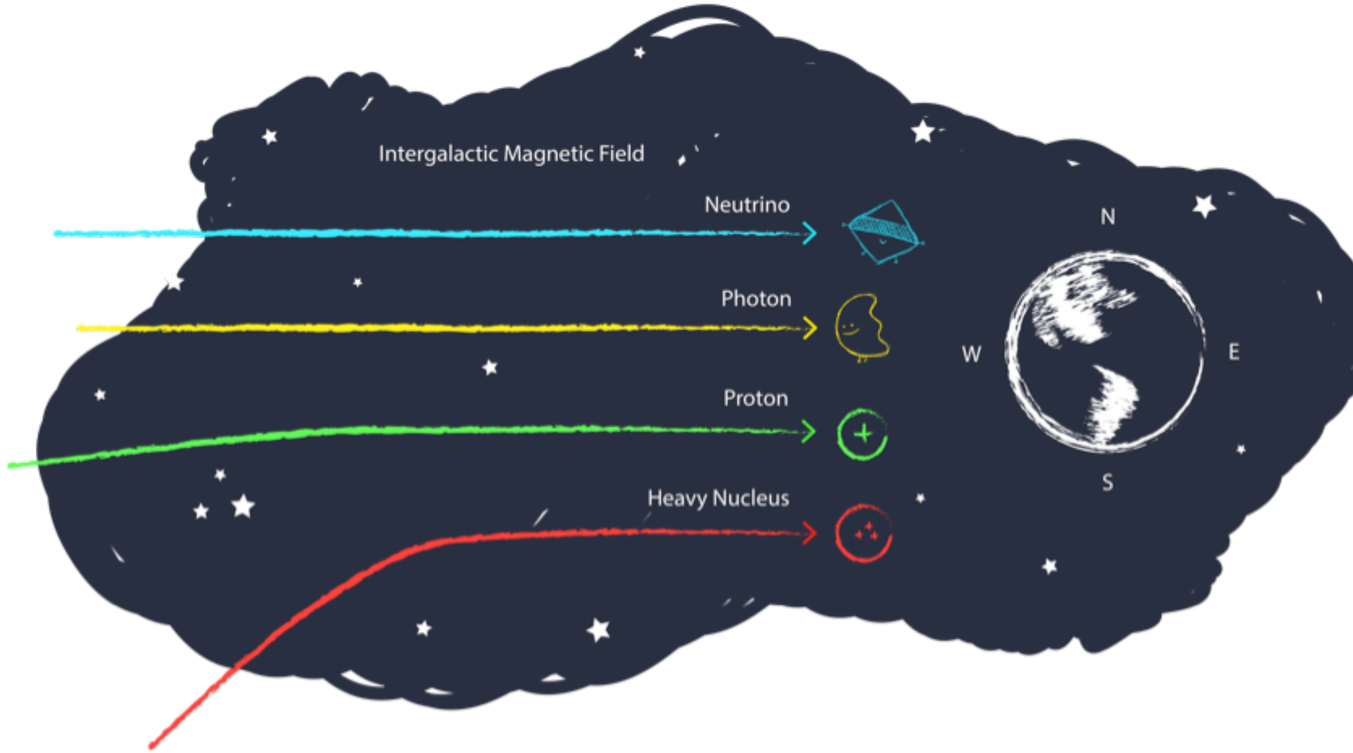
IN·DEI·NOMINE·FELICITER

28th Symposium on Astroparticle Physics in the Netherlands

27-28 June 2024

Why ultra-high energy neutrinos ?

- Key role in understanding -> origin of the ultra-high energy cosmic rays (UHECR).
- In the EeV range, neutrinos are expected to be produced in the same sources where the UHECR are thought to be accelerated.
- Neutrinos are not deviated by the magnetic field -> point back to the sources.



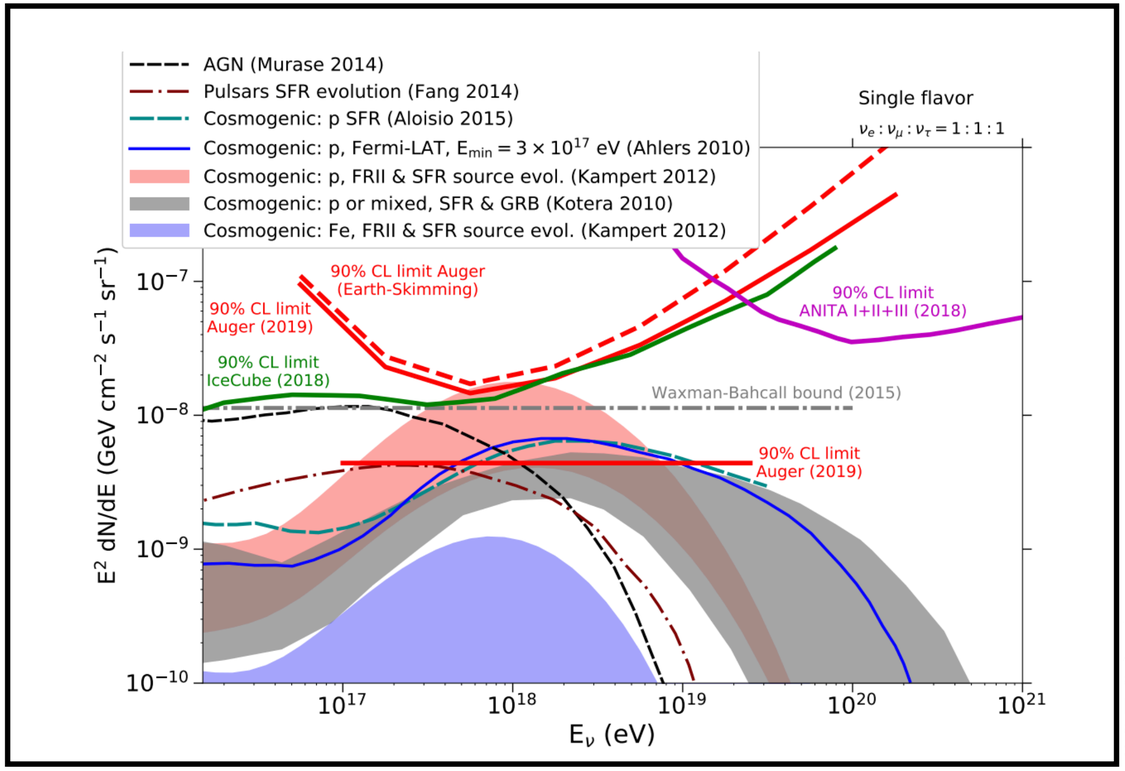
Source: https://subarutelescope.org/subaru20anniv/assets/files/Shigeru_Yoshida.pdf

Contact: a.khakurdikar@astro.ru.nl
Radboud Universiteit

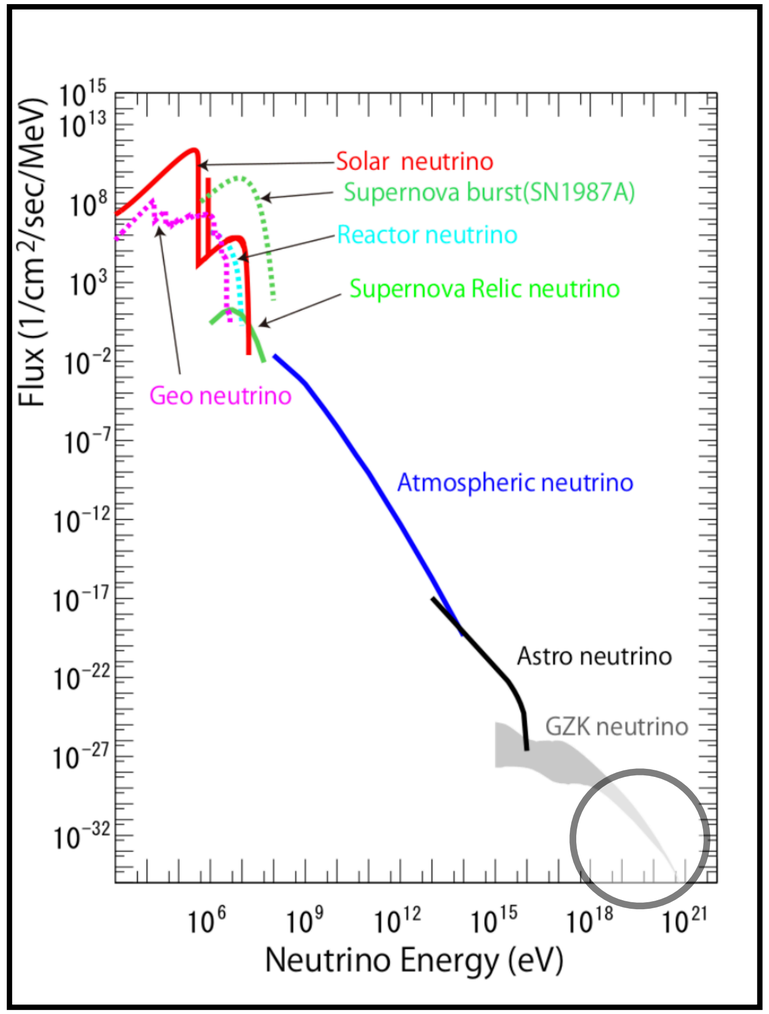
Search for ultra-high energy neutrinos

- A big challenge for the ultra-high energy neutrinos is that their flux is very low.

- We need a large detector for the detection and to get significant statistics.



The expected neutrino fluxes for Pierre Auger Observatory, several cosmogenic and astrophysical models of neutrino production,
as well as the Waxman-Bahcall bound.
Source: JCAP10(2019)022

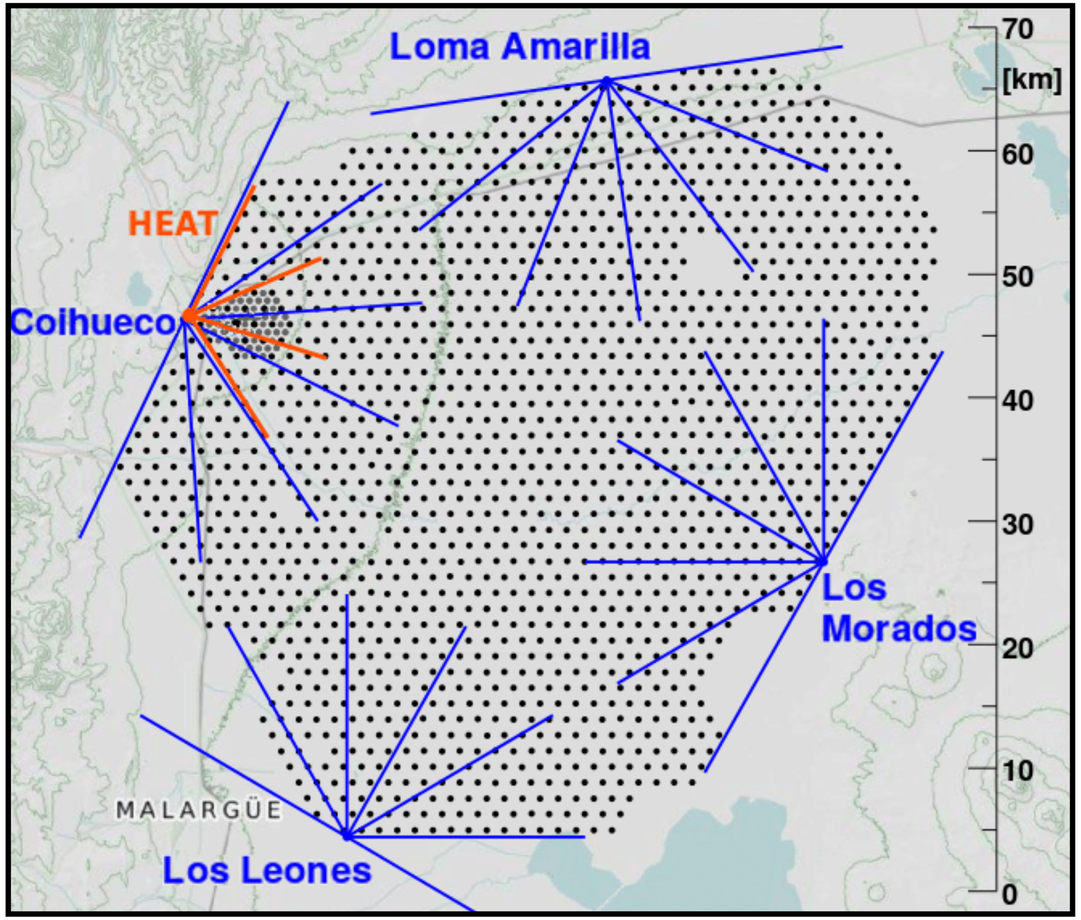


Source: https://subarutelescope.org/subaru20anniv/assets/files/Shigeru_Yoshida.pdf

Contact: a.khakurdikar@astro.ru.nl

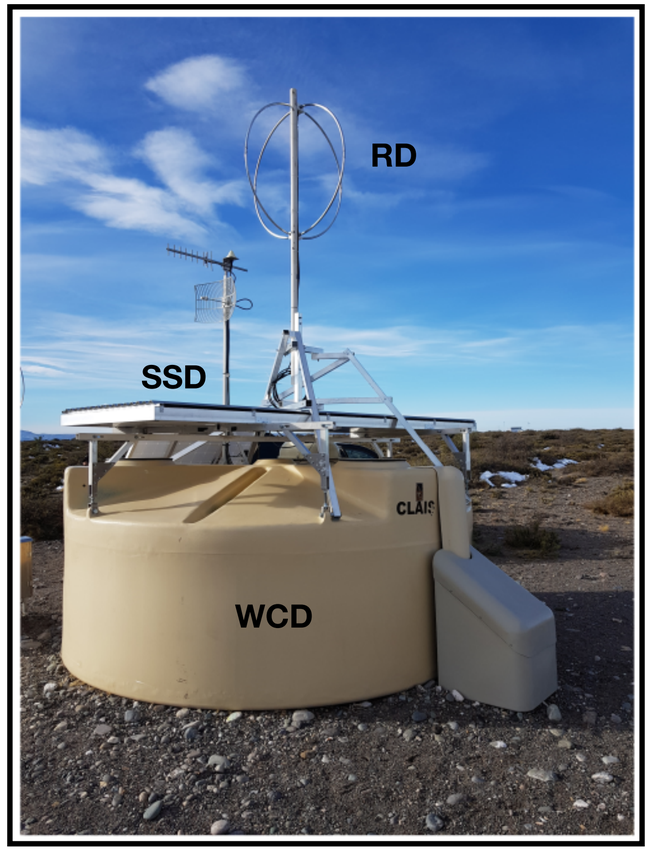
Radboud Universiteit

Pierre Auger Observatory



A schematic of the Pierre Auger Observatory where each black dot is a water Cherenkov detector.

Surface Detector



Talks (28th June 2024):
Dr. C.Galea
(Detection of UHE particles using RD)

M. Ismaiel
(AugerPrime and Radio Trigger)

A. Bwembya
(UCHER Composition using RD)

- Located in Mendoza Province, Argentina
- Surface area - 3000 km^2
- **Upgrade**: Radio antenna (**RD**) on top of each particle detector.

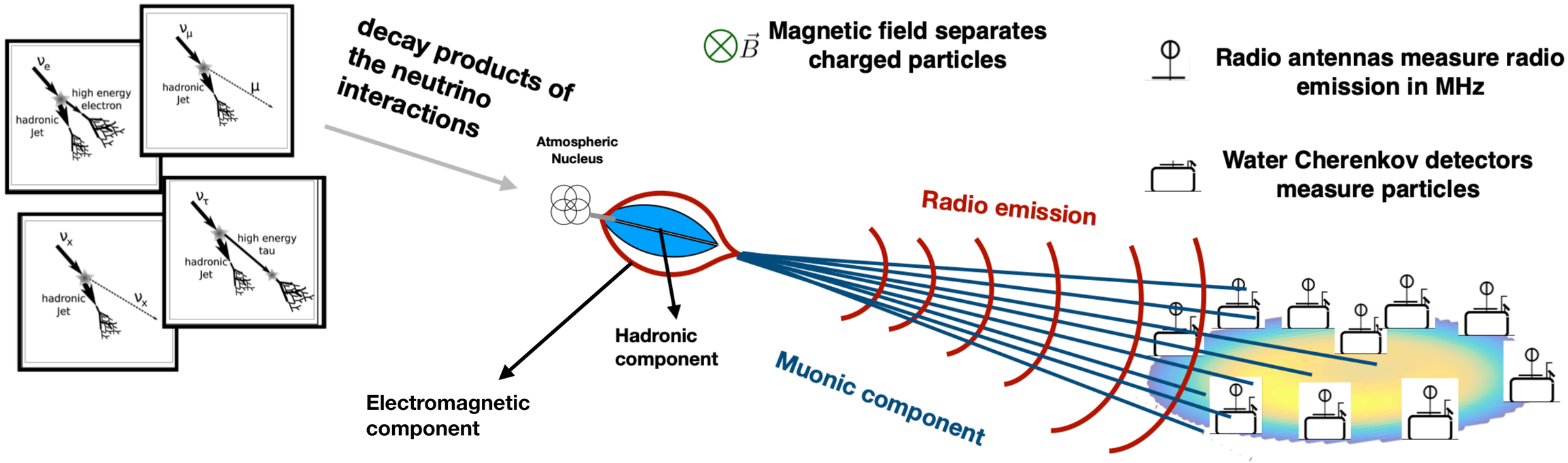
Contact: a.khakurdikar@astro.ru.nl

Radboud Universiteit

Ultra-high energy neutrinos

- Neutrinos with Energy E > 0.1 EeV

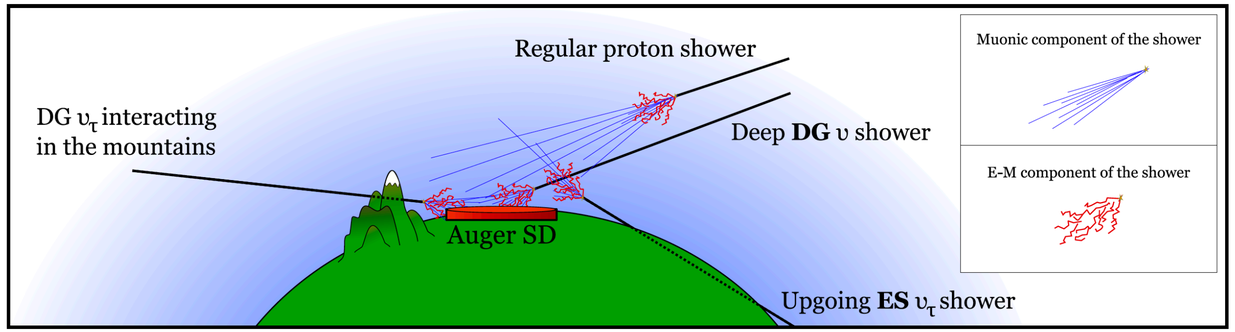
- Highly inclined extensive air showers (EAS) induced by the **neutrinos** : $75^\circ \lesssim \theta \lesssim 85^\circ$



Contact: a.khakurdikar@astro.ru.nl

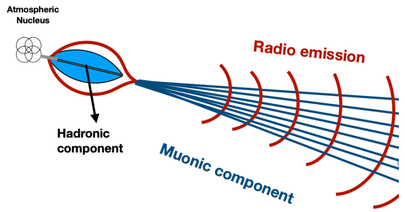
Radboud Universiteit

EAS RD simulations

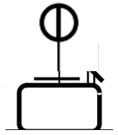


Simulate the ν-nucleon interactions to obtain the decay products

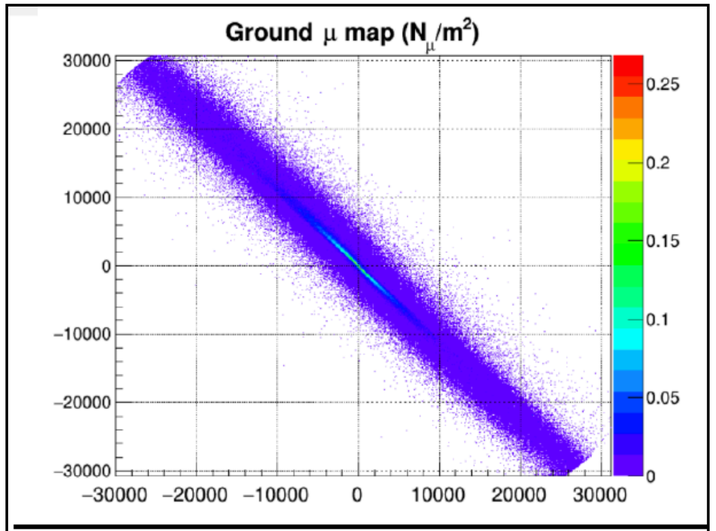
Simulate the air showers
induced by the decay products



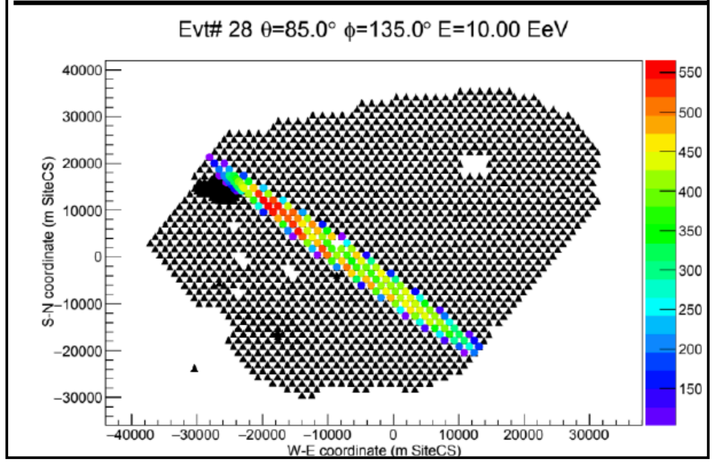
Detector simulation
(Hybrid reconstruction - WCD + RD)



Muonic part measured by WCD



Radio part measured by RD



EAS footprint after the detector simulation using the RDSim framework

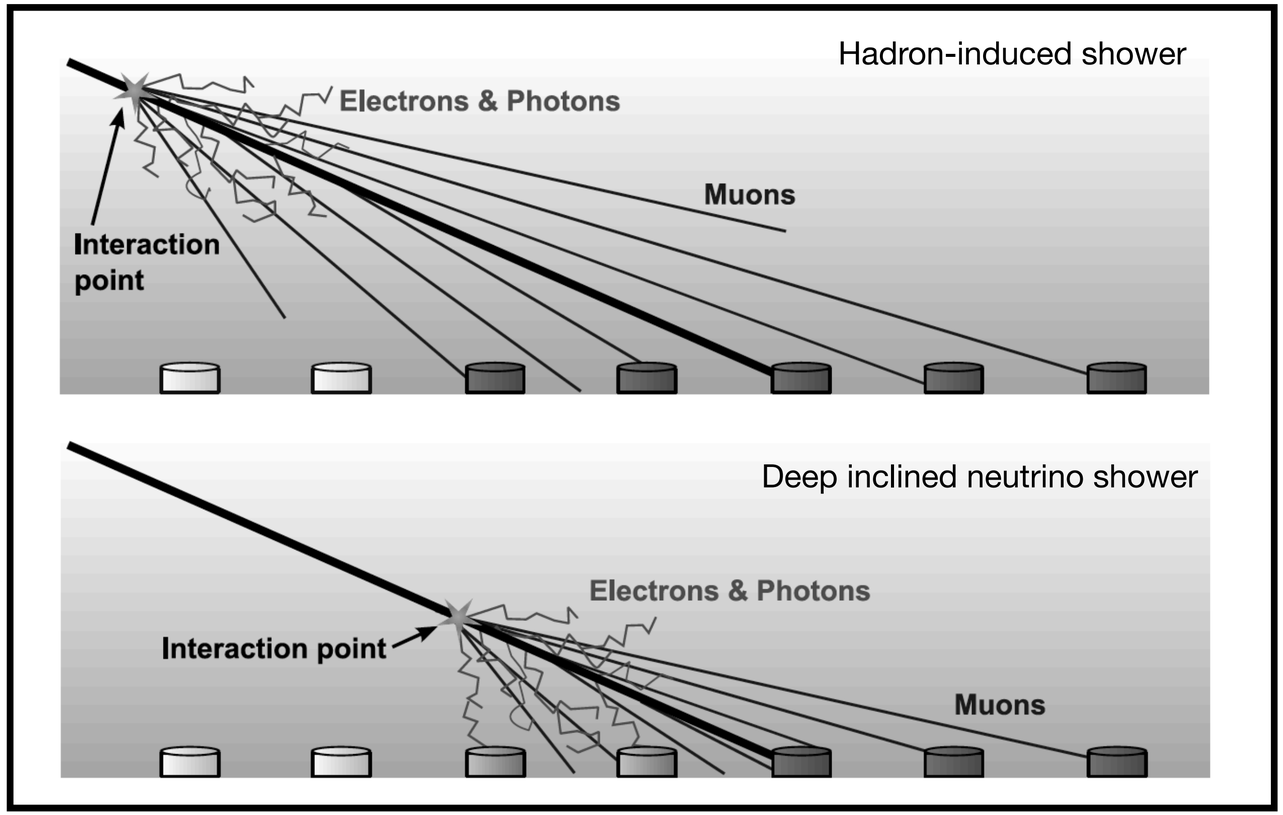
Contact: a.khakurdikar@astro.ru.nl

Radboud Universiteit

Identify the neutrino induced EAS

- The main challenge in detecting the UHE neutrinos is to identify a neutrino-induced shower (signal) in the background of showers initiated by UHERCRs.

EM component is absorbed and only the muons reach the detector



Significant EM component at the detector level

- RD upgrade will increase the sensitivity of the neutrino detection due to the significant EM component.

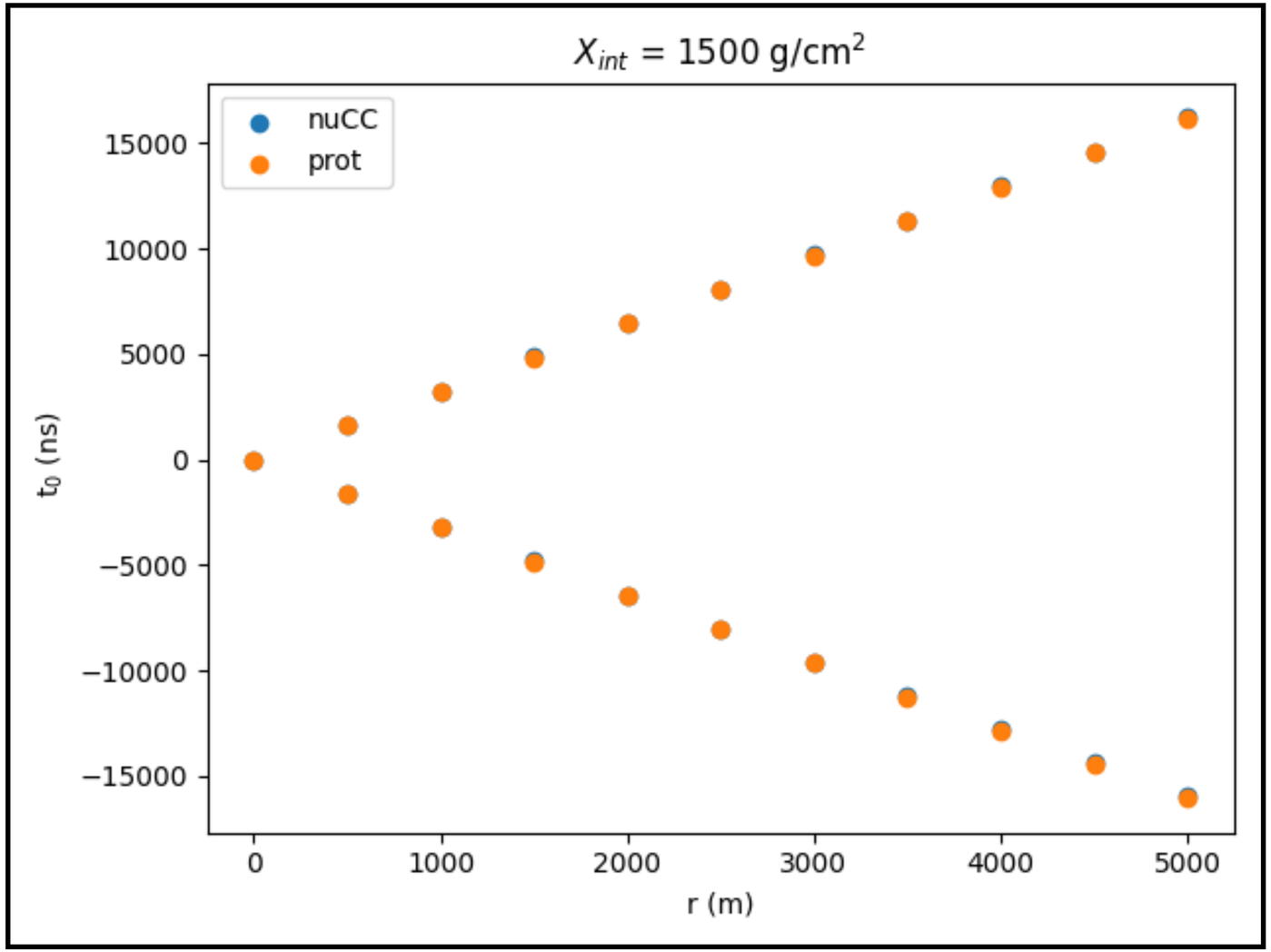
Source: arXiv:1202.1493

Contact: a.khakurdikar@astro.ru.nl
Radboud Universiteit

Features for identification

- Time at which the signal was found, relative to the event start time

- Due to geometrical reasons, the arrival of the first particles at lateral distance r from the axis is expected to be delayed with respect to planar shower front.

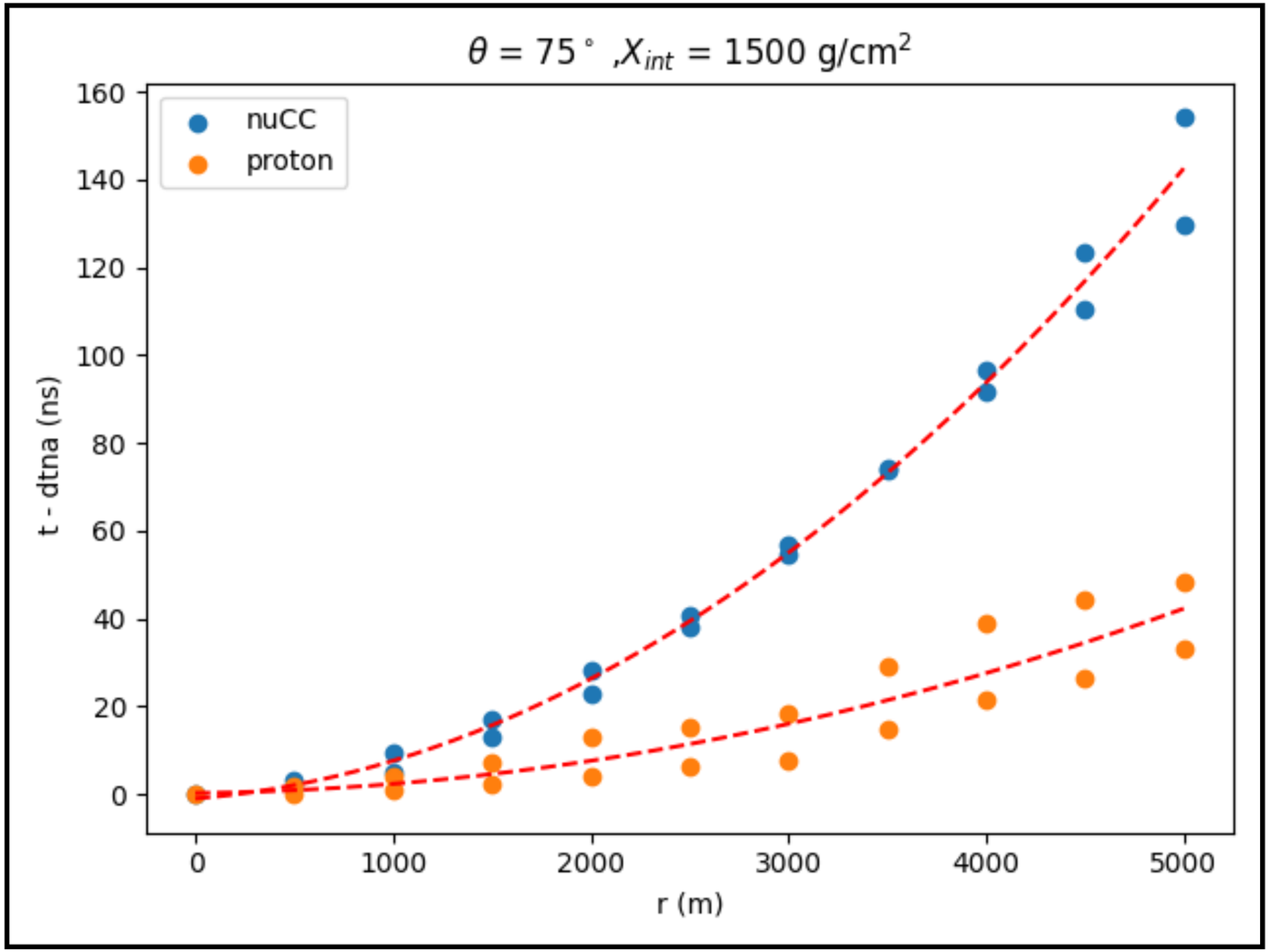


Signal time measured with respect to the distance from the shower core for a single event.

Contact: a.khakurdikar@astro.ru.nl

Features for identification

- We calculate the time delay and get a structure of shower front from the corrected time signal.

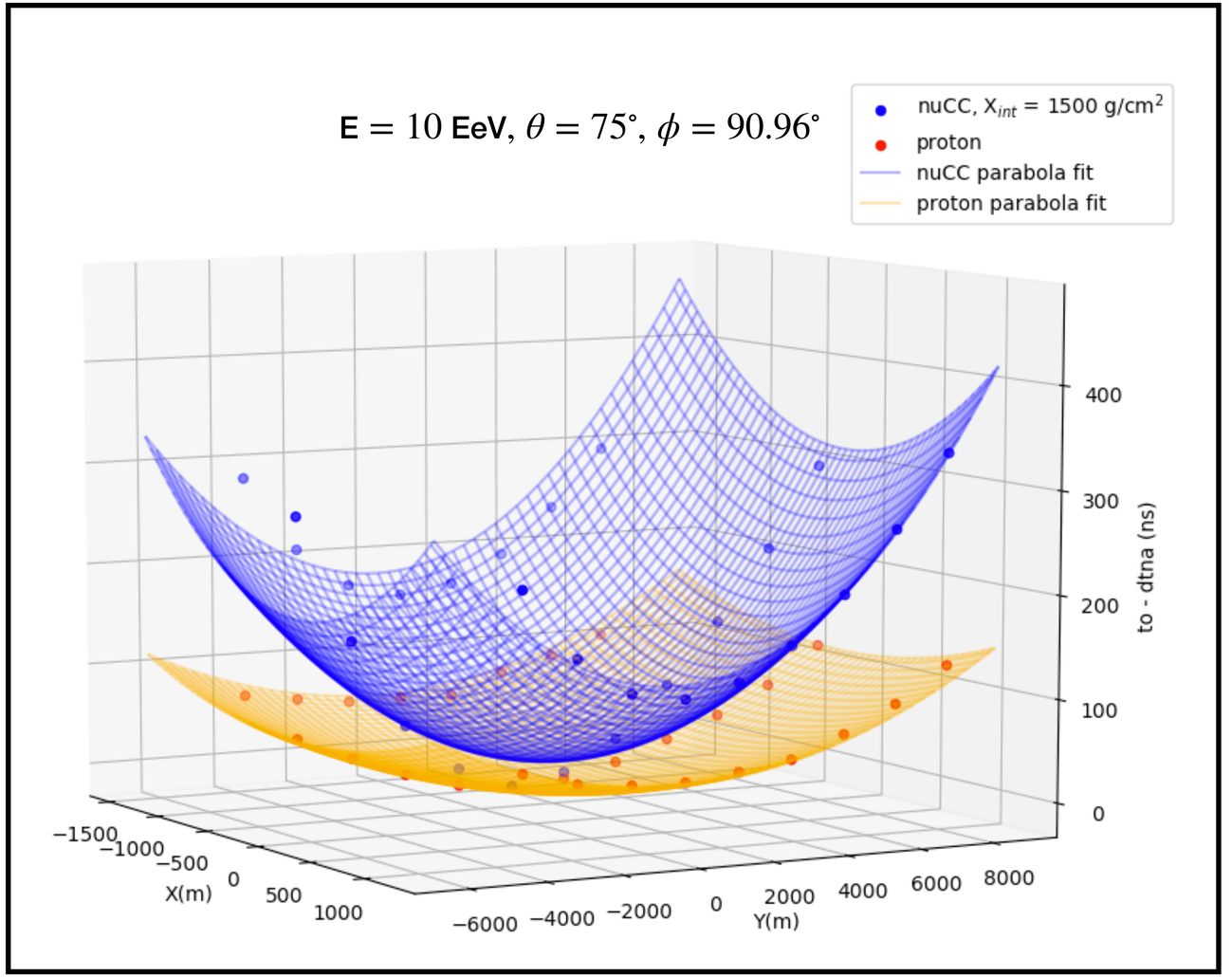


Corrected time signal with respect to the distance from the shower core

Contact: a.khakurdikar@astro.ru.nl

Features for identification

- Neutrino showers are in general much closer to the ground and hence the wavefront will be more curved as compared to hadron induced showers.

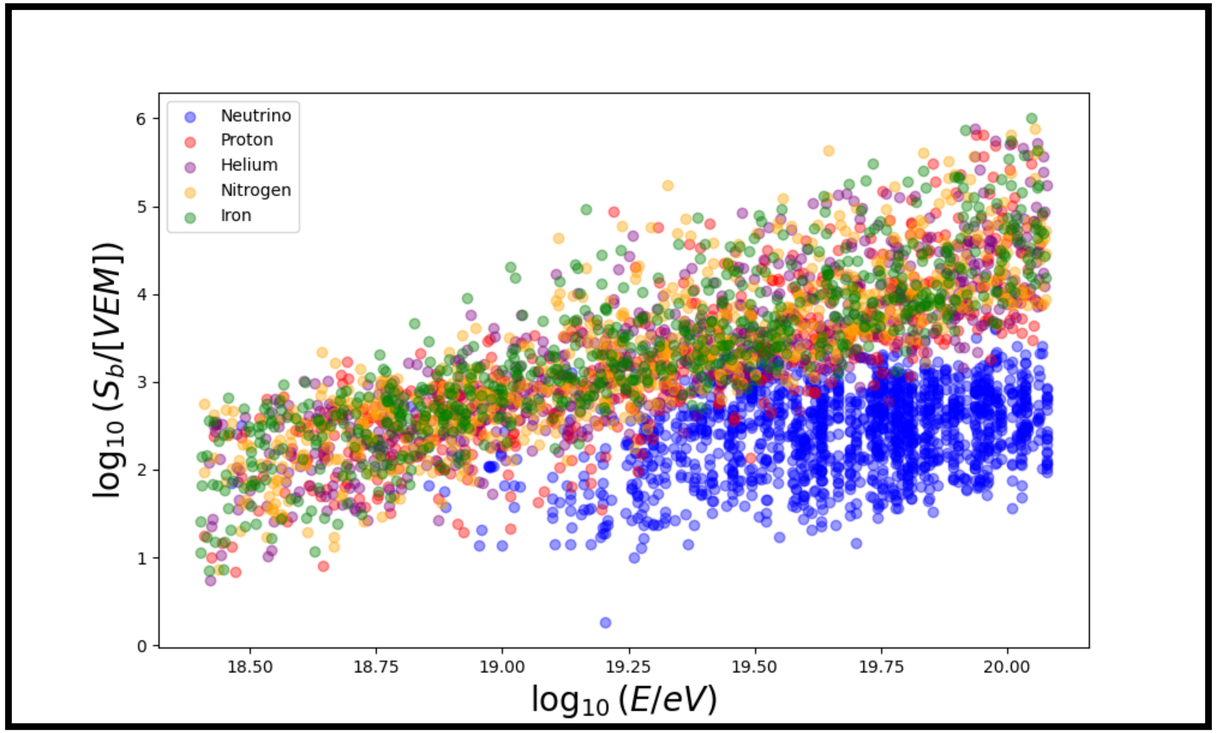


Corrected time signal with respect to the distance from the shower core

Contact: a.khakurdikar@astro.ru.nl

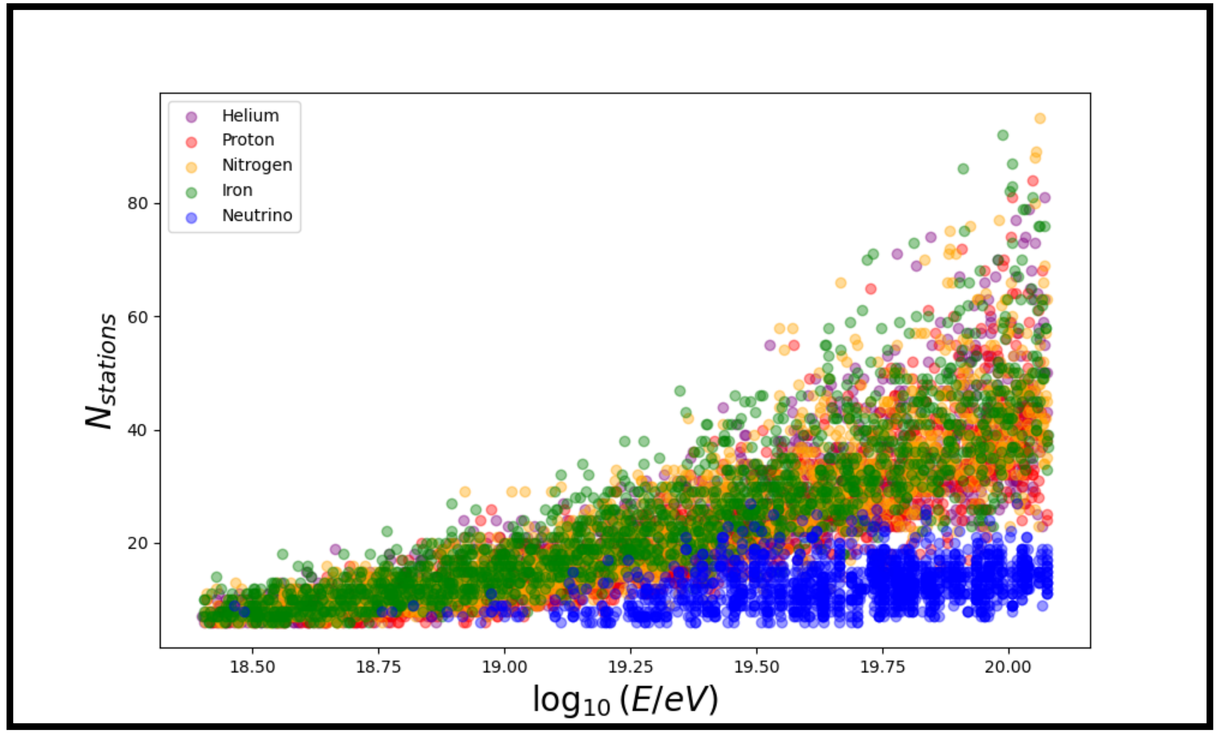
Radboud Universiteit

Features for identification



Total muon signal in an event

$$S_b = \sum_i S_i \times \left(\frac{R_i}{3500\ \text{m}}\right)^b$$, S_i is the measured muon signal at a distance R_i



Number of stations per event with detectable muon and radio signal

Contact: a.khakurdikar@astro.ru.nl

Radboud Universiteit

Separation of neutrinos from Background

Signal-like events:

- Downgoing neutrinos
- E = 1.0 - 120 EeV
- $\theta = 75^\circ - 85^\circ$
- Varying Interaction depth

Background-like events:

- Primaries: proton, helium, nitrogen and iron
- E = 1.0 - 120 EeV
- $\theta = 75^\circ - 85^\circ$, Uniform distribution in $\sin\theta\cos\theta$

- **Input Features (X) -> Fit parameters, Sb, Nstat**
- **Target variable (y) -> primary particle (Neutrino or Hadron)**
- **Train the model using the Random Forest Classifier***

* Detailed method in Back-up slides

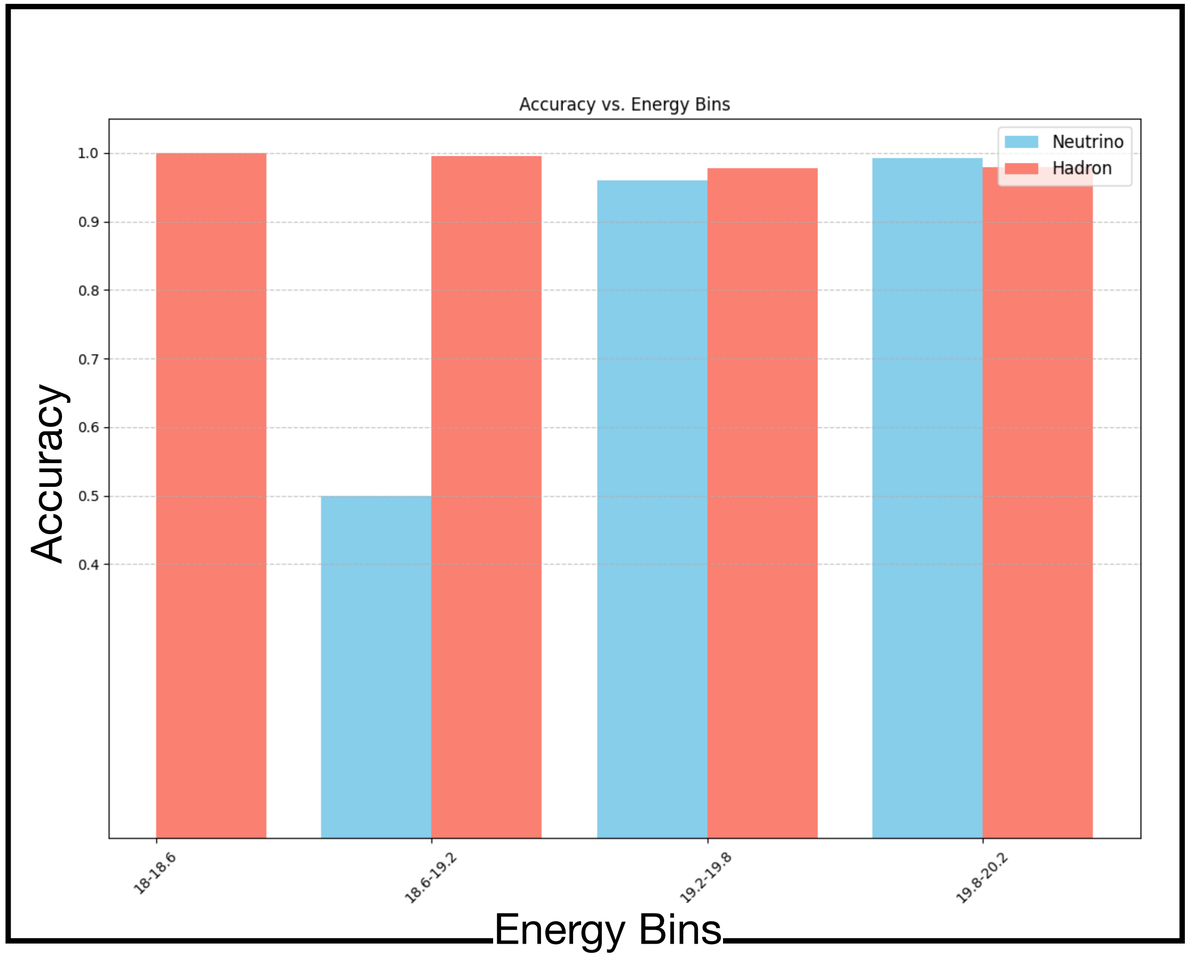
Contact: a.khakurdikar@astro.ru.nl

Separation of neutrinos from Background

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Elliptical parabola (Ground plane - $ax^2 + by^2 + cx + dy + exy + f$) : Accuracy = **0.963**

(Features used: a,b,c,d,e,f, Sb, Nstat (min = 6))



*final prediction is made by taking the ensemble of 100 individual trees and optimizing the hyper parameters

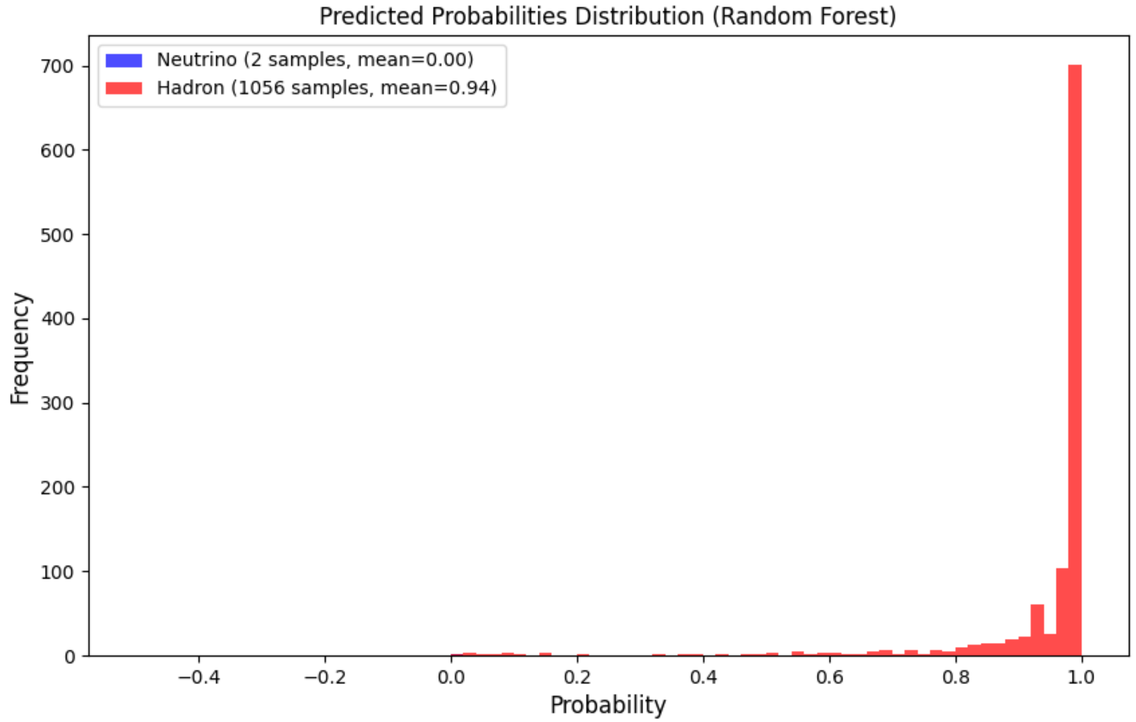
Contact: a.khakurdikar@astro.ru.nl

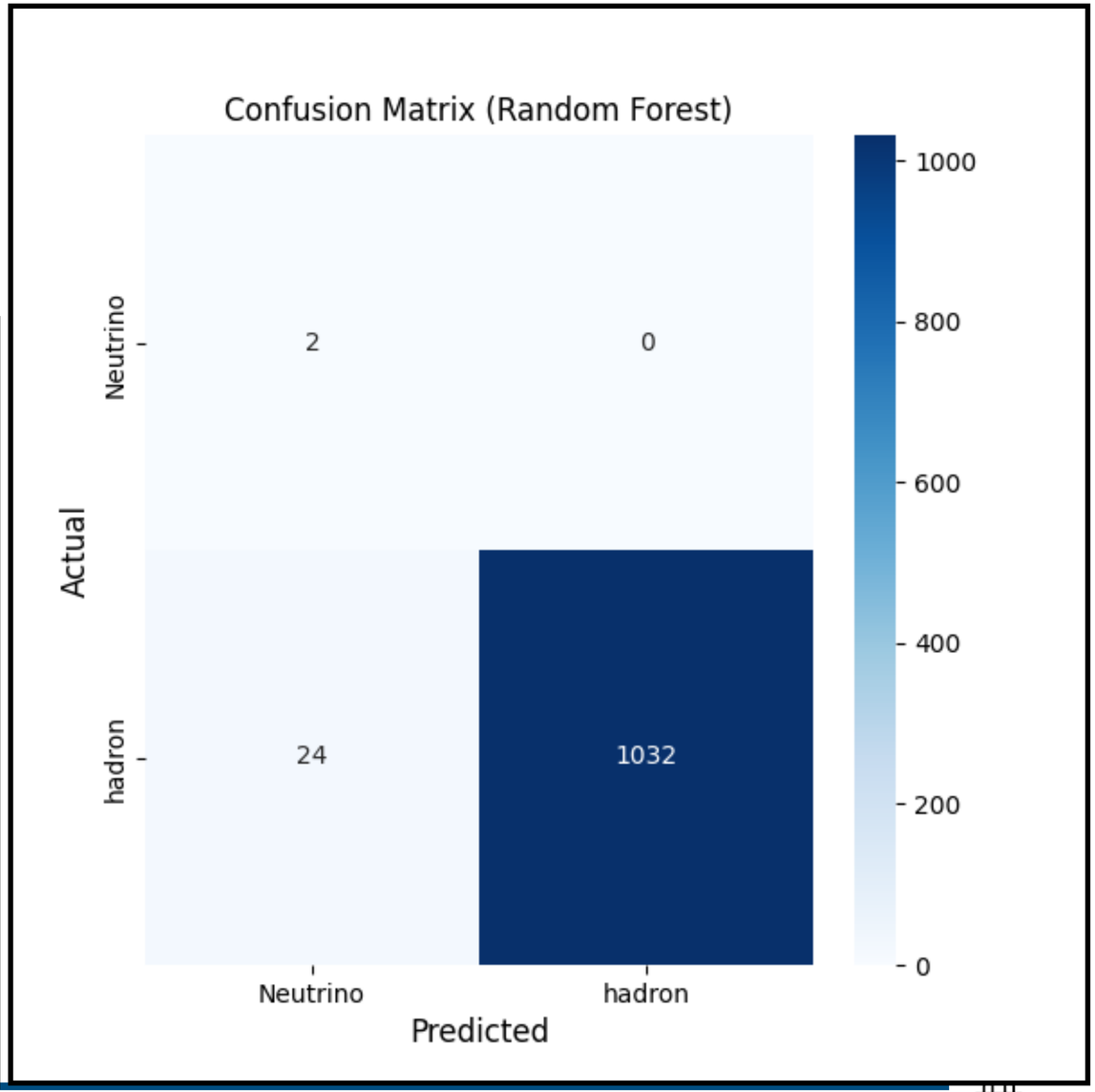
Radboud Universiteit

Separation of neutrinos from Background

Another Test set: hadrons (1-120 EeV) + 2 neutrinos (~ 10 EeV)

Accuracy Percentage for Neutrinos (Random Forest): 100.00%
Accuracy Percentage for Hadrons (Random Forest): 98.01%





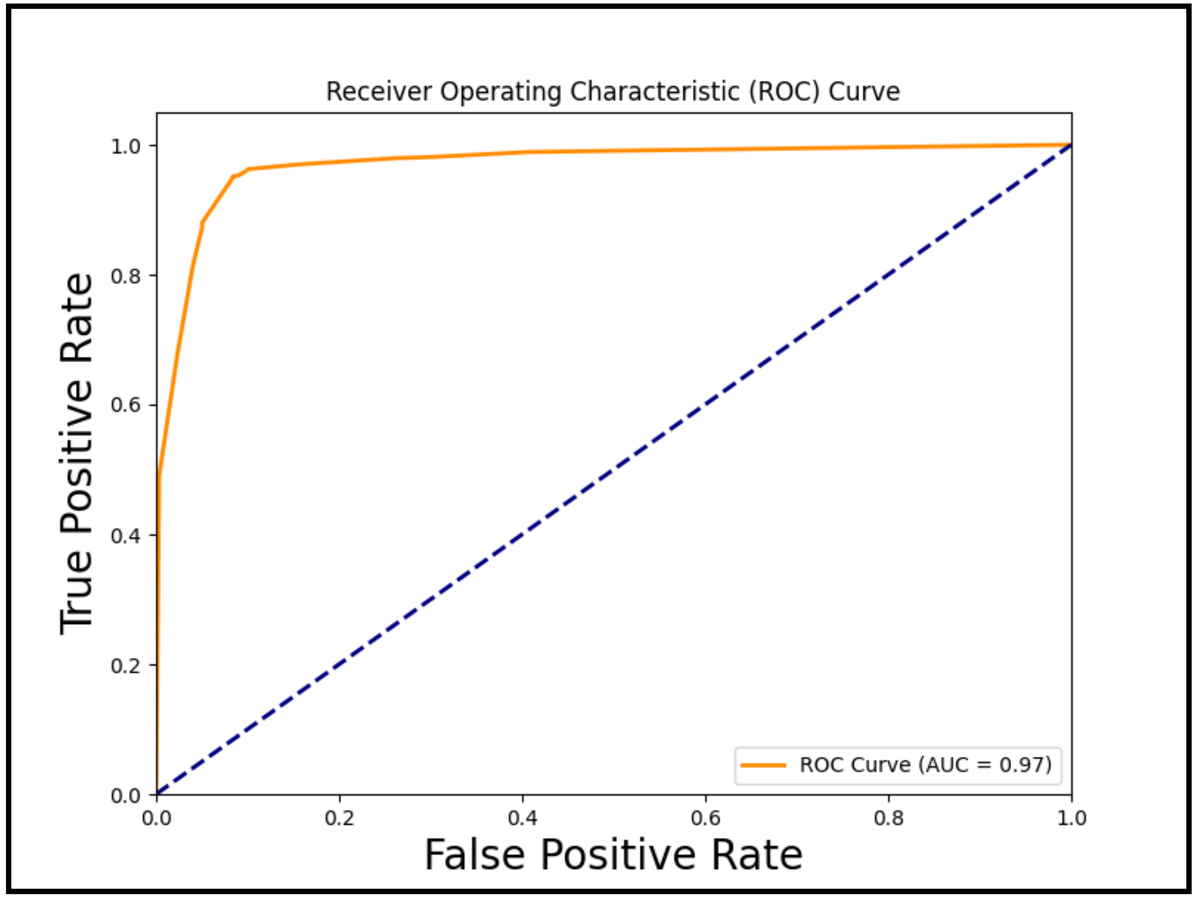
Contact: a.khakurdikar@astro.ru.nl

Radboud Universiteit

Separation of neutrinos from Background

- ROC curves are typically used for binary classification problems, where the target variable has two classes

- The area under the ROC curve (AUC) is a measure of the classifier's performance.



$$TPR = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \qquad FPR = \frac{\text{False Positives}}{\text{False Positives + True Negatives}}$$

Contact: a.khakurdikar@astro.ru.nl

Radboud Universiteit

Summary

- Ultra-high energy **neutrinos** play a key role in understanding the origin of **UHECRs**.
- The main **challenge** in detecting the UHE neutrinos is to identify a neutrino-induced shower (signal) in the **background** of showers initiated by UHERCRs.
- **Geometry of the air shower footprint** is the main **difference** in the hadron-induced showers and the neutrino-induced showers
- Supervised Learning tool, such as **Random Forest Classifier** help to **classify** the showers induced by **neutrinos** and **hadrons**.

Contact: a.khakurdikar@astro.ru.nl
Radboud Universiteit

Back-up

WCD + RD analysis

- Currently we can only use simulations for our background-like events. So, we can keep in mind the kinks in the background events.

1. The uncertainties on hadronic models,
2. The ignorance of unknown physical processes
3. The unpredictable detector effects may make very unreliable estimation of the background.
4. Also, the minimum number of simulated showers that would require to properly populate the tails of the distributions with a statistically significant number of entries was unaffordable.

Step 1: Background-like and Signal-like events

Simulations available:
Signal-like events: CoREAS showers

- Force Model: Sibyll2.3d
- Type: Electron Neutrino
- CC and NC interactions
- E = 1.0EeV, 10.0EeV
- $\theta = 66^\circ, 75^\circ$
- For varying Interaction lengths - 50 simulations per Xint

Background-like events: CoREAS showers

- Primaries: proton, helium, nitrogen and iron
- E = 1.0 - 120 EeV
- $\theta = 65^\circ - 85^\circ$, Uniform distribution in $\sin\theta\cos\theta$
- 2000 simulations per primary (Around 8000 simulations - there are some corrupt events)

Step 2: Reconstruction done with SdHAS + RD with 7.22 SNR threshold

Contact: a.khakurdikar@astro.ru.nl
Radboud Universiteit

Back-up

Simulations available: (1.5 km grid)
Signal-like events: CoREAS showers

- Force Model: Sibyll2.3d
- Type: Electron Neutrino
- CC and NC interactions
- E = 1.0 - 120 EeV
- $\theta = 75^\circ - 85^\circ$, Uniform distribution in $\sin\theta\cos\theta$
- For varying Interaction lengths starting from 100 g/cm^2
- $75^\circ : 100 - 3000$ g/cm^2
- $85^\circ : 100 - 7000$ g/cm^2

Background-like events: CoREAS showers

- Primaries: proton, helium, nitrogen and iron
- E = 1.0 - 120 EeV
- $\theta = 75^\circ - 85^\circ$, Uniform distribution in $\sin\theta\cos\theta$

Dataset contains - E, Zenith, fit parameters, Muon signal, Number of stations and primary particle (neutrino or background (p, fe, he, n))

1. Define the features and the target variable:,

Features (X) -> Fit parameters, Sb, Nstat

Target variable (y) -> primary particle (Neutrino or Hadron)

2. Split the data into training and testing sets: 80% training, 20% testing

3. Binary Tree Structure: The CART algorithm starts with the entire dataset and selects the feature and threshold that minimize the Gini impurity for a binary split

$X_i \leq$ **threshold -> left child node, else -> right child node**

Contact: a.khakurdikar@astro.ru.nl
Radboud Universiteit

Back-up

For $\theta = 75^{\circ}$

X(g/cm2)	Altitude(m)	DecayHeight (m)
100	24696	23296
500	14488	13088
1000	10111	8711
1500	7390	5990
2000	5332	3932
2500	3658	2258
3000	2239	839

For $\theta = 80^{\circ}$

X(g/cm2)	Altitude (m)	DecayHeight (m)
100	26806	25406
500	16671	15271
1000	12363	10963
1500	9827	8427
1600	9410	8010
2000	7931	6531
2100	7600	6200
2500	6392	4992
3000	5088	3688
3500	3952	2552
3600	3740	2340
4000	2942	1542
4100	2753	1353
4500	2032	632

For $\theta = 85^{\circ}$

X(g/cm2)	Altitude (m)	DecayHeight (m)
100	29400	28000
500	19513	18113
1000	15376	13976
1500	12970	11570
2000	11269	9869
2500	9937	8537
3000	8814	7414
3500	7840	6440
4000	6977	5577
4500	6201	4801
5000	5496	4096
5500	4848	3448
6000	4250	2850
6500	3692	2292
7000	3171	1771
7500	2681	1281
8000	2219	819
8100	2129	729
8500	1781	381

Contact: a.khakurdikar@astro.ru.nl
Radboud Universiteit

Back-up

t_o: time at which the signal was found, relative to the event start time

$$r_{ant} = \sqrt{x_{ant}^2 + y_{ant}^2}$$

$$\phi_{ant} = \text{atan2(y , x)}$$

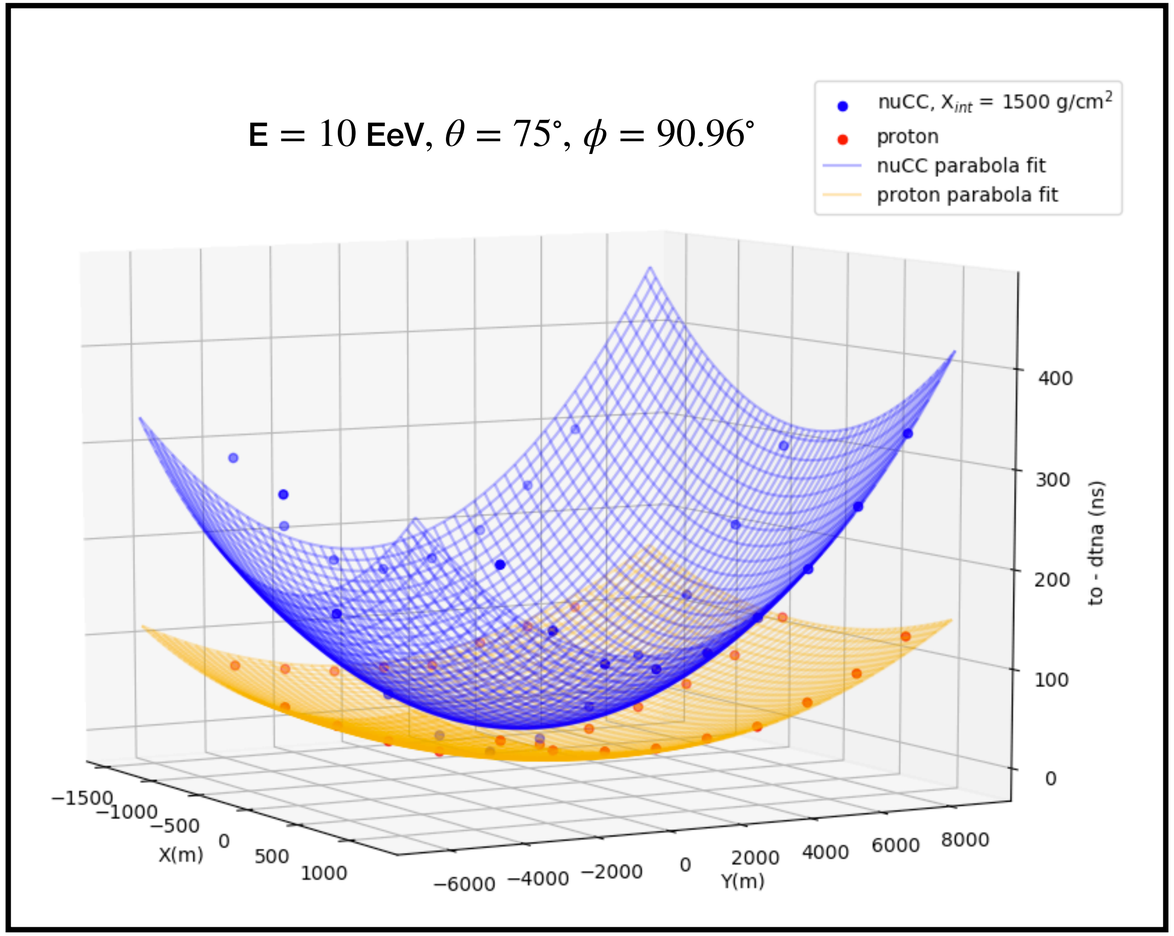
$$\text{angle} = |\phi_{MC} - \phi_{ant}|$$

$$r_{proj} = r_{ant} * \cos(\text{angle})$$

Delay in time due to the curvature of the air shower front

$$dtna = -(r_{proj} \sin(\theta_{MC})) / c$$

- We calculate the time delay and get a structure of shower front from the corrected time signal.



Corrected time signal with respect to the distance from the shower core

Contact: a.khakurdikar@astro.ru.nl

Radboud Universiteit

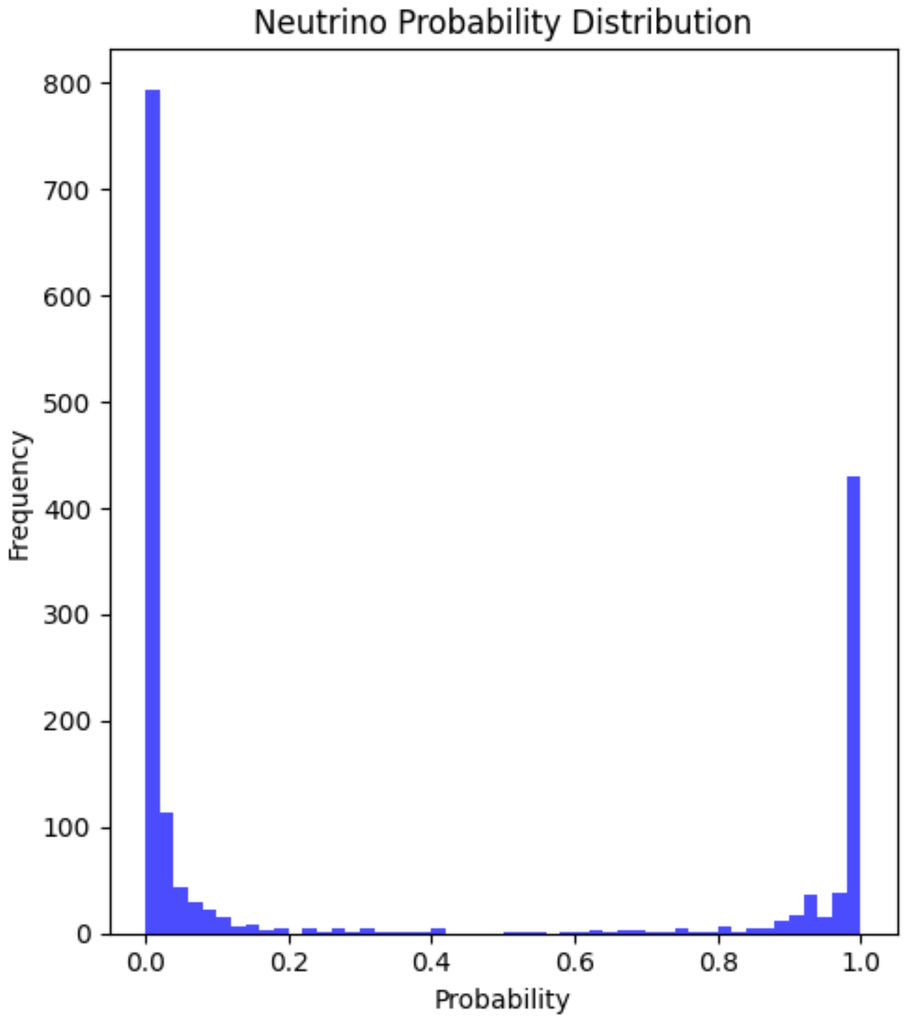
Back-up

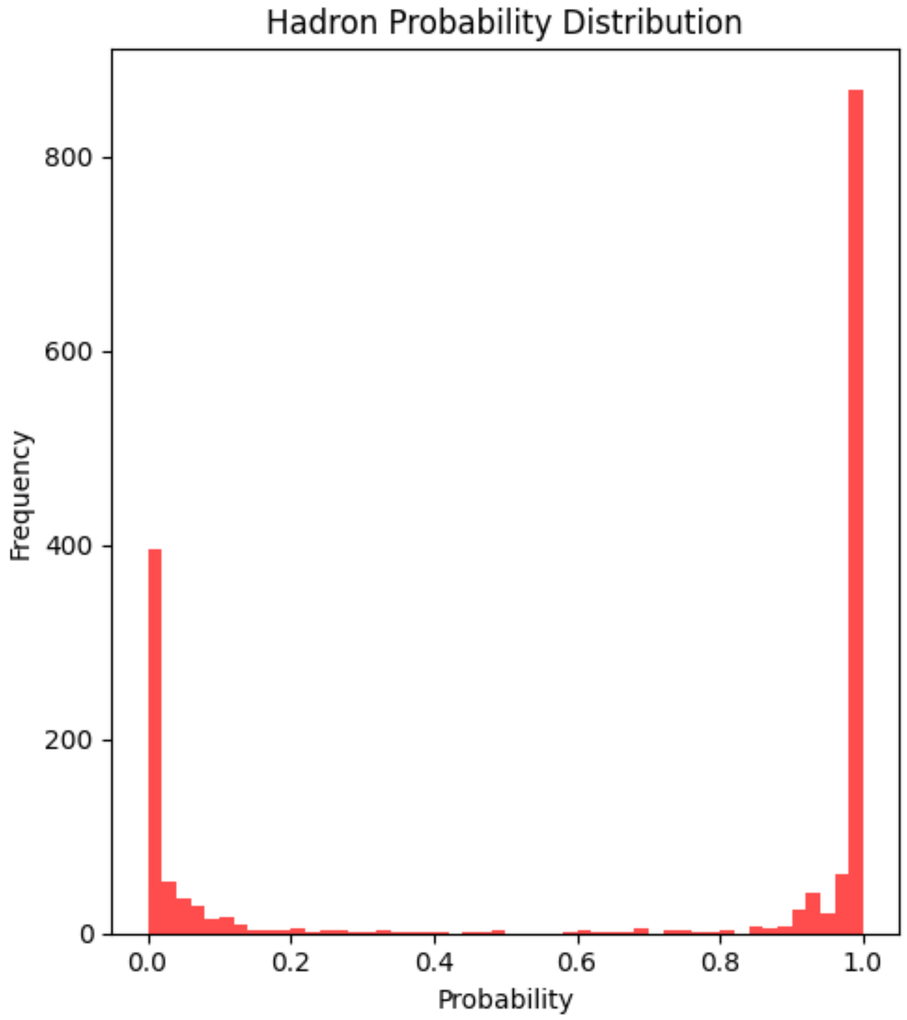
If an instance falls into a leaf node N, the predicted probability p of belonging to class 1 is:

$$p = \frac{\text{count}(N,\ \text{class} = 1)}{\text{count}(N)}$$

Where:

- count (N, class = 1) is the number of instances in node N that belongs to class 1
- count (N) is the total number of instances in node N

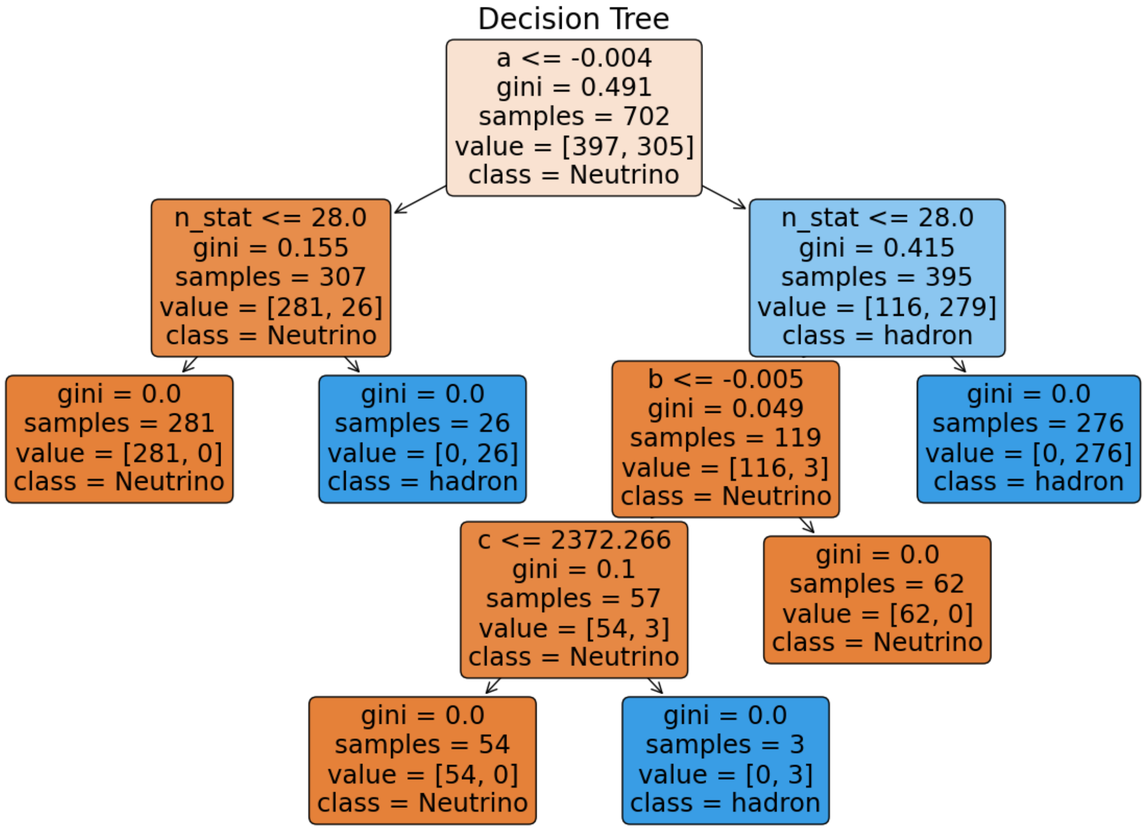


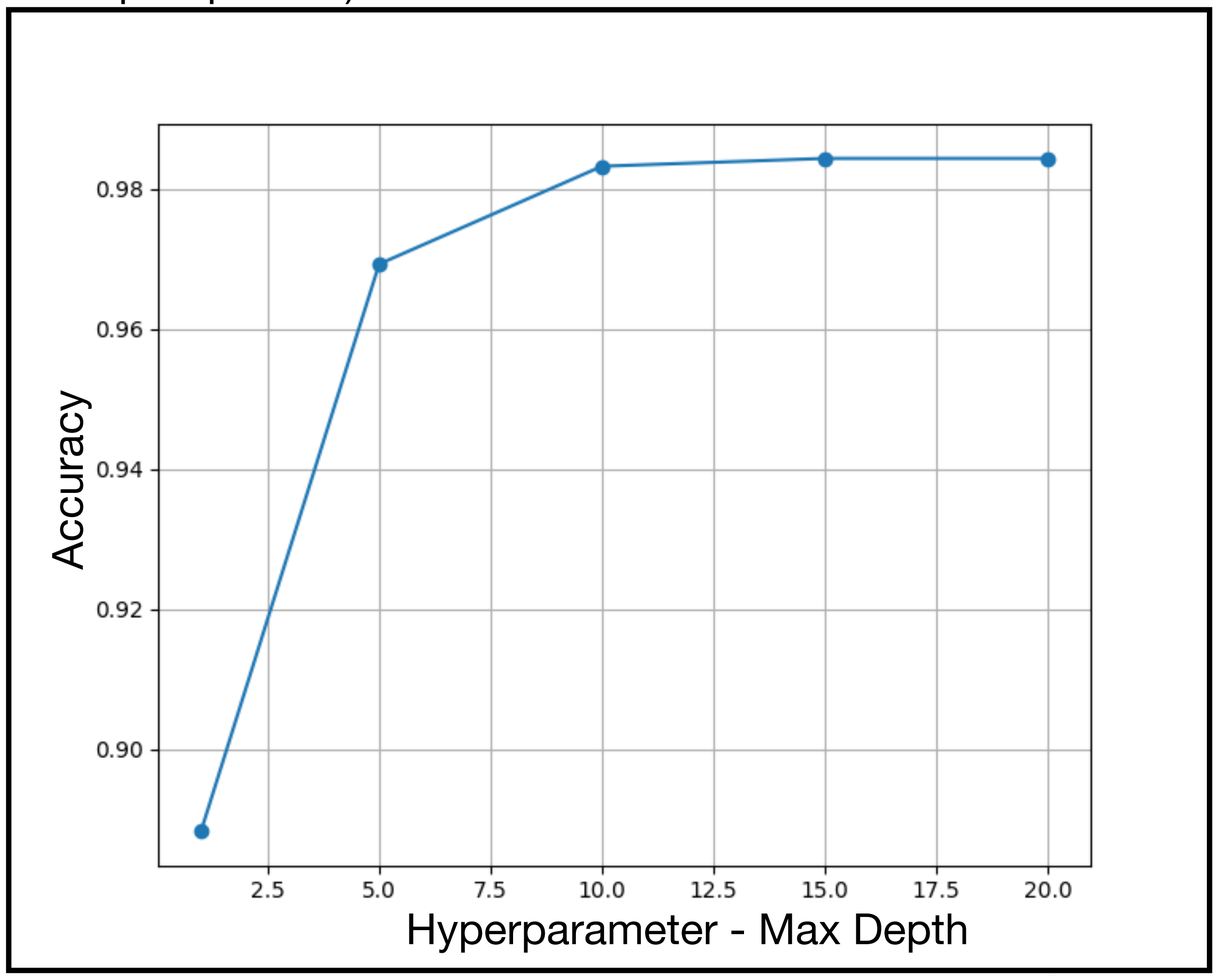


Contact: a.khakurdikar@astro.ru.nl

Back-up

4. **Tree Depth**: The process continues, creating decision nodes at each level of the tree. The depth of the tree depends on factors like the dataset and the stopping criteria (e.g., maximum depth, minimum samples per leaf).





Contact: a.khakurdikar@astro.ru.nl

Radboud Universiteit

Back-up

5. **Leaf Nodes**: When a stopping criterion is met (e.g., maximum depth reached), or if further splitting does not significantly reduce impurity, the algorithm creates leaf nodes. Each leaf node represents a predicted class label.

6. **Pruning**: After building the tree, the algorithm might prune branches that do not contribute significantly to impurity reduction. Pruning helps prevent overfitting, ensuring the model generalizes well to new data.

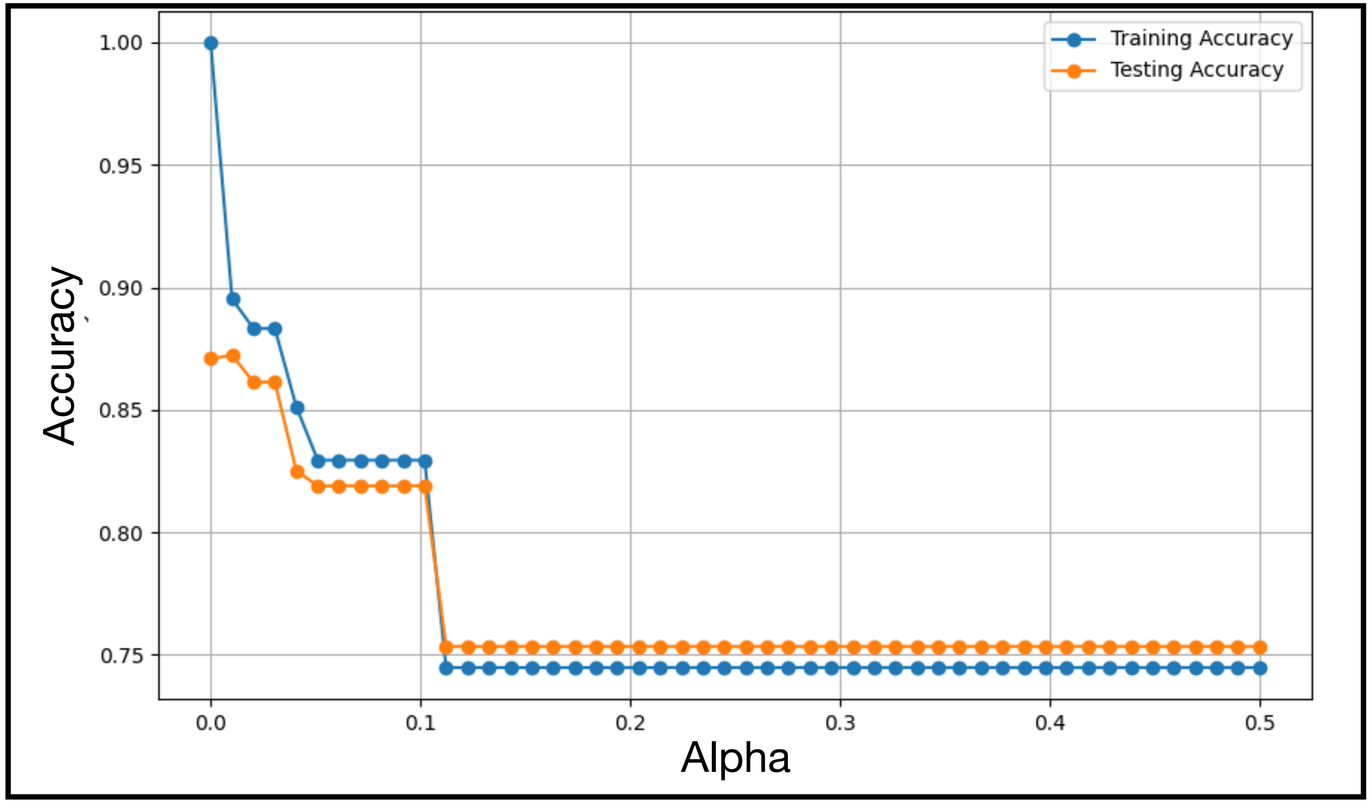
$$R_\alpha(T) = R(T) + \alpha \,.\, |T|$$

- $R_\alpha(T)$ is the cost complexity criterion of tree T with respect to parameter α

- $R(T)$ is the total impurity measure of tree T

- $|T|$ is the number if terminal nodes (leaves) of tree T

The alpha parameter controls the complexity of the decision tree. A higher value of alpha results in more aggressive pruning, leading to simpler trees with fewer nodes. Conversely, a lower value of alpha allows the tree to grow larger and potentially more complex.



Contact: a.khakurdikar@astro.ru.nl

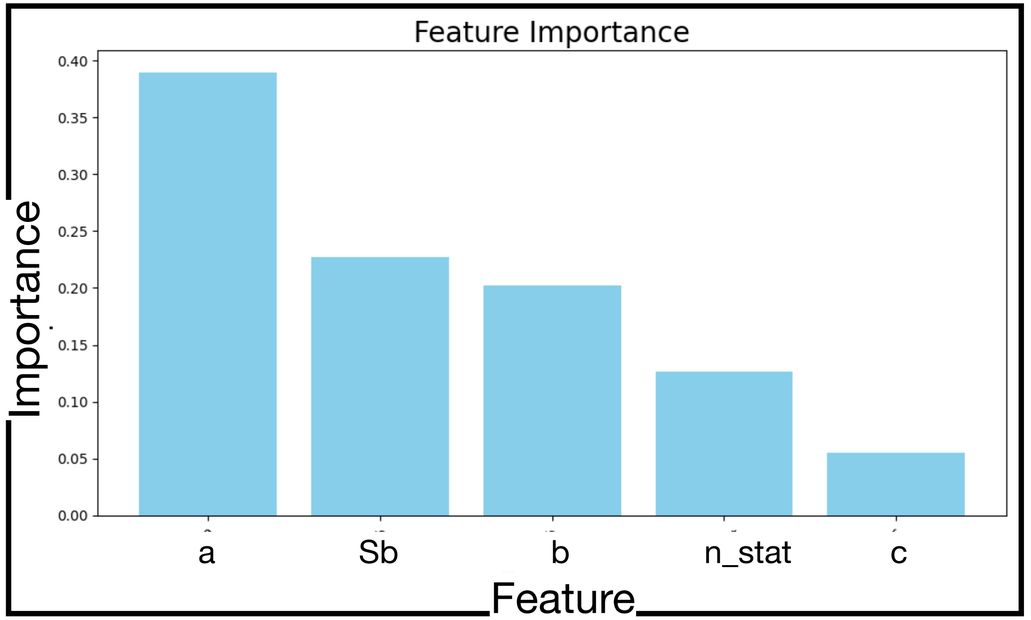
Radboud Universiteit

Back-up

7. **Feature Importance**: After training, the model can provide information about the importance of each feature. Feature importance is calculated based on how much each feature contributes to impurity reduction across all decision nodes. In this case, the final decision tree might have decision nodes based on different features and thresholds (not just 'dist'). The tree structure and decision rules would be determined by the specific features and conditions that minimize impurity during the training process. The resulting tree is a set of decision rules that collectively predict the 'primary_particle' class labels for new instances.

$$\text{Gini Importance (feature)} = \sum_{\text{nodes using feature}} \left(\frac{N_{node}}{N_{total}} \times \text{Gini}_{node} - \frac{N_{left}}{N_{node}} \times \text{Gini}_{left} - \frac{N_{right}}{N_{node}} \times \text{Gini}_{right} \right)$$

- N_{node} is the total number of samples in the node
- N_{left} and N_{right} are the number of samples in the left and right child nodes after the split.
- Gini_{node} is the Gini impurity of the node before the split
- Gini_{left} and Gini_{right} are the Gini impurities of the left and right child nodes after the split.



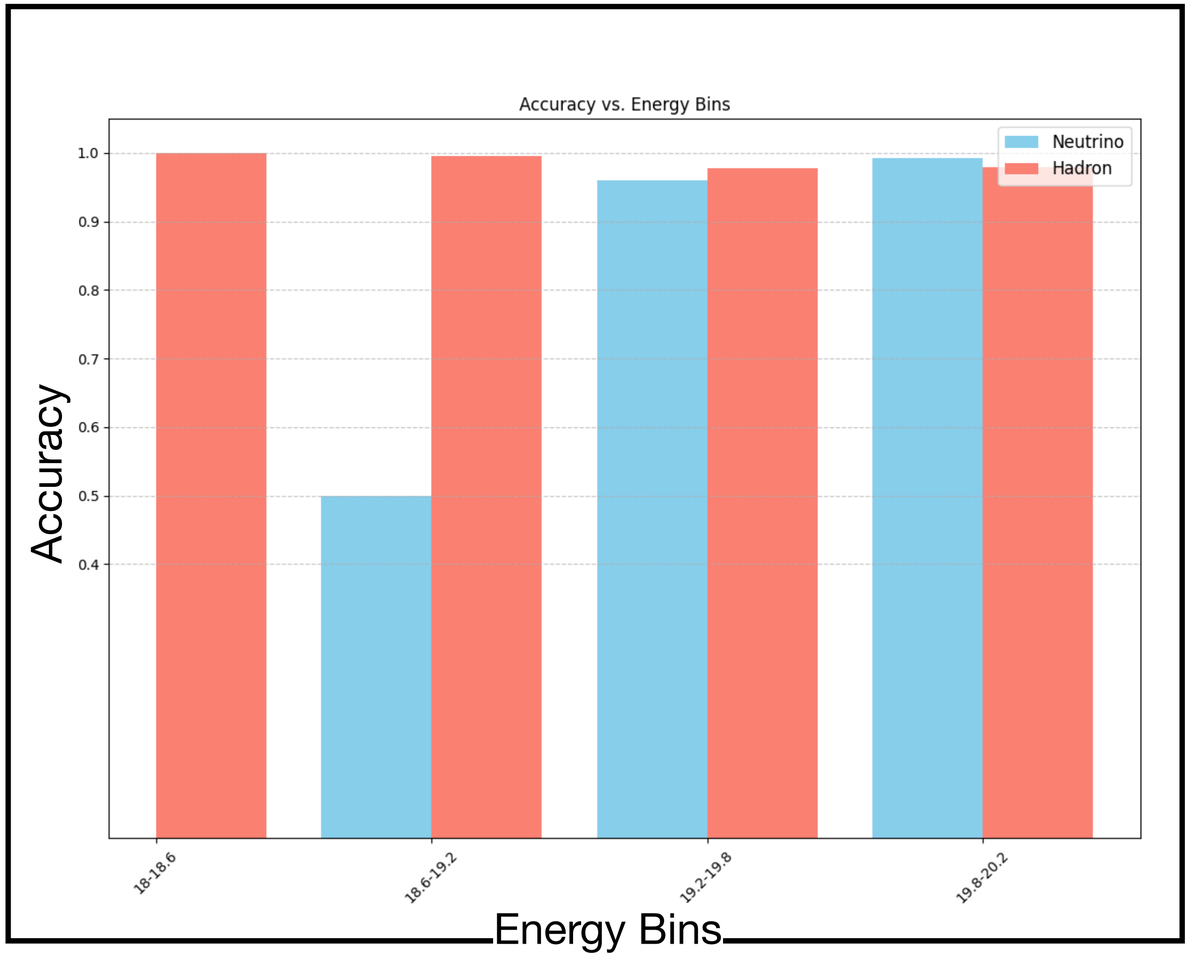
Contact: a.khakurdikar@astro.ru.nl

Back-up

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

1. **Elliptical parabola** (Ground plane - $ax^2 + by^2 + cx + dy + exy + f$) : Accuracy = **0.963**
(Features used: a,b,c,d,e,f, Sb, Nstat (min = 6))

2. **Hyperbola** (Shower plane - $a(x^2 - y^2) + bxy + c$):
Accuracy = **0.832**
(Features used: a,b,c, Sb, Nstat (min = 4))



*final prediction is made by taking the ensemble of 100 individual trees and optimizing the hyper parameters

Contact: a.khakurdikar@astro.ru.nl

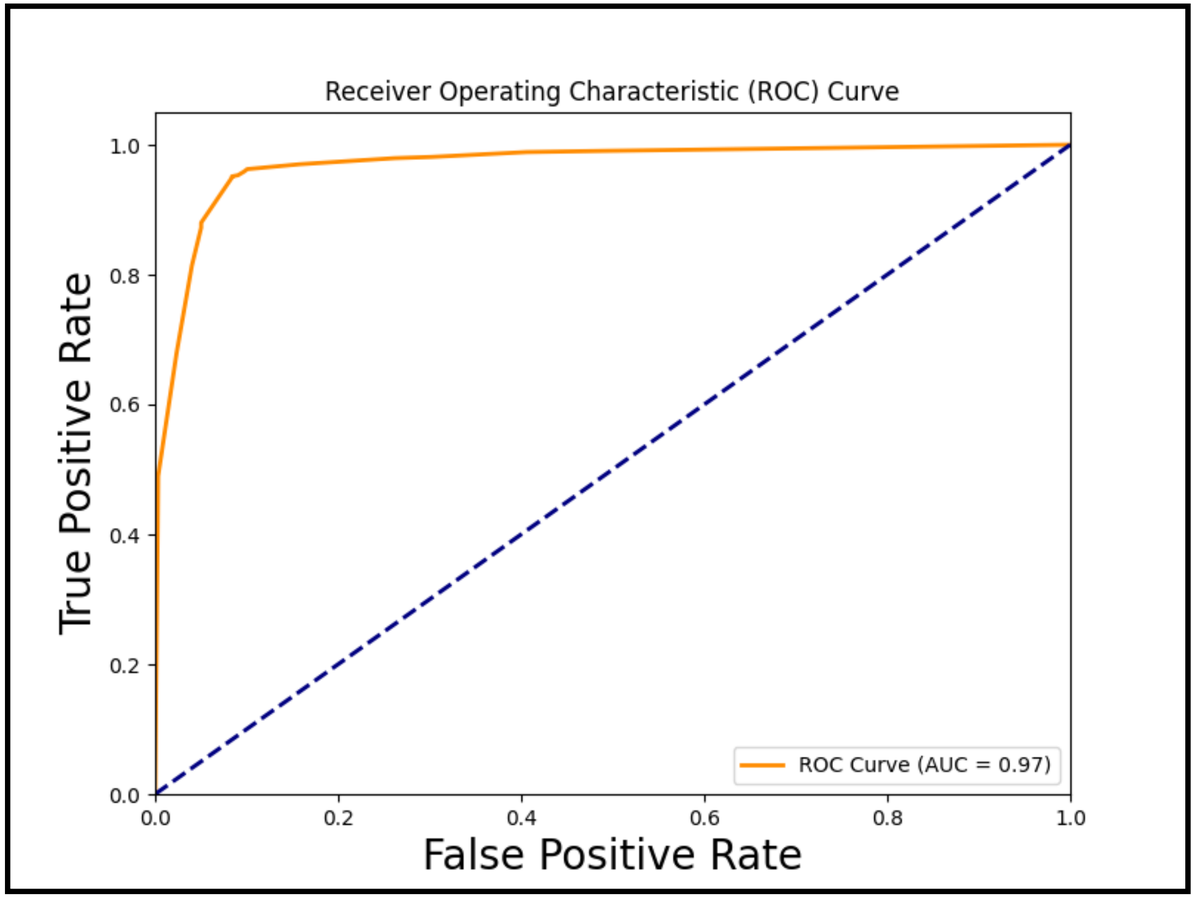
Back-up

- ROC curves are typically used for binary classification problems, where the target variable has two classes

- The area under the ROC curve (AUC) is a measure of the classifier's performance. A perfect classifier would have an AUC of 1.0, while a random classifier would have an AUC of 0.5. Generally, a higher AUC value indicates better classifier performance.

TPR - the ratio of true positives to the total number of actual positive instances. It measures the proportion of actual positive instances that are correctly identified by the classifier.

FPR - the ratio of false positives to the total number of actual negative instances. It measures the proportion of actual negative instances that are incorrectly classified as positive by the classifier.



$$TPR = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \qquad FPR = \frac{\text{False Positives}}{\text{False Positives + True Negatives}}$$

Contact: a.khakurdikar@astro.ru.nl

Radboud Universiteit

Back-up

Dataset contains - E, Zenith, fit parameters, Muon signal, Number of stations and primary particle (neutrino or background (p, fe, he, n))

1. Define the features and the target variable:,

Features (X) -> Fit parameters, Sb, Nstat

Target variable (y) -> primary particle (Neutrino or Hadron)

2. Split the data into training and testing sets: 80% training, 20% testing

3. Random Forest Classifier: trains on the training data.

4. Use the trained model to make predictions on the testing set.

5. Calculate the accuracy of the model

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Simulations available: (1.5 km grid)
Signal-like events: CoREAS showers

- Force Model: Sibyll2.3d
- Type: Electron Neutrino
- CC and NC interactions
- E = 1.0 - 120 EeV
- $\theta = 75^\circ - 85^\circ$, Uniform distribution in $\sin\theta\cos\theta$
- For varying Interaction lengths starting from 100 g/cm^2

Background-like events: CoREAS showers

- Primaries: proton, helium, nitrogen and iron
- E = 1.0 - 120 EeV
- $\theta = 75^\circ - 85^\circ$, Uniform distribution in $\sin\theta\cos\theta$

Contact: a.khakurdikar@astro.ru.nl
Radboud Universiteit

Back-up

The DecisionTreeClassifier in scikit-learn typically employs the CART (Classification and Regression Trees) algorithm.

Consider the features: 'a', 'b', 'c', ‘Sb’, ‘Nstat’ and the target variable ‘primary_particle’.

1. **Binary Tree Structure**: The CART algorithm starts with the entire dataset and selects the feature and threshold that minimize the Gini impurity for a binary split. Let's say the algorithm decides to split based on the ‘a' feature.

2. **Decision Node 1**: The decision node checks if 'a' is less than or equal to a certain threshold. If true, the instance goes to the left child node; otherwise, it goes to the right child node.

3. **Child Nodes**: The left child node might represent instances with 'a' <= Threshold, and the right child node represents instances with 'a' > Threshold. The algorithm repeats this process for each child node, selecting features and thresholds that minimize impurity.

$X_i \leq$ threshold -> left child node, else -> right child node

Contact: a.khakurdikar@astro.ru.nl

Back-up

Assuming two classes: ‘Class A’ and ‘Class B’, and considering a subset of the dataset:

1. Instance 1: a = 1, b =2, c = 0.5, d = 0.8, e = 1.2, f = 0.9, primary = A (Class A)
2. Instance 2: a = 0.8, b =1.5, c = 0.7, d = 1.2, e = 1.5, f = 1.1, primary = B (Class B)

Assuming Class A and Class B have uneven distribution:

$$p_{\text{class A}} = \frac{1}{2} + \epsilon$$

$$p_{\text{class B}} = \frac{1}{2} - \epsilon$$

Contact: a.khakurdikar@astro.ru.nl
Radboud Universiteit

Back-up

Now calculate the Gini impurity for the entire dataset:

$$\text{Gini(S)} = 1 - (p^2_{\text{class A}} + p^2_{\text{class B}})$$

$$\text{Gini(S)} = 1 - \left(\left(\frac{1}{2} + \epsilon\right)^2\right) + \left(\left(\frac{1}{2} - \epsilon\right)^2\right)$$

Source: https://thesai.org/Downloads/Volume11No2/Paper_77-Evaluating_the_Impact_of_GINI_Index.pdf

Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm

Lets consider a split based on a condition, a <= threshold value. The dataset splits into two subsets:

Subset 1 (Left child node): Instance 1 (Class A)

Subset 2 (Right child node): Instance 2 (Class B)

Contact: a.khakurdikar@astro.ru.nl

Back-up

Calculate Gini impurity for each subset:

$$\text{Gini}(S_1) = 1 - (1^2 + 0^2) = 0$$

$$\text{Gini}(S_2) = 1 - (0^2 + 1^2) = 0$$

Now, calculate the Gini gain:

$$\text{Gini\_gain} = \text{Gini}(S) - \left(\frac{|S_1|}{|S|} . \text{Gini}(S_1) + \frac{|S_2|}{|S|} . \text{Gini}(S_2) \right)$$

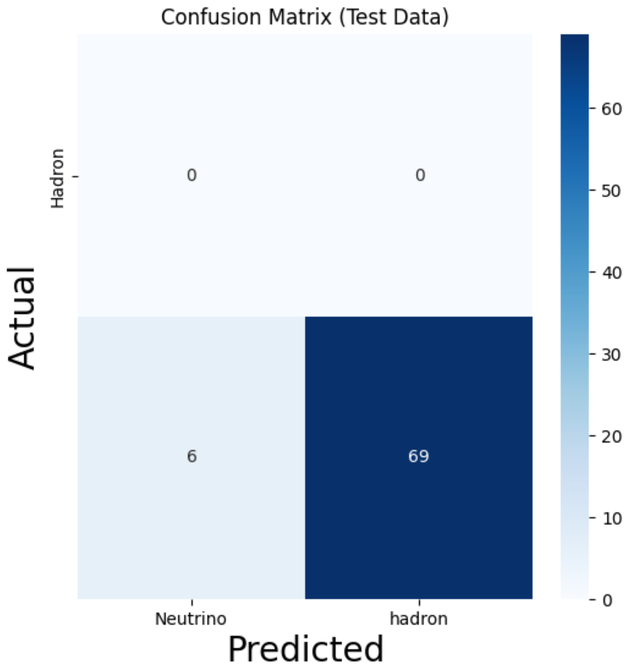
$$\text{Gini\_gain} = 1 - \left(\frac{1}{2}.0 + \frac{1}{2}.0 \right) = 1$$

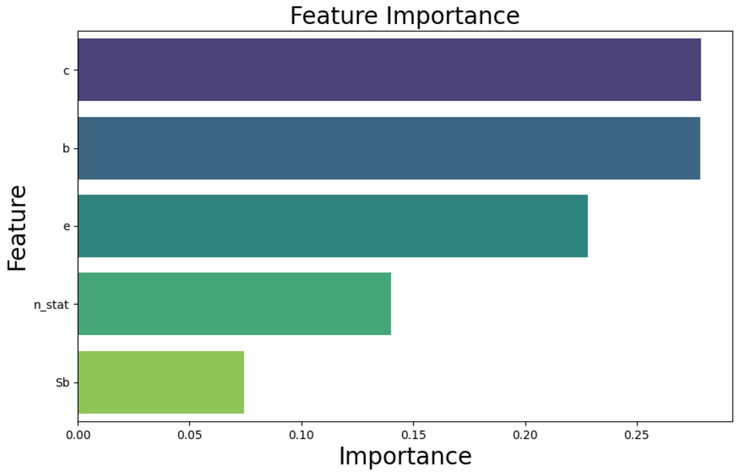
In this case, the Gini gain = 1, indicating that the split based on a <= threshold value is highly effective in reducing impurity. The decision tree algorithm would consider this split when constructing the tree, as it maximizes the Gini gain. This process continues for subsequent nodes in the tree.

Contact: a.khakurdikar@astro.ru.nl
Radboud Universiteit

Back-up

Preliminary






```
Training set - Simulations of hadrons and neutrinos
Testing set - Data (with cuts) labelled as hadrons


Best Hyperparameters:
'max_depth': 7,
'min_samples_leaf': 1,
'min_samples_split': 2,
'n_estimators': 50

Training Accuracy: 0.995427264932838
Test Accuracy: 0.9066666666666666
```

Testing Set:
N = 508
Jan 2023 - Mar 2024
Data cuts
- Energy: 18 - 20.2
- Zenith: 75 - 85 deg
- Min. # station = 4
- Fitting of the data

After cuts, N = 75

energy	theta	a	b	c	d	e	f	Sb	n_stat	predicted prob neutrino
3.654595E+1	81.076209	-0.00004	-0.000027	0.000073	-2.759059	-4.589458	-85577.081870	2.51466	14	0.616896
3.495700E+1	84.704008	-0.00275	-0.000017	0.000431	59.335315	-7.613110	-317490.13110	3.71454	14	0.622647
5.02516E+18	83.583287	-0.00001	-0.000013	0.000043	0.695728	-4.822097	-31416.499360	3.46026	6	0.600596
7.59305E+18	82.412615	0.001187	-0.004644	-0.003712	-61.842881	5.289445	392011.803257	2.77814	8	0.661316
4.065427E+1	78.607122	0.000342	0.000003	-0.000050	-11.205621	-2.481488	92230.715108	2.13199	6	0.649142
3.409763E+1	80.594984	0.000002	-0.000006	0.000031	1.567173	-5.296361	-111798.53600	2.20536	7	0.506253

Contact: a.khakurdikar@astro.ru.nl
Radboud Universiteit