PROCEEDINGS OF SCIENCE

A Multi-Stage Machine Learning Approach to Cosmic Ray Detection in the RNO-G Experiment

The RNO-G Collaboration
(a complete list of authors can be found at the end of the proceedings)

E-mail: martin01021207@huskers.unl.edu, ikravchenko2@nebraska.edu

The Radio Neutrino Observatory-Greenland (RNO-G), deployed at Summit Station, Greenland, aims to detect ultra-high-energy (UHE) neutrinos. To maximize sensitivity, RNO-G operates with low trigger thresholds, leading to data dominated by background noise, including thermal and anthropogenic sources. A potential additional background source is associated with cosmic ray signals coming from cosmic ray-induced air-showers and further sub-showers produced in the ice, which can mimic neutrino signals. Understanding these events is crucial for improving event classification in future neutrino searches.

To address this, two parallel approaches are being developed within the collaboration. The primary method employs a linear discriminant analysis, while an exploratory approach, which this presentation is focused on, utilizes a three-stage event filtering scheme. This scheme sequentially applies a cut-based thermal noise filter, followed by machine learning classifiers—a boosted decision tree (BDT) and a convolutional neural network (CNN)—trained on both simulated cosmic-ray candidates and real background-dominated data. The method effectively rejects background while preserving high signal efficiency.

The presentation showcases this powerful machine learning based analysis, highlighting its performance in distinguishing deep cosmic-ray candidates from other types of impulsive background. These results will inform future RNO-G neutrino searches, enhancing its capability to isolate astrophysical neutrino events.

Corresponding authors: Chao-Hsuan (Martin) Liu[1*], Ilya Kravchenko[1]

[1] *Dept. of Physics and Astronomy, University of Nebraska-Lincoln*
855 N. 16th St., Lincoln, NE 68588, USA

[*] Presenter

*39th International Cosmic Ray Conference (*ICRC2025*)*
15–24 July 2025
Geneva, Switzerland

ICRC 2025
The Astroparticle Physics Conference
Geneva July 15-24, 2025

ISSN 1824-8039. Published by SISSA Medialab.

https://pos.sissa.it/

PoS(ICRC2025)319

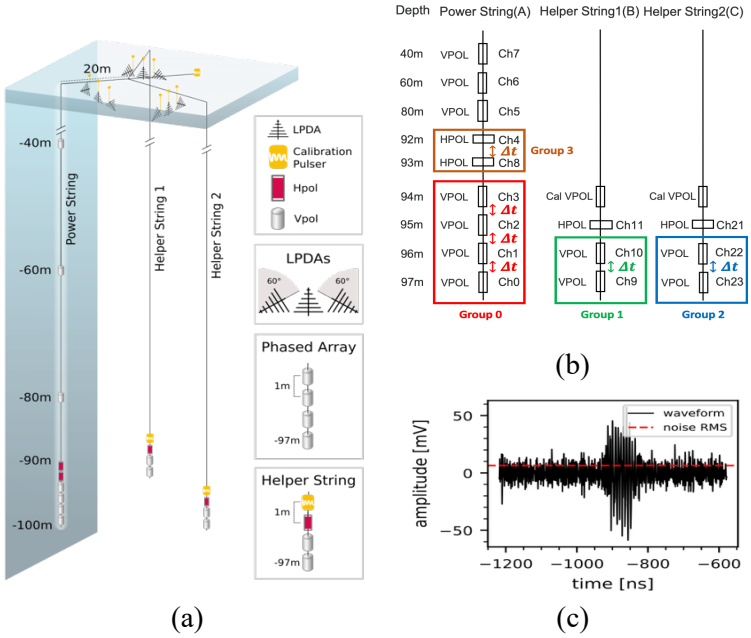


Figure 1: (a) Layout of an RNO-G station. Each station includes 9 LPDA surface antennas, 11 vertically polarized (VPol) in-ice antennas, and 4 horizontally polarized (HPol) in-ice antennas. The four deepest VPol channels on the power string form the phased array (PA), where low-threshold triggers occur. **(b)** Grouping of the 15 in-ice channels used in the hit filter. Coincidence is checked among channels of the same type and spacing (1 m apart) within a 10 ns window. To pass, events must show at least two coincident pairs—one in the PA and one in another group. If both are in the PA, their sequence must match expected signal propagation. **(c)** Example waveform from a single channel. The red dashed line shows the noise RMS.

1. Introduction

1.1 Radio Neutrino Observatory-Greenland (RNO-G)

The Radio Neutrino Observatory-Greenland (RNO-G) is located at Summit Station, Greenland [1]. It's the first ultra-high-energy (UHE) radio observatory in the northern hemisphere and is designed to detect UHE neutrinos through radio signals generated by the Askaryan effect. RNO-G includes 35 stations, each equipped with 24 antennas: 9 log-periodic dipole antennas (LPDAs) on the surface, 11 vertically polarized in-ice antennas, and 4 horizontally polarized in-ice antennas. Channels 0–3 make up the phased array (PA), where the low-threshold triggers are recorded. A diagram of the station layout is shown in Fig. 1(a).

1.2 Deep Cosmic Rays (CR)

Because RNO-G uses low trigger thresholds to improve sensitivity, most recorded events are dominated by background noise, such as thermal fluctuations or human-made interference. Another possible source of background is from cosmic rays—specifically, signals from air showers or secondary showers that develop in the ice. These can sometimes resemble neutrino signals, especially when they trigger the phased array. We refer to such events as *deep cosmic rays*.

1.3 Data Set and Simulation Set

For this analysis, 10% of the low-threshold data from Stations 13 and 23 is designated as a burn sample. This subset is used for training classifiers and estimating background rates. The remaining 90% remains blinded until the final stages of the analysis. Preprocessing includes voltage calibration, block offset correction, cable delay correction, resampling, bandpass filtering, and continuous wave (CW) filtering. For training and testing, we currently use a simplified simulation set based on NuRadioMC [2], where neutrino-induced showers act as stand-ins for in-ice cosmic ray events. A more realistic CR simulation is in development.

1.4 Exploratory Analysis for Deep CR Search

While the main analysis [3] uses a linear discriminant method, this work focuses on a more experimental approach. Here, we implement a three-stage filtering process: first, we apply a thermal noise cut; next, a boosted decision tree (BDT) classifier; and finally, a convolutional neural network (CNN). These models are trained on a mix of simulated CR events and real, background-dominated data. The goal is to improve event classification by efficiently reducing background while retaining potential deep CR candidates.

2. Multi-Stage Event Filtering and Event Selection

2.1 Thermal Noise and Hit Filter

Most radio-frequency (RF) triggered events in RNO-G data are caused by thermal noise. To remove a large portion of these background events before applying machine learning, we use a simple, cut-based "hit filter (HF)." The filter works by first applying a Hilbert transform to each waveform to produce its envelope, from which we identify the amplitude maximum—called a *hit*—and record the corresponding time. The filter then checks whether neighboring antennas registered hits within a 10 ns coincidence window. For an event to pass, there must be at least two such coincident pairs, with at least one pair located in the phased array (PA). As shown in Fig. 1(b), the 15 in-ice channels are grouped based on their positions:

Group 0	Group 1	Group 2	Group 3
Channels 0–3 (PA)	Channels 9 & 10	Channels 22 & 23	Channels 4 & 8 (HPol)

If a coincident pair is found in the PA (Group 0) and another in a different group, the event passes the filter. If both pairs are in the PA, the filter requires them to follow a time sequence consistent with a traveling signal, to further suppress random thermal noise. An additional condition allows an event to pass even without two coincident pairs: if any single channel's envelope peak exceeds 6.5 times its noise RMS—a level that thermal noise is unlikely to reach. This hit filter rejects about 94% of thermal events while preserving approximately 95% of signal events in simulation. Events that pass this stage undergo further filtering by machine learning classifiers.

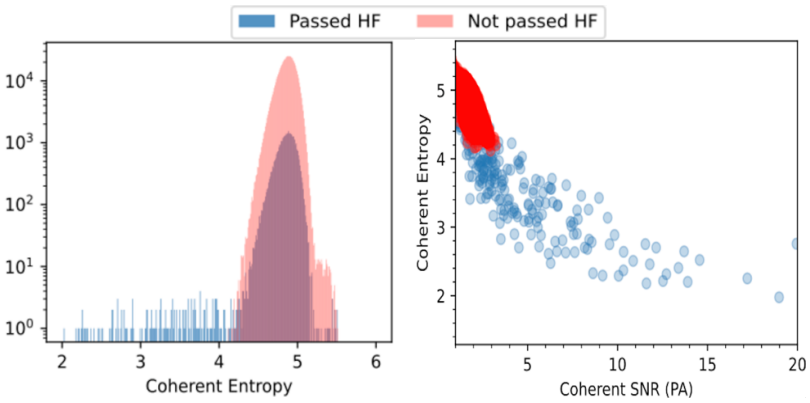
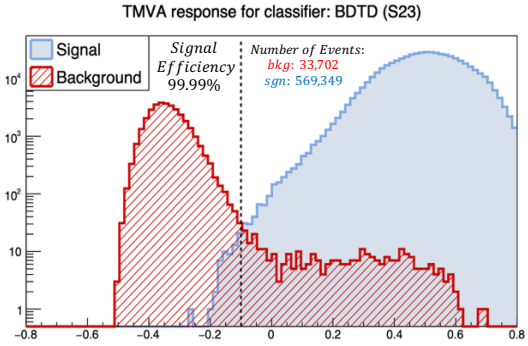


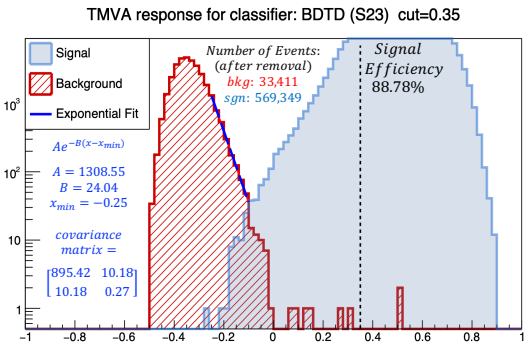
Figure 2: Distribution of coherent entropies for events. Noisy events (coherent entropy > 4.5) that pass the hit filter are randomly selected for the classifier background training set; the rest are used for testing. Events failing the hit filter are excluded from classification. The right panel shows a 2D distribution of coherent entropy vs. coherent SNR.

2.2 BDT Classification

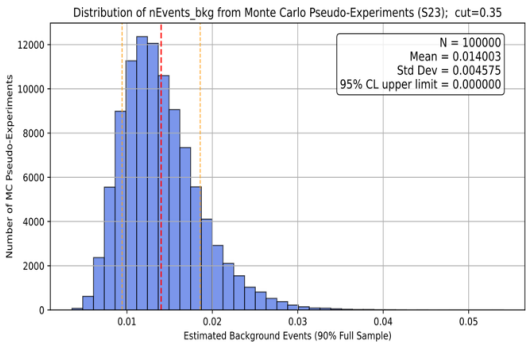
We use a Boosted Decision Tree (BDT) classifier for its speed and convenience, with training and evaluation performed using TMVA [4]. The BDT uses adaptive boosting with a maximum tree depth of 3 and 400 trees in total.



(a)



(b)



(c)

Figure 3: BDT classification results. **(a)** BDT score distributions from the testing sets. The dashed line indicates the score corresponding to 99.99% signal efficiency. **(b)** Score distributions after applying three event-removal criteria: 1. Same run. 2. High trigger rate (>2 Hz). 3. High wind speed (>10 m/s). A BDT score cut at 0.35 is applied. Due to the impact of the anthropogenic event removal cuts on the high-score tail, an intermediate region of the background distribution is fit to an exponential function. **(c)** Estimated number of background events in the blinded 90% dataset, obtained via Monte Carlo integration of the fitted exponential with uncertainties sampled from the fit covariance.

2.2.1 Analysis Variables

For events that pass the hit filter, we extract several analysis variables from the waveforms: signal-to-noise ratio (SNR), Shannon entropy, kurtosis, impulsivity [5], and two quantities derived from the hit filtering step. Waveforms are divided into two groups: one for the four phased array (PA) channels, and another for all 15 in-ice channels (including the PA channels). For each group, a coherently summed waveform (CSW) is constructed. The waveform with the lowest entropy is selected as the reference, and each other waveform is cross-correlated with it. Time lags corresponding to the peak correlations are used to align (or "roll") the waveforms, after which they are summed. This coherent sum enhances the signal pulse while suppressing uncorrelated noise. Each group provides 9 variables: average SNR, coherent SNR, average entropy, coherent entropy, average kurtosis, coherent kurtosis, coherent impulsivity, and the two hit-filter-based variables—number of coincident pairs and number of high hits. Together, the two groups contribute 18 variables per event.

2.2.2 Classification Data Sets

Events from the burn sample that pass the hit filter and have coherent entropy > 4.5 are considered noise-like. From these, 6000–7000 are randomly selected to form the BDT's background training set; the rest are used as the testing background set (see Fig. 2). Simulated signal events passing the hit filter are similarly split into training and testing sets, with 6000–7000 events randomly selected across energies for training.

2.2.3 Classification Results

The 18 input variables are passed to the BDT for training and testing. The BDT output score distributions for the testing set are shown in Fig. 3(a).

2.2.4 Impulsive Anthropogenic Event Removal

Impulsive anthropogenic noise can contaminate the data, so we apply additional event-level cuts after the initial BDT scoring. Given the low expected cosmic-ray rate (~10 events per station per year [6]), we take a conservative approach. We first identify the BDT score corresponding to 99.99% signal efficiency. Among events above this score, we apply three rejection criteria: 1. Remove events from the same run. 2. Remove events from runs with high trigger rates (>2 Hz). 3. Remove events from periods of high wind speed (>10 m/s). The resulting score distributions after these cuts are shown in Fig. 3(b).

2.2.5 BDT Score Cut and Background Estimation for Full Data

To select signal-like events, we apply a BDT score cut of 0.35, rejecting all events below this threshold. See Fig. 3(b). While working with the burn sample, we aim to estimate how many background events might remain in the 90% of blinded data after applying this cut. Since the highest-score tail of the background distribution is already significantly depleted by the anthropogenic event removal step, we instead fit an intermediate section of the right-hand side of the background histogram to an exponential function. This mid-range region provides a more reliable estimate of the underlying background distribution unaffected by the harsh cuts at the tail. Following the procedure in the ARA analysis [7], we: 1. Fit the selected mid section of the background histogram to an exponential function. 2. Use the fit covariance matrix to capture uncertainties. 3. Perform 100,000 Monte Carlo pseudo-experiments, sampling fit parameters from Gaussian distributions defined by the fit and its uncertainties. 4. Integrate each sampled exponential fit from the cut value (0.35) upward. 5. Scale the integrated results by a factor of 9 to estimate the expected background in the 90% of the data. The estimated background distribution and associated statistical uncertainties are shown in Fig. 3(c). While the expected background is low, any surviving events will be subject to a CNN classifier in the next stage, which is designed to reject remaining background and retain true cosmic-ray candidates.

2.3 CNN Classification

The Convolutional Neural Network (CNN) classifier is implemented using PyTorch and integrated with TMVA [4]. Its architecture is summarized as follows:

Layer	Kernel Size	Batch Normalization	Activation Function	Max Pooling
1st convolutional: (1,10)	3×3	Yes: (10)	ReLU	No
2nd convolutional: (10,10)	3×3	Yes: (10)	ReLU	Yes, kernel size 2×2
3rd convolutional: (10,10)	3×3	No	ReLU	Yes, kernel size 2×2
Flatten	–	No	Linear: (640,256) + ReLU	–
Output	–	No	Linear: (256,2) + sigmoid	–

2.3.1 Image Construction

Each input to the CNN is a 32×32 image constructed from waveform features, as shown in Fig. 4. The image consists of four segments:

Row 32 (top)	Hilbert envelope of the coherently summed waveform (CSW) for all 15 in-ice channels.
Row 31 (second from top)	Hilbert envelope of the CSW for the 4 PA channels.
Rows 16–30	Hilbert envelopes of the 15 in-ice channels (Channel 1 to 15, bottom to top).
Rows 1–15 (bottom)	Square root of the energy envelope ($\sqrt{\text{TKEO}}$) [8] for all 15 in-ice channels, ordered from Channel 1 at the bottom to Channel 15.

To fill each row, we compute the RMS of the envelope, divide the envelope into 32 equally sized time windows, and extract the local maximum in each window. These maxima are then divided

by the RMS, yielding 32 unitless values per row.

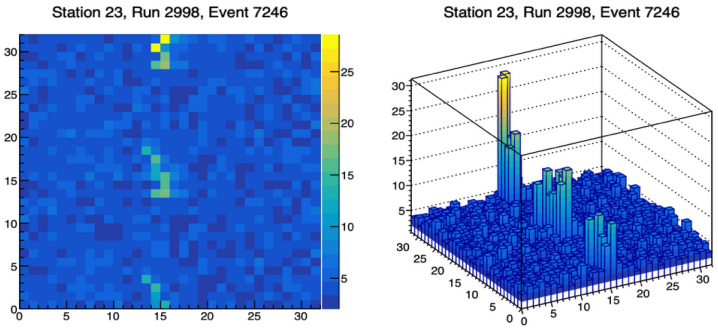


Figure 4: An example of a 32×32 image used in CNN classification. The bottom 15 rows show $\sqrt{\text{TKEO}}$ for all 15 in-ice channels. The next 15 rows display the Hilbert envelopes for those same channels. The top two rows show the Hilbert envelopes of the coherently summed waveform (CSW) for the PA channels and all in-ice channels, respectively. Each bin in a row represents the local maximum of the envelope in a time slice, divided by the envelope's RMS. The right panel provides a 3D visualization of the same image.

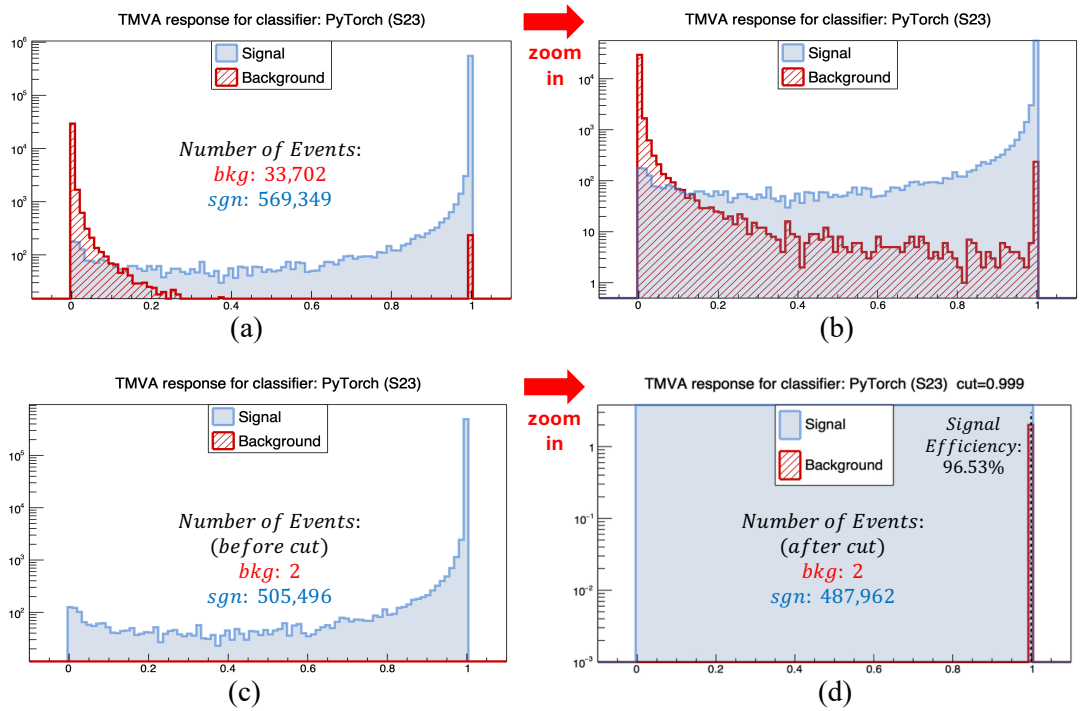


Figure 5: CNN classification results. **(a)** CNN score distributions when applied in parallel to the BDT. **(b)** Zoom-in of (a), highlighting the background distribution. **(c)** CNN score distributions after BDT classification and anthropogenic event removal. **(d)** Zoom-in of (c), showing that two events retain very high CNN scores.

2.3.2 Training and Testing

CNN training uses the same data sets as the BDT (see Section 2.2.2), with the difference that waveforms are converted into images for the CNN instead of using analysis variables. Figs. 5(a) and 5(b) show the CNN score distributions when applied in parallel to the BDT; Figs. 5(c) and 5(d) show the results when the CNN is applied after BDT classification and anthropogenic event removal. Although the CNN provides a powerful classification capability, it requires more computational resources: each event is represented by 1024 image pixels (32×32), compared to 18 input variables for the BDT. Therefore, in the full analysis of the 90% blinded sample, images are only constructed for events that survive the BDT and event removal stages, significantly reducing the processing load.

2.3.3 Classification Results

The CNN effectively eliminates thermal noise events when applying a high score threshold of 0.999. It also serves as a second stage of background rejection for impulsive events that survive the BDT. In Fig. 5(d), two events with high BDT scores also yield high CNN scores, reinforcing their classification as signal-like.

2.4 Directional Reconstruction and Reverse Matched Filtering

2.4.1 Directional Reconstruction

Events passing all three stages — hit filtering, BDT classification, and CNN classification — are subjected to directional reconstruction to determine their zenith θ and azimuth φ. These directionally reconstructed events are then compared to simulated cosmic-ray events using a reverse matched filter approach.

2.4.2 Reverse Matched Filtering

An unconventional approach is adopted at this stage: each surviving candidate event is used as a *template* for matched filtering, which is then applied to the simulated data set. The goal is not to further suppress backgrounds but to identify simulated events that closely resemble the observed event in waveform shape and arrival direction. This helps build confidence in the physical nature of the candidate. To qualify as a good match, we require:

1. Only impulsive channels in the template event must yield matched filtering scores above threshold; other channels should not.
2. Directional agreement with simulation:

$$|\theta_{sim} - \theta_{temp}| \leq 1^\circ \quad \text{and} \quad |\varphi_{sim} - \varphi_{temp}| \leq 2^\circ. \tag{1}$$

Among the Station 23 events from the previous stage, one candidate matches with some simulated events, while another has no matches. The matched filtering results for the event are as follows:

Sim Energy (10^x)	16.00	16.50	17.00	17.50	18.50	19.00
N sim total	92,326	147,993	14,055	237,102	26,621	58,246
N sim matched	0	12	3	0	0	25

2.4.3 Template vs. Simulation

Once matches are identified, the waveform of the real event template is directly compared with a well-matching simulated event. As shown in Fig. 6, the four panels on the left display the individual waveforms of the four PA channels, and the right panel shows the coherently summed waveform (CSW) constructed from the same channels.

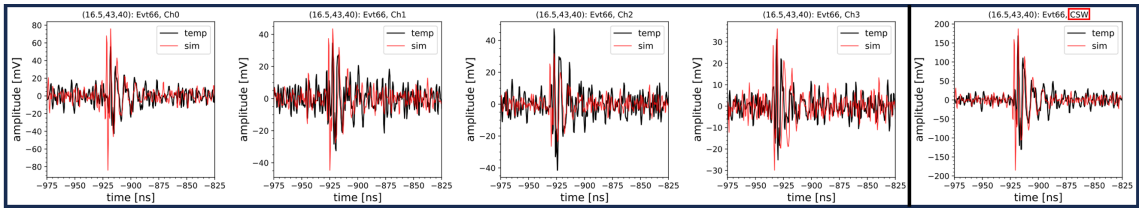


Figure 6: Comparison between a real event template ($\theta = 44^\circ, \varphi = 39^\circ$) and a simulated waveform ($E = 10^{16.5} eV, \theta = 43^\circ, \varphi = 40^\circ$). Left four panels show the four PA channels; right panel shows the CSW of the PA channels.

3. Conclusion

While the candidate event from the burn sample (10% of the full data) has not yet been

fully confirmed as a cosmic-ray (CR) event, it exhibits several features consistent with expectations for a deep in-ice CR event. It has passed all three stages of the event selection pipeline — hit filtering, BDT classification, and CNN classification — and has shown strong agreement with simulated events through directional reconstruction and reverse matched filtering. This makes it a promising candidate. However, further information is required to strengthen the case for its cosmic origin. Polarization measurements and Cherenkov ring identification will play a crucial role in validating the event. Additionally, the simulations used in this analysis are based on a simplified set of NuRadioMC simulations with neutrino-induced showers serving as proxies for CRs. A more realistic CR simulation set is currently under development and will include key features such as the correct interaction geometry and surface effects. Once the updated simulation set is available, classifiers will be retrained to reflect the improved modeling. The remaining 90% of the data from Stations 13 and 23 will then be unblinded, enabling a full-sample search for deep CR events in RNO-G. This will allow a final estimation of the sensitivity and event rate expectations. The current analysis results are summarized in Table 1. This multi-stage machine learning approach has demonstrated strong background rejection power and has provided a robust framework for identifying potential CR signals. These results will inform future RNO-G neutrino searches, enhancing its capability to isolate astrophysical neutrino events.

Stage	**Item**			**Station 13**		**Station 23**		**CR Simulation**		
0	N Events Burn Sample / Sim			609,154		720,413		605,704		
1	N Events After Hit Filtering			39,127		40,329		576,343		
–	N Events Training	N Events Testing		6,306	32,821	6,627	33,702	6,994	569,349	
2-0	BDT	B. Rej. / S. Eff. (%)		97.23		98.98		99.99		
		N False Positives		910		344		–		
2-1	After Event Removal Cuts	N False Positives Left		14		53		–		
		N Events Left Total		31,925		33,411		569,349		
2-2	BDT Score 0.35 Cut	90% Full Sample	Background Estimation	0.00112		0.014		*Signal Efficiency (%)*	86.7	88.8
			95% CL Upper Limit	0		0				
		N Events Left Total		1		2		493,626	505,496	
3	CNN Score 0.999 Cut	N Events Left Total		1		2		*Signal Efficiency (%)*	93.7	96.5
								462,632	487,962	
4	Matched Filtering	N Events Matched		1		1		–		

Table 1: Summary of current analysis results using the 10% burn sample and a simplified simulation set. The multi-stage machine learning approach — including hit filtering, BDT classification, and CNN classification — is applied to search for deep in-ice cosmic-ray events in the RNO-G experiment.

References

[1] **RNO-G** Collaboration, A. Nelles et al. PoS(ICRC2025)1129

[2] C. Glaser et al. *NuRadioMC* arXiv:1906.01670v2 [astro-ph.IM]

[3] **RNO-G** Collaboration, B. Hendricks et al. PoS(ICRC2025)1057

[4] A. Hoecker et al. *TMVA* arXiv:physics/0703039v5 [physics.data-an]

[5] P.W. Gorham et al. *Constraints on the diffuse high-energy neutrino flux from the third flight of ANITA* arXiv:1803.02719v3 [astro-ph.HE], pp. 9.

[6] Alan Coleman et al. *In-Ice Askaryan Emission from Air Showers* arXiv:2410.08615v2 [astro-ph.HE]

[7] P. Allison et al. *A low-threshold UHE neutrino search with ARA* arXiv:2202.07080v1 [astro ph.HE]

[8] Kaiser, J. F. (1990). *On a simple algorithm to calculate the 'energy' of a signal*. pp. 381–384

Full Author List: RNO-G (June 30th, 2025)

S. Agarwal[1], J. A. Aguilar[2], N. Alden[3], S. Ali[1], P. Allison[4], M. Betts[5], D. Besson[1], A. Bishop[6], O. Botner[7], S. Bouma[8], S. Buitink[9,10], R. Camphyn[2], J. Chan[6], S. Chiche[2], B. A. Clark[11], A. Coleman[7], K. Couberly[1], S. de Kockere[12], K. D. de Vries[12], C. Deaconu[3], P. Giri[13], C. Glaser[7], T. Glüsenkamp[7], H. Gui[4], A. Hallgren[7], S. Hallmann[14,8], J. C. Hanson[15], K. Helbing[16], B. Hendricks[5], J. Henrichs[14,8], N. Heyer[7], C. Hornhuber[1], E. Huesca Santiago[14], K. Hughes[4], A. Jaitly[14,8], T. Karg[14], A. Karle[6], J. L. Kelley[6], C. Kopper[8], M. Korntheuer[2,12], M. Kowalski[14,17], I. Kravchenko[13], R. Krebs[5], M. Kugelmeier[6], R. Lahmann[8], C.-H. Liu[13], M. J. Marsee[18], K. Mulrey[10], M. Muzio[6,5], A. Nelles[14,8], A. Novikov[19], A. Nozdrina[4], E. Oberla[3], B. Oeyen[20], N. Punsuebsay[19], L. Pyras[14,21], M. Ravn[7], A. Rifaie[16], D. Ryckbosch[20], F. Schlüter[2], O. Scholten[12,22], D. Seckel[19], M. F. H. Seikh[1], Z. S. Selcuk[14,8], J. Stachurska[20], J. Stoffels[12], S. Toscano[2], D. Tosi[6], J. Tutt[5], D. J. Van Den Broeck[12,9], N. van Eijndhoven[12], A. G. Vieregg[3], A. Vijai[11], C. Welling[3], D. R. Williams[18], P. Windischhofer[3], S. Wissel[5], R. Young[1], A. Zink[8]

[1] University of Kansas, Dept. of Physics and Astronomy, Lawrence, KS 66045, USA

[2] Université Libre de Bruxelles, Science Faculty CP230, B-1050 Brussels, Belgium

[3] Dept. of Physics, Dept. of Astronomy & Astrophysics, Enrico Fermi Inst., Kavli Inst. for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA

[4] Dept. of Physics, Center for Cosmology and AstroParticle Physics, Ohio State University, Columbus, OH 43210, USA

[5] Dept. of Physics, Dept. of Astronomy & Astrophysics, Center for Multimessenger Astrophysics, Institute of Gravitation and the Cosmos, Pennsylvania State University, University Park, PA 16802, USA

[6] Wisconsin IceCube Particle Astrophysics Center (WIPAC) and Dept. of Physics, University of Wisconsin-Madison, Madison, WI 53703, USA

[7] Uppsala University, Dept. of Physics and Astronomy, Uppsala, SE-752 37, Sweden

[8] Erlangen Centre for Astroparticle Physics (ECAP), Friedrich-Alexander-University Erlangen-Nürnberg, 91058 Erlangen, Germany

[9] Vrije Universiteit Brussel, Astrophysical Institute, Pleinlaan 2, 1050 Brussels, Belgium

[10] Dept. of Astrophysics/IMAPP, Radboud University, PO Box 9010, 6500 GL, The Netherlands

[11] Department of Physics, University of Maryland, College Park, MD 20742, USA

[12] Vrije Universiteit Brussel, Dienst ELEM, B-1050 Brussels, Belgium

[13] Dept. of Physics and Astronomy, Univ. of Nebraska-Lincoln, NE, 68588, USA

[14] Deutsches Elektronen-Synchrotron DESY, Platanenallee 6, 15738 Zeuthen, Germany

[15] Whittier College, Whittier, CA 90602, USA

[16] Dept. of Physics, University of Wuppertal D-42119 Wuppertal, Germany

[17] Institut für Physik, Humboldt-Universität zu Berlin, 12489 Berlin, Germany

[18] Dept. of Physics and Astronomy, University of Alabama, Tuscaloosa, AL 35487, USA

[19] Dept. of Physics and Astronomy, University of Delaware, Newark, DE 19716, USA

[20] Ghent University, Dept. of Physics and Astronomy, B-9000 Gent, Belgium

[21] Department of Physics and Astronomy, University of Utah, Salt Lake City, UT 84112, USA

[22] Kapteyn Institute, University of Groningen, PO Box 800, 9700 AV, The Netherlands

PoS(ICRC2025)319

Acknowledgments

We are thankful to the support staff at Summit Station for making RNO-G possible. We also acknowledge our colleagues from the British Antarctic Survey for building and operating the BigRAID drill for our project.

We would like to acknowledge our home institutions and funding agencies for supporting the RNO-G work; in particular the Belgian Funds for Scientific Research (FRS-FNRS and FWO) and the FWO programme for International Research Infrastructure (IRI), the National Science Foundation (NSF Award IDs 2112352, 2111232, 2111410, and collaborative awards 2310122 through 2310129), and the IceCube EPSCoR Initiative (Award ID 2019597), the Helmholtz Association, the Swedish Research Council (VR, Grant 2021-05449 and 2021-00158), the University of Chicago Research Computing Center, and the European Union under the European Unions Horizon 2020 research and innovation programme (grant agreements No 805486), as well as (ERC, Pro-RNO-G No 101115122 and NuRadioOpt No 101116890).