PROCEEDINGS OF SCIENCE

Machine learning in Baikal-GVD experiment

I. Kharuk,[a,b,*] **G. Safronov,**[a,c] **A. Matseiko**[a,b] **and A. Leonov**[a,b]

[a] *Institute for Nuclear Research of the Russian Academy of Sciences,*
60th anniversary of October Avenue, 7a, 117312, Moscow, Russia

[b] *Moscow Institute of Physics and Technology,*
Institutskiy Lane, 9, 141700, Dolgoprudny, Russia

[c] *Joint Institute for Nuclear Research,*
12 Joliot-Curie 6, Dubna, Moscow Region, 141980, Russia

E-mail: ivan.kharuk@phystech.edu, grigorybs@gmail.com, matseiko.av@phystech.edu, leonov.ayu@phystech.edu

Baikal-GVD is a large-volume underwater neutrino telescope located in Lake Baikal, Russia. We report on machine learning techniques used for the analysis of its data. Namely, we discuss neural networks developed for the following goals: (1) suppression of noise activations of Baikal-GVD's optical modules due to the natural luminescence of Baikal water; (2) identification of neutrino-induced events and estimation of their flux; (3) reconstruction of arrival direction of incoming neutrinos. It is shown that the accuracy of developed methods surpass that of analogous standard reconstruction techniques on Monte-Carlo simulated data.

38th International Cosmic Ray Conference (ICRC2023)
26th of July - 3rd of August, 2023
Nagoya, Japan

[*]Speaker

PoS(ICRC2023)1077

1. Introduction

Experimental neutrino astronomy is a relatively young and rapidly evolving branch of astrophysics. Such an interest in neutrino studies is due to their unique properties. First, unlike hadrons, neutrinos are not affected by cosmological magnetic fields. Hence their reconstructed arrival direction point directly to their source. Second, in contrast to high energy photons, the universe is almost transparent to neutrinos. Lastly, neutrinos are important in context of multi-messenger astronomy. Together with electromagnetic radiation, gravitational waves, and air shower, they form the four "messengers". Detection or non-detection of signals in any of these channels sheds light on the physics beyond astrophysical events.

Baikal-GVD is an underwater neutrino telescope located in Lake Baikal, Russia, and is the largest in Northern Hemisphere. It is aimed at studying cosmological neutrinos in 1 TeV – 100 PeV range, and, as of 2022, has an effective volume of 0.5 km^3 with respect to cascade-like neutrino-induced events. The target effective volume, 1 km^3, is planned to be reached by 2030. The combinatin of data from the three lagsest neutrino telescope, i.e. of Baikal-GVD, IceCube [1], and KM3NeT [2], allows for comprehensive analysis of neutrino astrophysics over the whole sky.

The interaction of a cosmological neutrino with water medium would result in the appearance of relativistic particles that produce Cherenkov radiation. To register this light, Baikal-GVD employs 25-cm optical modules with Hamamatsu R7081-100 photomultipliers. This allows to register neutrinos arriving solely from under the ground, since the sky emits substantial background noise from extensive air showers (EAS). The optical modules are grouped into vertical strings, having 36 modules on each string with a distance of 15m between them. The OMs are located at a depth of 750–1275 m below the surface. 8 strings, organized in approximately regular heptagon-shape in a horizontal projection (one string per edge and one at the center), form a cluster. As of 2022, there are 11 clusters with an average distance of 300 m between them [3].

The coordinates of each OM are measured in real-time leveraging an acoustic positioning system. Lasers are used to synchronize OM clocks, the synchronization error is within 3 ns.

The triggering condition for registering an *event* is the following: within 100 ns interval, two adjacent OMs registered signals with integral charge depositions of at least 4.5 and 1.5 photo-electrons (p.e.). If this condition is met, the data from all OMs with charge deposition exceeding 0.3 p.e. is read out and saved as well.

In the present report, we discuss neural network developed for data analysis of Baikal-GVD data. Application of machine learning techniques in other neutrino telescopes can be found in [4–8].

2. Background noise suppression

OMs are located underwater, which makes them subject to the natural luminescence of Baikal water. This results in the registration of *noise* hits, which constitute up to 90% of the data collected in an event. Before reconstructing neutrino's properties, one should suppress these noise hits. This is the first problem that we are going to address using machine learning techniques.

Noise hits are uncorrelated, with a frequency of 20-100 Hz, depending on depth and season, with a typical charge deposition of approximately 1 p.e. On the other hand, *signal* hits, stemming from the propagation of relativistic particle through the effective volume of the detector, have larger charge deposition and form patterns. One way to suppress noise hits would be to increase the threshold for data recording, from 0.3. p.e. to several p.e. However, this would result in a loss of some of the signal hits, thus hindering an accurate reconstruction of neutrino's properties. This necessitates an introduction of effective filtering algorithms. Since signal hits from complicated patterns, machine learning techniques are expected to have higher accuracy than algorithmic reconstruction procedures.

For training a neural network to distinguish between noise and signal hits, we used Mote-Carlo simulated data. For each of the simulated hits, we know whether it was a signal or a noise one. This allows us pose the problem of rejecting the noise as as a supervised segmentation task.

Mote-Carlo simulation includes muon neutrino-induced events, as well as air shower-induced ones. Some of the relativistic muons appearing during the evolution of an air showers might reach the detector, resulting in a registration of an event. Such events constitute majority of the data and cannot easily rejected without analysis. Note also that such events are indistinguishable from muon neutrino-induced events coming from above the horizon. Due to this reason, all neutrino-induced events were simulated as coming from under the horizon.

The simulation procedure includes full modelling of air shower evolution, detectors' responses, and takes into account photons scattering in the water. The noise hits are simulated in accordance with theirs experimentally estimated spectrum and frequency.

Signals recorded by OM are split into discrete impulses, called hits. Each hit is characterized by the following quantities: (1-3) corresponding OM coordinates; (4) OM activation time; (5) integral charge, disposed during the hit. For each event, this data is used as input to our neural network.

The simulated data consists of approximately 10^7 events of each type. We decided to take an equal number of air shower-induced and neutrino-induced events to avoid bias related to different signatures of these two types of events. It was verified that neural network's metrics are not sensitive to varying this proportion.The Monte-Carlo dataset has been split into training, test, and validation sets at a ratio of 8/1/1.

The signal hits of a given event happen close in time. This provides physical motivation for the following data representation — for a given event, all triggered OMs are ordered in a one-dimensional array, in accordance to their activation times. This facilitates the problem of identifying signal hits for the neural network, and efficiently leverages the temporal structure of the events. We found this data representation to be the most efficient for the task of noise rejection.

The developed neural network has a U-net-like architecture [9]. For a making prediction for a given hit, it leverages the information from both neighboring and long-range hits, which improves the accuracy of the predictions. With the motivation to make use of a temporal structure of the events, additional LSTM layers has been appended before and after the U-net block. Figure 1 demonstrates a schematic representation of the neural network's architecture.

The following metrics were chosen to estimate the effectiveness of the developed neural network:

- Signal purity (precision) — the fraction of correctly predicted signal hits from all events, recognized by the neural network as signal ones;
- Survival efficiency (recall) — the fraction of correctly predicted signal hits from all signal hits in the data;

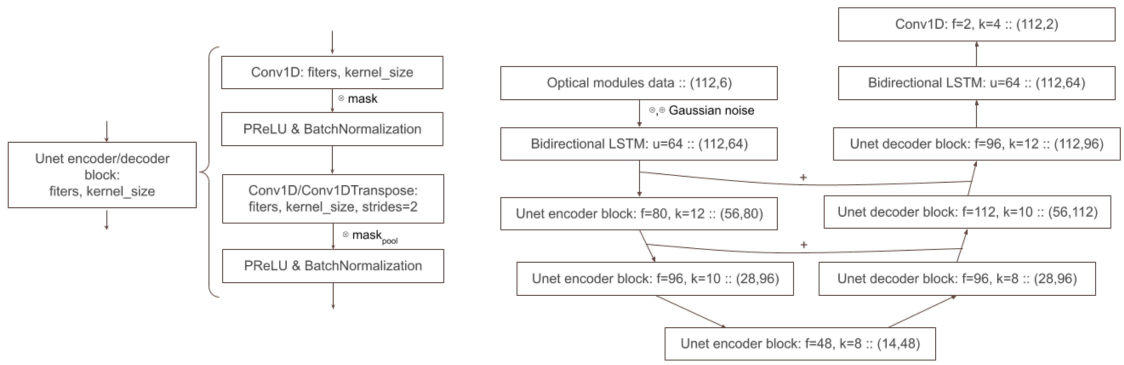

Figure 1: The architecture of the neural network, developed for the suppression of noise hits.

- Dispersion of the time residuals.

The time residuals is calculated as a difference between the expected time of the hit registration and the actual one, which stems from light scattering in the water. The importance of this metric manifests for the problem of neutrino arrival angle reconstruction, where only hits with small time delays are to be selected.

A special loss function has been implemented to minimize the time residual's dispersion. Apart from a penalty for predicting the wrong hit class, an additional penalty is given for identifying hits with large time residuals as signal ones. Such approach allowed to minimize the relative metric by 10%.

We have compared the metrics for the noise suppression using the neuarl network and the standard algorithm [10, 11]. For this purpose we have fixed neural network's classification threshold so that both methods have the same dispersion of time residuals, 5 ns. For the developed method, this yielded 99.5% precision and 96% recall, while the standard approach the metrics were 95.5% and 95%, accordingly.

3. Identifying neutrino-induced events

After suppressing noise hits in the data, it can be used for identification of neutrino-induced events. The proportion of neutrino-induced to air shower-induced events in the data is approximately 1:10^6–10^7. The majority of air-shower induced events can be easily discerned by reconstructing muon's direction of propagation in water. However, such method has low accuracy for cascade-like events. This necessitates the introduction of additional filters, which result in the reduction of effective neutrino exposure. This motivates for the development of a neural network for the task, which would not have this drawback.

The same Monte-Carlo simulation was leveraged as the training data. The neural network, described in the previous section, was employed to filter out noise hits. Since it is known, whether each given event is neutrino-induced or air shower-induced, the present task can be formulated as a binary classification problem. The one-dimensional time-ordered representation of the data, described in the previous section, was found to be the optimal one for the given task as well.

A standard convolution-based neural network was chosen as the basis. Then, task-specific modifications have been applied, such as a gradual decrease in the data dimension ("length" of the one-dimensional time-ordered array), as well as the skip connections. Moreover, appending a recurrent layer (namely, LSTM) before and after the convolutional block of the network allowed us to improve the metrics and minimize their dependence on the "length" of the input events.

Event identified by the neural network should have as small number of false-neutrino identification as possible, while preserving the majority of true-neutrino-induced events. For this purpose, we have made the following. First, we used special loss function, focal loss [12]. It facilitates learning for the events, that are difficult to classify, even if those occur rarely in the data. Second, we set weights for air shower-induced and neutrino-induced events to 10 and 1, correspondingly. This strongly penalizes network for identifying neutrino-induced event as air-shower induced one, and hence increases signal purity. Finally, we trained 3 neural networks and averaged their predictions. This allows us to further increase signal purity by leveraging the effect of a "collective" decision.

After training a neural network we observed that, starting from a certain classification threshold, there were no false-neutrino identifications, while approximately 90% of true-neutrino events were preserved. In general, this allows one to make catalogs of neutrino-induced events with a given false-positive rate by choosing the proper classification threshold.

We also developed a method for estimating neutrino flux based on the prediction of the above-described neural network. Let us call the fraction of neutrino- and air shower-induced events to the right of the classification threshold as exposure and suppression, accordingly. For a trained neural network, they can be evaluated as a function of the classification threshold on the test data set. The resulting functions can be considered as internal characteristics, inherent to the neural network. Then, for a give data set, the number of neutrino-induced events may be evaluated using the following formula:

$$N = \frac{n(\xi) - S(\xi)n(0)}{E(\xi) - S(\xi)} \; . \tag{1}$$

Here, ξ is the classification threshold, $E(\xi)$ and $S(\xi)$ are the estimated exposition and suppression, and $n(\xi)$ is the number of events to the right to the classification threshold.

The errors of estimating $E(\xi)$ and $S(\xi)$ are defined by the number of Monte-Carlo events, employed for their estimation. The identified number of neutrino-induced event candidates, $n(\xi)$, is a random parameter, distributed in accordance to the Poisson law. This allows to estimate the error (variance) of the reconstructed number of neutrino-induced events, given by formula 1. We fix the classification threshold by requiring the resulting error to be as small as possible.

Figure 2 demonstrates the number of reconstructed neutrino-induced events and the corresponding error as a function of the classification threshold on a particular data set. As the graph illustrates, the developed method yields a successful identification of 65 neutrino-induced events on top of 500 000 background air shower-induced events, with an error of around 10 events.

4. Neutrino's incoming direction reconstruction

Since neutrinos are not affected by cosmological magnetic fields, their arrival direction point directly to their source. Identifying their sources and spectrum are the primary goals of neutrino telescopes. Below we report on the neural network developed for this purpose.

PoS(ICRC2023)1077

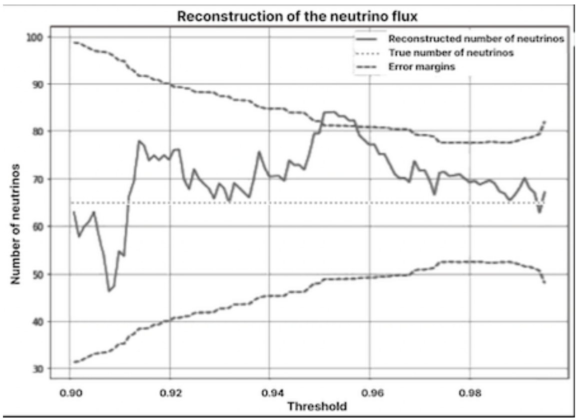

Figure 2: The architecture of the neural network, employed for noise suppression.

For this task, we used the same Monte-Carlo data set as before, additionally limited to solely muon neutrino-induced events. The free path of relativistic muons, produced when a muon neutrino interacts with the environment, have a free path length of approximately 1 kilometer. As a result, they produce well-distinguishable tracks, which allows to reconstruct the direction of muon's propagation vector with substantial accuracy. It almost coincides with the original neutrino's arrival direction, due to the momentum conservation law.

The standard algorithmic procedure was employed to remove noise hits from the data. Additionally, we considered events with at least 8 signal hits on at least 2 strings. This is a standard condition for the algorithmic method of reconstructing neutrino's arrival direction. Such choice of data processing steps allowed us to directly compare the resulting accuracy of neutrino's arrival direction reconstruction by the algorithmic approach and by the neural network.

Since the direction of neutrino's arrival is known for each event, the task can be formulated as the regression problem, i.e. reconstructing azimuth and polar angles. More precisely, we reconstruct the unit vector collinear with the vector of neutrino's arrival direction.

In comparison to the two previous tasks, in this case we employed graph neural network with dynamical edges [13]. The graph neural networks are an efficient solution for tasks with a complex internal data structure, which cannot be represented as regular sequences or arrays. Although one can still use the one-dimensional time-ordered representation of the data described above, graph representation of the data turned out to be more efficient.

We introduced OMs as graph vertices and initiated the adjacency matrix so that 4 closest (in time) OMs are connected. The graph neural network updates multiple times the values that characterize the graph's vertices and edges, leveraging the information from neighboring vertices and their neighbors. At the final stage, the information from entire graph is aggregated to predict the neutrino's arrival direction.

The accuracy of reconstructing neutrino's arriving direction is summarized in table 1. Graph neural network shows better performance than the standard approach, allowing to reduce 50% angular resolution by 0.5 °.

Standard reconstruction			
Metric	**Azimuth angle**	**Polar angle**	**Direction**
50% Angular resolution	5.42	0.53	2.62
68% Angular resolution	13.2	0.95	5.62
Graph neural network			
Metric	**Azimuth angle**	**Polar angle**	**Direction**
50% Angular resolution	4.16	0.53	2.10
68% Angular resolution	7.13	0.83	3.18

Table 1: Angular resolutions and directions of the neutrino arrival

5. Conclusion

The presented neural network constitute the first step toward introducing neural networks to the analysis of Baikal-GVD data. The performance of the designed networks is compatible to or superior to the standard algorithmic reconstruction procedures. Presently, the possibility of applying the developed methods to the analysis of real experimental data is under validation. The novelty of our approach is that we use temporal structure of events, rather than its geometric properties.

Acknowledgments

The work was supported by the Russian Science Foundation grant no. 22-22-20063.

References

[1] IceCube collaboration, M. G. Aartsen et al., *Evidence for High-Energy Extraterrestrial Neutrinos at the IceCube Detector*, *Science* **342** (2013) 1242856, [1311.5238].

[2] KM3NeT collaboration, S. Aiello et al., *Sensitivity of the KM3NeT/ARCA neutrino telescope to point-like neutrino sources*, *Astropart. Phys.* **111** (2019) 100–110, [1810.08499].

[3] Y. Malyshkin, B.-G. collaboration et al., *Baikal-gvd neutrino telescope: Design reference 2022*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1050** (2023) 168117.

[4] IceCube collaboration, N. Choma, F. Monti, L. Gerhardt, T. Palczewski, Z. Ronaghi, Prabhat et al., *Graph Neural Networks for IceCube Signal Classification*, 1809.06166.

[5] IceCube collaboration, M. Huennefeld, *Deep Learning in Physics exemplified by the Reconstruction of Muon-Neutrino Events in IceCube*, *PoS* **ICRC2017** (2018) 1057.

[6] M. Huennefeld, *Reconstruction techniques in icecube using convolutional and generative neural networks*, in *EPJ Web of Conferences*, vol. 207, p. 05005, EDP Sciences, 2019.

[7] KM3NeT collaboration, S. Reck, D. Guderian, G. Vermariën and A. Domi, *Graph neural networks for reconstruction and classification in KM3NeT*, *JINST* **16** (2021) C10011, [2107.13375].

[8] S. Aiello, A. Albert, S. A. Garre, Z. Aly, F. Ameli, M. Andre et al., *Event reconstruction for km3net/orca using convolutional neural networks*, *Journal of Instrumentation* **15** (2020) P10005.

[9] O. Ronneberger, P. Fischer and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[10] Baikal-GVD collaboration, I. Belolaptikov et al., *Neutrino Telescope in Lake Baikal: Present and Nearest Future*, *PoS* **ICRC2021** (2021) 002, [2109.14344].

[11] Baikal-GVD collaboration, A. D. Avrorin et al., *An efficient hit finding algorithm for Baikal-GVD muon reconstruction*, *PoS* **ICRC2021** (2021) 1063, [2108.00208].

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, *Focal loss for dense object detection*, in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[13] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein and J. M. Solomon, *Dynamic graph cnn for learning on point clouds*, *Acm Transactions On Graphics (tog)* **38** (2019) 1–12.