PROCEEDINGS OF SCIENCE

Dark rate reduction with machine learning techniques for the Hyper-Kamiokande experiment

Aurora Langella,[a,b,*] **Lucas N. Machado**[c] **and Bernardino Spisso**[d]

[a] *Università degli Studi di Napoli Federico II, Naples, Italy*
[b] *INFN - Sezione di Napoli*
[c] *University of Glasgow*
[d] *INFN - Sezione di Napoli, gruppo collegato di Salerno*
E-mail: alangella@na.infn.it, lucas.nascimentomachado@glasgow.ac.uk, spisso@na.infn.it

The next generation water-Cherenkov detector Hyper-Kamiokande, is currently under construction in Japan and it is expected to be ready for data taking in 2027. Thanks to its huge fiducial volume and high statistics, Hyper-Kamiokande will contribute to many investigations such as CP-violation, determination of neutrino mass ordering and potential observations of neutrinos from astrophysical sources. To increase the sensitivity of the detector, Hyper-Kamiokande will have a hybrid configuration of photo-detectors: thousands of 20-inch photomultipliers tubes will be combined with modules containing 3-inches photomultipliers arranged inside a pressure-resistant vessel, called *Multi Photomultipliers Tubes modules*. Many efforts are on-going to reduce the expected dark counts for a detector geometry which includes both photo-detector modules. Machine learning-based techniques are being developed to reduce the detector's overall dark rates, which could have a significant impact on Hyper-Kamiokande's sensitivity to low-energy neutrinos.

41st International Conference on High Energy physics - ICHEP2022
6-13 July, 2022
Bologna, Italy

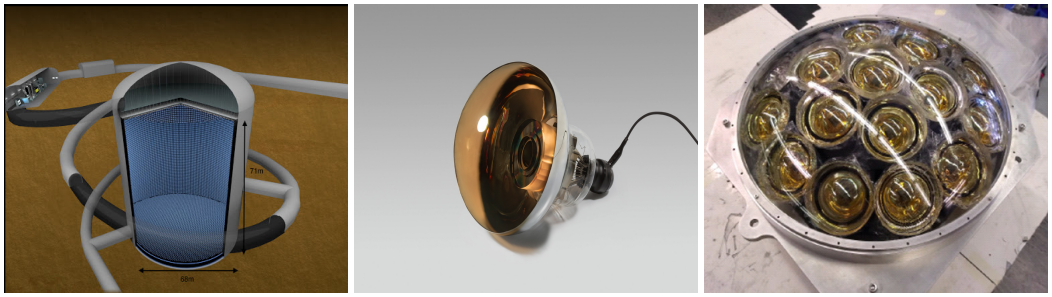
*Speaker

PoS(ICHEP2022)601

1. Introduction

Hyper-Kamiokande (Hyper-K) is a multi-purpose experiment under construction in Japan for the observation of atmospheric, solar and accelerator neutrino oscillations, for neutrino astrophysics, proton decay and physics beyond the Standard Model. In particular, it will observe neutrino beams produced at J-PARC accelerator complex, about 300 km far, to investigate leptonic CP violation. The tunnel construction started in May 2021 while the data taking is expected to start in 2027.
The detector consists of a cylindrical tank with a fiducial mass of 187 ktons and filled with highly transparent purified water which plays two roles: a target material for incoming neutrinos and a source of nucleons to decay. A schematic view of Hyper-K is shown in Fig. 1(a).
Photomultiplier Tubes (PMTs) will be used for the light detection in Hyper-K. These photo-sensors, characterized by a single-photon sensitivity, enable the reconstruction of the spatial and timing distributions of the Cherenkov photons which are emitted by secondary particles from neutrino interactions in water and nucleon decays. The PMTs are placed within a support structure that divides the tank in two regions. For the inner part of the tank, a hybrid configuration will be used, which involves a combination between 20-inches Hamamatsu Photonics R12860-HQE PMTs with Box-and-Line dynode type (B&L) and the so-called multi-PMT (mPMT) modules (Fig. 1(b) and 1(c)). These modules, originally designed for the KM3NeT experiment [1] and optimized for Hyper-K requirements, consist of cylindrical pressure-resistant vessels with 19 3-inches PMTs, full readout electronics and power supply inside. The PMTs are placed on a support structure that allows each photo-sensor to have a specific orientation inside the detector.
The outer part acts mainly as veto for entering particles and will be monitored by around 10k 3-inches PMTs.

Detailed simulation studies are on-going to investigate the impact of the photo-detection system on Hyper-K physics performances. Preliminary results show that one of the strengths of the high-efficient photo-detection system combining B&L PMTs and mPMTs lies in the possibility of improving vertex resolutions. This can lead to a larger fiducial volume as well as better energy resolution. Due to a high background near the detector walls, in fact, it's necessary to apply a fiducial volume cut when reconstructing events with a consequent statistics drop. At the same time, in order to reconstruct the energy of the incoming particle, it's important to identify where



(a) A schematic view of the Hyper-Kamiokande experiment (b) B&L prototype (c) mPMT prototype

Figure 1

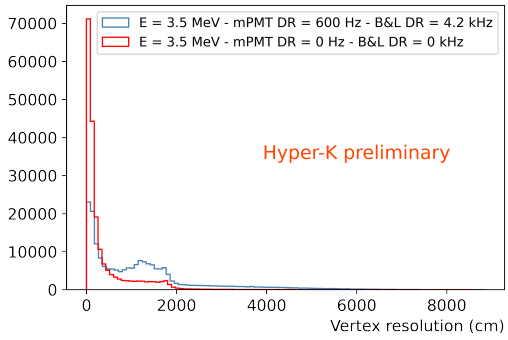


Figure 2: Comparison of the vertex resolution distribution for a sample of 3.5 MeV electrons without considering dark rate for both mPMT and B&L PMT (red line) and including dark rate nominal value for both photo-sensors (blue line).

the interaction was (due to the regressed energy dependence on the photon's path), thus a worse vertex resolution means also a worse energy resolution. An improvement of these performances would mean a great benefit in particular for the low energy neutrino detection (e.g. solar neutrinos, supernova neutrinos).

Simulation studies on Hyper-K performances demonstrate that the better vertex resolution comes from the overall lower dark rate. Fig.2 shows the comparison of the vertex resolution distribution, defined as the Euclidean distance between the reconstructed and true vertices, in two different cases: in the first one (red line) it is simulated a 3.5 electrons sample without dark rate for both B&L PMT and mPMT; in the second one (blue line) it is simulated a 3.5 MeV electron sample with dark rates of 600 Hz for the PMTs in the mPMT and of 4.2 kHz for the B&L PMT (which are the expected values for both photo-sensors). Thus, in the dark rate case, it is clearly visible a bump in the distribution for higher values of vertex resolution (around 1000 cm) and a much longer tail which in turn worsen the expected vertex resolution. For this reason, studies are being performed to further reduce the overall dark rates. By implementing multivariate statistical methods to reconstruction algorithms, backgrounds due to dark rates can be reduced, improving the sensitivity to low energy neutrinos.

The goal of this work, in particular, is to apply a machine learning classifier to reject events coming from dark rate or that heavily mis-reconstructed. We adopt the Boosted Decision Tree (BDT) [2] method implemented with *scikit-learn* machine learning libraries.

In the following a description of the analysis performed and of the first preliminary results will be given.

2. Hyper-Kamiokande event simulation

The datasets used for the training and validation of the BDT are simulated using the simulation package Water Cherenkov Simulator [3] (WCSim) and reconstructed with the Low Energy Analysis Framework (LEAF) [4].

WCSim is a very flexible Geant4 [5] based program, adopted by the T2K, Super-K and Hyper-K Collaboration for developing and simulating large water Cherenkov detectors.

LEAF is a tool developed by the Hyper-Kamiokande Collaboration for low energy event reconstruction, i.e. from few MeV to few tens MeV. Output variables of LEAF reconstruction are mostly

Table 1: LEAF output variables and relative importance for the most efficient BDT.

Variable	Importance	Definition	Variable	Importance	Definition
lf intime	0.242832	Nr. of hits within a time window used by LEAF	mPMT hits 50	0.030523	Nr. of hits in the mPMTs within 50 ns window
ID hits 200	0.162972	Nr. of hits in the ID within 200 ns window	mPMT hits 200	0.014775	Nr. of hits in the mPMTs within 200 ns window
lf NLL	0.146056	Negative Log Likelihood for best vertex candidate	mPMT hits 400	0.011890	Nr. of hits in the mPMTs within 400 ns window
ID hits 50	0.101106	Nr. of hits in the ID within 50 ns window	fromwall	0.009037	Distance from the detector wall
ID hits	0.069899	Nr. of hits in the ID	lf r2	0.003880	Squared radius of rec. vertex on the X-Y plane
ID hits 400	0.066533	Nr. of hits in the ID within 400 ns window	lf vertex[2]	0.003676	Reconstructed vertex, Z
mPMT hits	0.049383	Nr. of hits in the mPMTs	rawhit num	0.003179	Collection of hits, including dark noise
digithit num	0.048075	Nr. of digitized hits	lf vertex[1]	0.001658	Reconstructed vertex, X
lf vertx[3]	0.033012	Reconstructed vertex, T	lf vertex[0]	0.001514	Reconstructed vertex, Y

related to the reconstructed vertex and the number of hits in different time windows. A more detailed description of the variables is reported in Table 1.
For this work, we considered Hyper-K geometry (i.e. a cylindrical tank with a radius of 32.4 m and a height of 65.8 m) with 20k B&L PMTs with a dark rate of 4.2 kHz and 2k mPMTs with a dark rate for the single 3-inches PMT of 600 Hz.
For the training dataset we used 3 MeV electrons with a Gaussian profile (σ = 0.5 MeV) as signal events while background is composed by only dark rate events.
For validation we used electrons with energies in the range 1–18 MeV.

3. BDT Analysis and Results

The first step has been the tuning of the following BDT parameters, in order to get the most efficient classifier (for more details on BDT parameters, check [2]):

1. Ntrees (**NT**) = number of decision trees in the BDT;;
2. Minimum samples in leaf nodes (**NL**) = the percentage of the input sample that is required to make a new leaf;
3. AdaBoostBeta (**AB**) = the weight of AdaBoost for the classifier;

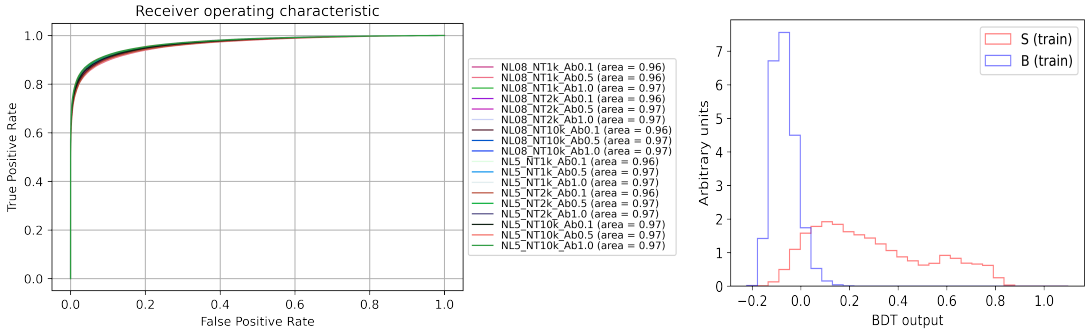
BDTs were trained using the following combinations of values for these three parameters:

$$\textbf{NT} = [1\text{k}, 2\text{k}, 10\text{k}] \text{ , } \textbf{NL} = [0.08\%, 5\%] \text{ , } \textbf{AB} = [0.1, 0.5, 1.0]$$

A comparison of the Receiver Operating Characteristic (ROC) curves, which shows the performances of the BDTs, is shown in Figure 3(a). All of the BDTs have similar performances and are efficient to discriminate signal (low energy electron events) from background (only dark rate events). For each of these BDTs, the signal/background separation has been evaluated. Even though the efficiencies are very similar, the best choice based on computation time for the classification and reduction efficiency for the DR-like events results to be:

$$\textbf{NT} = 1\text{k}, \textbf{NL} = 5\%, \textbf{AB} = 1.0$$

Fig. 3(b) reports the signal/background separation for the chosen BDT while in Table 1 is reported the importance of all the features used for training. Based on the BDT score distribution we defined a cut to reduce events that are only due to dark rate or are heavily mis-reconstructed.
To evaluate the efficiency of the BDT in reducing these events, we compared the vertex resolution



(a) ROC curve of all the BDTs trained.

(b) Signal/background separation plot of the most efficient BDT.

Figure 3

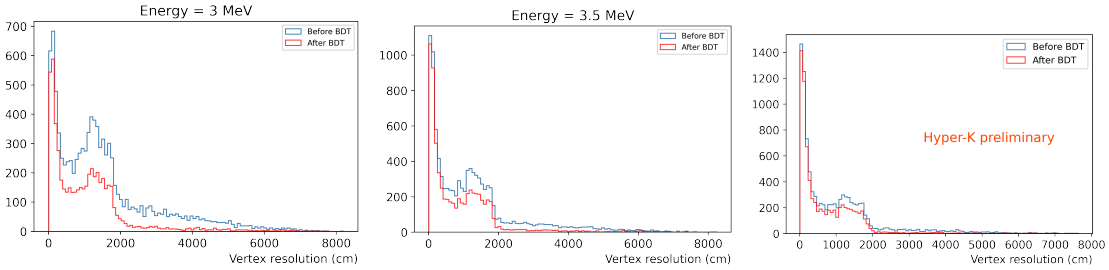


Figure 4: Comparison of the vertex resolution distribution with a without the BDT cut for three samples of 3, 3.5 and 4 MeV electrons.

distribution with and without applying the BDT cut for electrons of different energies. Fig. 4 shows the comparison for three samples of 3 MeV, 3.5 MeV and 4 MeV. These plots show how the BDT classifier is able to reduce the second peak and the longer tail (mainly due to dark rate events, as observed in Figure 2) while keeping most of the events of the first peak (mainly composed of electron events correctly reconstructed).

Based on these distributions, we defined as nominal value for the vertex resolution the point p such that the range 0–p contains the 68% of the distribution. We used this definition to compare the vertex resolution as a function of energy with and without applying the BDT cut.

As shown in Fig. 5, for 1 MeV electrons, we don't see any improvement as very low energy signals are usually mis-reconstructed, thus, there is no explicit improvement in the vertex resolution. For electrons samples with energies in the range 2–7 MeV a clear improvement is visible. For higher energies samples no improvements are found, which might be related to the bias in the BDT to the energy of the sample used as signal for training (3.5 MeV electrons with Gaussian profile). Further investigation is needed.

4. Conclusion

The Hyper-Kamiokande experiment, expected to start operations in 2027, aims to obtain many important results in several physics studies, thanks to its large fiducial volume and and high-efficient

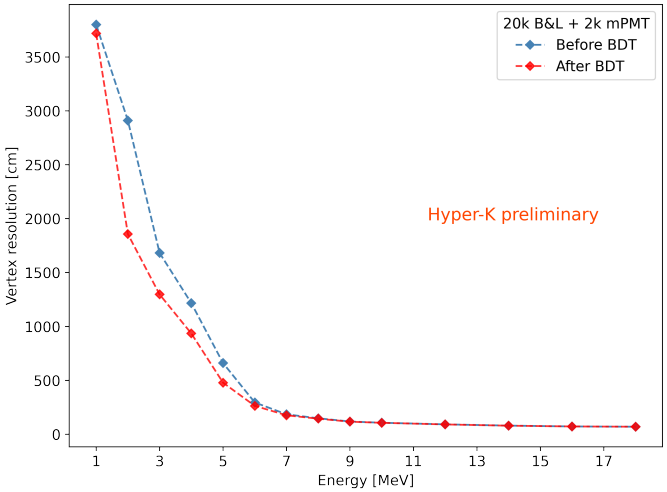


Figure 5: Vertex resolution as a function of energy with and without applying the BDT cut.

photo-detection system.

We applied a multivariate analysis techniques study to explore the possibility of reducing the background due to dark rate events, in order to improve the vertex resolution. This improvement can be a great benefit in particular for the reconstruction in the low energy region (e.g. solar neutrinos, supernova neutrinos). We trained a BDT classifier with samples of 3 MeV electrons (with a dark rate of 600 Hz for the PMTs in the mPMT and of 4.2 kHz for the B&L PMT) as signal and only dark rate samples as background in order to reject events that are due only to dark rate hits or are heavily mis-reconstructed. We obtained promising preliminary results, in particular we see an improvement for the vertex resolution in the energy range 2–7 MeV. For the future we plan to further optimize the BDT classifier and compare it with other machine learning methods. We also plan to train the classifier with electron samples of different energy to overcome a possible energy-bias, as well as with different particles. Next we'll extend this study to reduce other significant backgrounds in the low energy region, such as radioactive background, to further improve the resolution of low energy events.

References

[1] S. Adrián-Martínez et al. (KM3NeT), Eur. Phys. J. C74, 3056 (2014), arXiv:1405.0839 [astro-ph.IM].

[2] Coadou, Yann. (2013). Boosted Decision Trees and Applications. EPJ Web of Conferences. 55. 02004, 10.1051/epjconf/20135502004

[3] Web, https://github.com/WCSim/WCSim

[4] Web, https://github.com/bquilain/LEAF

[5] S. Agostinelli et al. (GEANT4), Nucl. Instrum. Meth. A506, 250 (2003).