# 6th International Workshop on Deep Learning in Computational Physics (DLCP-2022)

Wednesday, 6 July 2022 - Friday, 8 July 2022


# Book of Abstracts

# Contents

**Posters** / **21**

## Neuromorphic Improvement of the Weizsaecker Formula

**Author:** Mihai-Octavian Dima[1]

[1] *JINR - MLIT*

**Corresponding Author:** modima@cern.ch

Yearly nuclide mass data is fitted to improved versions of the Weizsaecker formula. The present attempt at furthering the precision of this endeavour aims to reach beyond just precision, and obtain predictive capability about the "Stability Island" of nuclides. The method is to perform a fit to a recent improved liquid drop model with isotonic shift. The residuals are then fed to a neural network, with a number of "feature" quantities. The results are then discussed in view of their perspective to predict the "Stability Island".

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Posters** / **27**

## NARX Neural Prediction of Oscilational Instability at the IBR-2M reactor

**Author:** Mihai-Tiberiu Dima[1]

[1] *JINR - MLIT*

**Corresponding Author:** mtdima@jinr.ru

During the start-up regime of the IBR-2M power fluctuations appear, which the AR system dampens. Their origin is not completely clear, however it is known that the major reactivity sources are from design - respectively the OPO and DPO reflectors (axial fluctuations towards the active zone and their relative phase of intersecting each other facing the center of the active zone).
A neuromorphic solution is sought to anticipate (5-10 s) such fluctuations. I present encouraging preliminary results obtained with a Non-linear Autoregressive Exogenous neural network, the main features of the fluctuations being anticipatable.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Posters** / **50**

## Sampling of Integrand for Integral Calculation Using Shallow Neural Network

**Authors:** Vladimir Papoyan[1]; Alexander Ayriyan[2]; Hovik Grigorian[1]

[1] *JINR*

[2] *Laboratory of Information Technologies, JINR*

**Corresponding Author:** papoyan8@gmail.com

We present the effect of using the Metropolis-Hastings algorithm for sampling the integrand on the accuracy of calculating the value of the integral. In addition, a hybrid method for sampling the integrand is proposed, in which part of the training sample is generated by applying the Metropolis-Hastings algorithm, and the other part includes points of a uniform grid. Numerical experiments show that when integrating in high-dimensional domains, sampling of integrands both by the Metropolis-Hastings algorithm and by a hybrid method is more efficient with respect to the points of a uniform grid.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Posters / 24**

# The Monte Carlo simulation of MiniSPD stand

**Authors:** Andrey Kutov[None]; Elena Kokoulina[1]; Nurlan Barlykov[None]; Vladimir Dudin[2]; Vladimir Nikitin[1]; Vsevolod Popov[1]

**Co-authors:** Artem Ivanov [3]; Evgeniy Martovitski [1]; Temur Enik [4]; Vitalii Burtsev [1]

[1] *JINR*

[2] *Dmitrievich*

[3] *Joint Institute for Nuclear Research*

[4] *Jinr*

**Corresponding Authors:** bar-nurlan@mail.ru, vladimirdudin95@mail.ru

The Spin Physics Detector, a universal facility for studying the nucleon spin structure and other spin-related phenomena with polarized proton and deuteron beams. It will be placed in one of the two interaction points of the NICA collider that is under construction at the Joint Institute for Nuclear Research (Dubna, Russia). The main objective of the proposed experiment is the comprehensive study of the unpolarized and polarized gluon content of the nucleon.

In accordance with the possible configuration of the SPD setup, our collaboration manufactured the MiniSPD stand. At present, this stand is used for testing SPD detector prototypes with cosmic muons, the Data Acquisition System (DAQ) and the Detector Control System (DCS).

Using GEANT4 software and ROOT framework, we have been carried out Monte Carlo simulation of three modules of two-sided silicon plates of MiniSPD stand for two cases: with and without taking into account operation of the scintillator triggers. We illustrate the solution of the alignment task which is the important part of any experiment. Our simulation silicon detectors are agreed well with experimental data on cosmic muons.

We are currently engaged in modeling and processing experimental data at the MiniSPD stand, so this device does not involve the use of machine learning methods. But when there is a transition to a full SPD installation, when processing data, machine learning methods will be needed and with our performance we would like to interest and attract specialists from this field.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Posters** / **62**

# Анализ данных в образовательной деятельности университета «Дубна»

**Author:** Кристина Жаткина[None]

**Co-authors:** Oksana Streltsova [1]; Mikhail Belov [2]

[1] *JINR*

[2] *Dubna State Univeristy*

**Corresponding Author:** zhatkina-96@mail.ru

В 2020 году университет «Дубна» столкнулся с вынужденным выходом на дистанционное обучение. Была сформирована цифровая среда университета, включающая в себя различные ресурсы для помощи в информировании студентов и организации образовательного процесса.

Цель данной работы провести анализ накопленных данных, проверить степень влияния дистанционного обучения на уровень успеваемости студентов и установить взаимосвязь между данными успеваемости студентов и их уровнем удовлетворенности образованием.

После первичной обработки (~ 70 тыс. строк) и анализа данных из 1С была установлена положительная динамика успеваемости студентов Института САУ университета «Дубна» в период дистанционного обучения. Однако, следует отметить, что данные не равномерно распределены по годам, связано это в первую очередь с цифровизацией образования, которая начала набирать обороты лишь несколько лет назад.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Posters** / **70**

# Application of machine learning methods to determine the state of the cardiovascular system based on the analysis of indicators of a quantum phase space of the rhythm of the cardiovascular system

**Authors:** V.P. Tsvetkov[1]; I.V. Tsvetkov[1]; E.K. Paramonova[2]; S.A. Mikheev[1]; A.I. Tsvetkov[3]

[1] *Tver State University*

[2] *Tver regional clinical hospital*

[3] *Tver state university*

**Corresponding Author:** czvetkov.1990@bk.ru

The most complete information about the state of the human cardiovascular system is provided by the analysis of the array of cardiointervals (RR-intervals) of 24-hour holter monitoring (HM).
The most important task of analyzing a large array of HM RR-intervals is the introduction of the

main parameters that most adequately reflect the properties of this array. One way to solve this problem is to construct an extended quantum phase space Seq of the instantaneous heart rate and search for the main parameters that describe its properties.
The Seq space provides a powerful tool for studying non-deterministic chaotic systems. In particular, it allows one to visualize HM data by presenting digital information in a form convenient for observation and analysis.
In particular, it allows one to visualize HM data by presenting digital information in a form convenient for observation and analysis. The forms of the obtained graphs differ significantly visually in their form.
It is proposed to use the profiles of graphs of the regularity I_r and irregularity I_nr^+,I_nr^- indexes respectively of the Seq space as markers of the state of the cardiovascular system, reflecting the properties of an array of cardiointervals, and to carry out their detailed analysis for a large number of patients using machine learning methods. To train the neural network, data arrays of cases of cardiac conditions with a known pathology or norm are used.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Posters / 63**

# Algorithmic block for behavioral tests in the BIOHLIT information system for radiobiological studies

**Author:** Yuri Butenko[1]

**Co-authors:** Alexey Stadnik [2]; Inna Kolesnikova ; Kristina Golikova ; Dina Utina [1]

[1] *JINR*

[2] *Dubna university*

**Corresponding Author:** gohas94@gmail.com

The BIOHLIT information system (IS) for analyzing behavioral and pathomorphological changes in the central nervous system when studying the effect of ionizing radiation on laboratory animals. Information system is being jointly developed by specialists from MLIT and LRB JINR.
The IS is necessary for storing data in a single information space, enhancing the detection of laboratory animals in the behavioral tests (Open Field, T-maze, etc.) arena, calculating individual behavioral patterns, as well as for reducing time and energy costs, minimizing the human factor when dealing with histological slides. This will allow processing experimental data in no time and defining qualitative and quantitative changes in the central nervous system after exposing to ionizing radiation. For these purposes, on the basis of modern technologies of computer vision and machine learning, an algorithm was developed; it enables to automate the analysis of the behavioral reactions of experimental animals through video files.
To solve this problem, it was necessary to develop several subgroups of algorithms: algorithms for the automated marking of the field of experimental setups, algorithms for tracking the animal's position in experimental setups of different types and algorithms for evaluating the animal's behavioral patterns that characterize its emotional status and orienting-exploratory reactions. As a result of the operation of this algorithm, the information obtained is stored in different forms: a visualized track of the laboratory animal's movement, a video file with tracking the laboratory animal's position, a heat map by sectors and a JSON file that stores all the information obtained from the video file for subsequent statistical analysis. The JSON file contains information on automated marking by the sectors and boundaries of the experimental setup, a track of the animal's movement by frame coordinates, the characteristics of the original video file (frame size, number of frames per second). These data are required for additional analysis (statistical, etc.).

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics / 61**

# Machine Learning in MLIT. History, Challenges, and Prospects

**Author:** Gennady Ososkov[1]

[1] *Joint Institute for Nuclear Research*

**Corresponding Author:** ososkov@jinr.ru

Machine learning (ML) methods began to be used in the MLIT laboratory from the very beginning of its organization in 1966, when one of the main tasks of the LVTA was the automation of film data processing used at that time in physics experiments. This included the problems of automating film measurements and calibration of the then-built scanning machines Spiral Reader and AELT (Automat on Elektron Tube) scanning tables, as well as the acquired HPD scanner. These tasks of event reconstruction, calibrating measuring devices and physical hypothesis testing were still relevant, although they were solved by classical methods of computational statistics, first on tube and later on transistor computers, not yet equipped with programming languages at the time.

With the advent of experiments with electronic data acquisition directly into the memory of more powerful computers equipped with programming languages such as Fortran, it became possible to apply robust fitting methods and wavelet analysis, the LVTA was one of the first among other physics centers in the 1990s to apply neural networks and cellular automata for tracking and other applied tasks in biology, medicine, earth physics and pattern recognition.

However, the real heyday of neural network methods application in physics and related fields came with new advances in computer technology, GPU cards, Supercomputers and the advent of the Big Data era. Big Data required radically new approaches, and new technologies made possible the application of deep neural networks with their amazing ability to solve most basic ML problems in classification, pattern recognition, prediction and hypothesis testing. The diversity of tasks coming to MLIT from JINR laboratories and collaborating institutes required the use of a wide range of different types of deep neural networks: multilayer perceptrons, autoencoders, convolutional, recurrent and graph neural networks, and recently, transformers with their attention mechanism used to increase weights of the most important features of objects under study. The depth of these neural networks, i.e. increase in the number of hidden layers allows to significantly strengthen the descriptive side of neuromodels, i.e. to expand the number of features of the phenomenon under study, though inevitably poses very complicated problems of their training and validation. These problems can be solved in many ways by using different program libraries written in interpretive Python language, using PyTorch framework. Nevertheless, to solve different event reconstruction problems in many High Energy Physics (HEP) experiments, a special library Ariadne, developed at MLIT with participation of performers from Dubna and SPbSU Universities, was required.

Among the problems solved in recent years in MLIT using deep learning methods, we can mention first of all the tasks of processing track data from the experiments BM@N, BES-III, SPD, modeling fine structures in mass distributions of nuclear reaction products, as well as such problems from related fields as atmospheric pollution monitoring using satellite imagery, detection and prevention of agricultural plant diseases.

Prospects for further development of machine learning methods are dictated primarily by new high luminosity and multiplicity HEP experiments, leading to exabyte streams of experimental information, requiring their online filtering and new approaches for ultrafast processing. Most applications from other fields are characterized by the lack of large databases required for training neural networks, which also requires the development of new approaches in the choice of neural network types, their structure and training methods.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **10**

## Welcome words

**Corresponding Author:** korenkov@jinr.ru

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **12**

## ML in particle astrophysics

**Corresponding Author:** kryukov@theory.sinp.msu.ru

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **54**

## Identification of similarities and differences in water bodies in terms of photosynthetic activity and response to toxicants using Machine Learning methods

**Authors:** Sergei Khruschev[1]; Roman Chervitsov[1]

**Co-authors:** Tatiana Plyusnina [1]; Taras Antal [2]; Galina Riznichenko [3]

[1] *Lomonosov Moscow State University*

[2] *Pskov State University*

[3] *Lomonosov Moscow State University, MSU*

**Corresponding Authors:** roman123qwe123@gmail.com, mce2000@mail.ru

During the experiment, 9 water bodies located in the Pskov region were studied: the pond of the Mirozhka River, the delta of the Velikaya River, the Kamenka River, lakes Kalatskoye, Teploe, Lesitskoye, Tiglitsy, Chudskoye (Peipsi), Pskovskoye. Water samples with phytoplankton were taken from each water body, and toxicants ($CdSO_4$ or $K_2Cr_2O_7$) were added at a concentration of 20 μM and 50 μM. These samples (together with control samples without toxicant addition) were incubated for several days. The chlorophyll fluorescence transients (OJIP-curves) were recorded using AquaPen–C 100 (Photon System Instruments, Czech Republic) fluorometer at actinic light intensity of 1000 μmol photons/$m^2$/s and wavelength of 650 nm every 3–5 hours of incubation (with breaks for the night). 465 fluorescence induction curves were obtained, for each of which 16 JIP test parameters ($V_J$, $V_I$, ABS/RC, $ET_0$/RC, $TR_0$/RC, $DI_0$/RC, $S_m$, $S_s$, $M_0$, N, $\Phi_{Po}$, $\Phi_{Eo}$, $\Phi_D$, $\Phi_{Pav}$, $\Psi_0$, $PI_{ABS}$). The resulting dataset was analyzed using the Python programming language. Cluster analysis together with data dimension reduction methods (PCA, *t*-SNE, UMAP) made it possible to put forward a hypothesis about the presence among the considered water bodies of 2 groups, the phytoplankton of which differ in the dynamics of resistance to the toxic effects of heavy metals. Machine learning methods have made it possible to identify the forms of induction curves typical for each type of water bodies and the dynamics of their change during toxic exposure. To reveal the biophysical mechanisms associated with such a different reaction, it is necessary to analyze the shape of the induction curves

using mathematical models based on the ideas about the structure of the photosynthetic electron transport chain.

This research was carried out as part of the Scientific Project of the State Order of the Government of Russian Federation to Lomonosov Moscow State University No. 121032500060-0 with partial support by the Russian Science Foundation (projects 20-64-46018 and 22-11-00009).

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics** / 32

# Machine learning approach to identify cores of EAS observed by the GRAPES-3 experiment

**Authors:** Medha Chakraborty[1]; S. Ahmad[2]; A. Chandra[2]; S.R. Dugad[3]; U.D. Goswami[4]; S.K. Gupta[3]; Hari Haran Balakrishnan[1]; Y. Hayashi[5]; P. Jagadeesan[3]; A. Jain[3]; P. Jain[6]; S. Kawakami[5]; H. Kojima[7]; S. Mahapatra[8]; P.K. Mohanty[3]; R. Moharana[9]; Y. Muraki[10]; P.K. Nayak[3]; T. Nonaka[11]; A. Oshima[7]; B.P. Pant[12]; S. Paul[3]; Diptiranjan Pattanaik[13]; G.S. Pradhan[14]; M. Rameez[3]; K. Ramesh[3]; L.V. Reddy[3]; R. Sahoo[14]; R. Scaria[14]; S. Shibata[7]; K. Tanaka[15]; F. Varsi[16]; M. Zuberi[3]

[1] *Tata Institute of Fundamental Research*

[2] *Aligarh Muslim University, Aligarh 202002, India*

[3] *TIFR, Mumbai*

[4] *Dibrugarh University, Dibrugarh 786004, India*

[5] *Graduate School of Science, Osaka City University, Osaka 558-8585, Japan*

[6] *IIT Kanpur*

[7] *College of Engineering, Chubu University, Kasugai, Aichi 487-8501, Japan*

[8] *Utkal University, Bhubaneshwar 751004, India*

[9] *Indian Institute of Technology Jodhpur, Jodhpur 342037, India*

[10] *Institute for Space-Earth Environmental Research, Nagoya University, Nagoya 464-8601, Japan*

[11] *Institute for Cosmic Ray Research, Tokyo University, Kashiwa, Chiba 277-8582, Japan*

[12] *IIT Jodhpur, India*

[13] *Tata Institute of Fundamental Research, Mumbai*

[14] *IIT Indore, India*

[15] *Graduate School of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan*

[16] *IIT Kanpur, India*

**Corresponding Author:** medha.chakraborty@tifr.res.in

The GRAPES-3 experiment located in Ooty consists of a dense array of 400 plastic scintillator detectors spread over an area of 25,000 $m^2$ and a large area (560 $m^2$) tracking muon telescope. Everyday, the array records about 3 million showers in the energy range of 1 TeV - 10 PeV induced by the interaction of primary cosmic rays in the atmosphere. These showers are reconstructed in order to find several shower parameters such as shower core, size, and age. High-energy showers landing far away from the array often trigger the array and are found to have their reconstructed cores within the array even though their true cores lie outside, due to reconstruction of partial information. These showers contaminate and lead to an inaccurate measurement of energy spectrum and composition. Such showers can be removed by applying quality cuts on various shower parameters, manually as well as with machine learning approach. The improvements achieved by the use of machine learning will be presented.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **20**

# Deep learning in the collider physics

**Author:** Lev Dudko[1]

[1] *SINP MSU*

**Corresponding Author:** dudko@sinp.msu.ru

Different aspects of deep learning applications in the collider physics will be discussed in the talk. The main topic of the talk is the methodology of data analysis optimizations with deep neural networks. Short overview of the methods to search for “new physics” with neural network technique will be presented.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **40**

# Application of convolutional neural networks for data analysis in TAIGA-HiSCORE experiment

**Authors:** Anna Vlaskina[1]; Alexander Kryukov[2]

[1] *Moscow State University*

[2] *SINP MSU*

**Corresponding Authors:** kryukov@theory.sinp.msu.ru, nina.vankalas@gmail.com

The TAIGA experimental complex is a hybrid observatory for high-energy gamma-ray astronomy in the range from 10 TeV to several EeV. The complex consists of such installations as TAIGA-IACT, TAIGA-HiSCORE and a number of others. The TAIGA-HiSCORE facility is a set of wide-angle synchronized stations that detect Cherenkov radiation scattered over a large area. With TAIGA-HiSCORE data provides an opportunity to reconstruct shower characteristics, such as shower energy, direction of arrival, and axis coordinates. The main idea of the work is to apply convolutional neural networks to analyze HiSCORE events, considering them as images. The distribution of registration times and amplitudes of events recorded by HiSCORE stations is used as input data. The paper presents the results of using convolutional neural networks to determine the characteristics of air showers. It is shown that even a simple model of convolutional neural network provides the accuracy of recovering EAS parameters comparable to the traditional method. Preliminary results of air shower parameters reconstruction obtained in a real experiment and their comparison with the results of traditional analysis are presented. The work was supported by the Russian Science Foundation, grant №22-21-00442

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics / 39**

## Energy reconstruction with machine learning techniques in JUNO: aggregated features approach

**Author:** Arsenii Gavrikov[1]

**Co-authors:** Yury Malyshkin [2]; Fedor Ratnikov [3]

[1] *HSE University, JINR*

[2] *JINR*

[3] *HSE University*

**Corresponding Author:** gavrikov@jinr.ru

The Jiangmen Underground Neutrino Observatory (JUNO) is a neutrino experiment under construction with a broad physics program. The main goals of JUNO are the determination of the neutrino mass ordering and the high precision measurement of neutrino oscillation properties. High quality reconstruction of reactor neutrino energy is crucial for the success of the experiment.

The JUNO detector is equipped with a huge number of photomultiplier tubes (PMTs) of two types: 17 612 20-inch PMTs and 25 600 3-inch PMTs. The detector is designed to provide an energy resolution of 3% at 1 MeV. Compared to traditional reconstruction methods, Machine Learning (ML) is significantly faster for the detector with so many PMTs.

In this work we studied ML approaches for energy reconstruction from the signal gathered by the PMT array and presented fast models using aggregated features: fully connected deep neural network and boosted decision trees. The dataset for training and testing is generated with full simulation using the official JUNO software.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics / 49**

## Decision trees as an alternative for particle identification with TPC and TOF detector system

**Author:** Vladimir Papoyan[1]

**Co-authors:** Alexander Ayriyan [2]; Hovik Grigorian [1]; Alexander Mudrokh [1]

[1] *JINR*

[2] *Laboratory of Information Technologies, JINR*

**Corresponding Authors:** papoyan8@gmail.com, ayriyan@jinr.ru

Machine Learning methods are wildly used for particle identification (PID) in experimental high energy physics nowadays. Particle identification plays an important role in high-energy physics analysis therefore determines the success of the performing an experiment. This determines importance of using machine learning to the PID problem. This report gives a preliminary status of application of decision tree to particle identification problem with TPC and TOF detector system.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **43**

## Using conditional variational autoencoders to generate images from atmospheric Cherenkov telescopes

**Authors:** Stanislav Polyakov[1]; Alexander Kryukov[1]; Andrey Demichev[1]; Yulia Dubenskaya[1]; Елизавета Гресь[None]; Anna Vlaskina[2]

[1] *SINP MSU*

[2] *Moscow State University*

**Corresponding Author:** s.p.polyakov@gmail.com

Monte Carlo method is commonly used to simulate Cherenkov telescope images of atmospheric events caused by high-energy particles. We investigate the possibility of augmentation the Monte Carlo-generated sets using other methods. One of these methods is variational autoencoders.
We trained conditional variational autoencoders (CVAE) using a set of Monte Carlo-generated images from one Cherenkov telescope of TAIGA experiment for atmospheric events caused by gamma quanta (gamma events). Images generated by the trained autoencoders are similar to the Monte Carlo images, in particular, an average score by a classifier trained to distinguish Monte Carlo generated images of gamma events is 0.982-0.986 for one of the autoencoders, compared to 0.99 for Monte Carlo images.
This work was funded by the Russian Science Foundation (grant No. 22-21-00442).

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **53**

## Using Conditional GAN to Control the Statistical Characteristics of the Generated Images from Imaging Atmospheric Cherenkov Telescopes

**Authors:** Yulia Dubenskaya[1]; Alexander Kryukov[1]; Andrey Demichev[1]; Stanislav Polyakov[1]; Елизавета Гресь[None]; Anna Vlaskina[2]

[1] *SINP MSU*

[2] *Moscow State University*

**Corresponding Author:** jdubenskaya@gmail.com

Currently, generative adversarial networks (GANs) are a promising tool for image generation in the astronomy domain. Of particular interest are conditional GANs (CGANs), which allow you to divide images into several classes according to the value of some property of the image, and then specify the required class when generating images. In the case of images from Imaging Atmospheric Cherenkov Telescopes (IACT), an important property is the total brightness of all image pixels (image size), which is directly connected to the energy of primary particles. We used a CGAN technique to

generate images of the Cherenkov telescope of the TAIGA-IACT experiment. As a training set, we used a sample of 2D images obtained using TAIGA Monte Carlo simulation software. We applied an artificial division of the images of the training set into 10 classes, sorting them by size and defining the boundaries of the classes so that the same number of images fall into each class. We then used these classes while training our CGAN. The paper shows that for each class, the size distribution of the generated images is close to normal with an average value located approximately in the middle of the corresponding class. We also show that for the generated images, the size distribution summed over all classes is close to the original distribution in the training set. The results obtained will be useful for data augmentation and more accurate generation of realistic synthetic images similar to the ones taken by IACT. This work was funded by the Russian Science Foundation (grant No. 22-21-00442).

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **36**

## Energy reconstruction in analysis of Cherenkov telescopes images in TAIGA experiment using Deep Learning methods

**Author:** Elizaveta Gres[1]

**Co-author:** Alexander Kryukov [2]

[1] *Irkutsk State University*

[2] *SINP MSU*

**Corresponding Author:** greseo@mail.ru

Imaging Atmospheric Cherenkov Telescopes (IACT) of TAIGA astrophysical complex allow to observe high energy gamma radiation helping to study many astrophysical objects and processes. TAIGA-ACT enables us to select gamma quanta from the total cosmic radiation flux and recover their primary parameters, such as energy and direction of arrival. The traditional method of processing the resulting images is an image parameterization - so-called the Hillas parameters method. At the present time Machine Learning methods, in particular Deep Learning methods have become actively used for IACT image processing.
This report presents the analysis of simulated Monte Carlo images by several Deep Learning methods for a single telescope (mono-mode) and multiple IACT telescopes (stereo-mode). The estimation of the quality of energy reconstruction was carried out and their energy spectra were analyzed using several types of neural networks. Using the developed methods the obtained results were also compared with the results obtained by traditional methods based on the Hillas parameters. The work was financially supported by the Russian Science Foundation, grant No. 22-21-00442.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **58**

## Deep neural network applications for particle tracking at the BM@N and SPD experiments

**Authors:** Pavel Goncharov[1]; Egor Shchavelev[2]; Gennady Ososkov[1]; Leonid Lubchenkov[None]; Daniil Rusov[3]

[1] *Joint Institute for Nuclear Research*

[2] *Saint Petersburg State University*

[3] *Dubna State University*

**Corresponding Author:** pgoncharov13@gmail.com

Particle tracking is an essential part of any high-energy physics experiment. Well-known tracking algorithms based on the Kalman filter are not scaling well with the amounts of data being produced in modern experiments. In our work we present a particle tracking approach based on deep neural networks for the BM@N experiment and future SPD experiment. We have already applied similar approaches for BM@N Run 6 and BES-III Monte-Carlo simulation data, which are relatively simpler and produce less data during the experiment. This work is the next step in our ongoing study of tracking with the help of machine learning —revised algorithms (combination of Recurrent Neural Network (RNN) and Graph Neural Network (GNN) for the BM@N Run 7 Monte-Carlo simulation data, and GNN for the preliminary SPD Monte-Carlo simulation data) are presented. Encouraging results in terms of track efficiency and processing speed for both experiments are demonstrated.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **31**

# Study of the VH(bb) production by MVA methods

**Author:** Faig Ahmadov[1]

[1] *JINR & IP ANAS*

**Corresponding Author:** fahmadov@jinr.ru

Taking into account that, at a Higgs boson mass of 125 GeV, the probability of its decay into bb is greater than the sum of the probabilities of all other decay channels, this channel makes a great contribution to the study of the Higgs boson. A more suitable channel for the production of the Higgs boson for studying it in bb decay is associative production with a vector boson. It was in this channel that the decay of the Higgs boson into a pair of b-quarks was observed for the first time. Therefore, the VH(bb) process is a very important channel for studying the properties of the Higgs boson.
In LHC experiments, multivariate VH(bb) analysis began to be used after 2013, before that, cut-based analysis was used. As a multivariate analysis, among the multivariate techniques, the Boosted Decision Tree (BDT) in ATLAS and the Deep Neural Network (DNN) in CMS were used. In this work, these two methods were compared to find out which one can achieve the best performance. The 2L channel (ZH(bb), where Z decays into 2 charged leptons) from the three lepton channels (0L, 1L, 2L) was chosen, which includes the VH(bb) analysis. The list of input variables for BDT or DNN is similar to those used in the analysis in the ATLAT experiment. Up to 0.4 million signals and the same number of background events were used for training. The settings used in the ATLAS analysis, which has the best performance, were chosen to tune the BDT hyperparameters. Various number of events (2K, 5K, 10K, 0.1M, 0.2M and 0.4M) are trained and different settings for NN are obtained, providing performance that exceeds that of BDT. It turns out that for any number of training events, it is possible to find corresponding NN settings with better performance than BDT. The only problem with NN training is that it is computationally intensive compared to BDT.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 1. ML in Particle Astrophysics and High Energy Physics** / **64**

## Model interpretability methods for high energy physics analysis

**Author:** Andrei Zaborenko[1]

**Co-authors:** Lev Dudko [2]; Petr Volkov ; Georgy Vorotnikov ; Emil Abasov

[1] *Moscow State University*

[2] *SINP MSU*

**Corresponding Author:** azaboren@cern.ch

Most modern machine learning models are known as black-box models. By default, these predictors don't provide an explanation as to why a certain event or example has been assigned a particular class or value. Model explainability methods aim to interpret the decision-making process of a black-box model and present it in a way that is easy for researchers to understand. These methods can provide *local* (figuring why a specific input has been assigned a specific output) an *global* (uncovering general dependencies between input features and the output of the model) explanations. In this talk we will cover several popular model-agnostic explainability methods and compare them in explaining the output of a neural network in the scope of high-energy physics analysis. We will also use a modern high accuracy glass-box machine learning model (Explainable Boosting Machine) and show how its predictions can be used to better understand the data.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 2. Modern Machine Learning Methods** / **13**

## Sponsor report - Softline

**Session 2. Modern Machine Learning Methods** / **29**

## Quantum end-to-end-IT of self-organized intelligent controller design based on quantum deep machine learning: Quantum neural networks and quantum genetic algorithms applications in quantum intelligent control of classical systems –quantum supremacy

**Author:** Sergey Ulyanov[1]

[1] *JINR*

**Corresponding Author:** ulyanovsv46_46@mail.ru

V.V. Korenkov, A.G. Reshetnikov, S.V. Ulyanov, P.V. Zrelov
MLIT, JINR

The physical interpretation of self-organization control process on quantum level is discussed based on the quantum information-thermodynamic models of the exchange and extraction of quantum (hidden) value information from/between classical particle's trajectories in particle swarm [1,2]. Main physics and information thermodynamics aspects of quantum intelligent control of classical control objects discussed and described from control Benchmark models viewpoint design on the basis of new laws of quantum Lagrange / Hamilton deep machine learning.
1. Physics of quantum hidden information phenomena. New types of quantum correlations (as behavior control coordinator with quantum computation by communication) and information transport (value information) between particle swarm trajectories (communication through a quantum link) are introduced.
2. Quantum logic of intelligent classical system control. The structure of developed quantum fuzzy inference (QFI) model includes necessary self-organization properties and realizes a self-organization process as a new quantum search algorithm (QSA). In particular case, in intelligent control system (ICS) structure, QFI system is a QSA block, which performs post-processing of the results of fuzzy inference of each independent fuzzy controller (FC) and produces the generalized control signal output. In this case the on-line output of QFI is an optimal robust control signal, which combines best features of each independent FC outputs (self-organization principle). For design of FC - KB original structures of quantum neural networks and quantum genetic algorithm developed and applied.
3. Quantum software engineering of quantum intelligent control physics law. Quantum soft computing optimizer toolkit of KB –design processes is described. Benchmarks of robust KB design from imperfect FC - KB as the new quantum synergetic information effects of extracted quantum information demonstrated. Moreover, the new force control law from quantum thermodynamic described: with extracted hidden quantum information from classical control signal states (on micro-level) possible to design in on-line new control force that can produce on macro-level more value work amount than the work losses on the extraction of this amount of hidden quantum information.
It is a new control law of physics-cybernetics open hybrid systems including port-Hamiltonian controlled dynamic objects [3].
4. Applications. Effective application of new quantum intelligent controller in mega-science project NICA, intelligent cognitive robotics and quantum drones for applications in project "Industry 5.0" demonstrated [4]. Perspective applications of quantum software engineering discussed.
Conclusions
Therefore, the operation area of such ICS can be expanded greatly as well as its robustness. Robustness of control signal is the background for support the reliability of control accuracy in uncertainty environments. The effectiveness of the developed QFI model is illustrated for important case - the application to design of robust intelligent control system of unstable essentially nonlinear control object in unpredicted control situations (autonomous mobile robots, robotic manipulators, swarm robotics with information exchange etc.).

References
1. Ulyanov S.V. System and method for control using quantum soft computing. - US patent No 6,578,018B1, 2003.
2. Ulyanov S.V. Self-organizing quantum robust control methods and systems for situations with uncertainty and risk. - Patent US 8788450 B2, 2014.
3. Ulyanov S. V. Quantum Algorithm of Imperfect KB Self-organization Pt I: Smart Control - Information-Thermodynamic Bounds // Artificial Intelligence Advances. - 2021. - Vol. 3. - No 2. - Pp. 13-36; https://doi.org/10.30564/aia.v3i2.3171.
4. Korenkov V.V., Reshetnikov A.G., Ulyanov S.V., Zrelov P.V. Intelligent cognitive robotics. Vol. 2, Pt. 2: Quantum supremacy of quantum intelligent control –quantum neural networks and quantum genetic algorithms in quantum deep learning. –M.: Kurs, 2022.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 2. Modern Machine Learning Methods / 26**

## Stochastic vs. BFGS Training in Neural Discrimination of RF-

# Modulation

**Author:** Maria Dima[1]

[1] *JINR - MLIT*

**Corresponding Author:** mmdima@jinr.ru

Neuromorphic classification of RF-Modulation type is an on-going topic in SIGINT applications. Neural network training approaches are varied, each being suited to a certain application. For exemplification I show the results for BFGS (Broyden-Fletcher-Goldfarb-Shanno) optimisation in discriminating AM vs FM modulation and of stochastic optimisation for the challenging case of AM-LSB vs. AM-USB discrimination. Although slower than BFGS, the stochastic training of a neural network avoids better local minima, obtaining a stable neurocore.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 2. Modern Machine Learning Methods** / **14**

# TAIGA Astrophysical Complex –status, results, plans

**Agreement to place**:

**Session 2. Modern Machine Learning Methods** / **35**

# Relation Extraction from Texts Containing Pharmacologically Significant Information on base of Multilingual Language Models

**Authors:** Anton Selivanov[1]; Roman Rybka[1]; Alexander Sboev[2]

[1] *NRC "Kurchatov Institute"*

[2] *NRC "Kurchatov Institute"; NRNU "MEPhI"*

**Corresponding Author:** aaselivanov.10.03@gmail.com

In this paper we estimate accuracy of solving the task of relation extraction from texts containing pharmacologically significant information on the set of corpora in two languages:
1) the expanded version of RDRS corpus, that contains texts of internet reviews on medications in Russian;
2) the DDI2013 dataset containing MEDLINE abstracts and documents from DrugBank database in English;
3) the PhaeDRA corpus containing MEDLINE abstracts in English.

Relation extraction accuracy for Russian and English was estimated with comparison of two multilingual Language models: XLM-RoBERTa-large and XLM-RoBERTa-sag-large. Additionaly we used the State-of-the-Art specialized models aimed at English language: bioBERT, bioALBERT, bioLinkBERT. Earlier research proved XLM-RoBERTa-sag-large to be the most efficient language model for the previous version of the RDRS dataset. We used the same approach to relation extraction included two

steps: named entity recognition and relation extraction on predicted entities. Each step was estimated separately.

As a result, it is shown, that multilingual XLM-RoBERTa-sag model achieves relation extraction macro-averaged f1-score equal to 85.42% on the ground-truth named entities, 53.83% on the predicted named entities on new version of RDRS corpus. Additionally, XLM-RoBERTa-sag was estimated on the datasets for relation extraction in English (DDI2013, PhaeDRA) and achieves accuracy comparable with the top specialized models.

Consequently, XLM-RoBERTa-sag model sets the state-of-the-art for considered type of texts in Russian, and achieves accuracy comparable with the SotA results in Engilsh.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 2. Modern Machine Learning Methods** / **57**

## A spiking neural network with fixed synaptic weights based on logistic maps for a classification task

**Authors:** Dmitriy Kunitsyn[1]; Alexander Sboev[2]; Alexey Serenko[3]

**Co-author:** Roman Rybka [4]

[1] *National Research Nuclear University MEPhI*

[2] *NRC "Kurchatov Institute"; NRNU "MEPhI"*

[3] *National Research Centre "Kurchatov Institute"*

[4] *NRC "Kurchatov Institute"*

**Corresponding Author:** selibrin@mail.ru

Spiking neural networks which model action potentials in biological neurons are increasingly popular for machine learning applications thanks to ongoing progress in the hardware implementation of spiking networks in low-energy-consuming neuromorphic hardware. However, obtaining a spiking neural network model that solver a classification task as accurately as a formal neural network remains a challenge.
We study a spiking neural network model with non-trainable synaptic weights preset on base of logistic maps, similarly to what was proposed recently in the literature for formal neural networks. We show that one layer of spiking neurons with such weights can transform input vectors preserving the information about the classes of the input vectors, so that this information can be extracted from the neuron's output spiking rates by a subsequent classifier, such as Gradient Boosting.
The accuracy obtained on the Fisher's Iris classification task is 95%, with the deviation range of 5% over the five cross-validation folds.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 2. Modern Machine Learning Methods** / **51**

## Decomposition of Spectral Contour into Gaussian Bands using Gender Genetic Algorithm

**Authors:** Gavriil Kupriyanov[1]; Igor Isaev[2]; Ivan Plastinin[3]; Tatiana Dolenko[4]; Sergei Dolenko[5]

[1] *Faculty of Physics, M.V.Lomonosov Moscow State University, Moscow, Russia*

[2] *D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University, Moscow, Russia; Kotelnikov Institute of Radio Engineering and Electronics, Russian Academy of Sciences, Moscow, Russia*

[3] *D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University, Moscow, Russia*

[4] *Faculty of Physics, M.V.Lomonosov Moscow State University, Moscow, Russia; D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University, Moscow, Russia*

[5] *D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University*

**Corresponding Author:** dolenko@srd.sinp.msu.ru

One of the methods for analysis of complex spectral contours (especially for spectra of liquid objects) is their decomposition into a limited number of spectral bands with physically reasonable shapes (Gaussian, Lorentzian, Voigt etc.). Consequent analysis of the dependencies of the parameters of these bands on some external conditions in which the spectra are obtained may reveal some regularities bearing information about the physical processes taking place in the object.
The problem with the required decomposition is that such decomposition is an inverse problem that is often ill-conditioned or even incorrect, especially in presence of noise in spectra. Therefore, this problem is often solved by advanced optimization methods less subject to be stuck in local minima, such as genetic algorithms (GA).
In the conventional version of GA, all individuals are similar regarding the probabilities and implementation of the main genetic operators (crossover and mutation) and the procedure of selection. In this study, we test a new version of GA –gender GA (GGA), where the individuals of the two genders differ by the probability of mutation (higher for the male gender) and by the procedures of selection for crossover. In this study, we compare the efficiency of gradient descent, conventional GA and GGA in solving the problems of decomposition of the Raman valence band of liquid water into Gaussian bands.
This study has been funded by the SINP MSU state budget topic 6.1 (01201255512).

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 2. Modern Machine Learning Methods** / **34**

# Deep learning approach to high dimensional problems of quantum mechanics

**Authors:** Vladimir Roudnev[1]; Stepanova Margarita[2]

[1] *St-Petersburg State University*

[2] *SPbSU*

**Corresponding Author:** v.rudnev@spbu.ru

Traditional linear approximation of quantum mechanical wave functions are not practically applicable for systems with more than 3 degrees of freedom due to the “the curse of dimensionality”. Indeed, the number of parameters required to describe a wave function in high-dimensional space grows exponentially with the number of degrees of freedom. Inevitably, strong model assumptions should be used when studying such systems numerically. There are, however, estimates of the complexity of a function reproduced by a deep neural network (DNN) that demonstrate the same exponential growth with respect to the number of the network layers. The number of parameters for DNN grows only linearly with the number of layers. This gives us a hope that application of DNN as an approximant for a wave function in high-dimensional space might moderate the computational requirements for

reproducing such systems and make 4- or higher-dimensional systems feasible for direct numerical modeling.

We present a study of DNN approximation properties for a multi-dimensional quantum harmonic oscillator. We demonstrate that the computational resources required to reproduce the wave function depend on the dimensionality of the problem and the quantum numbers of the state. Increasing the number of hidden layers in a fully-connected direct propagation DNN we can reproduce some excited states of a multidimensional system with computational resources comparable to low-dimensional cases. Using the DNN as an approximant for a wave function paves a way to developing a new class of computational schemes for solving the Schroedinger equation for high-dimensional systems.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 2. Modern Machine Learning Methods** / **38**

## Neural Networks Application to Classification of Credit Institutions

**Authors:** Elena Akishina[1]; Victor Ivanov[2]; Anastasiya Prikazchikova[None]

[1] *JINR*

[2] *JINR, LIT*

**Corresponding Author:** aska4.92@mail.ru

The paper presents the application of the methodology of machine learning (artificial neural networks) and the method of principal component analysis to the problem of classifying data on the base of credit institutions.
The feed-forward neural network (multilayer perceptron with hidden layers) was applied to specially prepared input data. As a result, the set of credit institutions was successfully splat to the groups: reliable and unreliable (the institutions whose licenses were revoked).
Principal component analysis (PCA) was applied to the input data aiming to reduce data dimension. Wherein, the result of classifying the reduced data with the neural network remained practically at the same level.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences** / **41**

## Underwater biotope mapping: automatic processing of underwater video data

**Authors:** Oleg Iakushkin[1]; Ekaterina Pavlova[None]; Anastasiya Lavrova[None]; Olga Sedova[1]

[1] *Saint-Petersburg State University*

**Corresponding Author:** oleg.jakushkin@gmail.com

The task of analysing the inhabitants of the underwater world is applicable to a wide range of applied problems: construction, fishing, and mining. Currently, this task is applied on an industrial scale by a rigorous review done by human experts in the field of underwater life. In this work, we present a tool that we have created that allows us to significantly reduce the time spent by a person on video analysis. Our technology offsets the painstaking video review task to AI, creating a shortcut that allows experts to only verify the accuracy of the results. To achieve this we have developed an observation pipeline by dividing the video into frames; assessing their degree of noise and blurriness; performing corrections via resolution increase; analysing the number of animals on each frame; building a report on the content of the video, and displaying the obtained data of the biotope on the map. This greatly reduces the time spent analysing underwater video data.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences / 15**

## IT ecosystem based on machine learning methods and data analysis technologies for radiobiological research

**Corresponding Author:** strel@jinr.ru

**Agreement to place**:

**Session 3. Machine learning in Biology and Other Natural Sciences / 28**

## Approximation of high-resolution surface wind speed in the North Atlantic using discriminative and generative neural models based on RAS-NAAD 40-year hindcast

**Authors:** Vadim Rezvov[1]; Mikhail Krinitskiy[2]

[1] *MIPT*

[2] *Shirshov Institute of Oceanology, Russian Academy of Sciences*

**Corresponding Author:** rezvov.vyu@phystech.edu

Surface wind is one of the most important fields in climate change research. Accurate prediction of high-resolution surface wind has a wide variety of applications, such as renewable energy and extreme weather forecasts. Downscaling is a methodology for high-resolution approximation of physical variables from low-resolution modeling outputs. Statistical downscaling methods allow to avoid computationally expensive high-resolution hydrodynamic simulations. Deep learning methods, including artificial neural networks (ANNs), are one of the typical machine-learning approaches approximating complex nonlinear functional relationships. In our study, we explored the capabilities of statistical 5x downscaling of surface wind over the ocean in the North Atlantic region. We applied several downscaling methods, including cubic interpolation as a reference solution, various convolutional ANNs, and generative adversarial networks (GANs). We compared downscaling results in terms of several quality measures including the ones representing the reconstruction of extreme winds. We evaluated the performance and the quality of different methods and reference solution to identify advantages and lacks of machine-learning downscaling. We consider GANs as the most promising ANN architectures for surface wind downscaling based on both performance and quality.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences / 23**

## Artificial neural networks for multi-label cloud types classification from all-sky optical imagery over the ocean.

**Authors:** Nikita Veremev[None]; Mikhail Krinitskiy[1]

[1] *Shirshov Institute of Oceanology, Russian Academy of Sciences*

**Corresponding Author:** nikita.veremiev@yandex.ru

Cloudiness plays an important role in the hydrological cycle of the atmosphere. Cloud types and other cloud spatial and temporal characteristics privide the ability to make short-term in situ weather forecasts. With the help of clouds, one may also track the content of various impurities in the air. Most importantly, clouds are the major obstacle on the pathway of incoming solar radiation, thus, classifying cloud types may be useful for solar energy plants. In our study, we use artificial neural networks for classifying cloud types in all-sky optical imagery of the visible hemisphere of the sky. The problem is soft classification due to presence of several cloud types at once in most of images. We constructed a convolutional neural network based on SE-ResNeXt101 architecture. The DASIO (Dataset of All-Sky Imagery over the Ocean) dataset for our study is collected in Indian, Atlantic and Arctic oceans from 2014 till 2021 and contains over 1.5 million images. For the 80‘000 of them, the visual observations of cloud types are provided by experienced meteorologists. Being trained with this dataset, our model achieved high performance: sample-averaged F1-score is 0.7 which is state of the art compared to contemporary studies.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences / 37**

## Предсказание матрицы контактов для коротких пептидов с использованием свёрточной нейронной сети

**Author:** Artem Maminov[1]

[1] *FRC CSC RAS*

**Corresponding Author:** artem_maminov@mail.ru

В данной работе предлагается рассмотреть метод предсказания матрицы контактов для пептидов. В данной статье были выбраны пептиды с длинной до 45 аминокислотных остатков для упрощения расчётов. Для предсказания использовались свёрточные нейронные сети (CNN) из-за схожести пространства признаков белков и изображений, к котором обычно успешно применяются свёрточные нейронные сети. Признаки были созданы с использованием инструмента SCRATCH (генерации вторичной структуры, растворимости и профиля белка PSSM). CNN реализована на языке программирования Python с применением библиотеки Keras. Для работы со структурами белков использовался модуль BioPython, позволяющий извлекать матрицу расстояний между атомами каркаса белка и на основе этой матрицы рассчитывать

матрицу контактов нативной структуры. В результате были сформированы обучающие, валидационные и тестовые выборки. Была построена многослойная свёрточная нейронная сеть для решения задачи мультивыходной бинарной классификации. Для оценки качества предсказания были построены матрицы неточностей для порога в 8 и 12˚, рассчитаны метрики F1-score (0.78), recall (0.73) и precision (0.86). Также был использован инструмент FT-COMAR для восстановления третичной структуры из предсказанной матрицы контактов и сравнения с нативной структурой по метрики RMSD. Среднее значение метрики RMSD по выборке белков равно 6.76 и 5.84˚ для порогов 8 и 12˚ соответственно.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences / 48**

## Visual clustering of ocean sediment grains using a combination of unsupervised machine learning methods.

**Authors:** Viktor Golikov[None]; Mikhail Krinitskiy[1]; Dmitrii Borisov[1]

[1] *Shirshov Institute of Oceanology, Russian Academy of Sciences*

**Corresponding Author:** golikov.va@phystech.edu

Quantitative, granulometric and classification-based distribution of oceanic sediment grains are important indicators in paleo-reconstruction of the characteristics of marine waters. Currently, the classification of grains is performed visually by an expert on a limited subset of a sediment sample using a binocular microscope. It is a highly time-consuming process in which geological expertise is required of the observer. In this study, we propose a method to automate and accelerate this kind of work using a combination of machine learning algorithms. We photograph sediment samples prepared for examination using a digital optical microscope. We then apply a clustering algorithm including classical and neural machine learning techniques. An experienced marine geologist then identifies the resulting clusters. Our method significantly reduces the time consumption of the expert. We demonstrate that the proposed method is able to divide sediment grains into homogeneous groups suitable for further accurate classification. This will allow further evaluation of important characteristics (paleoindicators), such as the ratio of biogenic carbonate grains and therigenic grains, as well as the ratio of whole shells and shell fragments. The clustering results obtained using our algorithm may be used to train a more accurate classification algorithm.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences / 19**

## Google Earth Engine and machine learning for Earth monitoring

**Author:** Alexander Uzhinskiy[1]

[1] *Dr.*

**Corresponding Author:** zalexandr@list.ru

Hyperspectral images are a unique source for obtaining many kinds of information about the Earth's surface. Modern platforms support users to perform complex analyses with a collection of images without the use of any specialized software. Google Earth Engine (GEE) is a planetary-scale platform for Earth science data & analysis. Atmospheric, radiometric, and geometric corrections have been made on number of image collections at GEE. While working with row data, it Is possible to use build-in GEE function to filter data and create composites to get cloud score threshold and the percentile. It is also possible to use custom algorithms for atmospheric corrections. There are over 100 satellite image collections and modeled datasets. Some collections have a spatial resolution of up to 15 meters. GEE has the JavaScript online editor to create and verify code and Python API for advanced applications. All that made GEE very convenient tool for different Earth monitoring projects. Over the last decades there has been considerable progress in developing a machine learning methodology for a variety of Earth Science applications involving trace gases, retrievals, aerosol products, land surface products, vegetation indices, fire and flood tracking, ocean applications, and many others. In this report, we will review basic GEE functions and practice, some examples of successful applications, and our experience in environmental monitoring.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences / 59**

## Taking into Account Mutual Correlations during Selection of Significant Input Features in Neural Network Solution of Inverse Problem s of Spectroscopy

**Authors:** Nickolay Shchurov[1]; Igor Isaev[None]; Sergei Burikov[2]; Tatiana Dolenko[2]; Kirill Laptinskiy[3]; Sergei Dolenko[4]

[1] *Faculty of Physics Lomonosov Moscow State University*

[2] *Faculty of Physics, M.V.Lomonosov Moscow State University, Moscow, Russia; D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University, Moscow, Russia*

[3] *D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University, Moscow, Russia*

[4] *D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University*

**Corresponding Author:** shchurov_no2@mail.ru

In neural network solutions to many physical problems, there is a need to reduce the dimension of the input data in order to achieve a more accurate and stable solution while reducing computational complexity.
When solving an inverse problem in spectroscopy, multicollinearity is often observed between the input features, making it necessary to use a selection method that takes into account the correlation between the input features.
The method used in this study is based on the iterative selection of features with the highest correlation with respect to the target variable and on the elimination of features with high mutual correlation.
The paper compares the quality of a neural network solution to the problem of determining the concentrations of heavy metal ions in water by Raman spectra on the complete set of input features and on its subsets compiled using the selection method under consideration, as well as using traditional methods of selecting significant input features.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences / 60**

## In situ wind speed nowcasting using data-driven approach.

**Authors:** Viktor Golikov[None]; Mikhail Krinitskiy[1]

[1] *Shirshov Institute of Oceanology, Russian Academy of Sciences*

**Corresponding Author:** golikov.va@phystech.edu

The weather forecast has a significant impact on a variety of human industries. In particular, knowledge of the short-term wind speed conditions is essential for fishery, energy management, surfing and others. One of the most effective neural network models for time series forecasting is LSTM (Long short-term memory), however, the accuracy of its forecast decreases significantly with increasing forecasting range. At the same time, numerical models based on physical laws make it possible to achieve accurate results even over long-term intervals. In an attempt to combine these 2 types of models, an LSTM-based neural network was developed using wind speed data at the forecast point and atmospheric parameters in a domain of size covering mesoscale wind events. Based on these weather characteristics and numerical model data, the neural network makes a short-term forecast of the wind speed module at the point. In the current state, our model outperforms the persistent forecast. In the future, the results can be improved by adjusting hyperparameters and introducing the ability to use the result of numerical models as a basis for the final forecast.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences / 47**

## Hazy Images Dataset With Localized Light Sources For Experimental Evaluation Of Dehazing Methods

**Authors:** Andrei Filin[1]; Andrei Kopylov[1]; Oleg Seredin[1]; Inessa Gracheva [1]

[1] *Tula State University*

**Corresponding Author:** andrewifilin@gmail.com

Recently, the haze removal methods have taken increasing attention of researchers. An objective comparison of haze removal methods struggles because of the lack of real data. Capturing pairs of images of the same scene with presence/absence of haze in real environment is a very complicated task. Therefore, the most of modern haze datasets contain artificial images, generated by some model of atmospheric scattering and known scene depth. Among the few real datasets, there are almost no datasets consisting of images obtained in low light conditions with artificial light sources, which allows evaluating the effectiveness of nighttime haze removal techniques. In this paper, we present such dataset, consisting of images of 2 scenes at 4 lighting levels and 4 levels of haze intensity. The scenes has vary “complexity” - the first scene consists of objects with a simpler texture and shape (smooth, rectangular and round objects); the second scene is more complex - it consists of objects with small details, protruding parts and localized light sources. All the images were taken indoor in controlled environment. An experimental evaluation of state-of-the art haze removal methods was carried out on the resulting dataset.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences / 22**

## Data-driven approximation of downward solar radiation flux based on all-sky optical imagery using machine learning models trained on DASIO dataset.

**Author:** Vasilisa Koshkina[1]

**Co-authors:** Mikhail Krinitskiy [2]; Nikita Anikin [3]; Mikhail Borisov [4]

[1] *Moscow Institute of Physics and Technology (National Research University)*

[2] *Shirshov Institute of Oceanology, Russian Academy of Sciences*

[3] *Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow, Russia*

[4] *Moscow Institute of Physics and Technology, Moscow, Russia*

**Corresponding Author:** koshkina.vs@phystech.edu

Cloud cover is the main physical factor limiting the downward shortwave (SW) solar radiation flux. In modern models of climate and weather forecasts, physical models describing radiative transfer through clouds may be used. However, this is a computationally expensive option. Instead, one may use parameterizations which are simplified schemes for approximating environmental variables. The purpose of our study is to assess the capability of machine learning models in the scenario of statistical approximation of radiation flux based on all-sky optical imagery. We applied various machine learning (ML) models within the assumption that an all-sky photo fully encapsulates information about the downward shortwave radiation. We examine several types of ML models: some classic ML models along with a convolutional neural network (CNN). These models were trained using the dataset of all-sky imagery accompanied by SW radiation flux measurements. The Dataset of All-Sky Imagery over the Ocean (DASIO) is collected in Indian, Atlantic and Arctic oceans during several expeditions from 2014 till 2021. When training our CNN, we applied heavy source data augmentation in order to force the CNN to become invariant to brightness variations and, thus, approximating a relationship between the visual structure of cloudiness and SW flux. We demonstrate that the CNN supersedes existing parameterizations known from literature in terms of RMSE of flux. Our results allow us to assume that one may acquire downward shortwave radiation flux directly from all-sky imagery. We also demonstrate that CCNs are capable of estimating downward SW radiation flux based on clouds' visible structure.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences / 52**

## Neural network recovery of missing data of one geophysical method from known data of another one in solving inverse problems of exploration geophysics.

**Authors:** Igor Isaev[1]; Ivan Obornev[2]; Eugeny Obornev[3]; Eugeny Rodionov[3]; Mikhail Shimelevich[3]; Sergey Dolenko[4]

[1] *D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University, Moscow, Russia; Kotelnikov Institute of Radio Engineering and Electronics, Russian Academy of Sciences, Moscow, Russia*

[2] *D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University, Moscow, Russia; S.Ordjonikidze Russian State Geological Prospecting University, Moscow, Russia*

[3] *S.Ordjonikidze Russian State Geological Prospecting University, Moscow, Russia*

[4] *D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University, Moscow, Russia*

**Corresponding Author:** isaev_igor@mail.ru

This study is devoted to the inverse problems of exploration geophysics, which consist in reconstructing the spatial distribution of the properties of the medium in the Earth's thickness from the geophysical fields measured on its surface. We consider the methods of gravimetry, magnetometry, and magnetotelluric sounding, as well as their integration, i.e. simultaneous use of data from several geophysical methods to solve the inverse problem. In their previous studies, the authors have shown that the integration of geophysical methods allows improving the quality of the solution of the inverse problem in comparison with the individual use of each of them.
One of the obstacles to using the integration of geophysical methods can be the situation when for some measurement points there is no data from one of the geophysical methods used. At the same time, the data spaces of different integrated geophysical methods are interconnected, and the values of the observed quantities (fields) for one of the methods can be possibly recovered from the known values of the observed quantities of another geophysical method by constructing a preliminary adaptive mapping of one of the spaces to another.
In this study, we investigate the neural network recovery of missing data of one geophysical method from the known data of another one and compare the quality of the solution of the inverse problem on full and on recovered data.
This study has been performed at the expense of the grant of the Russian Science Foundation no. 19-11-00333, https://rscf.ru/en/project/19-11-00333/.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences** / **56**

## Accuracy of COVID-19 evolution models for different forecast horizons

**Authors:** Saveliy Zavertyaev[1]; Ivan Moloshnikov[1]; Aleksandr Sboev[None]; Alexander Naumov[1]; Roman Rybka[2]

[1] *National Research Centre "Kurchatov Institute"*

[2] *NRC "Kurchatov Institute"*

**Corresponding Author:** zavertyaev.sv@phystech.edu

Currently, there are more than two years of statistics accumulated on COVID-19 for a large number of regions, which allows the use of algorithms that require large training sets, such as neural networks, to predict the dynamics of the disease.
The article provides a comparative analysis of various COVID-19 models based on forecasting for the period from 07/20/2020 to 05/05/2022 using statistics on the regions of the Russian Federation and US states The forecast target is the sum of confirmed cases over the forecast horizon.
Models based on the Exponential Smoothing (ES) and LSTM methods were considered.
The training set included data from all regions. The MAPE metric was used for model comparison, the evaluation of the effectiveness of the LSTM in the learning process was carried out using cross-validation and the MSE metric.
Comparisons were made with models from literature sources, as well as with the baseline model "tomorrow like today"(sum for forecast horizon equals today cases multiplied by forecast horizon).
It is shown that on small horizons (up to 28 days) the "tomorrow like today"and ES algorithms show

better accuracy than LSTM. In turn, on longer horizons (28 days or more), the preference should be given to the more complex LSTM model.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 3. Machine learning in Biology and Other Natural Sciences** / **42**

## Application of a neural network approach to the task of the arena marking for the behavioral test «Open Field»

**Author:** Anastasia Anikina[1]

**Co-authors:** Inna Kolesnikova ; Dmitry Podgainy [2]; Dmitry Savvateev [3]; Юрий Северюхин [4]; Alexey Stadnik [5]; Oksana Streltsova [2]

[1] *Igorevna*

[2] *JINR*

[3] *Dubna State University*

[4] *JINR LRB*

[5] *Dubna university*

**Corresponding Author:** asya_ani@mail.ru

In the framework of the joint project of LIT and LRB JINR, aimed to the creation of an information system for the tasks of radiation biology, a module is being developed to study the behavioral patterns of small laboratory animals exposed to radiation. The module for behavioral analysis automates the analysis of video data obtained by testing of the laboratory animals in the different test systems. The «Open Field» installation is one of the systems. The considered installation has a form of round arena with the chequered-marked sectors and holes. The observation procedure on the laboratory animals takes 3-6 minutes. The "Open Field"test-system allows to register the general activity of animals. To this aim, we fix the quantity of passed sectors together with the number of intersections of the marked center. Also, we check how many burrows, standings upright with/without supports, standings still and motions on one place the animals did.

Therefore, one of our tasks is to develop an algorithm for the installation field marking. The report presents the algorithms for the field marking of the «Open Field» test system based on computer vision methods together with the method of key points within the neural network approach.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 4. Machine Learning in Education** / **17**

## Sponsor report - Softline

**Session 4. Machine Learning in Education** / **67**

# Sponsor report - RSC group

**Session 4. Machine Learning in Education** / **55**

# Analytical platform for intellectual labour market analysis

**Authors:** Anna Ilina[1]; Vladimir Korenkov[2]; Petr Zrelov[3]; Sergey Belov[1]; Igor Pelevanyuk[1]; Ivan Kadochnikov[2]; Vitaly Tarabrin[4]; Javad Javadzade[2]; Daria Priakhina[5]; Irina Filozova[2]; Iuliia Gavrilenko[6]; Roman Semenov[2]

[1] *Joint Institute for Nuclear Research*

[2] *JINR*

[3] *LIT JINR*

[4] *PRUE*

[5] *ЛИИТ*

[6] *Research Assistant, Plekhanov Russian University of Economics, Moscow, Russia*

**Corresponding Author:** annailina@jinr.ru

The paper presents an analytical platform that implements automated monitoring and analysis of the labor market in the Russian Federation. The platform is based on Big Data solutions and technologies. End-to-end processing corresponds to the general scheme of step-by-step solving of the problem - from data collection, their transformation, analysis, and modeling to services for visualization of results and decision-making. The analytical core of the system is a module for intelligent analysis of texts of job advertisements in the labor market. Vacancy data is collected from the most extensive databases in Russia (HeadHunter, TrudVsem, and SuperJob). The matching of job descriptions and the official list of professions of the Ministry of Labor and Social Protection of the Russian Federation using semantic analysis based on neural models trained on large bodies of texts. The results are presented using several services developed within the MS Power BI business intelligence system. Data collection has been ongoing since 2015. About 500 thousand active vacancies per day are processed.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**Session 4. Machine Learning in Education** / **69**

# ML/DL/HPC Ecosystem of the HybriLIT Heterogeneous Platform (MLIT JINR): New Opportunities for Applied Research

**Authors:** Yuri Butenko[1]; Marko Ćosić[2]; Andrey Nechaevskiy[1]; Dmitry Podgainy[1]; Ilhom Rahmonov[3]; Oksana Streltsova[1]; Maxim Zuev[1]

[1] *JINR*

[2] *Vinca institute of nuclear sciencies*

[3] *BLTP, Joint Institute for Nuclear Research*

**Corresponding Author:** zuevmax@jinr.ru

The report presents the possibilities for using the ML/DL/HPC ecosystem deployed on the HybriLIT Heterogeneous Platform (MLIT JINR) on top of JupyterHub, which provides opportunities for solving tasks not only in the field of machine learning and deep learning, but also for the convenient organization of calculations and scientific visualization. The ecosystem allows one to develop and implement program modules in Python, as well as to carry out methodical computations. The relevance of deploying such an environment is primarily associated with the great demand for software modules that are provided to a group of researchers or the scientific community, when all stages of the study can be reproduced; the code has been modified and used by the scientific community. Using the example of solving a specific problem to study the dynamics of magnetization in a Phi-0 Josephson Junction (Superconductor-Ferromagnet-Superconductor structure), a methodology for developing software modules is presented; it enables not only to carry out calculations, but also to visualize the results of the study and accompany them with the necessary formulas and explanations. In addition, the possibility of parallel implementation of the algorithm for performing computations for various values of parameters of the model based on the Joblib Python library is shown, and the results of computational experiments demonstrating the efficiency of parallel data processing are presented.
Another example of the capabilities of the ML/DL/HPC ecosystem is the development of modules with the integration of the MATLAB code in the Jupyter Notebook, which allows one to effectively perform applied computations for image analysis.
This work was supported by Russian Science Foundation grant No 22-71-10022.

**Agreement to place**:

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

**16**

# Closing