PROCEEDINGS OF SCIENCE

Using Modern Machine Learning Methods on KASCADE Data for Outreach and Education

Victoria Tokareva,[a,*] **Dmitriy Kostunin,**[b] **Ivan Plokhikh**[c,d] **and Vladimir Sotnikov**[e]

[a] *Karlsruhe Institute of Technology, IAP, 76021 Karlsruhe, Germany*
[b] *DESY, 15738 Zeuthen, Germany*
[c] *Novosibirsk State University, 630090 Novosibirsk, Russia*
[d] *Institute of Thermophysics SB RAS, 630090 Novosibirsk, Russia*
[e] *JetBrains Research, 194100 St. Petersburg, Russia*

E-mail: victoria.tokareva@kit.edu

Modern astroparticle physics makes wide use of machine learning methods in such problems as noise suppression, image recognition, event classification. When using these methods, in addition to obtaining new scientific knowledge, it is important also to take advantage of their educational potential. In this work we present a demo version of the machine-learning based application we have created, which helps students and a broader audience to get more familiar with the cosmic ray physics, and shows how machine learning methods can be used to analyze data. The work discusses the prospects for expanding the application's functionality and methodological approaches to the development of interactive outreach materials in this area.

The 5th International Workshop on Deep Learning in Computational Physics
28-29 June, 2021
Moscow, Russia

*Speaker

PoS(DLCP2021)007

1. Introduction

Nowadays outreach is of big importance for any field of science. Alongside with informing a broader public about the most recent progress in modern research, it also contributes to increase of investigators' opportunities through attraction of additional funding and manpower—both in form of citizen science collaborators and new research personnel.

As the interest in advanced IT technologies (such as neural networks, cloud computing and virtual reality) grows, they find more and more applications in research activities. On the one hand usage of modern methods increases public interest in research, on the other hand one faces new challenges related to public scientific communication: educating people both about the new technologies and how they are used to achieve the research goals, with taking into account the interdisciplinary nature of this knowledge.

In this work, we study the issue of promoting both machine learning and astroparticle physics. In order to do so, we consider the experience of our colleagues in outreach projects and analyze their experience in Section 2, then present our datasets and methods in Section 3, and introduce the Streamlit [1]-based outreach application developed by us in Section 4. The conclusion is given in Section 6.

2. Information technologies in outreach projects

The importance of outreach for particle physics and astrophysics is so great in recent years that, besides the creation of a huge number of outreach and educational projects worldwide, this has led to the emergence of associations of educators, such as IPPOG [2] and Teilchennetzwerk [3], as well as special sections dedicated to such projects at major [astro]particle-physics conferences such as EPS HEP Spring Meeting [4], DPG Spring Meeting [5], and the International Cosmic Ray Conference [6].

Intensive scientific communication is becoming a popular trend, supported by such large collaborations as IceCube, Auger [7, 8], KASCADE [9], KARTIN [10], KM3NeT [11], etc., as well as by large scientific institutions such as DESY [12](Germany), and the National Institute for Nuclear Physics [13] (INFN, Italy).

A large share of projects in the field of outreach and education is occupied by the segment of activities aimed primarily at the development of school education (such as QuarkNet [14], HiSPARC [15], Showers of Knowledge [16], EEE [17]) and focused on high school students and intensive communication with students and educational institutions.

The involvement of wider audience in science is associated both with citizen science projects (such as CREDO [18], REINFORCE [11]), and with easier-to-understand formats of excursions, exhibitions and video presentations.

At the same time, quite often the formats used by projects have the classic form of lectures, seminars, publications in periodicals and social networks, and while each of the mentioned projects has social networks and a web-page, the use of modern IT technologies in the outreach area is still quite limited.

One of the breakthroughs in this regard was the release of Jupyter Notebooks [19] in 2015, which allowed scientists and communicators to share analysis code in a convenient way. Today,

this platform is used for outreach by such collaborations as Auger, KASCADE, Tunka-Rex [20] and many others, and the publication of open materials in this format has become de facto standard for open data and scientific communication.

Interactive web applications are another interesting example of creating modern interactive outreach materials. Their distinctive feature is their wide availability that allows to use them on any mobile device online, without being limited by place or geographic location. Another distinguishing feature of this group of materials is the focus on the most modern technologies such as virtual reality and neural networks.

For example, augmented and virtual reality applications [21], developed by IceCube collaboration, allow users to learn more either about the physics of neutrinos and the mechanisms of their detection, as well as about life of astroparticle researchers at the South Pole.

A gamified web application based on the use of convolutional neural networks (CNN) [22] was created for TAIGA [23] experiment in the framework of GRADLC initiative. In this work the CNN, taught on TAIGA IACT's simulated data is used to determine a particle's type by its imprint on the IACT detector.

Interesting examples of interactive high-tech applications are citizen science applications such as online tools of the Gravitational Wave Open Science Center [24] as well as Cosmic@Web [12], CREDO detector [18].

The main advantages of applications are their high availability and interactivity. Besides, they can demonstrate modern technologies "in action" without requiring programming knowledge. Thus, applications provide an opportunity to increase engagement of current audience and to attract new one.

An application may be used either individually, or in conjunction with supplementary classes in educational institutions, or at science festivals for introductory tutorials or as interactive exhibits.

For the stated reasons, we decided to develop our own ML-based application based on open data from the KASCADE experiment.

3. Materials and methods

3.1 KASCADE open data

KArlsruhe Shower Core and Array DEtector (KASCADE) [25] is an detector aimed to study the cosmic ray primary composition and hadronic interactions. It was represented by an extensive air shower array, which included 252 scintillator detectors stations on a rectangular grid measuring simultaneously the electronic, muonic and hadronic components of the showers and located at 110 m a.s.l., 49° N, 8° E at the 200×200 m^2 area. The detectors worked in the energy range 10^{14}–2×10^{16} eV. Later with KASCADE-Grande extension the energy range was extended to 10^{14}–10^{18} eV. KASCADE (including all extensions) was in the operation from 1996 to 2013.

The data of the KASCADE experiment is published open-access on the KCDC portal [26], created in 2013 and later expanded within the activities of the GRADLC initiative [27]. Data access is provided by the websites of both collaborations and by the API.

For this work, we used CORSIKA [28] simulations, generated individually for H, He, C, Si, Fe primaries employing three modern hadronic interaction models: QGSJet-II.04 [29], EPOS-LHC [30] and Sibyll 2.3c [31].

Table 1: The amounts of primary particles in the application datasets.

Dataset	primary type					total
	p	He	C&O	Si	Fe	
dataset1	900	250	500	700	550	2900
dataset2	1000	500	300	180	200	2180

Two mixed data samples were prepared, numbered 1 and 2 in the application, respectively. The amounts of primary particles comprising the datasets can be found in table 1.

Both datasets contain the following reconstructed shower components: decimal logarithm of primary energy $\lg E$, shower core coordinates $(X,\ Y)$, zenith angle Ze, azimuth angle Az, decimal logarithm of electron $\lg N_e$ and muon number $\lg N_\mu$ at observation level, and shower age Age. A table describing the structure of datasets can be found in Fig. 1 b).

3.2 Machine learning for solving astroparticle problems

The machine-learning models used in the application are described in detail in the work [32] devoted to mass reconstruction of primaries and determining spectra of individual mass-groups composition.

They are based on the random forest [33] algorithm, which is an ensemble machine-learning method, using sets of decision trees on various sub-samples of training data. It is a very well-know flexible and robust supervised learning algorithm, broadly used for both classification and regression tasks. The general idea of the method is that a combination of learning models improves the overall result. In accordance with the basic principle of ensembling, each tree is built on its own training sample and there is an element of randomness in the choice of splits to ensure the quality and variety of the underlying algorithms. Our study used an implementation of the random forest algorithm from the scikit-learn [34] library.

The classifier was trained to return one of the five mass groups based on available hadron-interaction models using the following quality cuts: $X^2 + Y^2 < 91$ m, $\lg N_\mu \geq 3.6$, $\lg N_e \geq 4.8$, $Ze < 18°$, $0.2 < Age < 2.1$.

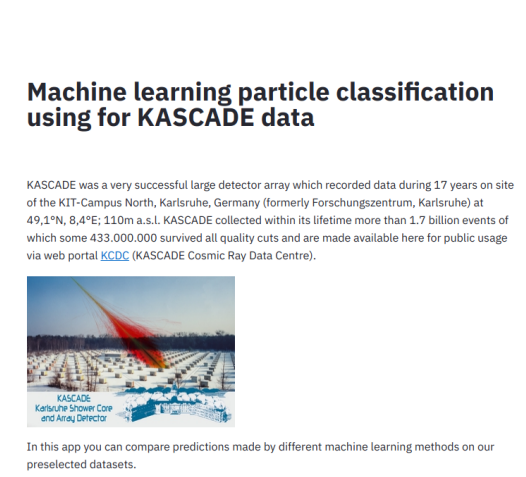
3.3 Approaches used for deploying neural networks models

There is a large number of solutions used to create custom web apps for machine learning and data science [35]. The most popular ones are the Dash [36] and Streamlit [1] libraries, and such solutions as Voila, Shiny, Panel can be used as well [37]. It is also a fairly common approach to use broader web frameworks such as Django, Tornado, or Flask [38].

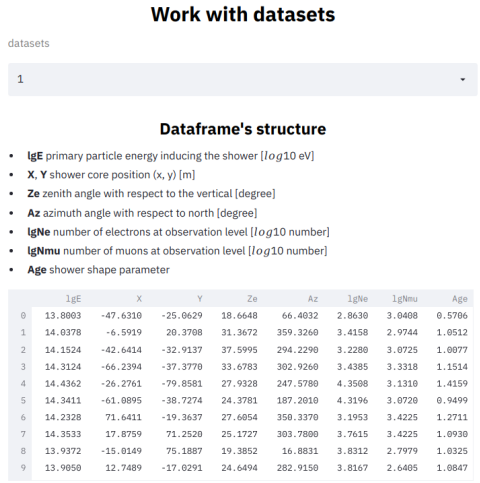
The mentioned solutions can be grouped according to such criteria as:

- **Supported programming language(s)**. Shiny only supports the R language, while all the other approaches work with the Python language, and some also with the Julia.

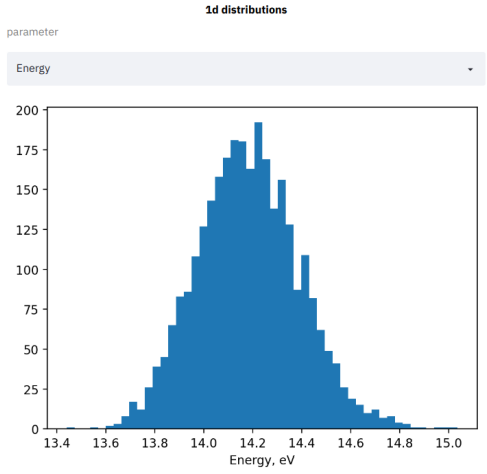
- **Simplicity**, which directly affects development speed. One of the leaders in this parameter is Streamlit. The Voila library also has a relatively simple API.



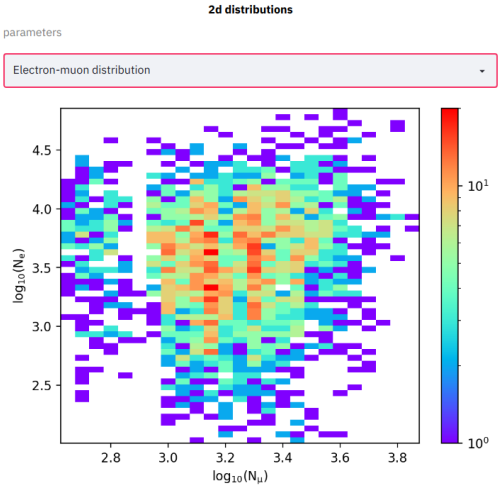
a)



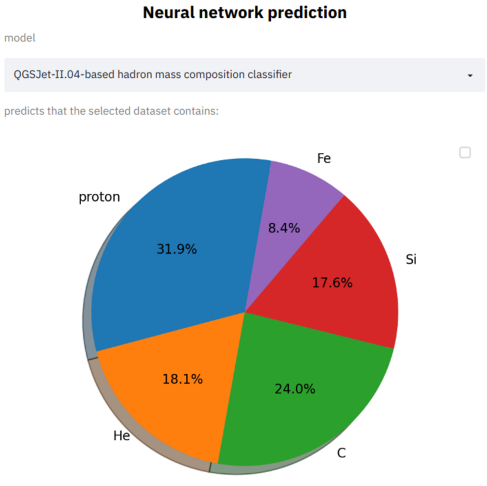
b)



c)



d)



e)

Figure 1: a) Interactive interface to explore individual shower parameters distributions; b) Interactive interface to explore some of 2d shower parameters distributions; c) Interactive interface to explore individual shower parameters distributions; d) Interactive interface to explore some of 2d shower parameters distributions; e) Interactive interface to explore some of 2d shower parameters distributions.

- **Purpose**. While solutions such as Streamlit and Voila are designed specifically for creating interactive applications in data science, and therefore have rather limited functionality, the use of general-purpose web frameworks allows one to create more flexible and adaptive solutions—which, nevertheless, affects development speed and support complexity.
- **Popularity**: according to GitHub ratings, Streamlit and Dash are currently the most popular solutions.
- **Open-source**: while while Streamlit or Flask are open source, Dash has proprietary functions.

Taking into account the above factors, for application development Streamlit library was choosen together with Docker and Kubernetes for the deployment.

4. Results

4.1 User Interface

The interface of the application we have developed consists of five parts, shown in Fig. 1. The **introductory part** provides a summary of the KASCADE experiment and a link to the KCDC site, where one can learn more about the experiment and it's data preservation.

The next part of the application is a screen for **working with datasets**. The top lines of the selected dataset are displayed in pandas dataframe format. The values of the data parameters are given. Streamlit allows us to sort the dataset strings by the values of certain parameters. It is also possible to display more records in the dataset or the entire dataset with a scroll bar.

Preliminary examination of the data can be done by constructing 1d or 2d histograms of parameter distributions. In this case, a 1d histogram can be built for any parameter at the user's choice. The following types of 2d histograms are available:

- shower footprint (X core to Y core distribution)
- N_e to N_μ distribution,

allowing one to make preliminary conclusions about the distributions of values in the presented datasets. Besides, graphs in Streamlit are interactive and allow actions like zooming in or out and changing the active area of the screen.

Comparison of the results of machine learning models, and the resulting comparison of hadronic interaction models is shown in the last section of the interface. The user can choose between the following particle classification models:

- QGSJet-II.04-based hadron mass composition classifier;
- Epos-LHC-based hadron mass composition classifier;
- Sibyll 2.3c-based hadron mass composition classifier.

The classification results for the selected dataset are shown as a pie chart of the number of primaries determined by the model. It can be seen that for classifiers trained on various models of hadronic interaction, some differences in predictions are observed.

4.2 Backend

The internal structure of the application corresponds to the diagram in Fig. 2.

The user interacts with the GUI, choosing a dataset and certain actions performed on the data. Further, preprocessing of events is possible, though it is not used in the current version of the application, since the data we use are simulated and thus do not require cleaning or handling of missing values, and decision trees are very robust machine learning models that are not sensitive to scaling and normalization of parameters.

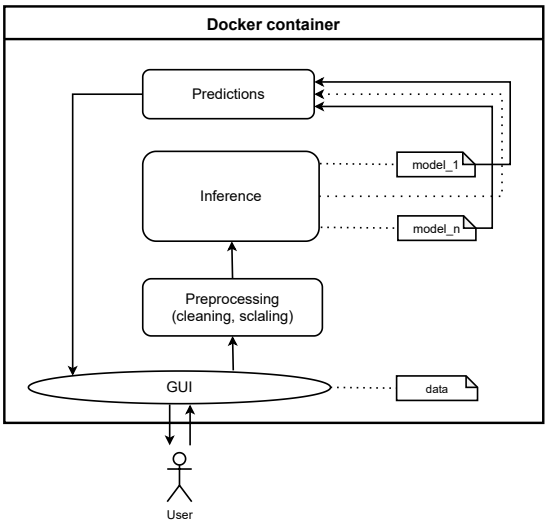


Figure 2: Application schema

However, when extending an application to work with other machine learning methods or to work with custom data, the preprocessing module becomes important.

To obtain predictions of particle types, we use inference working with pre-trained machine learning models. The application is containerized using Docker technology to improve the security and stability of its work.

4.3 Deployment

Development workflow is shown in Fig. 3. Application and deployment git repositories are being stored separately for security reasons. When new commits arrive to the server, they trigger TeamCity's pipelines: if the application repository was updated, the build pipeline will be triggered first, otherwise the deploy pipeline is launched.

In the case of repository updates, the build pipeline performs a checkout of the application repository, builds the Docker image and pushes it to the private image registry hosted by JetBrains Space. Otherwise, the deploy pipeline gets activated and renders the new version of Kubernetes configuration (e.g. with an updated version tag of the Docker image) from its template files and applies changes to the cluster.

Then Kubernetes updates its application deployment according to the new configuration (e.g. downloads the new Docker image).

5. Conclusion

In this work we discussed sharing knowledge on machine learning for astroparticle physics with a broad audience. In particular, we looked into sharing our results achieved in neural network-based analysis of KASCADE particle mass composition [32] within an interactive dedicated software for data exploration and visualization as well as for comparing predictions of neural networks, fit to different hadron mass compositions models.

In order to achieve this aim, the modern approaches to outreach in particle astrophysics were studied. Then, we prepared datasets, based on KASCADE simulations. Using Streamlit web

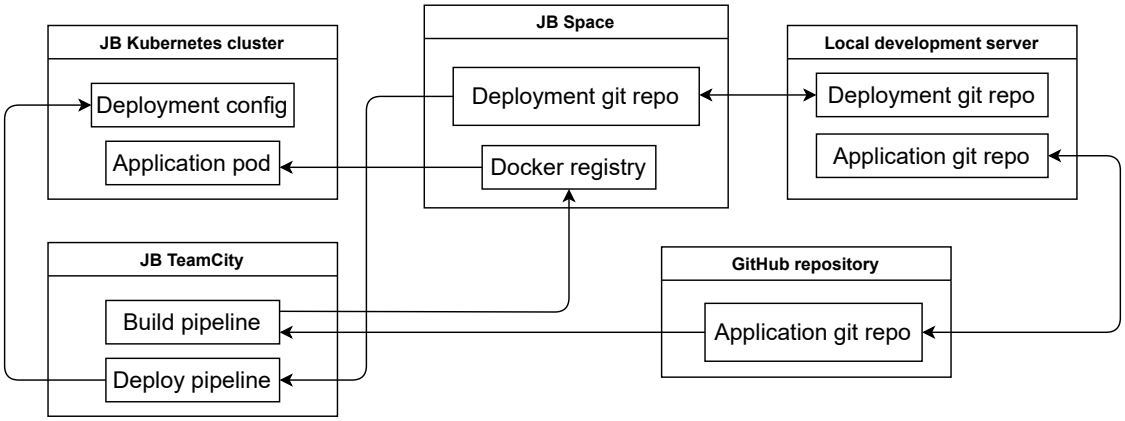


Figure 3: Deploy schema

framework, as well as Docker and Kubernetes, we have developed and deployed the first data-driven application based on open data from the KASCADE experiment and machine learning models. The result application is available online at https://kascade-streamlit.labs.jb.gg.

Further work in this area may include: expanding the range of available visualizations, enriching the data format, expanding the number of available machine learning models.

Acknowledgments

This work was supported by the Russian Science Foundation (Grant No. 18-41-06003) and Helmholtz Society (Grant No. HRSF-0027).

The authors aknowledge help of the colleagues from KCDC Team (esp. J. and D Wochele, A. Haungs) and from Astroparticle Physics Lab at JetBrains Research for fruitful discussions and interesting ideas.

References

[1] "Get started — Streamlit 0.86.0 documentation." https://docs.streamlit.io/en/stable/, 2021.

[2] B. Bruant Gulejova, A. Godinho, A. Ruiz Jimeno, A. Sharmazanashvili, A. Gorišek, A. Bens et al., *IPPOG Global Cosmic Rays Portal: Making Cosmic Rays Studies available to schools worldwide*, *PoS* **ICRC2021** (2021) 1362.

[3] M. Begalli and U. Bilow, *The German program Netzwerk Teilchenwelt*, in *Upgrading Physics Education to Meet the Needs of Society*, pp. 76–78, Springer, 2019.

[4] "The EPS-HEP2021 conference - Scientific Programme." https://indico.desy.de/event/28202/program, 2021.

[5] "DPG Frühjahrstagung Dortmund 2021, Outreach Methoden." https://indico.cern.ch/event/1017852/, 2021.

[6] "The 37th International Cosmic Ray Conference - Scientific Programme." https://icrc2021.desy.de/program/#e122019, 2021.

[7] V. Scherini, P. Abreu, M. Aglietta, J.M. Albury, I. Allekotte, A. Almela et al., *The 2021 Open-Data release by the Pierre Auger Collaboration*, *PoS* **ICRC2021** (2021) 1386.

[8] K.S. Caballero Mora, P. Abreu, M. Aglietta, J.M. Albury, I. Allekotte, A. Almela et al., *Outreach activities at the Pierre Auger Observatory*, *PoS* **ICRC2021** (2021) 1374.

[9] K. Link, V. Tokareva, A. Haungs, D. Kang, P. Koundal, F. Polgart et al., *Online Masterclass built on the KASCADE Cosmic ray Data Centre*, *PoS* **ICRC2021** (2021) 1378.

[10] M. Klein, C. Humm, L. Köllenberger, P. Niemann, Y. Scheuermann, P. Schrögel et al., *Virtual tours to the KATRIN experiment*, *PoS* **ICRC2021** (2021) 1376.

[11] R. Le Breton, V. Bertin, P. Coyle, G. de Wasseige, H. Glotin, C. Guidi et al., *The REINFORCE Project: Inviting Citizen Scientists to analyse KM3NeT data*, *PoS* **ICRC2021** (2021) 1392.

[12] "Cosmic@Web - Tools zur Online-Analyse." https://physik-begreifen-zeuthen.desy.de/angebote/kosmische_teilchen/cosmicweb/index_ger.html, 2021.

[13] C. Aramo, R. Antolini, V. Bocci, M. Buscemi, L. Caccianiga, A. Candela et al., *The online laboratories for OCRA - Outreach Cosmic Ray Activities INFN project*, *PoS* **ICRC2021** (2021) 1379.

[14] M. Bardeen, E. Gilbert, T. Jordan, P. Nepywoda, E. Quigg, M. Wilde et al., *The QuarkNet/grid collaborative learning e-Lab*, *Future Generation Computer Systems* **22** (2006) 700.

[15] K. van Dam, B. van Eijk, D.B.R.A. Fokkema, J.W. van Holten, A.P.L.S. de Laat, N.G. Schultheiss et al., *The HiSPARC experiment*, *Nucl. Instrum. Meth. Phys. Research A* **959** (2020) 163577.

[16] K. Gasnikova, *Distributed setup RUSALKA for detection of extensive atmospheric showers*, *Scientific Bulletin of Uzhgorod University. Series: Physics* **34** (2013) 217.

[17] "Extreme Energy Events (EEE)." https://eee.centrofermi.it/, 2021.

[18] P. Homola, D. Beznosko, G. Bhatta, L. Bibrzycki, M. Borczyńska, L. Bratek et al., *Cosmic-Ray Extremely Distributed Observatory*, *Symmetry* **12** (2020) 1835.

[19] T. Kluyver, B. Ragan-Kelley, F. Pérez, B.E. Granger, M. Bussonnier, J. Frederic et al., *Jupyter Notebooks-a publishing format for reproducible computational workflows*, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing*, pp. 87–90, 2016, DOI.

[20] V. Lenok, O. Kopylova, D. Wochele, F. Polgart, S. Golovachev, V. Sotnikov et al., *Tunka-Rex Virtual Observatory*, *PoS* **ICRC2021** (2021) 421.

[21] "Interactive IceCube Experiences." https://icecube.wisc.edu/outreach/interactive-experiences/, 2021.

[22] V. Tokareva, Y. Kazarina, A. Haungs, D. Kostunin, A. Kryukov, E. Postnikov et al., *Multi-messenger Astroparticle Physics for the Public via the astroparticle.online Project*, *PoS* **ICRC2021** (2021) 1373.

[23] N. Budnev, I. Astapov, P. Bezyazeekov, E. Bonvech, V. Boreyko, A. Borodin et al., *TAIGA—an advanced hybrid detector complex for astroparticle physics and high energy gamma-ray astronomy in the Tunka valley*, *Journal of Instrumentation* **15** (2020) C09031.

[24] A. Trovato, *GWOSC: Gravitational wave open science center*, in *The New Era of Multi-Messenger Astrophysics*, p. 082, 2019, DOI.

[25] T. Antoni, W.D. Apel, F. Badea, K. Bekk, A. Bercuci, H. Blümer et al., *The Cosmic ray experiment KASCADE*, *Nucl. Instrum. Meth. Phys. Res. A* **513** (2003) 490.

[26] A. Haungs, D. Kang, S. Schoo, D. Wochele, J. Wochele, W.D. Apel et al., *The KASCADE Cosmic-ray Data Centre KCDC: Granting Open Access to Astroparticle Physics Research Data*, *Eur. Phys. J. C* **78** (2018) 741.

[27] V. Tokareva, I. Bychkov, A. Demichev, J. Dubenskaya, O. Fedorov, A. Haungs et al., *German-Russian Astroparticle Data Life Cycle Initiative to foster Big Data Infrastructure for Multi-Messenger Astronomy*, *PoS* **ICRC2021** (2021) 938.

[28] D. Heck, J. Knapp, J. Capdevielle, G. Schatz, T. Thouw et al., *CORSIKA: A Monte Carlo code to simulate extensive air showers*, Tech. Rep. FZKA 6019, Forschungszentrum Karlsruhe (1998).

[29] S. Ostapchenko, *Monte Carlo treatment of hadronic interactions in enhanced Pomeron scheme: QGSJET-II model*, *Phys. Rev. D* **83** (2011) 014018.

[30] T. Pierog, I. Karpenko, J.M. Katzy, E. Yatsenko and K. Werner, *EPOS LHC: Test of collective hadronization with data measured at the CERN Large Hadron Collider*, *Phys. Rev. C* **92** (2015) 034906.

[31] F. Riehn, R. Engel, A. Fedynitch, T.K. Gaisser and T. Stanev, *Hadronic interaction model Sibyll 2.3 d and extensive air showers*, *Phys. Rev. D* **102** (2020) 063002.

[32] D. Kostunin, I. Plokhikh, M. Ahlers, V. Tokareva, V. Lenok, P. Bezyazeekov et al., *New insights from old cosmic rays: A novel analysis of archival KASCADE data*, in *37th International Cosmic Ray Conference (ICRC 2021)*, 2021, DOI.

[33] L. Breiman, *Random forests*, *Machine learning* **45** (2001) 5.

[34] G. Louppe, *Accelerating random forests in scikit-learn*, in *EuroScipy 2014*, 2014, http://hdl.handle.net/2268/171887.

[35] P. Singh, *Machine Learning Deployment as a Web Service*, in *Deploy Machine Learning Models to Production*, pp. 67–90, Springer (2021), DOI.

[36] "Dash Python User Guide." https://dash.plotly.com/, 2021.

[37] A. Sutchenkov and A. Tikhonov, *Active investigation and publishing of calculation web based applications for studying process*, in *Journal of Physics: Conference Series*, vol. 1691, p. 012096, IOP Publishing, 2020, DOI.

[38] N. Idris, C.F.M. Foozy and P. Shamala, *A generic review of web technology: Django and flask*, *International Journal of Advanced Science Computing and Engineering* **2** (2020) 34.