Форма «Т». Титульный лист заявки в Российский научный фонд
Конкурс 2024 года «Проведение фундаментальных научных исследований и поисковых научных исследований отдельными научными группами»

<table>
<tr><td rowspan="4">Название проекта
Методы машинного обучения для совместного анализа мультимодальных экспериментальных данных и извлечения редких событий в гамма-астрономии</td><td>Номер проекта</td><td>24-11-00136</td></tr>
<tr><td colspan="2">Код типа проекта: ОНГ(2024)</td></tr>
<tr><td colspan="2">Отрасль знания: 01</td></tr>
<tr><td colspan="2">Основной код классификатора: 01-202
Дополнительные коды классификатора:
01-210 01-213</td></tr>
<tr><td colspan="2">Код ГРНТИ 28.23.37</td></tr>
<tr><td>Фамилия, имя, отчество (при наличии) руководителя проекта:
Крюков Александр Павлович</td><td colspan="2">Контактные телефон и e-mail руководителя проекта:
+79163630991, kryukov@theory.sinp.msu.ru</td></tr>
<tr><td colspan="3">Полное и сокращенное наименование организации, через которую должно осуществляться финансирование проекта:
Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В.Ломоносова»
Московский государственный университет имени М.В.Ломоносова, Московский университет или МГУ</td></tr>
<tr><td>Объем финансирования проекта в 2024 г.:
7000 тыс. руб.</td><td>Год начала проекта:
2024</td><td>Год окончания проекта:
2026</td></tr>
<tr><td>Фамилии, имена, отчества (при наличии) основных исполнителей (полностью)</td><td colspan="2">Дубенская Юлия Юрьевна
Постников Евгений Борисович
(руководитель проекта в данной графе не указывается)</td></tr>
<tr><td colspan="3">Гарантирую, что при подготовке заявки не были нарушены авторские и иные права третьих лиц и/или имеется согласие правообладателей на представление в Фонд материалов и их использование Фондом для проведения экспертизы и для обнародования (в виде аннотаций заявок).</td></tr>
<tr><td colspan="2">Подпись руководителя проекта

_____ /А.П. Крюков/

Подпись руководителя организации*
* Либо уполномоченного представителя, действующего на основании доверенности или распорядительного документа. В случае подписания формы уполномоченным представителем организации (в т.ч. – руководителем филиала) к печатному экземпляру заявки прилагается копия распорядительного документа или доверенности, заверенная печатью организации. Непредставление копии распорядительного документа или доверенности в случае подписания формы уполномоченным представителем организации, а также отсутствие расшифровки подписи, является основанием недопуска заявки к конкурсу.

_____/_____/

Печать (при наличии) организации</td><td>Дата регистрации заявки
13.11.2023 г.</td></tr>
</table>

Форма 1. Сведения о проекте

1.1. Название проекта

на русском языке

Методы машинного обучения для совместного анализа мультимодальных экспериментальных данных и извлечения редких событий в гамма-астрономии

на английском языке

Machine learning methods for combined analysis of multimodal experimental data and extracting of rare events in gamma-ray astronomy

1.2. Приоритетное направление развития науки, технологий и техники в Российской Федерации, критическая технология

Указывается согласно перечню (Указ Президента Российской Федерации от 7 июля 2011 года №899) в случае, если тематика проекта может быть отнесена к одному из приоритетных направлений, а также может внести вклад в развитие критических технологий Российской Федерации.

3. Информационно-телекоммуникационные системы.

8. Нано-, био-, информационные, когнитивные технологии.

Направление из Стратегии научно-технологического развития Российской Федерации (утверждена Указом Президента Российской Федерации от 1 декабря 2016 г. № 642 «О Стратегии научно-технологического развития Российской Федерации») *(при возможности отнесения)*

H1 Переход к передовым цифровым, интеллектуальным производственным технологиям, роботизированным системам, новым материалам и способам конструирования, создание систем обработки больших объемов данных, машинного обучения и искусственного интеллекта

1.3. Ключевые слова *(приводится не более 15 терминов)*

на русском языке

глубокое обучение, разнородные данные из нескольких источников, мультимодальные данные, объединение данных, извлечение признаков, совместный анализ данных, редкие события

на английском языке

deep learning, multi-source heterogeneous data, multimodal data, data fusion, feature extraction, joint data analysis, rare events

1.4. Аннотация проекта *(объемом не более 2 стр.; в том числе кратко – актуальность решения указанной выше научной проблемы и научная новизна)*

Данная информация может быть опубликована на сайте Фонда в информационно-телекоммуникационной сети «Интернет».

на русском языке

В современную цифровую эпоху порождаются все увеличивающиеся объемы данных. Поэтому обработка и анализ получаемой информации является одной из наиболее важных и насущных задач. Часто эти данные поступают из различных источников, отражают различные стороны объектов или явлений и являются существенно неоднородными. Такие данные имеют разные типы и форматы, что очень сильно затрудняет их совместный анализ. В связи с этим, есть потребность в разработке новых эффективных и совершенствовании существующих методов совместного анализа больших потоков разнородных, мультимодальных данных, что является центральной научной проблемой, на решение которой направлен данный проект.

Актуальность этой проблематики обусловлена многочисленными примерами важности таких методов в конкретных прикладных областях, например, в медицине, управлении процессами жизнедеятельности городов, климатических и экологических исследованиях, естественных науках и многих других.

Конкретной задачей в рамках проблемы, на решение которой направлен проект, является разработка новых эффективных способов совместного анализа мультимодальных данных, которые будут протестированы на реальных данных из области гамма-астрономии, а именно данных, получаемых с помощью гибридного эксперимента TAIGA, регистрирующего широкие атмосферные ливни. Мультимодальность означает, что совокупный набор данных состоит

из нескольких подмножеств, каждое из которых содержит данные одного типа. Важно отметить, что мультимодальность характерна в целом для современной многоканальной астрономии, поскольку собираемая информация об изучаемых явлениях имеет большое разнообразие по своей природе и характеристикам. Новизна предлагаемого проекта обусловлена новаторским методологическим подходом для решения этой задачи, а именно осуществлением объединения и совместного анализа не на уровне сырых экспериментальных данных, а после извлечения с помощью нейросетевых технологий их существенных признаков, которые отражают сущность явления, а не конкретный метод его регистрации. Использование существенных признаков позволит выделить редкие явления, к которым относятся гамма частицы.

Первая часть проекта будет посвящена выбору и оптимизации методов извлечения существенных признаков, то есть преобразования входного пространства в подпространство меньшей размерности, которое сохраняет релевантную информацию, адекватную цели научного исследования. Важной частью исследования будет разработка методов интерпретации существенных признаков на языке прикладной области, основанных на машинном обучении. Вторая часть проекта будет посвящена разработке методов совместного анализа существенных признаков данных, полученных из разных источников, и извлечения с их помощью физической информации об исследуемом явлении. Для этого предполагается использовать различные методы: простую конкатенацию признаков, перенос или многозадачное обучения, а также совместное обучение.

Особое внимание будет уделено решению проблемы извлечения редких (аномальных) событий. Дело в том, что соотношение сигнальных событий от источников гамма-излучения во Вселенной к фоновым событиям составляет 1:10000. Для извлечения сигнальных событий предлагается использовать, например, недавно предложенные состязательные автоэнкодеры и метод нормализующих потоков.

Решение поставленных задач и разработанные методы обеспечат ученых инструментарием для совместного анализа больших мультимодальных данных. Его эффективность будет подтверждена на примере задач гамма-астрономии, что, в свою очередь, создаст хороший задел для лучшего понимания процессов, происходящих во Вселенной. В силу общего характера решаемых задач, разработанные методы могут быть применены в других областях науки и техники, требующих комплексного анализа данных, поступающих по нескольким каналам. Таким образом, задачи, поставленные в проекте, являются актуальными, инновационными, масштабными и носят мультидисциплинарный характер.

на английском языке

In today's digital age, huge and ever-increasing amounts of data are constantly being generated. Therefore, processing and analysis of the collected information is one of the most important and urgent tasks. Often the data come from diverse sources, concern different aspects of objects or phenomena, and, therefore, are significantly heterogeneous. Such data are presented in different types and formats, which makes their joint processing and analysis very difficult. Therefore, there is an urgent need to develop new efficient methods for joint processing and analysis of large flows of heterogeneous multimodal data, as well as to improve existing methods. This is the main scientific problem that this project aims to address. The relevance of this issue is due to numerous examples of the need for such methods in specific applied areas, such as medicine, the management of urban life processes, climate and environmental studies, natural sciences and many others.

The specific task within the problem addressed by the project is the development of new effective ways to isolate rare events through joint processing of multimodal data. The proposed methods will be tested on model and real data from the field of gamma-ray astronomy, namely on data obtained using a hybrid system of detectors recording extensive air showers in the TAIGA experiment. Multimodality means that the complete data set consists of several subsets, each of which contains the same type of data, and the data types in different subsets differ significantly. It is important to note that multimodality is a typical feature of data collected in experiments in modern multi-channel astronomy, since the information about the phenomena being studied is not only very large, but also very diverse in nature and characteristics. The novelty of the proposed project is due to an innovative methodological approach to solving this problem, namely, the integration and joint analysis not raw experimental data, but data processed by neural network models, in order to extract those essential features that reflect the fundamental nature of the phenomenon, and not the specific method of its registration. Also, the additional task will be to harmonize as much as possible the formats of features obtained from different data sets.

Thus, one part of the project will be devoted to the selection and optimization of methods for extracting essential features, that is, transforming the input space into a lower-dimensional subspace that preserves most of the relevant information adequate to the purpose of the study. An important part of the research will be the development of machine learning-based

methods for interpreting essential features in terms of an application domain, which will make it possible to control these features using parameters from that application domain. The second part of the project will be devoted to the development of methods for joint analysis of essential features of data obtained from different sources. To do this, it is expected to use methods such as various types of autoencoders, transfer and multi-task learning, normalizing flows and others. Particular attention will be paid to methods for identifying rare, so-called anomalous events. The fact is that the ratio of events received from gamma radiation sources in the Universe to background events is 1:10000. In particular, it is proposed to use recently developed adversarial autoencoders as well as normalizing flow techniques to extract events of interest.

Solving the problems and the developed methods will provide scientists with tools for joint analysis of large multimodal data. Its effectiveness will be confirmed by the example of gamma-ray astronomy problems, which, in turn, will create a good foundation for a better understanding of the processes occurring in the Universe. Due to the general nature of the problems being solved, the developed methods can be applied in other areas of science and technology that require complex analysis of data received through several channels. Thus, the tasks set in the project are relevant, innovative, large-scale and multidisciplinary in nature.

1.5. Ожидаемые результаты и их значимость *(указываются результаты, их научная и общественная значимость (соответствие предполагаемых результатов мировому уровню исследований, возможность практического использования ожидаемых результатов проекта в экономике и социальной сфере, в том числе для создания новой или усовершенствования производимой продукции (товаров, работ, услуг), создания новых или усовершенствования применяемых технологий))*

Данная информация может быть опубликована на сайте Фонда в информационно-телекоммуникационной сети «Интернет».

на русском языке

Основным результатом предлагаемого проекта будут новые эффективные нейросетевые методы и компьютерные модели на основе глубокого обучения, предназначенные для обработки и совместного анализа разнородных мультимодальных данных, полученных из различных источников, и выделения редких событий в потеке экспериментальных данных. Важной частью предложенных методов будут методы отбора существенных признаков экспериментальных данных с помощью машинного обучения с последующей интерпретацией полученных признаков в терминах предметной области для обеспечения возможности их дальнейшего качественного анализа. Актуальность и значимость этого результата связана с тем, что хотя в сравнительно простых случаях выбор величин, характеризующих исследуемое явление, может оказаться естественным и даже очевидным, при исследовании и моделировании сложных систем сам выбор существенных признаков изучаемых явлений является очень сложным и неоднозначным. Второй частью этих методов и соответствующих программных реализаций будут предложены методики обучения нейросетей на основе совместного использования тренировочных наборов признаков разнородных данных и последующего совместного анализа разнородных экспериментальных
данных, полученных из различных источников. Все это должно позволить выделить интересующие исследователей характеристики явлений, которые не могут быть получены из анализа данных отдельных экспериментальных установок. Заметим, что решение таких задач методами машинного обучения является инновационным и будет реализовано впервые в мире.
В рамках этих задач будут разработаны методы извлечения редких (аномальных) событий, с использованием современных нейросетевых моделей на основе современных моделей автоэнкодеров таких как состязательные автоэнкодеры.

Результаты, полученные в ходе выполнения проекта, будут апробированы на реальных данных из области гамма-астрономии, в первую очередь на данных эксперимента TAIGA (Tunka Advanced Instrument for cosmic ray physics and Gamma Astronomy; https://taiga-experiment.info). В эксперименте TAIGA такой подход будет применен впервые. Массив анализируемых данных проекта TAIGA является гетерогенным и состоит из данных, получаемых с черенковских телескопов (Imaging Atmospheric Cherenkov Telescope; IACT), с широкоугольных детекторов с фиксацией времени прихода и временной развертки сигнала TAIGA-HiSCORE, а также с мюонных детекторов. Данные черенковских телескопов TAIGA-IACT и детекторов TAIGA-HiSCORE являются существенно разнотипными: в первом случае это изображения, получаемые камерами телескопов, а во втором – пространственно-временные характеристики сигналов с регистрацией их интенсивностей.

Ключевой особенностью анализируемых данных является значительное, на 4 порядка, превышение количества фоновых событий по сравнению с сигнальными гамма событиями. Поэтому методы извлечения таких редких событий

будут играть ключевую роль в анализе экспериментальных данных.

Таким образом, успешное применение разработанных в рамках проекта методов для анализа разнородных данных эксперимента TAIGA убедительно продемонстрирует их возможности и практическую применимость. Необходимо подчеркнуть, что разработанные методы и их программная реализация, полученные для этого конкретного приложения, будут представлять большой самостоятельный научный интерес. В связи с этим в рамках проекта будет осуществлено всестороннее сравнительное исследование самих разработанных методов, а результаты представлены в виде значений соответствующих метрик и общих выводов.

Хотя разработанные методы будут апробированы для задач астрофизики, они могут быть с успехом применены в других областях фундаментальной и прикладной науки, а также в высокотехнологичных отраслях экономики. В частности, такие методы найдут применение в медицине, управлении процессами жизнедеятельности городов, климатических и экологических исследованиях, в энергетике, анализе финансовых рынков, при дистанционном зондировании Земли, материаловедении и многих других областях.

на английском языке

The main result of the proposed project will be new effective neural network methods and computer models based on deep learning, designed both for processing and joint analysis of heterogeneous multimodal data obtained from various sources, and for identifying rare events in a stream of experimental data. An important part of the proposed methods will be methods for selecting essential features of experimental data using machine learning, followed by interpretation of the selected features in terms of the application domain to enable their further qualitative analysis. The relevance and significance of this result is due to the fact that although in relatively simple cases the choice of parameters characterizing the phenomenon under study may turn out to be natural and even obvious, when studying and modeling complex systems, the choice of essential features of the phenomena being studied is very complex and ambiguous. The second part of these methods will be techniques for training neural networks based on the joint use of training sets of features of heterogeneous data and subsequent joint analysis of heterogeneous experimental data obtained from various sources, as well as software implementations of these techniques. All this should make it possible to identify the characteristics of phenomena that interest researchers, but cannot be obtained from the analysis of data from individual experimental setups. Note that solving such problems using machine learning methods is innovative and will be implemented for the first time in the world. As part of these tasks, methods will be developed for extracting rare (anomalous) events using modern neural network models based on modern autoencoder models such as adversarial autoencoders.

The results obtained during the project will be tested on real data from the field of gamma-ray astronomy, primarily on data from the TAIGA experiment (Tunka Advanced Instrument for cosmic ray physics and Gamma Astronomy; https://taiga-experiment.info). This approach will be applied for the first time in the TAIGA experiment. The data set to be analyzed in the TAIGA project is heterogeneous and consists of data obtained from the Imaging Atmospheric Cherenkov Telescope (IACT), from wide-angle detectors with recording the arrival time and time sweep of the TAIGA-HiSCORE signal, as well as from muon detectors . Data from the TAIGA-IACT Cherenkov telescopes and TAIGA-HiSCORE detectors are of significantly different types: in the first case, these are images obtained by telescope cameras, and in the second, they are the spatiotemporal characteristics of signals with registration of their intensities.

The key feature of the analyzed data is a significant, namely 4 orders of magnitude, excess in the number of background events compared to the gamma events that interest researchers. Methods for extracting such rare events will therefore play a key role in the analysis of experimental data.

Thus, the successful application of the methods developed within the framework of the project for the analysis of heterogeneous data from the TAIGA experiment will convincingly demonstrate their capabilities and practical applicability. However, it must be emphasized that the developed methods and their software implementation obtained for this particular application will be of great independent scientific interest. In this regard, within the framework of the project, a comprehensive comparative study of the developed methods themselves will be carried out, and the results will be presented in the form of values of the corresponding metrics and general conclusions.

Although the developed methods will be tested in astrophysics, they can be successfully applied in other areas of fundamental and applied science, as well as in high-tech sectors of the economy. In particular, such methods will find application in medicine, urban management, climate and environmental research, energetics, financial market analysis, Earth

remote sensing, material sciences, and many others.

1.6. В состав научного коллектива будут входить (указывается планируемое количество исполнителей (с учетом руководителя проекта) в течение всего срока реализации проекта):

Несоответствие состава научного коллектива (в том числе отсутствие информации в соответствующих полях формы) требованиям пункта 12 конкурсной документации является основанием недопуска заявки к конкурсу.

7 исполнителей проекта (включая руководителя),

В соответствии с требованиями пункта 12 конкурсной документации от 4 до 10 человек вне зависимости от того, в трудовых или гражданско-правовых отношениях исполнители состоят с организацией.

в том числе:

- **4** исполнителя в возрасте до 39 лет включительно;
- **2** аспиранта (адъюнкта) очной формы обучения;
- **1** студент очной формы обучения.

1.7. Планируемый состав научного коллектива с указанием фамилий, имен, отчеств (при наличии) членов коллектива, их возраста на момент подачи заявки, ученых степеней, должностей и основных мест работы, формы отношений с организацией (трудовой договор, гражданско-правовой договор) в период реализации проекта

1. Крюков Александр Павлович, 69 лет, к.ф.-м.н., МГУ имени М.В.Ломоносова, заведующий лабораторией, трудовой договор, руководитель.
2. Постников Евгений Борисович, 49 лет ,к.ф.-м.н., МГУ имени М.В.Ломоносова, с.н.с., трудовой договор, ответственный исполнитель.
3. Дубенская Юлия Юрьевна, 41 год, МГУ имени М.В.Ломоносова, н.с., трудовой договор, ответственный исполнитель.
4. Журов Дмитрий Павлович, 30 лет, младший научный сотрудник НИИПФ ИГУ, трудовой договор, исполнитель.
5. Власкина Анна Александровна, 23 года, магистр, физический факультет МГУ имени М.В.Ломоносова, трудовой договор, исполнитель.
6. Гресь Елизавета Олеговна, 25 лет, аспирант, физический факультет ИГУ, трудовой договор, исполнитель.
7. Волчугов Павел Андреевич, 27 лет, аспирант, физический факультет МГУ имени М.В.Ломоносова, трудовой договор, исполнитель.

Соответствие профессионального уровня членов научного коллектива задачам проекта

Предложенный проект посвящен информационным технологиям, одному из самых его бурно развивающихся разделов — интеллектуальным методам анализа больших данных. Проект имеет также сильно выраженный междисциплинарный характер, а его результаты ориентированы на одну из самых передовых областей научных исследований — астрофизику частиц, которая занимается изучением самых глубинных основ строения Вселенной. Поэтому в коллективе участвуют специалисты в области как ИТ, так в области гамма-астрономии, которя является площадкой для аппробации разрабатываемых методов.

Руководитель коллектива А.П.Крюков, к.ф.-м.н., заведующий лабораторией Символьных вычислений в физике высоких энергий Отдела теоретической физики высоких энергий НИИЯФ МГУ.
А.П.Крюков много лет занимается развитием и применением современных информационных технологий в физике. Им были развиты многочисленные математические методы в области компьютерной алгебры, в том числе для задач физики высоких энергий. Он являлся пионером внедрении грид технологии в России, выполнил многочисленные исследования в области распределенных вычислений. Имеет интересные физические результаты. А.П.Крюков является членом коллаборации CMS (ЦЕРН, Женева), которая в 2012 году открыла бозон Хиггса (в 2013 году нобелевская премия по физике была присуждена П.Хиггсу и Ф.Энглеру за теоретическое обоснование существования бозона Хиггса), что является одним из самых выдающихся научных результатов последних лет. Является членом международной коллаборации TAIGA, основной задачей которой является исследование в области физики космических лучей и гамма-астрономии, международной коллаборации Hyper-Kamiokande, изучающей физику нейтрино. За последние 5 лет им опубликовано 47 работ, индексируемых WoS или Scopus, из них 7 в высокорейтинговых журналах Q1. Его индекс Хирша равен 11 (Scopus). Руководил несколькими проектами РФФИ, РНФ, в том числе международными. В последние годы А.П.Крюков активно развивает методы машинного обучения и их применения в физике, в частности в гамма-астрономии, возглавляет исследования по применению сверточных нейронных сетей для классификации космических частиц высоких энергий, определения их параметров, а также быстрой генерации изображений атмосферных

черенковских телескопов с помощью генеративных нейросетей. В настоящее время является руководителем проекта РНФ 22-21-00442, тема "Моделирование выборок случайных событий с учетом априорной информации в астрофизических экспериментах методами машинного обучения". Имеет многочисленные свидетельства о регистрации программ. Под его руководством была защищена кандидатская диссертация в области ИТ, множество дипломных работ студентов. Является членом программных комитетов нескольких престижных конференций, среди которых серия конференций «International Workshop on Advanced Computing and Analysis Techniques in Physics Research» (ACAT), International Conference "Distributed Computing and Grid Technologies in Science and Education" (GRID, Дубна), "International congress Russian Supercomputing Days" (Москва), сопредседатель международного совещания "Deep Learning in Computational Physics" (DLCP). А.П.Крюков является экспертом РАН в области информационных технологий.

Ответственный исполнитель Е.Б.Постников, старший научный сотрудник, к.ф.-м.н.
Е.Б.Постников много лет занимается анализом данных и моделированием эксперимента в астрофизике. Он принимал участие в таких спутниковых экспериментах по изучению космических лучей, как российский NUCLEON и российско-итальянский PAMELA, а также российско-американском аэростатном эксперименте ATIC. С момента основаниямеждународной коллаборации TAIGA, членом которой является, активно работает в сфере наземной гамма-астрономии, решая такие задачи, как моделирование и оптимизация установки TAIGA, планирование наблюдений, анализ и физическая интерпретация данных установки. В 2005-2006 гг. проходил стажировку в Национальном институте ядерной физики (Триест, Италия), а в 2017-2018 гг. работал приглашенным исследователем в Институте физики Общества Макса Планка (Мюнхен, Германия) и исследовательском центре по физике частиц DESY (Гамбург, Германия) по теме моделирования и анализа данных установки TAIGA. За последние 5 лет им опубликована 61 работа, индексируемая WoS или Scopus, получены 2 свидетельства о государственной регистрации программ. Индекс Хирша равен 13 (Scopus).

Ю.Ю.Дубенская, научный сотрудник НИИЯФ МГУ, является опытным исследователем, принимала участие во многих проектах по распределенным вычислениям и использованию технологии виртуализации в области высокопроизводительных вычислений. В настоящее время занимается развитием методов генерации состояний физических систем с помощью генеративно-состязательных нейронных сетей. Ю.Ю.Дубенская получила ряд интересных результатов по моделированию событий на черенковских телескопах с использованием генеративных состязательных нейронных сетей (GAN). Важным аспектом предложенного подхода является то, что поток модельных событий имеет такие же статистические распределения параметров, как и экспериментальные данных. Это позволяет использовать полученные нейросетевые модели вместо ресурсоемких программ на основе методов Монте-Карло. Результаты докладывались на ряде международных конференций и опубликованы в журналах, индексируемых WoS и Scopus. Число публикаций - 15, индекс Хирша равен 3.

Д.П. Журов, младший научный сотрудник НИИПФ ИГУ, г.Иркутск, в 2017 году закончил магистратуру, в 2022 - аспирантуру по направлению "Математическое моделирование, численные методы и комплексы программ". Трехкратный участник чемпионата мира по программированию ACM ICPC, Северо восточный Европейский регион. С 2017 года член коллаборации TAIGA. Проходил курс дополнительного образования по специализации "Машинное обучение и анализ данных" от Яндекс и Московского Физико-Технического Института. С октября 2018 по март 2019 проходил стажировку в Научно-исследовательском центре DESY, г. Цойтен, Германия по программе DAAD «Михаил Ломоносов», где занимался моделированием наведения телескопов IACT и анализом экспериментальных данных. В 2021 году проходил стажировку в университете Сириус в рамках научной школы "Современные методы планирования и управления движением неполноприводных механических систем". Участвовал в гранте РНФ «Карлсруэ-Российская инициатива по работе с астрофизическими данными на протяжении их жизненного цикла», где занимался методами машинного обучения. Является автором 66(55 за последние 5 лет) работ, индексируемых WoS или Scopus, получены 7(4 за последние 5 лет) свидетельства о государственной регистрации программ для ЭВМ. Индекс Хирша -- 9 (Scopus). Научные интересы: системы управления, автоматизация эксперимента, машинное обучение и анализ данных.

Е.О.Гресь — аспирант, младший научный сотрудник (1/2) ИГУ, г.Иркутск. Несмотря на свой возраст уже имеет опыт обработки данных эксперимента TAIGA. В своей квалификационная работа магистранта «Методы глубокого обучения для анализа данных телескопов в эксперименте TAIGA» она разработала программу обработки данных телескопа на основе параметров Хилласа, что позволило заметно улучшить качество отбора гамма-событий в эксперименте TAIGA. В настоящее время она занимается задачей классификации событий методом сверточных нейронных сетей. Полученные результаты доложены на международных конференциях и опубликованы.

А.А.Власкина — магистрант физического факультета МГУ имени М.В.Ломоносова, которая только начинает свой путь в науке. Тем не менее, уже сейчас она добилась хороших результатов в области машинного обучения. Ее бакалаврская работа "Метод сверточных нейронных сетей для анализа широких атмосферных ливней в эксперименте HiSCORE" посвящена использованию сверточных нейросетей для улучшения отношения согнал/шум для изображений широких атмосферных ливней, порожденных гамма квантами от галактических и внегалактических источников. По материалам исследований по методам анализа данных с установки HiSCORE сверточными нейронными сетями были сделаны доклады на международных конференциях и опубликована статья.

П.А. Волчугов - аспирант 4 года обучения физического факультета МГУ. В настоящее время им была разработана методика выделения гамма-квантов регистрируемых атмосферными черенковскими телескопами TAIGA в стерео режиме, когда используются данные с нескольких телескопов одновременно. Предложенный подход позволил улучшить эффективность выделения гамма-квантов. Им была разработана математическая модель работы регистрирующих камер телескопов TAIGA-IACT, что позволило выполнить Монте-Карло моделирование работы установки с учетом ее индивидуальных особенностей. Он активно занимается Монте-Карло моделированием событий для черенковских телескопов. С учетом того, что модельные данных являются основой для получения размеченных тренеровочных наборов, то участие П.А.Волучова будет играть важную роль в проекте. Волчугов П.А. вляется автором 44 работ (Scopus). Индекс Хирша - 5 (Scopus). Научные интересы: гамма-астрономия, физика космических лучей, машинное обучение, анализ и моделирование данных.

Таким образом, коллектив имеет в своем составе опытных исследователей, аспирантов и студентов, сочетая в себе опыт и знания как в области информационных технологий, включая машинное обучение, так и в области физики высоких энергий, включая гамма-астрономию, методы для которой будут разрабатываться, апробироваться и исследоваться. Все это является залогом успешного выполнения программы исследований и получение заявленных результатов, предусмотренных проектом.

1.8. Планируемый объем финансирования проекта Фондом по годам *(указывается в тыс. рублей)*:

Несоответствие планируемого объема финансирования проекта (в том числе отсутствие информации в соответствующих полях формы) требованиям пункта 10 конкурсной документации является основанием недопуска заявки к конкурсу.

2024 г. - 7000 тыс. рублей,
2025 г. - 7000 тыс. рублей,
2026 г. - 7000 тыс. рублей.

1.9. Научный коллектив по результатам выполнения проекта в ходе его реализации предполагает опубликовать** в ведущих рецензируемых*** российских и зарубежных научных изданиях**** не менее

** **Приводятся данные за весь период выполнения проекта. Уменьшение количества публикаций (в том числе отсутствие информации в соответствующих полях формы) по сравнению с порогом, установленным в пункте 16.2 конкурсной документации является основанием недопуска заявки к конкурсу.**

*** **Издания, индексируемые в библиографических зарубежных базах данных публикаций и/или Russian Science Citation Index (RSCI).**

**** **Фонд вправе устанавливать (изменять) перечень международных баз данных, в которых индексируются научные издания, и/или научных изданий, публикации в которых будут учитываться с повышающим коэффициентом.**

В случаях принятия органами власти Российской Федерации или органами управления Фондом соответствующего решения Фонд вправе не менее чем за 8 месяцев до наступления отчетного периода в одностороннем порядке установить или изменить перечень международных баз данных, в которых индексируются научные издания, и/или научных изданий путем направления победителям конкурса соответствующего письменного уведомления.

12 публикаций,
из них

12 в изданиях, индексируемых в базах данных «Сеть науки» (Web of Science Core Collection) или «Скопус» (Scopus);

12 в изданиях, индексируемых в Russian Science Citation Index;

0 в изданиях, индексируемых в иных зарубежных библиографических базах данных.

Информация о научных изданиях, в которых предполагается опубликовать результаты проекта, в

том числе следует указать в каких базах индексируются данные издания - «Сеть науки» (Web of Science Core Collection), «Скопус» (Scopus), RSCI, РИНЦ, иные базы, а также указать тип публикации - статья, обзор, монография, иной тип

-- Astronomy and Computing (Q1), статья.
– Neural Networks (Q1), статья.
– MSU Bulletine. Physics series (Q4), статья.
-- Proceedings of Science (Q4), статья.
-- Программирование (Q4), статья.
– Вычислительные методы и программирование, статья.

Иные способы обнародования результатов выполнения проекта

Выступления на ведущих всероссийских и международных конференциях, в том числе на таких регулярно проводящихся конференциях как Russian Supercomputer Days, International Workshop on Deep Learning in Computational Physics (DLCP), International Conference "Distributed Computing and Grid Technologies in Science and Education" (GRID), International Cosmic Ray Conference (ICRC), Parallel computational technologies (PCT). Полный список конференций будет определен в ходе выполнения проекта..

1.10. Число публикаций членов научного коллектива, опубликованных в период с 1 января 2019 года до даты подачи заявки,

83, из них

67 – опубликованы в изданиях, индексируемых в Web of Science Core Collection или в Scopus,

51 – опубликованы в изданиях, индексируемых Russian Science Citation Index,

5 – опубликованы в изданиях, индексируемых в иных зарубежных библиографических базах данных.

1.11. Планируемое участие научного коллектива в международных коллаборациях (проектах) (при наличии)

TAIGA

1.12. Информация о возможности использовании результатов выполнения проекта в осуществлении хозяйственной деятельности предприятий Российской Федерации, в том числе о способе использования, о намерениях по внедрению на основании прогнозируемых результатов проекта новой или усовершенствованию производимой продукции (товаров, работ, услуг), новых или усовершенствованных применяемых технологий; о формировании по итогам реализации проекта научных и технологических заделов, обеспечивающих экономический рост и социальное развитие Российской Федерации (с приложением подтверждающих документов, при наличии)

Разработанные методы могут быть использованы в рамках национального проекта "Наука" не только для задач астрофизики, они могут быть с успехом применены в других областях фундаментальной и прикладной науки, а также в высокотехнологичных отраслях экономики. Результаты проекта будут обладать высоким мультидисциплинарным потенциалом. Они могут быть с успехом применены в фундаментальной и прикладной науке, а также в высокотехнологичных отраслях экономики. В частности, такие методы найдут применение в медицине, управлении процессами жизнедеятельности городов, климатических и экологических исследованиях, в энергетике, анализе финансовых рынков, при дистанционном зондировании Земли, материаловедении и многих других областях.

Руководитель проекта подтверждает, что

- все члены научного коллектива (в том числе руководитель проекта) удовлетворяют пунктам 6, 7, 13 конкурсной документации;
- на весь период реализации проекта руководитель проекта будет состоять в трудовых отношениях

с организацией, при этом трудовой договор с организацией не будет предусматривать возможность осуществления трудовой деятельности за пределами территории Российской Федерации (в том числе, путем направления работника в служебную командировку, значительная длительность которой не обусловлена целями проекта);

- при обнародовании результатов любой научной работы, выполненной в рамках поддержанного Фондом проекта, руководитель проекта и научный коллектив будут указывать на получение финансовой поддержки от Фонда и организацию, а также согласны с опубликованием Фондом аннотации и ожидаемых результатов проекта, соответствующих отчетов о выполнении проекта, в том числе в информационно-телекоммуникационной сети «Интернет», с использованием Фондом в некоммерческих целях представляемых в Фонд материалов, в том числе, содержащих результаты выполнения проекта, с предоставлением указанных материалов органам власти Российской Федерации, институтам развития;
- помимо гранта Фонда проект не будет иметь других источников финансирования в течение всего периода практической реализации проекта с использованием гранта Фонда;
- проект не является аналогичным по содержанию проекту, одновременно поданному на конкурсы научных фондов и иных организаций;
- проект не содержит сведений, составляющих государственную тайну или относимых к охраняемой в соответствии с законодательством Российской Федерации иной информации ограниченного доступа;
- доля членов научного коллектива в возрасте до 39 лет включительно в общей численности членов научного коллектива будет составлять не менее 50 процентов в течение всего периода практической реализации проекта;
- в установленные сроки будут представляться в Фонд ежегодные отчеты о выполнении проекта и о целевом использовании средств гранта.

Подпись руководителя проекта _____ /А.П. Крюков/

4.1. Научная проблема, на решение которой направлен проект

В настоящее время наблюдается чрезвычайно быстрый прогресс в технологии сбора разнообразных данных из многих источников в различных областях человеческой деятельности. Часто эти большие объемы данных поступают из различных источников, отражают различные стороны объектов или явлений, и, поэтому, имеют разные типы и форматы, что очень сильно затрудняет их совместную обработку и анализ. Другими словами, такие данные характеризуются значительной и сложной гетерогенностью (мультимодальностью). Поэтому традиционные алгоритмы их обработки часто оказываются недостаточно эффективными из-за сложности совместного всестороннего анализа и интерпретации таких сильно разнородных данных. Это вызывает повышенный спрос на альтернативные инструменты - с мощными возможностями совместной обработки. Разработка новых быстрых и эффективных, а также совершенствование существующих методов совместной обработки и анализа больших потоков разнородных данных является общей научной проблемой, на решение которой направлен данный проект.

Конкретной научной проблемой проекта в рамках указанной общей проблемы является разработка подходов и методов совместного анализа разнородных данных на основе искусственного интеллекта, в нашем конкретном случае - глубокого обучения. Передовая технология глубокого обучения обеспечила значительные прорывные решения многочисленных задач благодаря впечатляющим возможностям представления, реконструкции и анализа данных. Поэтому методы машинного обучения являются естественными и многообещающими технологиями в области совместного анализа мультимодальных (то есть, состоящих из нескольких типов/мод) данных.

4.2. Научная значимость и актуальность решения обозначенной проблемы

Целью объединения данных, полученных от нескольких источников (датчиков, детекторов и другой измерительной аппаратуры), является их совместная обработка и синтез информации, относящейся к исследуемому объекту или явлению, для получения более точной, более полной, более надежной и последовательной интерпретации и описания такого объекта или явления, чем при использовании одного источника. Чтобы этого добиться, необходима разработка новых быстрых и эффективных методов, а также совершенствование существующих методов совместной обработки и анализа больших потоков разнородных данных. Это обуславливает научную значимость и актуальность решения проблемы, на решение которой направлен данный проект. Актуальность этой научной проблематики также подтверждается многочисленными примерами важности и востребованности таких подходов и методов в конкретных прикладных областях, например, в медицине, управлении процессами жизнедеятельности городов, климатических и экологических исследованиях, в энергетике, анализе финансовых рынков, дистанционном зондировании Земли, материаловедении и многих других областях науки, производства и социальной жизни страны (см. п.4.5). При этом разработка общих теоретических подходов, а также экспериментальных исследований и практических реализаций являются весьма важными для лучшего понимания и совершенствования процесса совместного анализа неоднородных мультимодальных данных.

В рамках проекта будет осуществлено как общее исследование и разработка подходов и методов совместного анализа мультимодальных данных, так и тщательная практически значимая апробация полученных результатов в одной из наиболее фундаментальных и быстро развивающихся в последнее время областей — астрофизике частиц. Современная астрофизика исследует наиболее фундаментальные вопросы эволюции Вселенной от момента возникновения во времена Большого взрыва до современного состояния. Эти исследования требуют проведения большого числа экспериментов, в ходе которых собирается огромный объем данных, составляющий десятки петабайт. Экспериментальные установки, которые используются для исследования процессов во Вселенной, могут быть как наземными обсерваториями, разбросанными по всему земному шару, так и расположенными на космических аппаратах. Важной особенностью этих экспериментов, получивших общее название "мультиканальная астрофизика" (multi-messenger astrophysics), является то, что собираемая информация об изучаемых явлениях имеет большое разнообразие по своей природе (типу). Это может быть электромагнитное излучение от радиодиапазона до гамма-квантов сверхвысоких энергий, различные типы частиц - нейтрино, протоны, ядра различных атомов. В последние годы к ним добавились данные, получаемые с помощью детекторов гравитационных волн. Все это позволяет взглянуть на изучаемые процессы с разных сторон, и, следовательно, построить более точную картину мира. Таким образом, задача комплексного совместного анализа мультимодальных данных является одной из наиболее приоритетных задач современной многоканальной астрофизики, а разработка методов анализа, основанных на новых принципах и

обладающих высокой эффективностью и точностью, является актуальной задачей, требующей как теоретического решения, так и практической реализации в виде реальных программных продуктов, которые в данном проекте будут основаны на нейросетевых моделях (подробнее см. п.4.6).

Решение поставленных задач и разработанные в ходе работы над проектом методы обеспечат ученых инструментарием для совместного анализа больших данных. Его эффективность будет апробирована на задачах гамма-астрономии (астрофизики), что, в свою очередь, создаст хороший задел для лучшего понимания процессов, происходящих во Вселенной. В конечном счете это позволит пролить свет на множество вопросов современной астрофизики, в том числе об источниках космических лучей высоких энергий и механизме их образования, источниках нейтрино высоких энергий, а также поиске темной материи.

Подчеркнем, что в силу общего характера решаемых задач, разработанные методы могут быть применены в разнообразных областях науки и техники, требующих комплексного анализа разнородных данных, поступающих по нескольким каналам – как в случае перечисленных выше чисто прикладных задач, так и в фундаментальной науке. Таким образом, проблемы, поставленные в проекте, имеют большую научную значимость, являются актуальными, инновационными и масштабными и носят мультидисциплинарный характер.

4.3. Конкретная задача (задачи) в рамках проблемы, на решение которой направлен проект, ее масштаб и комплексность

Конкретной задачей в рамках общей проблемы, на решение которой направлен проект, является разработка новых и совершенствование существующих способов совместной обработки разнородных мультимодальных данных на основе выделения их существенных признаков методами глубокого обучения. Полученные результаты будут апробированы на модельных и реальных мультимодальных данных из области гамма-астрономии, а именно данных, получаемых с помощью гибридной установки, регистрирующих широкие атмосферные ливни (ШАЛ) в эксперименте TAIGA.

Проект включает в себя три части. Одна часть проекта будет посвящена выбору и оптимизации методов извлечения существенных признаков и их интерпретации в терминах предметной области. Пространство существенных признаков имеет заметно меньшую размерность, чем исходные данные. Понижение размерности является широко распространенным способом предварительной обработки многомерных данных при их анализе, визуализации и моделировании. Один из очевидных способов уменьшить размерность заключается в отборе только тех свойств входных данных, которые содержат релевантную информацию для решения конкретной проблемы (feature selection). Выделение (извлечение) признаков (feature extraction) - это более общий метод, в котором пытаются найти преобразование входного пространства в подпространство меньшей размерности, которое сохраняет большую часть релевантной информации, адекватной цели исследования. Важной новизной проекта, будет разработка методов интерпретации выделенных существенных признаков на языке прикладной области, основанных на машинном обучении, а также возможность управления этими признаками с помощью параметров из этой прикладной области. Будет проведено исследование свойств пространства существенных признаков на данных из прикладной области и проведен поиск оптимальной структуры этого пространства. Актуальность и значимость этого результата связана с тем, что хотя в сравнительно простых случаях выбор величин, характеризующих исследуемое явление, может оказаться естественным и даже очевидным, при исследовании и моделировании сложных систем сам выбор существенных признаков изучаемых явлений является очень сложным, неоднозначным и возможным только при помощи самых современных и должным образом адаптированных методов глубокого обучения.

Вторая часть проекта посвящена разработке методов совместного анализа (объединения) существенных признаков данных, полученных из разных источников. При этом возможны различные варианты, например, простая конкатенация, перенос обучения (transfer learning) или многозадачное обучение (multitask learning), использование сверточных и других типов сетей, и даже совместное обучение при удачном выделении признаков разных данных. Все это должно позволить выделить интересующие исследователей характеристики явлений, которые могут быть получены только при совместном анализе данных с нескольких экспериментальных установок. Заметим, что решение таких задач методами машинного обучения является инновационным и будет реализовано впервые в мире.

Третья часть проекта будет посвящена решению проблемы извлечение редких (аномальных) событий на основе методов, разработанных в двух первых частях проекта. Дело в том, что соотношение сигнальных событий от источников гамма-излучения во Вселенной к фоновым событиям составляет 1:10000. Для извлечения сигнальных событий предлагается использовать, например, недавно предложенные состязательные автоэнкодеры и метод

нормализующих потоков.

Результаты, полученные в ходе выполнения проекта, будут апробированы как на модельных так и на реальных данных эксперимента TAIGA (Tunka Advanced Instrument for cosmic ray physics and Gamma Astronomy; https://taiga-experiment.info). Массив анализируемых данных проекта TAIGA является мультимодальным и состоит из изображений широких атмосферных ливней (ШАЛ), порождаемых космическими лучами, с черенковских телескопов, и данных более сотни широкоугольных детекторов, которые регистрируют время прихода сигнала TAIGA-HiSCORE. При необходимости могут быть использованы данные мюонных детекторов. Данные черенковских телескопов TAIGA-IACT и детекторов TAIGA-HiSCORE являются существенно разнотипными: в первом случае это изображения, получаемые камерами телескопов, а во втором – пространственно-временные характеристики сигналов с регистрацией их интенсивностей. Таким образом, успешное применение разработанных в рамках проекта методов для анализа разнородных данных эксперимента TAIGA убедительно
продемонстрирует возможности этих методов и их практическую применимость.

Необходимо подчеркнуть, что разработанные методы и их программная реализация, полученные для этого конкретного приложения, будут представлять большой самостоятельный научный интерес. В связи с этим в рамках проекта будет осуществлено всестороннее сравнительное исследование самих разработанных методов, а результаты представлены в виде значений соответствующих метрик и общих выводов.

Таким образом, практическими задачами проекта будут:
• исследование существующих и разработка новых методов машинного обучения для выделения и отбора существенных признаков для наборов экспериментальных (обучающих) данных;
• разработка методики интерпретации полученных существенных признаков в терминах предметной области, к которой относится исследуемая система;
• исследование пространства существенных параметров и поиск его оптимального размера;
• выбор способа или комбинации способов совместного анализа мультимодальных данных на основе глубокого обучения (перенос или многозадачное обучение, совместное обучение и т. п.);
• разработка методов и алгоритмов совместного анализа гетерогенных мультимодальных экспериментальных данных на основе выделенных существенных признаков методами глубокого обучения;
• разработка методов поиска редких (аномальных) явлений — сигнальных гамма событий с помощью состязательных автоэнкодеров и нормализующих потоков.
• программная реализация нейросетевых моделей с использованием современного инструментария - таких как PyTourch, TensorFlow, Keras - для практической реализации разработанных методов и алгоритмов;
• практическая апробация разработанных методов на примере задачи в области астрофизики частиц, а именно, совместный анализ данных черенковских телескопов TAIGA-IACT и детекторов TAIGA-HiSCORE.

Как видно из вышесказанного, в ходе осуществления проекта будет проведен полный цикл исследований и решения поставленных задач — от теоретической разработки подхода до практической программной реализации соответствующих нейросетевых моделей. Результаты проекта обладают высоким мультидисциплинарным потенциалом. Разработанные методы будут апробированы на примере задач астрофизики. Тем не менее они могут быть с успехом применены в других областях фундаментальной и прикладной науки, а также в высокотехнологичных отраслях экономики. В частности, такие методы найдут применение в медицине, управлении процессами жизнедеятельности городов, климатических и экологических исследованиях, в энергетике, анализе финансовых рынков, при дистанционном зондировании Земли, материаловедении и многих других областях. Это обуславливает масштаб и комплексность задач проекта.

4.4. Научная новизна исследований, обоснование достижимости решения поставленной задачи (задач) и возможности получения предполагаемых результатов

Большая часть существующих работ по данной тематике посвящены конкретным узкоспециализированным областям применения. В соответствии с реальными прикладными задачами, в этих работах устанавливаются интуитивные критерии и на этой основе создаются соответствующие схемы слияния. В целом этой области исследований свойственна предметная направленность, которая не формирует необходимых теоретических основ и обобщенной системы алгоритмов. Отсутствие базовой теоретической основы и обобщенной системы алгоритмов не только препятствует глубокому пониманию процесса слияния разнородных данных, но также мешает успешному переносу методов, разработанных для одной предметной области, на другие.

Основным новаторским общим методологическим подходом для решения задачи совместного анализа мультимодальных гетерогенных данных в рамках проекта является осуществление их объединения не на уровне сырых данных, а после извлечения (с помощью нейросетевых технологий типа автоэнкодеров или сверточных сетей) существенных признаков, которые должны в большей степени отражать сущность явления, а не конкретный метод его регистрации. В ходе реализации соответствующих нейросетевых моделей будет проведено глубокое исследование свойств пространств существенных параметров и определены их оптимальные характеристики.

В настоящее время делаются первые, но весьма успешные шаги, направленные на применение методов машинного обучения в астрофизике. Главным образом применяются методы, основанные на деревьях решений и искусственных нейронных сетях, а также ряд других. Новизна предлагаемого проекта состоит в том, что впервые будут разработаны методы машинного, а точнее – глубокого, обучения для совместного анализа разнородных данных об одном или нескольких взаимосвязанных явлениях и объектах в области гамма-астрономии, которые включают изображения, получаемые с черенковских телескопов, а также пространственно-временные распределения и амплитуды сигналов, регистрируемые многоканальными детекторами.

В проекте впервые предполагается использовать нейросетевые модели, основанные на состязательных автоэнкодерах и нормализующих потоках, для поиска редких (аномальных) событий, к которым относятся гамма события от Галактических источников. Успешное решение поставленной задачи позволит ученым существенно продвинуться в понимании физики процессов.

Достижимость решения поставленных задач и получения предполагаемых результатов основана на следующем:
- высокий уровень развития теории и практики использования машинного обучения, обширные результаты по качественным и количественным характеристикам сетей различного типа, в том числе сверточным нейронным сетям (convolutional neural networks; CNN), обычным и вариационным автокодировщикам, рекуррентным сетям и т.п., сетям с различным выбором гиперпараметров и функций потерь;
- результаты по методам выделения и отбора существенных признаков изображений на основе машинного обучения;
- результаты по "ручному" подбору массива существенных признаков, основанных на физических моделях, для широких атмосферных ливней в астрофизике, исследуемых с помощью черенковских детекторов, который может служить референсным набором при апробации разработанных методов в прикладной области;
- высокая квалификация руководителя и основных исполнителей, имеющийся у коллектива исполнителей проекта научный задел в области информационных технологий и в области машинного обучения, опыт совместной успешной реализации проектов, в том числе международных.

4.5. Современное состояние исследований по данной проблеме, основные направления исследований в мировой науке и научные конкуренты

Для решения поставленных в рамках проекта конкретных задач будут проводится исследования по следующим направлениям:
- выделение существенных признаков входных экспериментальных данных средствами глубокого обучения и исследование свойств пространства существенных параметров;
- методы объединения и совместного анализа разнородных данных на основе глубокого обучения;
- методы выделения редких событий методами глубокого обучения.
- апробация и исследование этих методов для приложений в области обработки разнотипных/мультимодальных данных гибридной системы детекторов для гамма-астрономии.
Соответственно, ниже представлено современное состояние исследований по вышеуказанным направлениям с ссылками на литературные источники.

1. Выделение существенных признаков входных данных

Понижение размерности является широко распространенным способом предварительной обработки многомерных данных при их анализе, визуализации и моделировании. Один из очевидных способов уменьшить размерность заключается в отборе только тех измерений входных данных, которые содержат релевантную информацию для решения конкретной проблемы (feature selection). Выделение (извлечение) признаков (feature extraction) - это более общий метод, в котором пытаются найти преобразование входного пространства в подпространство меньшей размерности, которое сохраняет большую часть релевантной информации [1.1]. Методы выделения и отбора признаков

используются независимо или совместно с целью улучшения методов машинного обучения (в частности, точности (accuracy)), для визуализации и, что важно для данного проекта, для осознания и интерпретации полученных знаний о данных [1.2]. Как правило, признаки можно разделить на следующие категории: релевантные, нерелевантные или избыточные. В процессе выбора характеристик для алгоритма обучения стараются отобрать подмножество с наименьшим количеством измерений, которые больше всего способствуют точности обучения [1.3]. Преимущество отбора признаков заключается в том, что важная информация, относящаяся к одному признаку, не теряется, но если требуется создать небольшой набор признаков, а исходные признаки очень разнообразны, существует вероятность потери информации, поскольку при отборе некоторые признаки должны быть опущены. С другой стороны, выделение признаков, в принципе, позволяет уменьшить размер пространства признаков без потери информации об исходном пространстве признаков. Недостатком выделения признаков является тот факт, что комбинация исходных признаков часто не поддается простой интерпретации, а информация о том, какой вклад вносит исходный признак, иногда теряется [1.4]. Для разработки лучших методов отбора и выделения признаков было приложено много усилий, в частности были развиты такие подходы, как mRMR [1.58], RELIEF [1.59], CMIM [1.60], метод, основанный на коэффициентах корреляции [1.61], BW-ratio [1.62], INTERACT [1.63], Genetic Algorithm [1.64], SVM-REF [1.65], PCA (метод главных компонент) [1.7], нелинейный анализ главных компонент [1.66], независимый анализ компонент [1.67], гибридный подход [1.5], алгоритмы взвешивания [1.6] и др. Ввиду значительного количества существующих алгоритмов выбора и извлечения признаков возникает необходимость выработать критерии, которые позволяют решить, какой алгоритм использовать в определенных ситуациях. Самым популярным и широко используемым подходом к выделению признаков является метод главных компонент (Principle Component Analysis; PCA) [1.7]. PCA является непараметрическим методом, используемым для извлечения наиболее релевантной информации из набора избыточных или зашумленных данных. По существу это линейное преобразование данных, которое минимизирует избыточность (измеряемую с помощью ковариации) и максимизирует информацию (измеряемую с помощью дисперсии) [1.8]. Было предложено также много модификаций PCA. В работах [1.9,1.10] были предложены двухуровневые (гибридные) методы уменьшения размерности, которые объединяют методы отбора и выделения признаков с целью повышения эффективности классификации: на первом уровне уменьшения размерности характеристики выбираются на основе взаимной корреляции, а на втором уровне выбранные признаки используются для выделения новых признаков с помощью PCA.

Автокодировщик (автоэнкодер) был впервые представлен в конце 1980-х годов [1.11] как метод линейного извлечения признаков. Автокодировщик стремится изучить более простое представление данных путем сопоставления исходных данных с пространством низкой размерности. Основной принцип работы автокодировщика следует из названия. "Авто" означает, что этот метод не требует обучения с учителем, а "кодировщик (энкодер)" означает, что он порождает другое представление данных. В частности, автокодировщик изучает закодированное представление, минимизируя потери между исходными данными и данными, декодированными из этого представления. В 1989 г. была исследована [1.12] взаимосвязь между однослойным автокодировщиком и методом главных компонент (PCA). Было обнаружено, что новые сжатые признаки, полученные линейным автокодировщиком, аналогичны главным компонентам. Позже, с использованием нелинейных функций активации, автокодировщик стал нелинейным и способным порождать больше полезных признаков [1.13], чем линейные методы извлечения признаков. Однако возрождение интереса к автокодировщикам связано с успехом эффективного обучения глубоких архитектур. Так в работе [1.14] удалось добиться успеха в обучении составного автокодировщика на наборе данных MNIST с помощью жадного (greedy) алгоритма и послойного подхода. Последующие исследования показали, что составная модель автокодировщика может изучать значимые, абстрактные признаки и, таким образом, достигать лучших результатов классификации в данных большой размерности, таких как изображения и тексты [1.15–1.18]. Эффективность обучения многослойного автокодировщика дополнительно повышается за счет изменения инициализации весов [1.19].

Было также обнаружено, что по мере увеличения числа слоев веса более глубоких слоев резко увеличиваются так, что значения этих весов становятся больше, чем исходные входные характеристики. Эта проблема переобучения приводит к тому, что представления признаков глубоких слоев более вероятно зависят от сетевой структуры, а не от исходных входных характеристик. Поэтому в работе [1.20] была выдвинута идея увеличения разреженности сетевой структуры, чтобы ограничить увеличение веса, в работе [1.21] был добавлен член регуляризации в функцию потерь автокодировщика, чтобы наложить штраф на большие веса. Для решения этой же проблемы в [1.22] был предложен автокодировщик с шумоподавлением (Denoising Autoencoder; DAE) путем добавления шумов на входе. Предложенная модель DAE направлена не на восстановление исходных входных данных, а на восстановление этих данных, искаженных гауссовским шумом. Затем был предложен вариационный автокодировщик (Variational Autoencoder; VAE) [1.23] для генерации желаемых распределений представлений в скрытых слоях. В целом по мере повышения

эффективности обучения автокодировщики в настоящее время становятся все более популярным инструментом.

Автокодировщики являются инструментом понижения размерности данных на основе машинного обучения без учителя. При этом сжатые признаки на высоких уровнях обычно содержат только основную информацию об исходных данных. Это делает автоэнкодеры нечувствительными к небольшим изменениям. Чтобы сделать их чувствительными к незначительным изменениям, в [1.24] предложили сужающий автокодировщик (Contractive Autoencoder; CAE). В [1.25] был предложен обобщенный автокодировщик (Generalized Autoencoder; GAE), нацеленный на восстановление взаимосвязей данных, а не признаков данных. Серия приложений [1.26–1.30] GAE подтверждает, что нахождение взаимосвязей данных может привести даже к лучшим результатам, чем порождение признаков. В последнее время автокодировщики различного типа находят все большее применение в различных прикладных областях, например в акустике [1.31], медицине [1.32], в области компьютерной безопасности [1.33], и др.

Важно отметить, что, вообще говоря, в качестве кодировщика могут выступать не только автокодировщики, но и другие нейросети. Так в работе [1.68] в качестве кодировщика для извлечения признаков используется сверточная сеть (CNN). Выделенные признаки используются для эффективного и быстрого поиска семантического сходства среди объектов в очень больших наборах данных (в указанной работе метод непосредственно применяется для поиска среди 42 миллионов изображений галактик).

Как следует из вышесказанного, в настоящее время существуют разнообразные и хорошо развитые методы отбора и выделения существенных признаков в наборах данных, которые применялись для различных практических задач. Однако возможность использования выделенных таким образом из экспериментальных данных признаков для последующего использования в целях совместного анализа сильно неоднородных данных (в том числе, изображений и пространственно-временных данных), в частности, на примере задач многоканальной астрофизики, ранее не рассматривалась и является новаторской. Также будет исследован вопрос оптимального размера пространства существенных признаков, а также вопрос их интерпретации в терминах прикладной области.

2. Объединение и совместный анализ неоднородных мультимодальных данных на основе глубокого обучения

Неоднородность является одной из основных особенностей больших данных, и это приводит к проблемам при их интеграции и совместном анализе. Различают следующие виды неоднородности данных [2.1]:
- синтаксическая неоднородность возникает, когда два источника данных выражены на разных языках;
- концептуальная неоднородность, также известная как семантическая неоднородность или логическое несоответствие, обозначает различия в моделировании одной и той же интересующей области;
- терминологическая неоднородность означает различия в названиях при обращении к одним и тем же объектам из разных источников данных;
- семиотическая неоднородность, также известная как прагматическая неоднородность, означает различную интерпретацию сущностей людьми.

Такие неоднородные данные, получаемые из различных источников еще называют мультимодальными большими данными (multimodal big data). Они содержат обширную внутримодальную и кросс-модальную информацию и создают огромные проблемы для традиционных методов объединения (слияния; fusion) и совместного анализа данных. Методы слияния данных используются для сопоставления и объединения разнородных наборов данных для улучшения качества информации о реальных объектах и явлениях, что способствует более точному интеллектуальному анализу данных [2.2-2.5].

В настоящее время общие методы объединения данных изучают ценность данных с разных точек зрения. Например, простейший метод слияния данных — это непосредственное объединение (конкатенация) двух одномерных наборов данных, имеющих одинаковые области значений в этом измерении. Кроме того, объединение данных может быть выполнено с точки зрения извлечения существенных признаков разных измерений. Основываясь на существующем семантическом понимании текстовых данных, слияние данных также можно проводить на основе семантики данных. Методы объединения данных разделяются на три категории [2.6]: основанные на признаках (feature-based level), подразделенные на этапы (stage-based) и основанные на семантике (semantic meaning-based). В методах слияния данных на основе признаков признаки одного и того же измерения обычно извлекаются из разных данных, а затем эти признаки напрямую объединяются (конкатенация) [2.7,2.8] или используются для методов глубокого

обучения [2.9-2.11]. Для метода прямого объединения признаков необходимо отметить несколько проблем: во-первых, при непосредственном объединении данных необходимо удалить повторяющиеся признаки; во-вторых, некоторые функции разных размерностей, которые обеспечивают хорошую производительность модели, могут быть потеряны из-за прямого слияния; в-третьих, прямое слияние признаков может привести к переобучению.

Метод слияния данных, основанный на глубоком обучении, позволяет модели глубокого обучения достигать хороших результатов как в извлечении признаков, так и в их анализе.
Метод слияния данных на основе этапов [2.12 – 2.14] делит проблему на разные этапы, затем анализирует проблемы каждого этапа с помощью данных этого этапа и, наконец, объединяет результаты каждой задачи этапа. Для поэтапного метода слияния данных необходимо обратить внимание на следующие проблемы: во-первых, разделение целевой задачи на разные этапы приведет к потере связей между задачами на разных этапах; а во-вторых, при объединении решений каждого этапа возникает проблема оптимизации их комбинации. Вообще говоря, разные методы объединения данных обычно по-разному влияют на объединенные результаты.

Методы объединения данных на основе семантического значения основаны на сходстве и корреляции знаний, содержащихся в данных и измеряемых разными способами. Они подразделяются на четыре категории: метод слияния данных на основе нескольких представлений [2.15,2.16], метод слияния данных на основе сходства [2.17,2.18], метод слияния данных на основе вероятностных зависимостей [2.19,2.20] и метод слияния данных на основе переходного обучения [2.21,2.22]. С точки зрения процесса исследования объекта метод слияния данных на основе нескольких представлений делится на метод совместного обучения [2.23,2.24], метод многоядерного обучения [2.15,2.25] и метод обучения на основе подпространств [2.26,2.27].

Метод совместного обучения использует знания из разных представлений для одновременного обучения модели. Метод многоядерного обучения основан на машинном обучении, которое использует разные ядра в разных методах машинного обучения. Метод обучения на основе подпространств изучает потенциальное подпространство из разных представлений, предполагая, что входные представления генерируются из этого скрытого подпространства. Метод объединения данных на основе сходства обычно предназначен для измерения степени корреляции данных из нескольких источников, количественной оценки степени сходства и построения матрицы сходства для изучения. Связанная матричная факторизация [2.17,2.28] и нелинейная редукция размерности (изучение многообразий; manifold learning — русскоязычный термин пока не устоялся) [2.18,2.29] — это два классических метода слияния данных, основанных на сходстве данных. В работе [2.18] анализировались данные с помощью от двух или более источников, а затем использовался метод, основанный на manifold learning, для выявления внутренней структуры данных, что сделало предлагаемую модель мало зависимой от минимального предварительного знания о данных. Вероятностный метод объединения данных на основе зависимостей использует структурные графы. В этом методе разные данные отображаются в качестве узлов, а взаимосвязь между данными, например причинно-следственная связь, соответствует ребрам (ребра делятся на направленные и ненаправленные, которые определяются в соответствии с используемым структурным графом). После такого построения используются методы теории графов для объединения данных. Например, в работе [2.20] использовались пространственно-временные данные из разных районов для обнаружения городских коллективных аномалий. Существует метод слияния данных, основанный на концепции переноса обучения (transfer learning) [2.40], которая применяет полученные знания к другим проблемам. Метод слияния данных на основе переноса обучения делится на перенос обучения между наборами данных одного типа и перенос обучения между несколькими разными наборами данных. В первом случае данные могут быть перенесены из одного домена в другой, например, если во втором из них обучающих данных недостаточно много. Во втором - знания о нескольких наборах данных могут быть переданы из исходной задачи в целевую.

Различные подходы и технические вопросы объединения гетерогенных данных на основе машинного обучения рассмотрены в работах [2.30], [2.31],[2.32], [2.33]. Совместный анализ разнородных данных в узкоспециализированных прикладных областях используется, например, в работах [2.34] (молекулярная биология/протеомика), [2.35] (анализ транспортных потоков), [2.36] (энергетика), [2.37] (медицина), [2.38] (анализ фондовых рынков), [2.39] (материаловедение) и других.

Как видно, проблема объединения разнородных/мультимодальных данных интенсивно исследуется в мировой литературе. Подход, предлагаемый в данном проекте, отличается тем, что направлен на совместный анализ сильно неоднородных данных (изображения и пространственно-временные данные), основан на тщательном анализе и оптимизации методов выделения существенных признаков из данных, полученных из разных источников, чтобы

сделать их максимально приближенным к однородным наборам данных, а также тем, что впервые будет применен для решения задач астрофизики частиц.

3. Анализ мультимодальных данных в гамма-астрономии

За последнее время методы машинного обучения стали все чаще применяться для анализа данных в астрофизике частиц. Машинное обучение применялось для идентификации частиц и реконструкции параметров, извлечения энергетических спектров, реконструкции треков частиц. С ростом числа экспериментальных установок разного типа многоканальная астрономия и связанная с ней необходимость совместного анализа данных, поступающих с различных установок, будет все более востребованной. Она незаменима для исследований астрофизики таких явлений, как слияние нейтронных звезд и/или черных дыр, гамма-всплески и появление сверхновых звезд [3.1]. Телескопы, регистрирующие электромагнитное излучение космического и наземного базирования, детекторы нейтрино и космических лучей, а также детекторы гравитационных волн предоставляют обширные возможности исследования Вселенной путем тщательного и всестороннего анализа приходящих сигналов разной природы [3.1-3.9].

Электромагнитное излучение от различных космических источников излучается во всем диапазоне от радио до гамма квантов высоких энергий, охватывает различные временные шкалы от секунд до месяцев и лет. Широкий диапазон излучения и различные временные рамки требуют глобальной сети наземных телескопов и спутников, способных улавливать электромагнитные сигнатуры источников, а также соответствующих методов наблюдений и анализа данных, поступающих от нескольких различных источников [3.1]. Другим типом высокоэнергичных космических частиц являются нейтрино (см., например, [3.10 — 3.21], [3.39 — 3.50]).

В настоящее время является общепризнанным, что для обработки огромного объема данных получаемых в области многоканальной астрофизики критически важно ускорить разработку и внедрение новых методов обработки сигналов, причем в первую очередь — на основе методов машинного/глубокого обучения (см., например, [3.22]). В частности было показано [3.23], что алгоритмы машинного обучения (ML) могут использоваться как для классификации, так и для регрессии данных, анализа временных рядов. Эти методы также могут быть применены и для анализа данных от детекторов гравитационных волн LIGO и Virgo [3.23-3.29] . Эти и другие исследования вызвали интерес научного сообщества к новым методам на основе машинного обучения, что привело к нескольким интересным разработкам в области машинного обучения для анализа данных и моделирования источников, например, гравитационных волн [3.30–3.38].

Гамма-кванты очень высоких энергий (> 100 ГэВ) играют важную роль в общей картине многоканальной астрофизики. Важным преимуществом гамма-квантов является то, что они не отклоняются галактическими магнитными полями и, следовательно, с их помощью можно определить источник излучения. Спектр галактических гамма квантов раскрывает информацию об окружающей среде и напряженности магнитного поля как внутреннего, так и внешнего по отношению к источнику. Недавние открытия очень высокоэнергичных гамма-всплесков (Gamma Ray Bursts; GRB) предоставляют доказательства их связи с объектами, излучающими гравитационные волны. Одним из распространенных методов наземной регистрации является использование атмосферных черенковских телескопов с анализом изображений (IACT). Разделение широких атмосферных ливней, регистрируемых наземными черенковскими телескопами IACT [3.51], на вызванные гамма-излучением и космическими лучами (адронами) является одной из наиболее важных задач в технологии наземной гамма-астрономии. Также представляет большой интерес оценка таких параметров первичных частиц как энергия и направление прихода. Традиционный подход к решению этих задач основан на так называемых параметрах Хилласа [3.52], имеющих полуэмпирический характер, оценивающих длину, ширину, ориентацию и другие характеристики изображения в виде эллипса. Эмпирически подобранные диапазоны (ограничения) значений этих параметров позволяют задать дискриминатор, разделяющий события на "гамма-подобные" и "адрон-подобные". Подбор подходящих диапазонов значений параметров может производиться с использованием смоделированных алгоритмом Монте-Карло изображений телескопов [3.53,3.54].

В [3.55] был предложен дискриминатор на основе методов машинного обучения, позволяющий вместо простых диапазонов значений находить более сложные области в многомерном пространстве характеризующих изображение параметров (включая параметры Хилласа, но не обязательно ограничиваясь ими). В частности, метод случайного леса (random forest) применялся для телескопа проекта MAGIC [3.56], а после введения в строй второго телескопа был адаптирован и для стереорежима [3.57,3.58]. Метод расширяемых деревьев решений (boosted decision trees) применялся в работах [3.59] (телескопы проекта H.E.S.S., стереорежим) и [3.60] (телескопы проекта VERITAS,

стереорежим). В последние годы наилучшие результаты среди подходов для распознания изображений показывают сверточные нейронные сети (convolutional neural networks) [3.61]. В работе [3.62] этот подход предложен для классификации типов порождающих частиц в проекте CTA, где планируется установка десятков черенковских телескопов. В [3.63] для того же проекта предложены нейросети, позволяющие помимо классификации типов порождающих частиц оценивать энергию гамма-событий, направление прихода гамма-квантов и глубину первого взаимодействия. В [3.64] сверточные нейронные сети применены для поиска мюонных событий на основании изображений с одного из телескопов проекта VERITAS. В работе [3.65] сверточные рекуррентные нейросети использованы для классификации событий и оценки направления широкого атмосферного ливня для изображений четырех из пяти телескопов проекта H.E.S.S. В [3.66-3.68] они использованы для классификации событий и оценки энергии гамма-квантов для черенковского телескопа проекта TAIGA.

Следует отметить две сложности, связанные с использованием сверточных сетей. Во-первых, они до настоящего времени использовались для случая стереорежима при условии однородности изображений телескопов. Например, телескоп CT-5 проекта H.E.S.S. имеет другие характеристики по сравнению с первыми четырьмя, и его данные не использовались в анализе событий в работе [3.65]. Массив данных пилотного гибридного комплекса TAIGA-1 (1 кв.км) является мультимодальным: предполагается использование пяти (однотипных) черенковских телескопов, более сотни широкоугольных детекторов TAIGA-HiSCORE, которые кроме амплитуды сигнала регистрируют временную развертку и время прихода сигнала, и мюонных детекторов. Сверточные нейросети не допускают прямого объединения информации из таких разнородных источников. Обработка "сырых" экспериментальных данных, то есть данных, которые непосредственно приходят из экспериментального оборудования, является одним из важнейших этапов любого физического эксперимента. Разработке методов обработки таких данных на основе машинного обучения, в том числе в области астрофизики, уделяется в настоящее время большое внимание [3.70-3.72]. Важной особенностью задачи выделения гамма-событий является их чрезвычайная редкость по сравнению с фновыми адронными событиями. Даже при наблюдении Крабовидной туманности — самого яркого источника гамма-излучения, доступного для изучения черенковскими телескопами — гамма-события встречаются в тысячи раз реже адронных. Для более слабых источников разница может быть еще на 2-3 порядка выше. Задачи с несбалансированными данными достаточно широко распространены, и методам глубокого обучения для работы с такими задачами посвящено большое количество статей [3.73-3.75].

Для проведения физического анализа экспериментальных данных необходимо из сырых данных извлечь физические параметры объекта или процесса. Методы выделения и отбора признаков (параметров) используются независимо или совместно с целью улучшения методов машинного обучения (в частности, точности (accuracy)), для визуализации и интерпретации полученных знаний о данных [3.76-3.78]. В настоящее время также разворачиваются интенсивные исследования в области машинного обучения для моделирования выборок событий [3.79-3.83] как быстрая замена Монте-Карло моделирования.

4. Поиск редких событий.

Данная проблема связана с тем, что в общем потоке регистрируемых событий только 1:10000 часть — это сигнальные события от гамма частиц, которые представляют особый интерес в астрофизике.

Для выделения редких событий, которые можно рассматривать как аномалии в общем потоке, методами машинного обучения используется несколько подходов. Общий озор проблематики можно найти в работах V. Chandola, A. Banerjee, and V. Kumar [4.1] и M. Goldstein and S. Uchida [4.2]. Более современный обзор по проблеме можно найти в обзорах B. Ghojogh с соавторами [4.3] и G Papamakarios[4.4].

Одни из них является подход основанный на методе оценки плотности ядра (KDE). Идея этого метода заключается в построении приближения плотности ядра распределения входных данных путем представления ее в виде суммы базовых плотностей. В качестве таких базовых плотностей обычно рассматриваются константные функции на отрезке, треугольные, функция Епаничнекова. Но наиболее популярным ядром является гауссиан. Подробно с математической теорией KDE можно ознакомиться по монографии M.P. Wand,M.C. Jones [4.5]. К сожалению, прямое применение данного подхода оказывается не очень эффективным для случая больших размерностей пространства скрытых параметров.

Развитие этого метода явился метод нормализирующих потоков [4.9]. Основная идея данного подхода — это

построение цепочки взаимно однозначных преобразований плотности, которая переводит простое начальное распределение (обычно это нормальное распределение) в распределение переменных в скрытом пространстве [4.10]. Если преобразования обусловлены наблюдениями, нормализующие потоки можно обучить возвращать байесовские апостериорные оценки вероятности для любого наблюдения.

Вторым направлением, используемым для поиска аномалий, - это применение состязательных автоэнкодеров. Одним из первых, кто предложил данный тип генеративной сети, был Alireza Makhzani с соавторами [4.5]. Проблема обычных (вариационных) автоэнкодеров для целей выделения аномалий состоит в том, что продолжительное обучение автоэнкодеров неизбежно снижает ошибку реконструкции выбросов, что является характерным признаком аномалии, и, следовательно, ухудшает качество обнаружения аномалий.

Основная идея предложенного состязательного автоэнкодера состоит в добавлении дискриминатора по аналогии с GAN сетями. Задача дискриминатора состоит в том, чтобы наложить априорное распределение на скрытое представление. Это позволяет разместить аномалии в области с низкой вероятностью и, тем самым, выделить подобные события [4.7]. С учетом особенностей астрофизических данных, такой подход обещает заметно повысить качество выделения сигнальных гамма событий на общем фоне.

Заметим, что данный метод обобщается и на вариационные автоэнкодеры, см. например работу [4.8]. Как правило, состязательные автоэнкодеры и нормализирующие потоки используются совместно [4.12].

Работы связанные с состязательными автоэнкодерами и нормализационными потоками стали появляется в последнее время и в области астрофизики. Например, в работе [4.11] авторы исследуют аномальные особенности галактических спектров, а в работе [4.14] используется метод нормализирующих потоков для моделирования гало Галактики.
В работе [4.13] похожий подход используется для анализа солнечной и звездных атмосфер.

В настоящее время результатов по применению этих передовых методов для задач гамма астрономии авторам не известно. Таким образом, прямые конкуренты отсутствуют. Однако, так как данное направление стало активно развиваться, то мы ожидаем появления аналогичных исследований, например, в коллаборации H.E.S.S [4.15] и CTA [4.16]. В предлагаемом проекте будут разработаны ряд новых методов, а также оптимизированы имеющиеся подходы. Будет разработан метод совместного анализа экспериментальных данных, полученных от установок разного типа, в частности для черенковских телескопов TAIGA-IACT и детекторов с фиксацией времени прихода сигнала TAIGA-HiSCORE эксперимента TAIGA. Кроме того, будут исследованы различные варианты выделения существенных признаков получаемых экспериментальных данных и их интерпретации в терминах данной предметной области, а именно в терминах параметров широких атмосферных ливней. Все это вместе позволит решить задачу поиска аномальных событий в потоке космических частиц, что, в свою очередь, будет хорошей основой для физического анализа данных эксперимента TAIGA.

Литература

1.1. N. Chumerin and V. Hulle, M. M, “Comparison of Two Feature Extraction Methods Based on Maximization of Mutual Information” In: Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, pp. 343–348, 2006.
1.2. H. Motoda and H. Liu, “Feature selection, extraction and construction” In: Towards the Foundation of Data Mining Workshop, Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2002), Taipei, Taiwan, pp. 67–72, 2002.
1.3. L. Ladla and T. Deepa, “ Feature Selection Methods And Algorithms”, International Journal on Computer Science and Engineering (IJCSE), vol.3(5), pp. 1787-1797, 2011.
1.4. A. G. K. Janecek and G. F. Gansterer et al, “On the Relationship between Feature Selection and Classification Accuracy”, In: Proceeding of New Challenges for Feature Selection, pp. 40-105, 2008.
1.5. Veerabhadrappa, L. Rangarajan, “Bi-level Dimensionality Reduction Methods Using Feature Selection and Feature Extraction”, International Journal of Computer Applications, vol. 4(2), pp. 33-38, 2010.
1.6. L. Yu, and H. Liu, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution”, In: Proceedings of the 20th international conference on machine learning (ICML-03) 2003 (pp. 856-863).
1.7. Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics. New York: Springer-Verlag.
1.8. S. Cateni, et al, “Variable Selection and Feature Extraction through Artificial Intelligence Techniques”, Multivariate

Analysis in Management, Engineering and the Science, chapter 6, pp.103-118, 2012.
1.9. Veerabhadrappa, L. Rangarajan, “Bi-level Dimensionality Reduction Methods Using Feature Selection and Feature Extraction”, International Journal of Computer Applications, vol. 4(2), pp. 33-38, 2010.
1.10. Veerabhadrappa, L. Rangarajan, “Multilevel Dimensionality Reduction Methods Using Feature Selection and Feature Extraction”, International Journal of Artificial Intelligence and Applications, vol. 1(4), pp. 54-58, 2010.
1.11. D. Rumerhart, G. Hinton, and R. Williams, “Learning representations by back-propagation errors,” Nature, vol. 323, pp. 533–536, 1986.
1.12. P. Baldi and K. Hornik, “Neural networks and principal component analysis: Learning from examples without local minima,” Neural networks, vol. 2, no. 1, pp. 53–58, 1989.
1.13. N. Japkowicz, S. J. Hanson, and M. A. Gluck, “Non-linear autoassociation is not equivalent to pca,” Neural computation, vol. 12, no. 3, pp. 531–545, 2000.
1.14. H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, “An empirical evaluation of deep architectures on problems with many factors of variation,” in Proceedings of the 24th international conference on Machine learning. ACM, 2007, pp. 473–480.
1.15. K. Jarrett, K. Kavukcuoglu, Y. Lecun et al., “What is the best multi-stage architecture for object recognition?” in 2009 IEEE 12th International Conference on Computer Vision. IEEE, 2009, pp. 2146–2153.
1.16. L. Vincent, J. Rello, J. Marshall, E. Silva, A. Anzueto, C. D. Martin, R. Moreno, J. Lipman, C. Gomersall, Y. Sakr et al., “International study of the prevalence and outcomes of infection in intensive care units,” Jama, vol. 302, no. 21, pp. 2323–2329, 2009.
1.17. W. W. Ng, G. Zeng, J. Zhang, D. S. Yeung, and W. Pedrycz, “Dual autoencoders features for imbalance classification problem,” Pattern Recognition, vol. 60, pp. 875–889, 2016.
1.18. H.-C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, “Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1930–1943, 2013.
1.19. D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, “Why does unsupervised pre-training help deep learning?” Journal of Machine Learning Research, vol. 11, no. Feb, pp. 625–660, 2010.
1.20. C. Poultney, S. Chopra, Y. L. Cun et al., “Efficient learning of sparse representations with an energy-based model,” in Advances in neural information processing systems, 2006, pp. 1137–1144.
1.21. I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng, “Measuring invariances in deep networks,” in Advances in neural information processing systems, 2009, pp. 646–654.
1.22. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” The Journal of Machine Learning Research, vol. 11, pp. 3371–3408, 2010.
1.23. Kingma, Diederik P.; Welling, Max "Auto-Encoding Variational Bayes". ArXiv:1312.6114, 2013
1.24. S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, “Contractive auto-encoders: Explicit invariance during feature extraction,” in Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 833–840.
1.25. W. Wang, Y. Huang, Y. Wang, and L. Wang, “Generalized autoencoder: a neural network framework for dimensionality reduction,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 490–497.
1.26. Z. Camlica, H. Tizhoosh, and F. Khalvati, “Autoencoding the retrieval relevance of medical images,” in Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on. IEEE, 2015, pp. 550–555.
1.27. S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, “Deep feature learning for eeg recordings,” arXiv preprint arXiv:1511.04306, 2015.
1.28. L. Gao, J. Song, X. Liu, J. Shao, J. Liu, and J. Shao, “Learning in high-dimensional multimedia data: the state of the art,” Multimedia Systems, pp. 1–11, 2015.
1.29. Y. Wang, H. Yao, S. Zhao, and Y. Zheng, “Dimensionality reduction strategy based on auto-encoder,” in Proceedings of the 7th International Conference on Internet Multimedia Computing and Service. ACM, 2015, p. 63.
1.30. L. Meng, S. Ding, and Y. Xue, “Research on denoising sparse autoencoder,” International Journal of Machine Learning and Cybernetics, pp. 1–11, 2016.
1.31. Bonet-Sola D, Alsina-Pages RM., "A comparative survey of feature extraction and machine learning methods in diverse acoustic environments". Sensors. 2021 Jan;21(4):1274.
1.32. Yao R, Liu C, Zhang L, Peng P. Unsupervised anomaly detection using variational auto-encoder based feature extraction. In2019 IEEE International Conference on Prognostics and Health Management (ICPHM) 2019 Jun 17 (pp. 1-7). IEEE.
1.33. Zavrak S, Iskefiyeli M. Anomaly-based intrusion detection from network flow features using variational autoencoder. IEEE Access. 2020 Jun 10;8:108346-58.

1.34. Perret-Gallix, Denis. "Computational particle physics for event generators and data analysis." Journal of Physics: Conference Series. Vol. 454. No. 1. IOP Publishing, 2013.
1.35. Papaefstathiou, Andreas. "How-to: write a parton-level Monte Carlo particle physics event generator." The European Physical Journal Plus 135.6 (2020): 1-19.
1.36. van Leeuwen, Caspar, et al. "Deep-learning enhancement of large scale numerical simulations." SURF Whitepaper, arXiv preprint arXiv:2004.03454 (2020).
1.37. Otten, Sydney, et al. "Event generation and statistical sampling for physics with deep generative models and a density information buffer." arXiv preprint arXiv:1901.00875 (2019).
1.38. Butter, Anja, Tilman Plehn, and Ramon Winterhalder. "How to GAN LHC events." SciPost Phys. 7, 075 (2019), https://arxiv.org/abs/1907.03764
1.39. Alanazi, Yasir, et al. "Simulation of electron-proton scattering events by a Feature-Augmented and Transformed Generative Adversarial Network (FAT-GAN)." arXiv preprint arXiv:2001.11103 (2020).
1.40. de Oliveira, Luke, Michela Paganini, and Benjamin Nachman. "Controlling physical attributes in GAN-accelerated simulation of electromagnetic calorimeters." 18th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2017) Seattle, WA, USA. 2017. https://arxiv.org/pdf/1711.08813
1.41. de Oliveira, Luke, Michela Paganini, and Benjamin Nachman. "Learning particle physics by example: location-aware generative adversarial networks for physics synthesis." Computing and Software for Big Science 1.1 (2017): 4. https://arxiv.org/pdf/1701.05927
1.42. Derkach, D., et al. "Cherenkov Detectors Fast Simulation Using Neural Networks", Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 952 (2020): 161804; arXiv preprint arXiv:1903.11788.
1.43. Farrell, Steven, et al. "Next Generation Generative Neural Networks for HEP." EPJ Web of Conferences. Vol. 214. EDP Sciences, 2019.
1.44. Musella, Pasquale, and Francesco Pandolfi. "Fast and accurate simulation of particle detectors using generative adversarial networks." Computing and Software for Big Science 2.1 (2018): 8.
1.45. Di Sipio, Riccardo, et al. "DijetGAN: a Generative-Adversarial Network approach for the simulation of QCD dijet events at the LHC." Journal of High Energy Physics 2019.8 (2019): 110.
1.46. Hashemi, Bobak, et al. "LHC analysis-specific datasets with Generative Adversarial Networks." arXiv preprint arXiv:1901.05282 (2019).
1.47. Paganini, Michela, Luke de Oliveira, and Benjamin Nachman. "Accelerating science with generative adversarial networks: an application to 3D particle showers in multilayer calorimeters." Physical review letters 120.4 (2018): 042003.
1.48. Kai Zhou, Gergely Endrodi, Long-Gang Pang, and Horst Stöcker, "Regressive and generative neural networks for scalar field theory", Phys. Rev. D100, 011501(R) (2019)
1.49. Yang Zeng, Jin-Long Wu , Heng Xiao, "Enforcing Deterministic Constraints on Generative Adversarial Networks for Emulating", Physical Systems, arXiv 1911.06671
1.50. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
1.51. Jin-Long Wu et al., "Enforcing Statistical Constraints in Generative Adversarial Networks for Modeling Chaotic Dynamical Systems", Journal of Computational Physics 406 (2020): 109209, arXiv 1905.06841
1.52. M. Mirza, S. Osindero, "Conditional generative adversarial nets", arXiv:1411.1784.
1.53. Eric Heim, "Constrained Generative Adversarial Networks for Interactive Image Generation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019, arXiv 1904.02526
1.54. Zhiting Hu et al., "Deep Generative Models with Learnable Knowledge Constraints", 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.
1.55. Giorgio Gnecco ET AL., "Learning with hard constraints as a limit case of learning with soft constraints", ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016
1.56. Pablo Márquez-Neila Mathieu Salzmann Pascal Fua, "Imposing Hard Constraints on Deep Networks: Promises and Limitations", CVPR Workshop on Negative Results in Computer Vision, Hawaii, HI, 2017, arXiv 1706.02025.
1.57. von Rueden, Laura, et al. "Informed Machine Learning--A Taxonomy and Survey of Integrating Knowledge into Learning Systems." arXiv preprint arXiv:1903.12394 (2019)
1.58. Peng, H. C.; Long, F.; Ding, C. (2005). "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy". IEEE Transactions on Pattern Analysis and Machine Intelligence. 27 (8): 1226–1238.
1.59. Urbanowicz, Ryan J.; Meeker, Melissa; LaCava, William; Olson, Randal S.; Moore, Jason H. (2018). "Relief-Based Feature Selection: Introduction and Review". Journal of Biomedical Informatics. 85: 189–203.

1.60. Fleuret F. Fast binary feature selection with conditional mutual information. Journal of Machine learning research. 2004 Nov 1;5(9).
1.61. Hsu HH, Hsieh CW. Feature Selection via Correlation Coefficient Clustering. J. Softw.. 2010 Dec;5(12):1371-7.
1.62. S. Dudoit, J. Fridlyand and T. P. Speed, “Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data”, Journal of the American Statistical Association, 2002,Vol. 97, No. 457, pp. 77-87
1.63. Z. Zhao and H. Liu, “Searching for Interacting Features”, Proceedings of International Joint Conference on Artificial Intelligence, 2007, pp. 1156-1161
1.64. M.J Martin-Bautista and M-A Vila, “A Survey of Genetic Feature Selection in Mining Issues”, Proceedings of the Congress on Evolutionary Computation, 1999, Vol. 2, pp. 1314-1321
1.65. I. Guyon, J. Weston, S. Barnhill and V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines”, Machine Learning, 2002,Vol. 46, No. 1-3, pp. 389-422
1.66. Hastie, T.; Stuetzle, W. (June 1989). "Principal Curves" (PDF). Journal of the American Statistical Association. 84 (406): 502–506.
1.67. Lee, T.-W. (1998): Independent component analysis: Theory and applications, Boston, Mass: Kluwer Academic Publishers,
1.68. Stein G, Harrington P, Blaum J, Medan T, Lukic Z. Self-supervised similarity search for large scientific datasets. ArXiv preprint arXiv:2110.13151. 2021

2.1. Jirkovský, V., & Obitko, M. (2014). Semantic Heterogeneity Reduction for Big Data in Industrial Automation. ITAT, 1214.
2.2. Data, O. B. (2015). Transport Understanding and assessing options Corporate Partnership Board Report Corporate Partnership Board CPB. Technical report, International Transport Forum. http://www. internationaltransportforum.org 2.3. Viña A. Data Virtualization Goes Mainstream, White Paper, Denodo Technologies, Inc, USA, 2015, 1-18.
2.4. P. Pavlidis, J. Weston, J. Cai, W.S. Noble , Learning gene functional classifications from multiple data types., J. Comput. Biol. 9 (2) (2002) 401–411.
2.5. P. Maragos, P. Gros, A. Katsamanis, G. Papandreou, Cross-modal integration for performance improving in multimedia: a review, in: Proceedings of the IEEE International Conference on Image Processing, 2008, pp. 3412–3416.
2.6. Y. Zheng, Methodologies for cross-domain data fusion: an overview, IEEE Trans. Big Data 1 (1) (2015) 16–34.
2.7. Y. Fu, Y. Ge, Y. Zheng, Z. Yao, Y. Liu, H. Xiong, J. Yuan , Sparse real estate ranking with online user reviews and offline moving behaviors, in: Proceedings of the 2014 IEEE International Conference on Data Mining, IEEE, 2014, pp. 120–129.
2.8. Z. Wang, D. Zhang, X. Zhou, D. Yang, Z. Yu, Z. Yu, Discovering and profiling overlapping communities in location-based social networks., IEEE Trans. Syst. Man Cybern. 44 (4) (2014) 499–509.
2.9. S. Du, T. Li, X. Gong, Z. Yu, S.-J. Horng, A hybrid method for traffic flow forecasting using multimodal deep learning, arXiv preprint arXiv:1803.02099 (2018).
2.10. M. Pratama, J. Lu, S. Anavatti, E. Lughofer, C.-P. Lim, An incremental meta-cognitive-based scaffolding fuzzy neural network, Neurocomputing 171 (2016) 89–105.
2.11. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 689–696
2.12. L. Zhu, F. Guo, J.W. Polak, R. Krishnan, Urban link travel time estimation using traffic states-based data fusion, IET Intell. Transp. Syst. 12 (7) (2018) 651–663.
2.13. Y. Zheng, Y. Liu, J. Yuan, X. Xie, Urban computing with taxicabs, in: Proceedings of the 13th international conference on Ubiquitous computing, ACM, 2011, pp. 89–98.
2.14. B. Pan, Y. Zheng, D. Wilkie, C. Shahabi, Crowd sensing of traffic anomalies based on human mobility and social media, in: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2013, pp. 344–353.
2.15. Y. Zheng, X. Yi, M. Li, R. Li , Z. Shan, E. Chang, T. Li, Forecasting fine-grained air quality based on big data, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 2267–2276.
2.16. A. Kumar, H. Daumé, A co-training approach for multi-view spectral clustering, in: Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 393–400.
2.17. J. Shang, Y. Zheng, W. Tong, E. Chang, Y. Yu, Inferring gas consumption and pollution emission of vehicles throughout a city, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 1027–1036.
2.18. O. Katz, R. Talmon, Y.-L. Lo, H.-T. Wu, Alternating diffusion maps for multimodal data fusion, Inf. Fusion 45 (2018) 346–360.
2.19. N.J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, H. Xiong, Discovering urban functional zones using latent activity trajectories, IEEE Trans. Knowl. Data Eng. 27 (3) (2015) 712–725.

2.20. Y. Zheng, H. Zhang , Y. Yu, Detecting collective anomalies from multiple spatio-temporal datasets across different domains, in: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2015, p. 2.
2.21. W. Dai, Y. Chen, G.-R. Xue, Q. Yang, Y. Yu, Translated learning: Transfer learning across different feature spaces, in: Advances in Neural Information Processing Systems, 2009, pp. 353–360.
2.22. P. Yang, W. Gao , Multi-view discriminant transfer learning., in: International Joint Conference on Artificial Intelligence, 2013, pp. 1848–1854.
2.23. Y. Zheng, X. Chen, Q. Jin, Y. Chen, X. Qu, X. Liu, E. Chang, W.-Y. Ma, Y. Rui, W. Sun, A cloud-based knowledge discovery system for monitoring fine-grained air quality, preparation, Microsoft Tech Report, 2014. http://research.microsoft.com/apps/pubs/default.aspx.
2.24. A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the 11th Annual Conference on Computational Learning Theory, ACM, 1998, pp. 92–100.
2.25. M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, J. Mach. Learn. Res. 12 (7) (2011) 2211–2268.
2.26. Z. Li, J. Liu, J. Tang, H. Lu, Robust structured subspace learning for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2085–2098.
2.27. N. Chen, J. Zhu, E.P. Xing, Predictive subspace learning for multi-view data: a large margin approach, in: Proceedings of the Advances in Neural Information Processing Systems, 2010, pp. 361–369.
2.28. V.W. Zheng, Y. Zheng, X. Xie, Q. Yang, Collaborative location and activity recommendations with gps history data, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 1029–1038.
2.29. Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, E. Chang, Diagnosing new york city's noises with ubiquitous data, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2014, pp. 715–725.
2.30. Blasch, E., Pham, T., Chong, C. Y., Koch, W., Leung, H., Braines, D., & Abdelzaher, T. (2021). Machine learning/artificial intelligence for sensor data fusion–opportunities and challenges. IEEE Aerospace and Electronic Systems Magazine, 36(7), 80-93.
2.31. Chen, G., Liu, Z., Yu, G., & Liang, J. (2021). A New View of Multisensor Data Fusion: Research on Generalized Fusion. Mathematical Problems in Engineering, 2021.
2.32. Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. Neural Computation, 32(5), 829-864.
2.33. Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review. arXiv preprint arXiv:2205.01380.
2.34. Breckels, L. M., Holden, S. B., Wojnar, D., Mulvey, C. M., Christoforou, A., Groen, A., ... & Gatto, L. (2016). Learning from heterogeneous data sources: an application in spatial proteomics. PLoS computational biology, 12(5), e1004920.
2.35. Cvetek, D., Muštra, M., Jelušić, N., & Tišljarić, L. (2021). A survey of methods and technologies for congestion estimation based on multisource data fusion. Applied Sciences, 11(5), 2306.
2.36. Gilanifar, M. (2019). Heterogeneous data fusion for performance improvement in electric power systems (Doctoral dissertation, The Florida State University).
2.37. Hssayeni, M. D., & Ghoraani, B. (2021). Multi-modal physiological data fusion for affect estimation using deep learning. IEEE Access, 9, 21642-21652.
2.38. Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2021). A novel multi-source information-fusion predictive framework based on deep neural networks for accuracy enhancement in stock market prediction. Journal of Big Data, 8(1), 1-28.
2.39. Zhou, J., Hong, X., & Jin, P. (2019). Information fusion for multi-source material data: Progress and challenges. Applied Sciences, 9(17), 3473.
2.40. Fuzhen Zhuang et al., "A Comprehensive Survey on Transfer Learning", 2020, arXiv: 1911.02685
3.1. Dorner, D., Mostafa, M. and Satalecka, K., 2021. High-Energy Alerts in the Multi-Messenger Era. Universe, 7(11), p.393. Meszaros P, Fox D, Hanna C, Murase K. Multi-messenger astrophysics. Nature Reviews Physics. 2019, 1(10):585-99.
3.2. Nakar E 2007 Physics Reports 442 166
3.3. Berger E 2014 Annual Rview of Astronomy and Astrophysics 52 43
3.4. Tanaka M et al 2014 ApJ 780 9
3.5. Barnes J and Kasen D 2013 ApJ 775 9
3.6. Piran T Nakar E and Rosswog S 2013 MNRAS 430 2121
3.7. Hotokezaka K and Piran T 2015 MNRAS 450 1430
3.8. Hjorth J and Bloom J 2012 Cambridge Astrophysics Series 51 169
3.9. Mereghetti, S., 2008. The strongest cosmic magnets: soft gamma-ray repeaters and anomalous X-ray pulsars. The Astronomy and Astrophysics Review, 15(4), pp.225-287.
3.10. Ando S et al 2013 Reviews of Modern Physics 85 1401

3.11. Abbasi R et al 2009 Nuclear Instruments and Methods in Physics Research A 601 294
3.12. Ageron M et al 2011 Nuclear Instruments and Methods in Physics Research A 656 11
3.13. IceCube Collaboration 2013 Science 342
3.14. Aartsen M 2013 Physical Review Letters 111 021103
3.15. Aartsen M et al 2016, Physical Review D 93 022001
3.16. Hirata K et al 1988 Phys. Rev. D 38 448
3.17. Bionta R et al 1987 Phys. Rev. Lett. 58 1494
3.18. Alekseev E et al 1987 J. Exp. Theor. Phys. Lett. 45 589
3.19. Waxman E 1995 PhRvL 75 386
3.20. Vietri M 1995 ApJ 453 883
3.21. Adrián- Martínez S et al 2013, ApJ 774 19
3.22. Allen, G., Andreoni, I., Bachelet, E., Berriman, G.B., Bianco, F.B., Biswas, R., Kind, M.C., Chard, K., Cho, M., Cowperthwaite, P.S. and Etienne, Z.B., 2019. Deep learning for multi-messenger astrophysics: a gateway for discovery in the big data era. arXiv preprint arXiv:1902.00522.
3.23. D. George and E. A. Huerta, Phys. Rev. D 97, 044039 (2018), arXiv:1701.00008.
3.24. D. George and E. A. Huerta, Physics Letters B 778, 64 (2018), arXiv:1711.03121 [gr-qc].
3.25. H. Shen, D. George, E. A. Huerta, and Z. Zhao, ArXiv e-prints (2017), arXiv:1711.09919 [gr-qc].
3.26. W. Wei and E. A. Huerta, arXiv e-prints , arXiv:1901.00869 (2019), arXiv:1901.00869 [gr-qc].
3.27. A. Rebei, E. A. Huerta, S. Wang, S. Habib, R. Haas, D. Johnson, and D. George, arXiv e-prints , arXiv:1807.09787
3.28. (2018), arXiv:1807.09787 [gr-qc].
3.29. D. George, H. Shen, and E. A. Huerta, Phys. Rev. D 97, 101501 (2018).
3.30. A. J. K. Chua, C. R. Galley, and M. Vallisneri, ArXiv e-prints (2018), arXiv:1811.05491 .
3.31. H. Gabbard, M. Williams, F. Hayes, and C. Messenger, Physical Review Letters 120, 141103 (2018), arXiv:1712.06041.
3.32. X. Fan, J. Li, X. Li, Y. Zhong, and J. Cao, ArXiv e-prints (2018), arXiv:1811.01380
3.33. J. A. González and F. S. Guzmán, Phys. Rev. D 97, 063001 (2018), arXiv:1803.06060 .
3.34. Y. Fujimoto, K. Fukushima, and K. Murase, Phys. Rev. D 98, 023019 (2018), arXiv:1711.06748.
3.35. X. Li, W. Yu, and X. Fan, ArXiv e-prints (2017), arXiv:1712.00356.
3.36. H. Nakano et al., ArXiv e-prints (2018), arXiv:1811.06443.
3.37. T. Springenberg, A. Klein, S. Falkner, and F. Hutter, in Advances in Neural Information Processing Systems 29 ,edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016) pp. 4134–4142.
3.38. Y. Gal and Z. Ghahramani, In international conference on machine learning (2016) pp. 1050–1059.
3.39. Huerta, E.A., Allen, G., Andreoni, I., Antelis, J.M., Bachelet, E., Berriman, G.B., Bianco, F.B., Biswas, R., Kind, M.C., Chard, K. and Cho, M., 2019. Enabling real-time multi-messenger astrophysics discoveries with deep learning. Nature Reviews Physics,1(10), pp.600-608.
3.40. Aurisano, A. et al. A Convolutional Neural Network Neutrino Event Classifier. JINST 11, P09001 (2016).
3.41. Choma, N. et al. Graph neural networks for icecube signal classification. In Wani, M., Sayed-Mouchaweh, M., Lughofer, E., Gama, J. & Kantardzic, M. (eds.) Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, 386–391 42. Brice, S.J., 1996. The Results of Neural Network Statistical Event Class Analysis.
3.43. Psihas, F., Groh, M., Tunnell, C. and Warburton, K., 2020. A review on machine learning for neutrino experiments. International Journal of Modern Physics A, 35(33), p.2043005.
3.44. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
3.45. O. Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015.
3.46. J. Renner et al. Background rejection in NEXT using deep neural networks. Journal of Instrumentation, 12(01):T01004–T01004, jan 2017.
3.47. A. Aurisano et al. A convolutional neural network neutrino event classifier. Journal of Instrumentation, 11(09):P09001–P09001, Sep 2016.
3.48. A. Gando et al. Search for Majorana Neutrinos near the Inverted Mass Hierarchy Region with KamLAND-Zen. Phys. Rev. Lett., 117(8):082503, 2016.
3.49. Abi, B., Acciarri, R., Acero, M.A., Adamov, G., Adams, D., Adinolfi, M., Ahmad, Z., Ahmed, J., Alion, T., Monsalve, S.A. and Alt, C., 2020. Neutrino interaction classification with a convolutional neural network in the DUNE far detector. Physical Review D, 102(9), p.092003.
3.50. Zhenghao Fu. Detection of cosmic muon spallation background in ls-detector using machine learning. In: 29 the

International Conference on Neutrino Physics and Astrophysics, 2020. https://zenodo.org/record/4122953/files/ZhenghaoFuNeutrino2020_poster_Fu_v4.pdf?download=1 51. Weekes, T. C., et al. "Observation of TeV gamma-rays from the Crab nebula using the atmospheric Cherenkov imaging technique." Astrophysical Journal 342 (1989): 379-395.
3.52. Hillas, A.M.: Cerenkov light images of EAS produced by primary gamma rays and by nuclei. In: Proc. 19th Int. Cosmic Ray Conf., La Jolla, 1985, p. 445. NASA, Washington, D.C. (1985).
3.53. Reynolds, P. T., et al. "Survey of candidate gamma-ray sources at TeV energies using a high-resolution Cerenkov imaging system-1988-1991." The Astrophysical Journal 404 (1993): 206-218.
3.54. Mohanty, G., et al. "Measurement of TeV gamma-ray spectra with the Cherenkov imaging technique." Astroparticle Physics 9.1 (1998): 15-43.
3.55. Bock, R. K., et al. "Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope." Nuclear Instruments and Methods in Physics Research Section A. 516.2-3 (2004): 511-528.
3.56. J. Albert et al., Implementation of the Random Forest method for the Imaging Atmospheric Cherenkov Telescope MAGIC, Nuclear Instruments and Methods in Physics Research A 588 (Apr., 2008) 424–432, POL[arXiv:0709.3719].
3.57. Colin, P., et al. "Performance of the MAGIC telescopes in stereoscopic mode." arXiv preprint arXiv:0907.0960 (2009).
3.58. Sharma, M., et al. "Gamma/hadron segregation for a ground based imaging atmospheric Cherenkov telescope using machine learning methods: Random Forest leads." Research in Astronomy and Astrophysics 14.11 (2014): 1491.
3.59. S. Ohm, C. van Eldik, and K. Egberts, γ/hadron separation in very-high-energy γ-ray astronomy using a multivariate analysis method, Astroparticle Physics 31 (June, 2009) 383–391, POL[arXiv:0904.1136].
3.60. M. Krause, E. Pueschel, and G. Maier, Improved γ/hadron separation for the detection of faint γ-ray sources using boosted decision trees, Astroparticle Physics 89 (2017) 1–9.
3.61. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.
3.62. Nieto, D. et al. for the CTA Consortium: Exploring deep learning as an event classification method for the Cherenkov Telescope Array. Proceedings of Science 301, PoS(ICRC2017)809 (2017)
3.63. Mangano, S., Delgado, C., Bernardos, M. I., Lallena, M., Vázquez, J. J. R., & CTA Consortium. (2018, September). Extracting gamma-ray information from images with convolutional neural network methods on simulated Cherenkov Telescope Array data. In IAPR Workshop on Artificial Neural Networks in Pattern Recognition (pp. 243-254). Springer, Cham.
3.64. Feng, Q., Jarvis, J. on behalf of the VERITAS Collaboration: A citizen-science approach to muon events in imaging atmospheric Cherenkov telescope data: the Muon Hunter. Proceedings of Science 301, PoS(ICRC2017)826 (2017)
3.65. Shilon, I., et al. "Application of deep learning methods to analysis of imaging atmospheric Cherenkov telescopes data." Astroparticle Physics 105 (2019): 44-53.
3.66. Deep learning for energy estimation and particle identification in gamma-ray astronomy / E. Postnikov, A. Kryukov, S. Polyakov, D. Zhurov // CEUR Workshop Proceedings. — 2019. — Vol. 2406. — P. 90–99. (http://ceur-ws.org/Vol-2406/paper11.pdf).
3.67. Gamma/hadron separation in imaging air Cherenkov telescopes using deep learning libraries TensorFlow and PyTorch /E. B. Postnikov, A. P. Kryukov, S. P. Polyakov et al. // Journal of Physics: Conference Series. — 2019. — Vol. 1181. — P.012048. DOI: 10.1088/1742-6596/1181/1/012048 (https://iopscience.iop.org/article/10.1088/1742- 596/1181/1/012048).
3.68. Nieto Castaño, D., et al. "Studying Deep Convolutional Neural Networks With Hexagonal Lattices for Imaging Atmospheric Cherenkov Telescope Event Reconstruction." 36th International Cosmic Ray Conference (ICRC2019). Vol. 36. 2019.
3.69. E. Hoogeboom, J. W. T. Peters, T. S. Cohen, and M. Welling, “HexaConv”, (2018), arXiv:1803.02108.
3.70. Meng, T., Jing, X., Yan, Z. and Pedrycz, W., 2020. A survey on machine learning for data fusion. Information Fusion, 57, pp.115-129.
3.71. Travassos, Xisto L., Sérgio L. Avila, and Nathan Ida. "Artificial neural networks and machine learning techniques applied to ground penetrating radar: A review." Applied Computing and Informatics (2020).
3.72. Garofalo, Mauro, Alessio Botta, and Giorgio Ventre. "Astrophysics and big data: Challenges, methods, and tools." Proceedings of the International Astronomical Union 12, no. S325 (2016): 345-348.
3.73. Johnson, Justin M., and Taghi M. Khoshgoftaar. "Survey on deep learning with class imbalance." Journal of Big Data 6.1 (2019): 1-54.
3.74. Fajardo, Val Andrei, David Findlay, Charu Jaiswal, Xinshang Yin, Roshanak Houmanfar, Honglei Xie, Jiaxi Liang, Xichen She, and D. B. Emerson. "On oversampling imbalanced data with deep conditional generative models." Expert Systems with Applications 169 (2021): 114463.
3.75. Spelmen, Vimalraj S., and R. Porkodi. "A review on handling imbalanced data." In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1-11. IEEE, 2018.
3.76. H. Motoda and H. Liu, “Feature selection, extraction and construction” In: Towards the Foundation of Data Mining

Workshop, Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2002), Taipei, Taiwan, pp. 67–72, 2002.
3.77. L. Ladla and T. Deepa, " Feature Selection Methods And Algorithms", International Journal on Computer Science and Engineering (IJCSE), vol.3(5), pp. 1787-1797, 2011.
3.78. A. G. K. Janecek and G. F. Gansterer et al, "On the Relationship between Feature Selection and Classification Accuracy", In: Proceeding of New Challenges for Feature Selection, pp. 40-105, 2008.
3.79. van Leeuwen, Caspar, et al. "Deep-learning enhancement of large scale numerical simulations." SURF Whitepaper, arXiv preprint arXiv:2004.03454 (2020).
3.80. Otten, Sydney, et al. "Event generation and statistical sampling for physics with deep generative models and a density information buffer." arXiv preprint arXiv:1901.00875 (2019).
3.81. Alanazi, Yasir, et al. "Simulation of electron-proton scattering events by a Feature-Augmented and Transformed Generative Adversarial Network (FAT-GAN)." arXiv preprint arXiv:2001.11103 (2020).
3.82. de Oliveira, Luke, Michela Paganini, and Benjamin Nachman. "Controlling physical attributes in GAN-accelerated simulation of electromagnetic calorimeters." 18th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2017) Seattle, WA, USA. 2017. https://arxiv.org/pdf/1711.08813
3.83. de Oliveira, Luke, Michela Paganini, and Benjamin Nachman. "Learning particle physics by example: location-aware generative adversarial networks for physics synthesis." Computing and Software for Big Science 1.1 (2017): 4. https://arxiv.org/pdf/1701.05927
3.77. L. Ladla and T. Deepa, " Feature Selection Methods And Algorithms", International Journal on Computer Science and Engineering (IJCSE), vol.3(5), pp. 1787-1797, 2011.
3.78. A. G. K. Janecek and G. F. Gansterer et al, "On the Relationship between Feature Selection and Classification Accuracy", In: Proceeding of New Challenges for Feature Selection, pp. 40-105, 2008.
3.79. van Leeuwen, Caspar, et al. "Deep-learning enhancement of large scale numerical simulations." SURF Whitepaper, arXiv preprint arXiv:2004.03454 (2020).
3.80. Otten, Sydney, et al. "Event generation and statistical sampling for physics with deep generative models and a density information buffer." arXiv preprint arXiv:1901.00875 (2019).
3.81. Alanazi, Yasir, et al. "Simulation of electron-proton scattering events by a Feature-Augmented and Transformed Generative Adversarial Network (FAT-GAN)." arXiv preprint arXiv:2001.11103 (2020).
3.82. de Oliveira, Luke, Michela Paganini, and Benjamin Nachman. "Controlling physical attributes in GAN-accelerated simulation of electromagnetic calorimeters." 18th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2017) Seattle, WA, USA. 2017. https://arxiv.org/pdf/1711.08813
3.83. de Oliveira, Luke, Michela Paganini, and Benjamin Nachman. "Learning particle physics by example: location-aware generative adversarial networks for physics synthesis." Computing and Software for Big Science 1.1 (2017): 4. https://arxiv.org/pdf/1701.05927

4. Поиск редких событий.

4.1. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey,"
ACM Comput. Surv., Bd. 41, 2009.
4.2. M. Goldstein and S. Uchida, "A Comparative Evaluation of Unsupervised
Anomaly Detection Algorithms for Multivariate Data," PLOS ONE, 2016.
4.3. B. Ghojogh et al. «Generative Adversarial Networks and Adversarial Autoencoders:
Tutorial and Survey», arXiv:2111.13282v1
4.4. GUANSONG PANG, CHUNHUA SHEN, LONGBING CAO, A. VAN DEN HENGEL // Deep Learning for Anomaly Detection: A Review. arXiv:1912.02762v2
4.5. M.P. Wand,M.C. Jones // Kernel Smoothing. CRC Press, 230p, 1994. ISBN : 0412 552701
4.6. Alireza Makhzani et al. // Adversarial Autoencoders. arXiv:1511.05644v2
4.7. Laura Beggel, Michael Pfeiffer, and Bernd Bischl // Robust Anomaly Detection in Images using Adversarial Autoencoders. arXiv:1901.06355v1
4.8. Seung Yeop Shin and Han-joon Kim // Extended Autoencoder for Novelty Detection with
Reconstruction along Projection Pathway. Appl. Sci. 2020, 10, 4497; doi:10.3390/app10134497
4.9. Danilo Jimenez Rezende, Shakir Mohamed // Variational Inference with Normalizing Flows. arXiv:1505.05770v6
4.10. George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, Balaji Lakshminarayanan //
Normalizing Flows for Probabilistic Modeling and Inference. Journal of Machine Learning Research 22 (2021) 1-64
4.11. Vanessa Böhm, Alex Kim and Stéphanie Juneau // Fast and efficient identification of anomalous galaxy spectra with

neural density estimation. arXiv:2308.00752v1
4.12. Vanessa Böhm, Uroš Seljak // Probabilistic Autoencoder. Published in Transactions on Machine Learning Research (09/2022)
4.13. C. J. Díaz Baso , A. Asensio Ramos, and J. de la Cruz Rodríguez // Bayesian Stokes inversion with normalizing flows. Astronomy & Astrophysics 659, A165 (2022). https://doi.org/10.1051/0004-6361/202142018
4.14 Christopher C. Lovell et al. // A Hierarchy of Normalizing Flows for Modelling the Galaxy–Halo Relationship. arXiv:2307.06967v1
4.15. The H.E.S.S. Collaboration. https://www.mpi-hd.mpg.de/HESS/pages/collaboration/
4.16. CTA Consortium. https://www.cta-observatory.org/about/cta-consortium/

4.6. Предлагаемые методы и подходы, общий план работы на весь срок выполнения проекта и ожидаемые результаты *(объемом не менее 2 стр.; в том числе указываются ожидаемые конкретные результаты по годам; общий план дается с разбивкой по годам)*

Для решения задач, указанных в п. 4.3, будут использоваться следующие методы и подходы.
Общим подходом к решению задачи совместного анализа сильно разнородных данных (в том числе, изображений и данных пространственно-временного типа) будет предварительное извлечение из этих наборов данных таких существенных признаков, которые максимально отражают сущность описываемых объектов и явлений. Совместный анализ будет происходить на уровне полученных существенных признаков.

Общий подход к решению задачи выделения (извлечения) (feature extraction) и отбора (feature selection) существенных признаков для наборов экспериментальных данных будет основан на создании нейросетевой модели, состоящей из двух частей: энкодера и декодера. При этом в качестве этих частей не обязательно будут выступать соответствующие части традиционных автокодировщиков (обычных или вариационных). Будут исследованы варианты и с другими подходящими нейронными сетями. В частности, в качестве энкодера могут выступать сверточные сети (convolutional network; CNN), а в качестве декодера — генеративно-состязательная сеть (Generative adversarial network; GAN), на вход которой подается априорная информация о выявленных существенных признаках исследуемой системы. Возможно использование шумоподавляющего автокодировщика (Denoising autoencoder), — в частности, чтобы избежать проблем оверфиттинга. При этом входные данные восстанавливаются по некоторому его зашумленному варианту. Таким образом, автокодировщик должен будет не просто сжать полученный пример, но еще и частично выделить утраченные в процессе зашумления данные, то есть обучиться не тождественной, а довольно сложной функции (отображение зашумленных данных в исходные), которая может описывать многие интересные свойства (признаки) поступающих на вход данных. Другой важный вариант — разреженные автокодировщики (Sparse autoencoder), которые в некоторых случаях демонстрируют более точное выделение существенных признаков и игнорируют второстепенные детали структуры входных данных. Будут использоваться сети как с уже известной и хорошо зарекомендовавшей себя архитектурой, так и с архитектурой и гиперпараметрами, разработанными в процессе выполнения данного проекта.

Для успешного объединения и совместного анализа весьма желательно иметь возможность интерпретации полученных существенных признаков в терминах предметной области, к которой относится исследуемая система. Методика такой интерпретации также будет основана на использовании нейросетевых моделей. Предполагается, что существует исходная (возможно (полу)эмпирическая) физическая модель рассматриваемого явления и соответствующие величины, в терминах которых она описывается. Примерами таких переменных, на которых основаны соответствующие модели описания явлений, являются: в астрофизике частиц (точнее в экспериментах по исследованию космических гамма-квантов с помощью черенковских телескопов) - так называемые параметры Хилласа (Hillas M. In NASA. Goddard Space Flight Center // 19th Intern. Cosmic Ray Conf. 1985. V. 3. P. 445-448), в ускорительной физике высоких энергий — параметры адронных струй (Salam G. P. "Towards jetography". The European Physical Journal C. 67 (3): 637–686). Используя размеченный (обучающий) набор входных данных можно установить соответствие между существенными признаками, полученными с помощью энкодера, и величинами, в рамках которых формулируется исходная физическая модель (значения для величин находятся либо также с помощью соответствующей нейросети, либо с помощью другой методики, существующей в рамках исходной модели). Используя это соответствие как новый обучающий набор для обучения дополнительных нейросетей, можно установить (прямое и обратное) соответствие между выделенными существенными признаками и величинами исходной модели. Тем самым найденные существенные параметры будут проинтерпретированы в интуитивно понятных терминах предметной области, а конкретные величины предметной области могут быть пересчитаны в значения выделенных существенных параметров. Разработка методов и алгоритмов установления возможных взаимосвязей между полученными признаками может осуществляться на основе адаптации

существующих методов, используемых для отбора признаков , которые пригодны для обнаружения связей между признаками (Venkatesh B, Anuradha J. A review of feature selection and its methods. Cybernetics and Information Technologies. 2019, 19(1):3-26). Могут также использоваться методы эмпирического исследования (взаимо)влияния полученных признаков на генерацию аналогов исходных
данных.
Способность глубоких нейронных сетей (deep neural networks; DNN) обучаться иерархическому (послойному) представлению входных данных делает их особенно подходящими для решения задач обучения на разнородных, мультимодальных данных. Проблема того, как найти маргинальные и совместные представления разнородных модальностей таким образом, чтобы обеспечить их эффективную комбинацию, является центральной для объединения гетерогенных данных. "Маргинальное представление" определяется как результат преобразования унимодальных входных данных таким образом, чтобы обнаруживались скрытые существенные признаки. "Совместное представление" состоит из признаков, представляющих скрытые свойства, основанные на нескольких модальностях, таким образом кодируя информацию, которая может быть дополнительной, избыточной или совместной.

В целом, методы объединения разнородных данных можно классифицировать в зависимости от местоположения уровня (слоя) объединения на (1) раннее, (2) промежуточное и (3) позднее слияние (см., например, P. Pavlidis , J. Weston, J. Cai , W.S. Noble , Learning gene functional classifications from multiple data types., J. Comput. Biol. 9 (2) (2002) 401–411 ;
Ramachandram D, Taylor GW. Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process Mag 2017;34(6):96–108) (см. рис. 1 - 3 в файле 1). При раннем варианте объединяются исходные входные данные, а результирующий вектор обрабатывается как одномодальный вход; таким образом, эта архитектура глубокого обучения не различает модальности, от которых возникают те или иные признаки. Совместные представления мультимодального ввода изучаются напрямую, а маргинальные представления явно не изучаются. Кроме того, полезно различать раннее слияние, основанное на непосредственном моделировании "унимодальных" входных данных с помощью глубоких нейросетей (deep neural network; DNN), и методы, основанные на предварительном использовании автоэнкодеров, которые сначала порождают более низкоразмерные совместные представления (существенные признаки), используемые затем для дальнейшего моделирования с помощью DNN. В рамках проекта будут исследованы различные варианты объединения и выбраны оптимальные.

Очевидным преимуществом раннего слияния разнородных данных является его простота, потому что не нужно делать выбор в отношении того, как извлекать маргинальные представления (рис. 1, файл 1). Один из подходов к раннему слиянию заключается в простом объединении входных данных различных модальностей. Результирующий конкатенированный вектор вводится на первый уровень нейросети. Нейронная сеть не различает признаки из разных модальностей. В этом подходе кросс-модальные и внутримодальные корреляции изучаются одновременно на низком уровне абстракции. Объединенный входной вектор можно смоделировать с помощью полностью связанного входного слоя (конкатенация), если порядок признаков не имеет отношения к задаче обучения. Если порядок входных признаков имеет значение (например, последовательность прихода черенковского излучения от ШАЛ на разные телескопы), к конкатенированному вектору можно применить рекуррентные слои и/или сверточные слои. Представляется возможным использовать в ходе работы по проекту опыт применения рекуррентных сетей в других прикладных областях для решения сходных задач объединения, например, в работе [Mao, J., et al. (2014). "Deep captioning with multimodal recurrent neural networks (m-RNN)". ArXiv:1412.6632] была предложена мультимодальная рекуррентная нейронная архитектура.

Однако подход раннего объединения может быть не в состоянии идентифицировать взаимосвязь между модальностями, когда они становятся очевидными только на более высоких уровнях абстракции, потому что маргинальные представления явно не изучаются. Кроме того, методы раннего слияния могут приводить к искаженным результатам в случае разных объемов наборов сэмплов для разных модальностей.

При промежуточном варианте слияния изучаются и объединяются маргинальные представления в форме векторов существенных признаков вместо исходных мультимодальных данных. Таким маргинальным представлениям можно обучать нейросеть какого-либо одного типа (полносвязаная, сверточная и т. д.) или включающую различные типы подсетей (см. рис. 2, файл 1). Можно ожидать, что первый вариант хорошо работает в случае, когда модальности сравнительно однородны, а второй вариант должен лучше справляться с неоднородностью мультимодальных данных. Эти методы классифицируются как промежуточное слияние, потому что входными данными для слоев слияния являются функции (признаки, полученные на промежуточных слоях нейросетей), тогда как позднее слияние определяется как

слияние решений по подмоделям (см. рис. 3, файл 1). Интересной возможностью, которая будет исследована в рамках проекта, является постепенное слияние данных разных типов, когда сильно коррелированные модальности объединяются раньше, а другие модальности позже в архитектуре. Примером такой ситуации в астрофизике частиц, которая будет использоваться как базовая прикладная область для тестирования разработанных методов, является объединение одинаковых по типу данных нескольких черенковских телескопов (в частности, телескопов проекта TAIGA) и данных существенно другого типа - от массива детекторов (HiSCORE).

Преимущества промежуточных стратегий слияния заключаются в их гибкости — в частности, в поиске нужной глубины и последовательности слияния маргинальных представлений. Возможно, при этом удастся найти более точное отражение истинных отношений между модальностями. Архитектуры глубокого обучения особенно хорошо подходят для промежуточного слияния, поскольку они легко позволяют объединять маргинальные представления, объединяя их в общий слой и сопоставляя иерархические представления глубокой нейросети с природой самого изучаемого явления. В то время как раннее слияние не учитывает - из какой модальности происходит функция, методы промежуточного слияния используют это предварительное знание. Маргинальные репрезентации каждой модальности изучаются для обнаружения корреляций внутри модальности, прежде чем использовать их либо для совместного обучения (см. рис. 4), либо для прямого предсказания (рис. 5).

При позднем слиянии вместо объединения исходных данных или изученных признаков сами решения (выходные данные) отдельных одномодальных подмоделей объединяются в окончательное решение (Baltrusaitis T, Ahuja C, Morency LP. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans Pattern Anal Mach Intell 2019;41(2):423–43; см. также рис. 3 в файле 1). Это позволяет осуществлять хорошее обучение каждому отдельному маргинальному представлению, поскольку каждая модель может быть адаптирована к конкретной модальности. Однако окончательная модель не может выявить мультимодальные эффекты на уровне данных или функций. Различные стратегии агрегирования при позднем слиянии могут использоваться для объединения очень сильно разнородных модальностей. Самый простой подход к агрегированию решений из отдельных подмоделей состоит в том, чтобы взять среднее значение отдельных выходных данных. Для задачи классификации это может быть усреднение вероятностей функций softmax для каждого класса.

Другой возможный подход к обучению на основе конкатенации входных векторов состоит в том, чтобы найти совместное скрытое представление более низкой размерности, содержащее необходимую информацию для восстановления исходного ввода, например на основе автоэнкодеров (АЭ) (см. рис. 6). После получения совместного представления существенных признаков с помощью АЭ, их можно использовать для дальнейшего моделирования. При этом может оказаться полезным использование некоторых разновидностей АЭ, в частности АЭ с шумоподавлением (denoising autoencoders; DAE) и многоярусных автокодировщиков (stacked autoencoders; SAE).

Помимо использования АЭ в методах раннего слияния, они также могут применяться при промежуточном варианте: полученные признаки отдельных модальностей могут быть объединены конкатенацией в единый вектор и использованы в качестве входных данных для дальнейшего моделирования (см. рис. 7) или непосредственно в качестве входных данных для классификатора. Чтобы учесть внутримодальные и кросс-модальные корреляции, маргинальные и совместные представления могут быть изучены в одном АЕ (см. рис. 8): в этом случае первоначально АЭ состоит из ветвей, связанных с отдельными модальностями, а затем маргинальные представления, полученные в этих ветвях, сливаются в слое объединения. Такие совместные представления также можно изучить с помощью вариационных автоэнкодеров (variational autoencoders; VAE).

Метод раннего объединения на основе АЭ также можно использовать для инициализации слоев другой нейронной сети, как продемонстрировано в работе [Jaroszewicz A, Ernst J. An integrative approach for fine-mapping chromatin interactions. Bioinformatics 2020;36(6):1704–11.] в рамках общего подхода по переносу обучения (transfer learning; в русскоязычной литературе также используется термин "перенос знаний"). Такая инициализация может значительно улучшить процедуру обучения. При переносе обучения результаты обучения модели, решающей некую задачу, называемую исходной, используются при обучении другой модели, решающей родственную задачу, называемую целевой. В процессе решения задач проекта предполагается исследование эффективности применения переноса обучения для объединения разнородных данных различных типов, в том числе:
- inductive transfer learning (задачи обучения не совпадают, а пространства входных данных (домены в терминологии теории переноса обучения) могут быть как одинаковыми, так и различными);
- transductive transfer learning (задачи совпадают, а домены в которых они заданы различаются).

- instance-based transfer learning (перенос обучения на основе экземпляров: используются некоторые части тренировочного множества из исходного домена в ходе обучения целевой задачи);
- transfer learning for relational domains (перенос обучения для родственных доменов предполагает, что некоторые отношения между данными в исходном и целевом доменах подобны; в этом случае передаваемые знания – отношения между данными).
- feature representation transfer (перенос признакового представления; цель – обучить "хорошее" представление признаков целевого домена).
На данном этапе представляется, что задачам проекта больше всего соответствует последний из перечисленных выше методов. В частности, возможно использование модели, построенной для исходной задачи, в качестве фиксированного метода извлечения признаков при построении модели, решающей целевую задачу. При этом из исходной сети удаляется классификатор (последние полностью связанные слои), а начальную часть сети используют как инструмент выделения признаков. Взамен старого классификатора обучается новый - на признаках, построенных начальной частью сети (рис. 9 в файле 1). Таким образом, реализуется перенос признакового описания.

В другом случае осуществляется тонкая настройка параметров модели, построенной для решения исходной задачи, с целью решения целевой. Последние слои глубокой модели, соответствующие классификатору, который решает исходную задачу, заменяются новым классификатором (например, набором полностью связанных слоев с другим количеством выходов), и полученная модель обучается как единая система (рис. 10 в файле 1). В этом случае реализуется перенос обучения на основе экземпляров. При очень удачном выделении существенных признаков на первом этапе, что означает, что они для двух разнотипных источников данных оказываются (почти) однородны (однотипны), возможен вариант просто совместного обучения и анализа таких данных, то есть в обучающих мини-батчах можно пробовать просто смешивать полученные экземпляры данных-признаков.

Важной частью проекта является разработка методов поиска редких событий. Актуальность этого вопроса связана с важностью изучения гамма источников во Вселенной, а отношение числа гамма событий к фону редко превышает 1:10000.

Для решения этой проблемы предполагается использовать подходы, основанные на состязательных автоэенкодерах и нормализирующих потоках. Состязательные автоэнкодеры были предложены сравнительно недавно. Их применение особенно эффективно в случае обучения без учителя или когда получение размеченной выборки достаточного размера затруднено. В случае гамма астрономии у нас как раз такой случай, когда имеется огромный объем не размеченных данных, причем практически все события являются фоновыми, которые должны быть отсеяны. Метод нормализующих потоков хорошо дополняет предыдущий, а также задачи проекта, связанные с работой с существенными признаками. Он, в частности, позволяет стартуя с простого распределения параметров, например, нормальное, сгенерировать истинное распределение существенных параметров задачи, что также будет использовано в поиске редких событий.

Практическая апробация разработанных методов будет выполнена на примере реальных задач моделирования, анализа и генерации (аугментации данных) изображений широких атмосферных ливней (ШАЛ) в атмосферных черенковских телескопах, в частности, будут использованы реальные данные для телескопов эксперимента TAIGA (Tunka Advanced Instrument for cosmic ray physics and Gamma Astronomy; https://taiga-experiment.info). Проект предусматривает обширный план по применению, апробации и исследованию разработанных методов совместного анализа разнородных данных для решения задач экспериментальной астрофизики частиц, а именно для анализа данных эксперимента TAIGA. Для проведения физического анализа экспериментальных данных необходимо из сырых данных извлечь физические параметры объекта или процесса. Например, для задач гамма астрономии — это восстановление типа первичной частицы, направления, откуда она пришла, ее энергия. Совместная обработка данных из различных источников — экспериментальных установок (изображения в камерах атмосферных черенковских телескопов, пространственно-временные данные детекторов HiSCORE) должны существенно улучшить отношение сигнала к шуму и точность определения параметров первичных космических частиц. Учитывая специфику прикладной области, необходимо провести большой объем вычислительных экспериментов, который позволит определить наиболее эффективные методы объединения разнородных данных, представленных выше, а также архитектуры нейросетей как для решения задач классификации (идентификация типа первичных частиц), так и задач регрессии (например, определение энергии частиц). Одной из инновационных идей этого проекта является использование нейронных сетей для преобразования начального набора данных (изображения) в размеченные наборы, в которых используются существенные признаки, выделенные при помощи нейронных сетей — энкодеров. Такой подход

позволит не только сократить размер обучающих выборок, но повысить эффективность обучения. Важной особенностью такого подхода является разработка методов интерпретации существенных параметров в физических терминах. Разработка подходов с использованием нейронных сетей для решения этой задачи также предусмотрена в данном проекте.

Программная реализация инструментария для практической реализации разработанных методов и алгоритмов будет основана на открытых библиотеках для построения и тренировки нейросетей таких как PyTourch, TensorFlow, Keras, PyTorch с использованием высокопроизводительных графических процессоров компании NVIDIA и широкого спектра архитектур нейросетей, наборов гиперпараметров и функций ошибок.

Общий план работы на весь срок выполнения проекта и ожидаемые результаты.

В 2024 году на первом этапе проекта основное внимание будет уделено теоретическим исследованиям и разработке методов и алгоритмов для совместного анализа разнородных экспериментальных данных на основе выделения существенных признаков методами глубокого обучения. Планируется выполнить следующие работы:
- выполнить аналитический обзор современной научно-технической литературы по теме проекта , в том числе, по подходам и методам машинного обучения для:
а) выделения и отбора существенных признаков для наборов экспериментальных (обучающих) данных;
б) совместного анализа разнородных данных, поступающих из нескольких источников, включая, методами, основанные на использовании различных типов автоэнкодеров и сверточных нейросетей;
в) поиска аномальных событий на основе состязательных автоэнкодеров и нормализующих потоках;
г) обработки данных наземных экспериментов в области астрофизики частиц;
- создать выборки данных на основе методов Монте-Карло для тестирования разрабатываемых методов, алгоритмов и их программных реализаций;
- по результатам теоретического исследования адаптировать существующие и/или разработать новые методы и алгоритмы выделения и отбора существенных признаков для наборов данных;
- разработать предварительные версии нейросетевых моделей на основе разработанных методов и алгоритмов для выделения и отбора существенных признаков для наборов данных;
- разработать методики интерпретации полученных существенных признаков в терминах предметной области, к которой относится исследуемая система, установления взаимосвязей между полученными существенными признаками и физическими величинами, описывающими данное явление;
- провести теоретическое исследование и предварительный отбор методов машинного обучения для совместного анализа разнородных данных с учетом разработанных методов выделения и отбора существенных признаков для наборов экспериментальных (обучающих) данных; особое внимание должно быть уделено совместному анализу данных изображений и пространственно-временных данных;
- подготовить доклады на международные конференции и 2 статьи для публикации в изданиях, индексируемых WoS, Scopus или RSCI;
- подготовить промежуточный отчет.

Результаты работы на первом этапе:

- выводы по результатам анализа современной научно-технической литературы по теме проекта , в том числе, по подходам и методам машинного обучения для выделения и отбора существенных признаков данных, совместного анализа разнородных данных, поступающих из нескольких источников, обработки данных наземных экспериментов в области астрофизики частиц;
- наборы данных для тестирования в процессе разработки методов, алгоритмов и их программных реализаций;
- обоснованный выбор нейросетевых моделей поиска аномальных событий на основе состязательных автоэнкодеров и нормализующих потоках;
- результаты теоретического исследования и предварительного отбора методов машинного обучения для выделения и отбора существенных признаков для наборов экспериментальных (обучающих) данных;
- рабочие алгоритмы и нейросетевые модели выделения и отбора существенных признаков для наборов данных;
- программная реализация предварительной версии нейросетевых моделей для выделения и отбора существенных признаков для наборов данных методами машинного обучения;
- методики интерпретации полученных существенных признаков в терминах предметной области, установления взаимосвязей между полученными существенными признаками и физическими величинами, описывающими данное

явление;
- предварительный отбор методов машинного обучения для совместного анализа разнородных данных с учетом разработанных методов выделения и отбора существенных признаков данных, в том числе данных изображений и пространственно-временных данных;
- 2 статьи, отражающих промежуточные результаты работы по проекту;
- промежуточный отчет.

В 2025 году на втором этапе будет осуществлена программная реализация разработанных подходов, методов и алгоритмов. Планируется выполнить следующие работы с использованием инструментария PyTourch, TensorFlow/Keras:
- разработать программную реализацию методики интерпретации полученных существенных признаков в терминах предметной области, к которой относится исследуемая система (для выбранных наборов данных);
- провести углубленное теоретическое исследование и отбор методов машинного обучения для совместного анализа разнородных данных с учетом разработанных методов выделения и отбора существенных признаков;
- по результатам теоретического исследования адаптировать существующие и/или разработать новые методы и алгоритмы совместного анализа разнородных данных;
- адаптировать существующее и создать новое программное обеспечение на основе разработанных методов и алгоритмов для совместного анализа существенно разнородных данных, включая совместный анализ изображений и пространственно-временных данных;
- разработать нейросетевые модели для поиска редких событий в гамма астрономии на основе состязательных автоэнкодеров и нормализующих потоков;
- осуществить всестороннее тестирование разработанных методов, алгоритмов и их программных реализаций;
- осуществить государственную регистрацию разработанного программного обеспечения;
- подготовить доклады на международных конференциях и опубликовать результаты в 4 статьях в изданиях, индексируемых WoS, Scopus ил RSCI;
- подготовить промежуточный отчет.

Результаты работы на втором этапе:

- программная адаптация существующих и реализация новых алгоритмов:
- интерпретации полученных существенных признаков в терминах предметной области;
- для совместного анализа разнородных мультимодальных данных, включая совместный анализ изображений и пространственно-временных данных;
- установления взаимосвязей между полученными существенными признаками и физическими величинами, описывающими данное явление;
- нейросетевые модели для поиска редких событий в гамма-астрономии;
- качественные и количественные результаты тестирования разработанных методов, алгоритмов и их программных реализаций;
- государственная регистрация разработанного программного обеспечения;
- публикации 4 статей, отражающих промежуточные результаты работы по проекту
- промежуточный отчет.

В 2026 (заключительном) году на третьем этапе будут осуществлены исследования и вычислительные эксперименты для анализа и оптимизации разработанных методов и алгоритмов, их программной реализаций, а также их апробация на данных реальных экспериментов. Планируется выполнить следующие работы:
- осуществить вычислительные эксперименты и практическую апробацию разработанных методов на реальном примере задачи в области астрофизики частиц, а именно:
- совместный анализ широких атмосферных ливней (ШАЛ), используя существенно разнородные данные атмосферных черенковских телескопов TAIGA-IACT и детекторов TAIGA-HiSCORE для эксперимента TAIGA;
- провести поиск редких гамма событий по экспериментальным данным TAIGA;
- сравнить полученные результаты с результатами других подходов, не использующих машинное обучение;
- осуществить анализ, полученных результатов и на их основе провести оптимизацию программной реализации разработанных методов по результатам экспериментальных исследований, в том числе, выполнить сравнительный анализ
- провести общий анализ, полученных результатов, в том числе:
а) обобщение результатов исследований;

б) сопоставление анализа научно-информационных источников и результатов теоретических и экспериментальных исследований;
в) оценка эффективности полученных результатов в сравнении с современным научно-техническим уровнем;
- осуществить государственную регистрацию разработанного программного обеспечения;
- подготовить доклады на международных конференциях и опубликовать результаты в 6 статьях в изданиях, индексируемых WoS, Scopus ил RSCI;
- подготовить итоговый отчет.

Результаты работы в 2026 г.:

- аналитический отчет о практической значимости полученных результатов для реального крупномасштабного эксперимента в области астрофизики (TAIGA);
- выводы о проведенных вычислительных экспериментах, качестве и эффективности работы разработанных методов, алгоритмов и их программных реализаций;
- итоговый анализ полученных результатов;
- государственная регистрация разработанного программного обеспечения;
- публикации 6 статей, отражающих результаты работы по проекту (в том числе, в журналах, входящих в первый квартиль (Q1));
- итоговый отчет по проекту.

В ходе выполнения проекта члены коллектива примут участие в российских и международных конференциях, а также рабочих совещаниях с целью представления и обсуждения полученных результатов.

Планируются эксперименты с участием лабораторных животных:
нет

4.7. Имеющийся у научного коллектива научный задел по проекту, наличие опыта совместной реализации проектов *(указываются полученные ранее результаты, разработанные программы и методы)*

Предпосылкой успешного выполнения предлагаемых работ является существующий высокий уровень владения участниками проекта методами машинного обучения и их приложения в области физики элементарных частиц и астрофизики. В частности, участники коллектива исполнителей являлись участниками крупнейших международных проектов в области численной обработки экспериментальных данных: “Enabling Grids for E-sciencE” (EGEE, http://eu-egee.org) и European Grid Initiative (EGI, http://www.egi.eu), в проекте Европейского центра ядерных исследований (ЦЕРН, Женева, Швейцария) "The Worlwide LHC Computing Grid Project" (WLCG, http://www.cern.ch/WLCG). Указанные проекты направлены на сбор, обработку и анализ данных в области физики высоких энергий - передовой области науки, в которой моделирование событий является ключевым моментом.
Коллектив имеет очень большой опыт совместной реализации проектов. В частности, коллектив исполнителей участвовал в выполнении большого числа работ по развитию грид-технологии, а именно, в исследованиях в рамках многочисленных грантов РФФИ и европейской программы INTAS, ФЦП «Исследования и разработки по приоритетным направлениям развития научно- технологического комплекса России на 2014 – 2020 годы».

В последние годы, коллектив активно занимается исследованием методов машинного обучения, рядом вопросов по применению технологии распределенных реестров для естественных наук, а также вопросами использования технологии виртуализации для целей физики высоких энергий. Так, члены коллектива выполнили ряд проектов РНФ:
- проект 18-11-00075, тема "Разработка принципов и алгоритмов управления метаданными провенанса больших научных данных с использованием блокчейн-технологии", руководитель к.ф.-м.н. А.П.Демичев; исполнители А.П.Крюков, Ю.Ю.Дубенская;
- проект 18-41-06003, тема "RSF-Helmholtz: Карлсруэ-Российская инициатива по работе с астрофизическими данными на протяжении их жизненного цикла.", руководитель к.ф.-м.н. А.П.Крюков; исполнители к.ф.-м.н. С.П.Поляков, Ю.Ю.Дубенская.
Также выполнен проект, поддержанный РФФИ 18-37-00502, тема «Разработка и исследование методов повышения производительности суперкомпьютеров на основе миграции заданий с использованием контейнерной виртуализации», руководитель к.ф.-м.н. С.П.Поляков. исполнитель Ю.Ю.Дубенская.

В настоящее время под руководством А.П.Крюкова успешно осуществляется проект РНФ 22-21-00442, тема "Моделирование выборок случайных событий с учетом априорной информации в астрофизических экспериментах методами машинного обучения". Ответственный исполнитель Ю.Ю.Дубенская, исполнители А.А.Власкина, Е.О.Гресь. Результаты, полученные в ходе выполнения гранта РНФ 22-21-00442 будут использованы при выполнении данного проекта.

Основной задел непосредственно по предлагаемой тематике членами коллектива был выполнен в рамках сотрудничества с проектом TAIGA (Tunka Advanced Instrument for cosmic ray physics and Gamma Astronomy). Среди задач, которые решались в рамках проекта методами машинного обучения, были вопросы классификации (распознавания типа первичных частиц космических лучей) по изображениям, полученных с черенковских телескопов, задача регрессии (определение физических параметров первичных частиц таких как энергия, направление оси широкого атмосферного ливня и другие), а также задача генерации искусственных изображений широких атмосферных ливней для дальнейшего их использования в качестве быстрой альтернативы Монте-Карло моделированию. В рамках этой работы были разработаны новые подходы к распознаванию типа первичных частиц космических лучей и их характеристик на основе сверточных нейронных сетей. Было показано, что использование методов машинного обучения позволяет увеличить качество предсказания типа частиц и их характеристик на 20-30% по сравнению со традиционной методикой обработки данных, применяемой физиками. Было также показано, что использование стерео режима регистрации частиц (одновременная регистрация одних и тех же широких атмосферных ливней двумя и более телескопами) позволяет увеличить этот разрыв до 2 раз в ряде случаев. Это показывает, что объединение данных с нескольких экспериментальных установок (в данном случае – одного типа) позволяет существенно улучшить результаты их анализа и является существенным заделом и обоснованием для заявляемого проекта. В целом работы коллектива исполнителей данного проекта убедительно демонстрируют, что в настоящее время открываются широкие перспективы использования современных методов машинного обучения для анализа экспериментальных данных, в том числе в области астрофизики.

Работы в области применения машинного обучения в астрофизике, а также большой опыт членов коллектива в области физики высоких энергий, математической физике и информационных технологиях (руководитель коллектива А.П.Крюков имеет более 161 статей, входящих в SCOPUS, индекс Хирша 13) позволил сформулировать одну из ключевых задач моделирования событий в физике частиц методом машинного обучения, а именно проблему совместного анализа мультимодальных данных полученных из разных источников (экспериментальных установок) на уровне выделенных существенных признаков. Членами коллектива был разработан ряд программ по машинному обучению, которые успешно используются в области гамма астрономии для эксперимента TAIGA. На некоторые из них получены свидетельства о государственной регистрации программ для ЭВМ (см. ниже).

Результаты работы по данному направлению были представлены на ряде международных конференций, в частности, XIV International School on Neutrino Physics and Astrophysics, Саров, Россия, 18-23 июля 2022; The 6th International Workshop on Deep Learning in Computational Physics, ОИЯИ, Дубна, Россия, 6-8 июля 2022, The 7th International Conference on Deep Learning in Computational Physics, СПбУ, С.-Петербург, Россия, 21-23 июля 2023 (и конференциях этой серии в 2020-2021 гг.); 37-я Всероссийская конференция по космическим лучам, НИИЯФ МГУ, Россия, 27 июня - 2 июля 2022; 21st International Symposium on Very High Energy Cosmic Ray Interactions (ISVHECRI 2022), Индия, 23-28 мая 2022; 20th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2021, 29.11-03.12.2021, Daejeon, South Korea)), Daejeon, Корея, Республика, 29 ноября - 3 декабря 2021; 37th International Cosmic Ray Conference (ICRC2021); International Conference on Computer Simulation in Physics and beyond (CSP 2020), Москва, Россия, 12-15 октября 2020, The 27th International Symposium Nuclear Electronics and Computing (NEC'2019), Черногория, 30 сентября - 4 октября 2019; The 6th International Workshop on Deep Learning in Computational Physics (DLCP-2022), Россия, 6-8 июля 2022; 37 Всероссийская конференция по космическим лучам (ВККЛ-2022), Россия, 27 июня - 2 июля 2022 и ряде других.

Членами коллектива были получены свидетельства о государственной регистрации программ для ЭВМ, в том числе:
- "Программа для определения типа и параметров первичных частиц на основании данных атмосферного черенковского телескопа методом машинного обучения" (№2020618844, 05.08.2020);
- "Программа идентификации первичных частиц космических лучей по изображениям с атмосферных черенковских телескопов методом машинного обучения" (№2019666634, 12.12.2019);
- "Программа извлечения метаданных из файлов формата TAIGA-IACT: MDE IACT" (№2019664787, 13.11.2019)

- "Программа чтения бинарного формата данных TAIGA-IACT: IACT Reader" (No2019664196, 01.11.2019);
- «Пользовательский интерфейс для удаленного взаимодействия с системой управления распределенными хранилищами» (свидетельство о государственной регистрации программы для ЭВМ No 2019616518 от 23.05.2019);
- «Сервис управления данными и метаданными провенанса в распределенных хранилищах» (свидетельство о государственной регистрации программы для ЭВМ No 2019616702 от 29.05.2019);
- «Бизнес-логика управления распределенными хранилищами на основе блокчейн-технологии» (свидетельство о государственной регистрации программы для ЭВМ No 2019616519 от 23.05.2019).

Основные публикации за последние 5 лет:

Demichev A., Dubenskaya Ju, Kryukov A., Polyakov S.P., Gres E., Vlaskina A., "Machine learning methods for the analysis of taiga experiment data", 2022, Proceedings of Science, том 429
Demichev A., Dubenskaya Ju, Kryukov A., Polyakov S.P., Gres E., Vlaskina A., "Using Conditional GAN to Control the Statistical Characteristics of the Generated Images from Imaging Atmospheric Cherenkov Telescopes", 2022, Proceedings of Science, том 429
Demichev A., Dubenskaya Ju, Kryukov A., Polyakov S.P., Gres E., Vlaskina A., "Using conditional variational autoencoders to generate images from atmospheric Cherenkov telescopes", 2022, Proceedings of Science, том 429
Vlaskina A. A., Kryukov A.P. , "Application of convolutional neural networks for data analysis in TAIGA-HiSCORE experiment", 2022, Proceedings of Science, том 429 (будет опубликовано)
Gres E.O., Kryukov A.P., "Energy reconstruction in analysis of Cherenkov telescopes images in TAIGA experiment using deep learning methods", 2022, Proceedings of Science, том 429
Polyakov S., Demichev A., Kryukov A., Postnikov E., "Processing Images from Multiple IACTs in the TAIGA Experiment with Convolutional Neural Networks", 2022, Proceedings of Science, том 410, с. 016 DOI: http://dx.doi.org/10.22323/1.410.0016
Dubenskaya J., Kryukov A., Demichev A., "Modeling Images of Proton Events for the TAIGA Project Using a Generative Adversaria Network: Features of the Network Architecture and the Learning Process", Proceedings of Science, 2022, том 410, с. 011 DOI: http://dx.doi.org/10.22323/1.410.0011
Vlaskina A., Kryukov A., "Analysis of the HiSCORE Simulated Events in TAIGA Experiment Using Convolutional Neural Networks" Proceedings of Science, 2022, том 410, с. 018; DOI: http://dx.doi.org/10.22323/1.410.0018
Gres E., Kryukov A., "The Preliminary Results on Analysis of TAIGA-IACT Images Using Convolutional Neural Networks", Proceedings of Science, 2022, том 410, с. 015 DOI: http://dx.doi.org/10.22323/1.410.0015
Astapov I.I.,..., Kryukov A.P. и др., "Cosmic-Ray Research at the TAIGA Astrophysical Facility: Results and Plans", Journal of Experimental and Theoretical Physics, 2022, том 134, No 4, с. 469-478 DOI: http://dx.doi.org/10.1134/s1063776122040136
Vasyutina M., ..., Kryukov A.P. и др.,"Gamma/Hadron Separation for a Ground Based IACT in Experiment TAIGA Using Random Forest Machine Learning Methods", Proceedings of Science, 2022, том 410, с.008 DOI: http://dx.doi.org/10.22323/1.410.0008
Astapov I., ..., Kryukov A.P. и др. "Identification of electromagnetic and hadronic EASs using neural network for TAIGA scintillation detector array", Journal of Instrumentation, 2022, том 17, No 05, с. P05023 DOI: http://dx.doi.org/10.1088/1748-0221/17/05/p05023
Astapov I.,..., Kryukov A.P. и др., "Optimisation studies of the TAIGA-Muon scintillation detector array", Journal of Instrumentation, 2022, том 17, No 06, с. P06022 DOI: http://dx.doi.org/10.1088/1748-0221/17/06/p06022
Bogdanova G.,..., Kryukov A. и др., MPD Collaboration, "Status and initial physics performance studies of the MPD experiment at NICA", European Physical Journal A, 2022, том 58, No 7, с. 140 DOI: http://dx.doi.org/10.1140/epja/s10050-022-00750-6
Budnev N.,..., Kryukov A.P. и др., "TAIGA—A hybrid array for high energy gamma-ray astronomy and cosmic-ray physics", Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2022, том 1039, с. 167047-167047 DOI: http://dx.doi.org/10.1016/j.nima.2022.167047
Астапов И.И.,..., Крюков А.П. и др., "Изучение космических лучей на Астрофизическом комплексе TAIGA: результаты и планы", Журнал экспериментальной http://dx.doi.org/10.31857/S0044451022040095
Dubenskaya J., Kryukov A., Demichev A., "Fast simulation of gamma/proton event images for the TAIGA-IACT experiment using generative adversarial networks", Proceedings of Science, 2021, том 395, с.874; DOI: http://dx.doi.org/10.22323/1.395.0874
Polyakov S., Demichev A., Kryukov A., Postnikov E., "The use of convolutional neural networks for processing images from multiple IACTs in the TAIGA experiment", Proceedings of Science, 2021, том 395, с.753; DOI: http://dx.doi.org/10.22323/1.395.0753
Tokareva V., Bychkov I., Demichev A., Dubenskaya J., Fedorov O., Haungs A., Kang D., Kazarina Y., Korosteleva E., Kostunin D,

Kryukov A., "German-Russian Astroparticle Data Life Cycle Initiative to foster Big Data Infrastructure for Multi-Messenger Astronomy", Proceedings of Science, 2021, том 395, с.938; DOI: http://dx.doi.org/10.22323/1.395.0938
Postnikov E., Kryukov A., Polyakov S., Zhurov D., "Deep learning for energy estimation and particle identification in gamma-ray astronomy", In: Proceedings of the 3d International Workshop on Data Life Cycle in Physics (DLC 2019), 2019, CEUR Workshop Proceedings, том 2406, с. 90-99; http://ceur-ws.org/Vol-2406/paper11.pdf
Postnikov E.B., Kryukov A.P., Polyakov S.P., Shipilov D.A., Zhurov D.P. , "Gamma/Hadron Separation in Imaging Air Cherenkov Telescopes Using Deep Learning Libraries TensorFlow and PyTorch", Journal of Physics: Conference Series, 2019, том 1181, с. 012048; DOI: http://dx.doi.org/10.1088/1742-6596/1181/1/012048
Demichev A., Kryukov A., Prikhod'ko N., "Business Process Engineering for Data Storing and Processing in a Collaborative Distributed Environment Based on Provenance Metadata, Smart Contracts and Blockchain Technology", Journal of Grid Computing (Q1). 2021, том 19(1), с.1-30; DOI: http://dx.doi.org/10.1007/s10723-021-09544-4
Postnikov Evgeny B., Bychkov Igor V., Dubenskaya Julia Y., Fedorov Oleg L., Kazarina Yulia A., Korosteleva Elena E., Kryukov Alexander P., Mikhailov Andrey A., Minh-Duc Nguyen, Polyakov Stanislav P., Shigarov Alexey O., Shipilov Dmitry A., Zhurov Dmitry P., "Particle identification in ground-based gamma-ray astronomy using convolutional neural networks", In: Proceedings of the International Workshop on Data Life Cycle in Physics (DLC 2018), 2018, CEUR Workshop Proceedings, том 2267, с. 431-435; http://ceur-ws.org/Vol-2267/431-435-paper-82.pdf
Polyakov S., Dubenskaya J., Fedotova E. "A Container-Based Job Management System for Utilization of Idle Supercomputer Resources", In: Proceedings of the 4th International Workshop on Data Life Cycle in Physics (DLC 2020), 2020, CEUR Workshop Proceedings, том 2679, с.118-124; http://ceur-ws.org/Vol-2679/short12.pdf
Dubenskaya J., Polyakov S., Nguyen Minh-Duc, Fedotova E., "Comparison of Container Virtualization Tools for Utilization of Idle Supercomputer Resources, In: Proceedings of the 4th International Workshop on Data Life Cycle in Physics (DLC 2020), 2020, CEUR Workshop Proceedings, том 2679, с. 66-70; http://ceur-ws.org/Vol-2679/short4.pdf
Demichev A, Kryukov A, Prikhod'ko N., "Metadata driven data management in distributed computing environments with partial or complete lack of trust between user groups.", In: 2019 Ivannikov Ispras Open Conference (ISPRAS), IEEE Xplore Digital Library, 2020, pp. 35-41. IEEE; DOI: http://dx.doi.org/10.1109/ISPRAS47671.2019.00011
Demichev A., Kryukov A., Nikolay P., "Access Rights Management in Decentralized Distributed Computing Systems", Proceedings of IV International Workshop "Data life cycle in physics" (DLC-2020),CEUR Workshop Proceedings, 2020, том 2679, с.59-65; http://ceur-ws.org/Vol-2679/short3.pdf
Demichev A., Kryukov A., Prikhod'ko N., "Blockchain-Based Delegation of Rights in Distributed Computing Environment", Proceedings of 15th International Conference on Parallel Computing Technologies (PaCT-2019; Almaty, Kazakhstan, August 19–23, 2019), Lecture Notes in Computer Science, 2019, v. 11657, pp. 408–418; DOI: https://doi.org/10.1007/978-3-030-25636-4_32.
Bychkov I., Dubenskaya J., Korosteleva E., Kryukov A., Mikhailov A., Nguyen M.D., Shigarov A. , "Metadata extraction from raw astroparticle data of TAIGA experiment", In: Proceedings of the 3d International Workshop on Data Life Cycle in Physics (DLC 2019), 2019, CEUR Workshop Proceedings, том 2406, с. 26-34; http://ceur-ws.org/Vol-2406/paper4.pdf
A.P. Demichev, J.Yu. Dubenskaya, E.Yu. Fedotova, A.P. Kryukov, S.P. Polyakov, N.V. Prikhod'ko, "Hyperledger-based data provenance in distributed computing environments", 2019, Труды конференции Суперкомпьютерные дни в России: Труды международной конференции (23-24 сентября 2019 г., г. Москва). – М.: Изд-во МГУ, 2019. стр. 24 — 32; DOI: 10.29003/m680.RussianSCDays.
Demichev A.P. , Dubenskaya J.Yu., Fedotova E.Yu., Kryukov A.P., Polyakov S.P., Prikhod'ko N.V., "Provenance metadata management in distributed storages using the Hyperledger blockchain platform", In: Proceedings of The III International Workshop "Data life cycle in physics experiments 2019" (2-7 April 2019 Irkutsk, Russia), CEUR Workshop Proceedings, 2019, pp. 35-42; http://ceur-ws.org/Vol-2406/paper5.pdf

4.8. Перечень оборудования, материалов, информационных и других ресурсов, имеющихся у научного коллектива для выполнения проекта *(в том числе – описывается необходимость их использования для реализации проекта)*

Члены коллектива через сервисы, предоставляемые сотрудникам МГУ, имеют доступ к архивам научных журналов, включая иностранные, а также ограниченный доступ к наукометрическим системам Web of Science и Scopus. Таким образом, члены коллектива имеют доступ к публикациям современных исследований в области машинного обучения и их приложениям
в физике.

Коллектив имеет доступ к сетям Интернет с высокоскоростным каналом до 4Гб/с. Также члены коллектива обеспечены персональными рабочими местами, включенными в 1 Гб/с локальную сеть, оргтехникой (цветной принтер, средства

проведения видео и аудио конференций), а также необходимыми расходными материалами, что обеспечит должное выполнение плана работ, предусмотренных программой исследований.

В настоящее время коллектив имеет в своем распоряжении вычислительный сервер с графическими процессорами, что позволяет проводить исследования моделей на основе машинного обучения для целей подтверждения правильности выбранных подходов на начальном этапе проекта. Однако для выполнения всех запланированных исследований на мировом уровне, предполагается дополнительная закупка сервера с высокопроизводительными графическими процессорами компании NVIDIA Tesla V100 или их аналогов.

4.9. План работы на первый год выполнения проекта *(в том числе указываются запланированные командировки (экспедиции) по проекту)*

В 2024 году на первом этапе проекта основное внимание будет уделено теоретическим исследованиям и разработке методов и алгоритмов для совместного анализа разнородных экспериментальных данных на основе выделения существенных признаков методами глубокого обучения. Планируется выполнить следующие работы:
- выполнить аналитический обзор современной научно-технической литературы по теме проекта , в том числе, по подходам и методам машинного обучения для:
а) выделения и отбора существенных признаков для наборов экспериментальных (обучающих) данных;
б) совместного анализа разнородных данных, поступающих из нескольких источников, включая, методами, основанные на использовании различных типов автоэнкодеров и сверточных нейросетей;
в) поиска аномальных событий на основе состязательных автоэнкодеров и нормализующих потоках;
г) обработки данных наземных экспериментов в области астрофизики частиц;
- создать выборки данных на основе методов Монте-Карло для тестирования разрабатываемых методов, алгоритмов и их программных реализаций;
- по результатам теоретического исследования адаптировать существующие и/или разработать новые методы и алгоритмы выделения и отбора существенных признаков для наборов данных;
- разработать предварительные версии нейросетевых моделей на основе разработанных методов и алгоритмов для выделения и отбора существенных признаков для наборов данных;
- разработать методики интерпретации полученных существенных признаков в терминах предметной области, к которой относится исследуемая система, установления взаимосвязей между полученными существенными признаками и физическими величинами, описывающими данное явление;
- провести теоретическое исследование и предварительный отбор методов машинного обучения для совместного анализа разнородных данных с учетом разработанных методов выделения и отбора существенных признаков для наборов экспериментальных (обучающих) данных; особое внимание должно быть уделено совместному анализу данных изображений и пространственно-временных данных;
- подготовить доклады на международные конференции и 2 статьи для публикации в изданиях, индексируемых WoS, Scopus или RSCI;
- подготовить промежуточный отчет.

4.10. Планируемое на первый год содержание работы каждого основного исполнителя проекта *(включая руководителя проекта)*

Руководитель проекта А.П.Крюков:
- выполнение аналитического обзора современной научно-технической литературы по теме проекта;
- проведение теоретического исследования и предварительного отбора методов машинного обучения для выделения и отбора существенных признаков для наборов данных;
- теоретическое исследование и предварительный отбор методов машинного обучения для совместного анализа разнородных данных с учетом разработанных методов выделения и отбора существенных признаков для наборов данных;
- адаптация существующих и разработка новых методов и алгоритмов выделения и отбора существенных признаков для наборов данных;
- общее руководство и координация работ по проекту;
- подготовка научный статей по результатам исследований;
- подготовка промежуточного отчета.

Е.Б.Постников:
- выполнение аналитического обзора современной научно-технической литературы по теме проекта;

- адаптация существующих и разработка новых методов и алгоритмов выделения и отбора существенных признаков для наборов данных;
- разработка методик интерпретации полученных существенных признаков в терминах предметной области, к которой относится исследуемая система, установления взаимосвязей между полученными существенными признаками и физическими величинами, описывающими данное явление;
- участие в подготовке научных статей по результатам исследований.

Ю.Ю.Дубенская:
- создание выборок данных на основе методов Монте-Карло для тестирования в процессе разработки методов, алгоритмов и их программных реализаций;
- теоретическое исследование и предварительный отбор методов машинного обучения для совместного анализа разнородных данных с учетом разработанных методов выделения и отбора существенных признаков для наборов данных;
- адаптация существующих и разработка новых методов и алгоритмов выделения и отбора существенных признаков для наборов данных;
- создание программных нейросетевых моделей на основе разработанных методов и алгоритмов для выделения и отбора существенных признаков для наборов данных;
- участие в подготовке научных статей по результатам исследований.

4.11. Ожидаемые в конце первого года конкретные научные результаты *(форма изложения должна дать возможность провести экспертизу результатов и оценить степень выполнения заявленного в проекте плана работы)*

- выводы по результатам анализа современной научно-технической литературы по теме проекта , в том числе, по подходам и методам машинного обучения для выделения и отбора существенных признаков данных, совместного анализа разнородных данных, поступающих из нескольких источников, обработки данных наземных экспериментов в области астрофизики частиц;
- наборы данных для тестирования в процессе разработки методов, алгоритмов и их программных реализаций;
- обоснованный выбор нейросетевых моделей поиска аномальных событий на основе состязательных автоэнкодеров и нормализующих потоках;
- результаты теоретического исследования и предварительного отбора методов машинного обучения для выделения и отбора существенных признаков для наборов экспериментальных (обучающих) данных;
- рабочие алгоритмы и нейросетевые модели выделения и отбора существенных признаков для наборов данных;
- программная реализация предварительной версии нейросетевых моделей для выделения и отбора существенных признаков для наборов данных методами машинного обучения;
- методики интерпретации полученных существенных признаков в терминах предметной области, установления взаимосвязей между полученными существенными признаками и физическими величинами, описывающими данное явление;
- предварительный отбор методов машинного обучения для совместного анализа разнородных данных с учетом разработанных методов выделения и отбора существенных признаков данных, в том числе данных изображений и пространственно-временных данных;
- 2 статьи, отражающих промежуточные результаты работы по проекту;
- промежуточный отчет.

4.12. Перечень планируемых к приобретению за счет гранта оборудования, материалов, информационных и других ресурсов для выполнения проекта *(в том числе – описывается необходимость их использования для реализации проекта)*

Для проведения полномасштабных исследований, моделирования наборов данных для обучения, тестирования и валидации нейронных сетей, проверки работоспособности разработанных моделей, предусмотренных программой гранта, планируется приобретение сервера с мощными графическими ЦПУ фирмы NVIDIA или аналогичным ГПУ, в следующей комплектации:
- вычислительный сервер со следующими минимальными характеристиками: 2 ЦПУ Intel с частотой 3 ГГц, оперативная память 256 ГБ, 2 жестких диска по 18 ТБ, возможность установки до 4 ГПУ;

4.13. Файл с дополнительной информацией 1

С графиками, фотографиями, рисунками и иной информацией о содержании проекта. Один файл в формате pdf, до 3 Мб.

Текст в файлах с дополнительной информацией должен приводиться на русском языке. Перевод на английский язык требуется в том случае, если руководитель проекта оценивает данную информацию существенной для эксперта.

скачать...

4.14. Файл с дополнительной информацией 2 (если информации, приведенной в файле 1 окажется недостаточно)

С графиками, фотографиями, рисунками и иной информацией о содержании проекта. Один файл в формате pdf, до 3 Мб.

скачать...

Подпись руководителя проекта _____ /А.П. Крюков/