MACHINE LEARNING
FOR ASTROPHYSICS
CATANIA, 30 MAY - 1 JUNE, 2022

Book of Abstracts

Contents

28

From Supervised to Unsupervised Machine Learning: lessons learned from learning machines

Invited Speaker: Kai Polsterer (Heidelberg Institute for Theoretical Studies)

The amount and size of astronomical data-sets was growing rapidly in the last decades. Now, with new technologies and dedicated survey telescopes, the databases are growing even faster. VO-standards provide uniform access to this data. What is still required is a new way to analyze and tools to deal with these large data resources. E.g., common diagnostic diagrams have proven to be good tools to solve questions in the past, but they fail for millions of objects in high dimensional features spaces. Besides dealing with poly-structed and complex data, the time domain has become a new field of scientific interest.

By applying technologies from the field of computer sciences, astronomical data can be accessed more efficiently. Machine learning is a key tool to make use of the nowadays freely available datasets. This talk provides an overview of what can be achieved with supervised and unsupervised learning techniques, discussed on examples that show, what we learned when using machine learning algorithms on real astronomical data-set.

Supervised Learning / 1

Stellar populations of massive young clusters using supervised machine learning

Speaker: Koraljka Muzic (CENTRA, University of Lisbon)

Massive young star clusters are fundamental building blocks of galaxies, and the most abundant reservoirs of newly born stars in the Milky Way. A robust membership assignment is fundamental to study their populations and physical properties, since in the plane of the sky cluster members are often significantly outnumbered by Galactic field stars, as well as by faint distant galaxies. Traditionally, various methods have been used to separate these, including the presence of X-ray emission, infrared excess, spectroscopic youth features, and, to a lesser extent, proper motions. With the advent of the Gaia mission and its precision astrometry, we are now able to study complete stellar populations in massive young clusters at kpc distances, over areas much larger than ever before. In this contribution, I will present an application of the Probabilistic Random Forest algorithm to study the stellar population in the Rosette Nebula, harbouring a young massive cluster NGC 2244. The new sample of candidate members, which doubles the number of previously identified ones, allows us to derive the most complete Initial Mass Function to date, study the spatial structure of the region, detect expansion of NGC 2244, and discuss potential scenarios for its formation.

Supervised Learning / 2

Youth analysis of low-mass stars and brown dwarfs using machine learning methods

Speaker: Victor Almendros (CENTRA, FCUL, Universidade de Lisboa)

The formation of low mass stars and brown dwarfs can be studied through comparison of the

low mass population statistics of young clusters across different environments. Robust low mass populations of young clusters need to be obtained through near infrared spectroscopy, where the low-mass and young nature of member candidates can be confirmed. Traditionally, the spectroscopic analysis of these objects is not performed in a uniform manner, and the assessment of youth generally relies on the visual inspection of youth features whose behavior is not so well understood.
In this contribution, we will present a data set containing almost 3000 near infrared spectra of low-mass stars and brown dwarfs divided into three age classes: young, mid-age and field. We will first present the performance of the traditional methods used for the assessment of low mass and youth, and how we derive a homogeneous set of parameters for the entire data set. We will then present the application of four different machine learning (ML) methods with the aim of producing the best separation between age classes. We compare the performance of applying these methods on spectral features (spectral index + spectral type) or the entire spectrum. When applying ML methods on the entire spectrum we obtain metrics over 98% for the separation of the young and field classes, meaning an almost complete separation between members of young clusters and field contaminants, that will be of special interest for upcoming multi-object facilities such as NIRSPEC and NIRISS/JWST and MOONS/VLT. Using Random Forests we are able to identify what are the most important features in the near infrared spectra of cool dwarfs for youth classification.

Supervised Learning / 4

Stellar dating using chemical clocks and Bayesian inference

Speaker: Andres Moya (University of Valencia)

Dating stars is a major challenge with a deep impact on many astrophysical fields. One of the most promising techniques for this is using chemical abundances. Recent space- and ground-based facilities have improved the number of stars with accurate observations. This has opened the door for using Bayesian inference tools to maximise the information we can extract from them. In this work, we present accurate and reliable stellar age estimates of FGK stars using chemical abundances and stellar parameters, thanks to one of the most flexible Bayesian inference techniques, a hierarchical Bayesian model. The core of the model is a prescription of certain abundance ratios as linear combinations of stellar properties including age. We gathered four different testing sets to assess the accuracy, precision, and limits of our model. We also trained a model using chemical abundances alone. With all this, we found that our age estimates and those used as reference agree well. The mean absolute difference of our estimates compared with those used as reference is 0.9 Ga, with a mean difference of 0.01 Ga. When using open clusters, we reached a very good agreement for Hyades, NGC 2632, Ruprecht 147, and IC4651. We also found outliers that are a reflection of chemical peculiarities and/or stars at the limit of the validity ranges of the training set. The model that only uses chemical abundances shows a slightly worse mean absolute difference (1.18 Ga) and mean difference (-0.12 Ga).

Supervised Learning / 5

Dust extinction inference from Random Forest regression of insterstellar spectral line widths

Speaker: Santiago Gonzalez (CENTRA, IST, University of Lisbon)

Dust extinction is ubiquitous in the Universe and a challenge in the correction of the brightness

and color in astronomical observations. Spectral absorption lines from abundant gas atoms in the interstellar medium (ISM) like sodium, potassium and calcium, or molecules like diffuse interstellar bands, among others, serve as dust indicators and have been used to estimate dust extinction. However, several caveats and limitations exist like line saturation at high optical depths.
We explore here new automated avenues to infer dust extinction from spectral lines with a large sample of supernova spectra. We first develop an automated techinque to measure equivalent widths with high accuracy when compared to high-resolution spectra and simulated spectra. Secondly, with the Milky Way extinction values of Schlafly & Finkbeiner (2011), we train machine learning techniques like random forest regression that use all available lines as input features finding much better predictions of dust extinction than traditional methods based on trends from measurements of single species.

Supervised Learning / 3

QSOs selection in highly unbalanced photometric datasets: The "Michelangelo" reverse-selection method.

Speaker: Giorgio Calderone (Istituto Nazionale di Astrofisica)

I will present a novel selection method aimed at efficiently identifying high-redshift QSOs in highly unbalanced photometric datasets, characterized by a very low number of QSOs with respect to other sources. The method relies on a gradient boosting algorithm, although it may be be used with any other machine learning method providing classification probabilities. I applied the selection method on a sample of photometric data obtained by PanSTARRS1 (DR2), DES (Gold Y3), Gaia (EDR3) and WISE, and I will discuss its performances, as well as a comparison to its basic, direct-selection method counterpart, showing that the former privileges the selection completeness, while the latter privileges the success rate.

Supervised Learning / 162

Modelling galaxy emission-line kinematics using self-supervised, physics-aware, Bayesian neural networks

Speaker: James Dawson (Rhodes University)

In the upcoming decades large facilities, such as the SKA, will provide resolved observations of the kinematics of millions of galaxies. In order to assist in the timely exploitation of these vast datasets we have explored the use of self-supervised, physics aware neural networks capable of Bayesian kinematic modelling of galaxies. I will present the network's ability to model the kinematics of cold gas in galaxies with an emphasis on recovering physical parameterisations and accompanying modelling errors. The models discussed are able to recover rotation curves, inclinations and disc scale lengths for both CO and HI data which match well with those estimated in the literature. The models are also able to provide modelling errors over learned parameters thanks to the application of quasi-Bayesian Monte-Carlo dropout. This work shows the promising use of machine learning and, in particular, self-supervised neural networks in the context of kinematically modelling galaxies observed using interferomers such as ALMA and VLA as well as IFU instruments like SDSS (MaNGA).

Poster Session Day 1 / 99

Planetary Markers in Stellar Spectra: Jupiter-host Star Classification

Speaker: Miguel Andrea Zammit (University of Malta)

Decades of observational & theoretical research has explored the relation between a star's chemical, physical and galactic properties with the presence of orbiting planetary companions. Certain sources suggest that observed correlations are indicators of the environment of the system's protoplanetary disc, and subsequently its proclivity to facilitate planetary formation. This project aims to use the predictive power of machine learning to develop a classifier that uses spectral data of labelled target stars, to learn to model subtle discriminating markers and predict a binary class (Jupiter host or non-host) for every instance. Two approaches were highlighted: The first method was to use raw high-resolution stellar spectra as inputs, in order to preserve any inherent information within the spectrum. The second method was to use homogenous elemental abundance data curated from a pre-existing catalog, and implement a system capable of separating the planetary hosts from comparison stars based solely on the abundance levels of certain elements. To determine whether using raw high-resolution stellar spectra leads to consistent learning and generalisation, several convolutional neural networks (CNN) were implemented in a stacked architecture. Every CNN model was assigned a particular spectral range to collectively cover the entire spectrum, the results of which were then fed into a meta learner to aggregate their votes. Several architectures were simultaneously trained using the elemental abundance feature data, and cross-examination of both approaches was conducted.

Poster Session Day 1 / 106

Using Convolutional Neural Networks to Detect and Confirm Exoplanets

Speaker: Amelia Yu (Palo Alto Unified School District)

Using deep learning (DL), I developed a Python software program with convolutional neural network (CNN) modules from TensorFlow to detect exoplanets through their transit signals in the National Aeronautics and Space Administration (NASA) Kepler space telescope data. My program first utilizes normalization of the light curves, trains deep learning models, tests and evaluates the models with sample data, then folds the light curves of real data to intensify the transit signals, and subsequently applies the DL models to find exoplanets. With this program, I detected new exoplanets and found confirming evidence for previously unconfirmed exoplanets with special astrophysical properties. Two of these detected exoplanets are not listed in the KOI (Kepler Object of Interest) list and more than 20 of the exoplanets are ultra-short period (USP) exoplanets, whose orbital periods are shorter than one day. USP exoplanets are important subjects of research in astrophysics because in order to orbit at such close distances from their stars, these USPs demonstrate a few special physical patterns such as tidal interactions and spin evolution, and also challenge some astrophysical explanations. In addition, transit signals of exoplanets with extremely long periods are difficult to find in the Kepler data because the Kepler mission lasted for only over nine years and observes each star for a selected period of time. For this reason, there are much more KOIs with shorter periods than those with long periods in the NASA database. However, my deep learning program detected a possible Jupiter-like exoplanet in long orbital period together with two other KOI exoplanet candidates in a star system 592.7110±12.3435 parsec away from Earth. This is the first detection of this Jupiter-like exoplanet. It has an orbital period longer than 1600 days, a radius of 10.637 Earth radii, and a planet-star radius ratio of 0.127314. Similarly, Jupiter also has a radius of 11.209 Earth radii and a planet-star radius ratio of 0.102668. Moreover, the size of the transit signal is ~2%, which is comparable to that of Jupiter. These similar stellar and planetary features all indicate that this newly detected exoplanet is a possible Jupiter-like exoplanet, and this multiplanetary system is a Solar-like system, in that a Solar-like system has at least one Jupiter-like or Saturn-like planet. According to NASA, Jupiter is perhaps the most important planet of our system because as the largest planet in the system, it distorts orbits of comets, knocks asteroids out of their orbits, and its gravity affects the orbits of other planets. This new Jupiter-like exoplanet can help expand our understanding about the impact of a Jupiter-like exoplanet with astrophysical significance in its multiplanetary system that has differences from our Solar system. All of the findings indicate that deep learning is an effective method to detect exoplanets and uncover important evidence on exoplanets in big data, and my program can be built upon and reused by other astronomers.

Poster Session Day 1 / 113

Classification of Evolved Stars with (Unsupervised) Machine Learning

Speaker: Jamie Welsh (University of Portsmouth)

The next-generation of observational astronomy instrumentation is expected to generate massively large and high complexity data volumes (big data) at rates of several gigabytes per second. Such enormous volumes impose extremely challenging demands on traditional approaches for data processing and analysis. Machine learning algorithms are playing an increasingly important role in detecting and classifying celestial objects in big data volumes. Our work is focused on analysing the effectiveness of unsupervised machine learning algorithms for classification of evolved stars based on multi-wavelength photometric measurements. The foundation is a custom made reference dataset compiled from available stellar catalogues for target sources - Asymptotic Giant Branch, Wolf Rayet, Luminous Blue variable and Red Supergiant stars. The dataset is composed of approximately 16,000 sources and features 8 independent colours retrieved from photometric catalogues - Wise, 2MASS and Gaia, spectral features were not considered within the dataset. Our experimental results indicate that the clustering algorithm HDBSCAN can utilise colours effectively to classify these sources, with the highest result having attained 65% accuracy. We further investigated the application of feature extraction methods to the dataset, including autoencoders and manifold learning algorithms UMAP and T-SNE. Our results show that these methods significantly improve clustering performance, most notably separating oxygen-rich and carbon-rich AGB stars, despite exhibiting very similar temperatures. Our best result was achieved by combining UMAP and HDBSCAN, attaining accuracy of 86%. We envisage that our findings can be replicated across other datasets containing photometric data, towards achieving even higher accuracies - to this extent we plan to perform a future systematic experimentation. We are also planning to make our ML pipeline available within the NEANIAS cloud-based science gateway to provide an easy-to-use interactive testbed environment, inviting domain scientists to design, realise, evaluate and optimise customised classification workflows for evolved stars.

Poster Session Day 1 / 117

Identification of ultracool dwarfs in J-PLUS DR2 using Virtual Observatory tools and Machine Learning techniques

Speaker: Pedro Mas (Centro de Astrobiologia, INTA-CSIC)

The Javalambre Photometric Local Universe Survey Data Release 2 (J-PLUS DR2) covers 2\,176 deg^2 using a unique filter system of 12 optical bands. This large coverage of the electromagnetic spectrum allows a more accurate determination of physical parameters such as, for instance, effective temperatures.

Current surveys like J-PLUS, and others to come in the near future, are causing a data avalanche in Astronomy. In this scenario, the Virtual Observatory makes the difference in what refers to the discovery, access and analysis of scientific data. Moreover, the huge volume of information generated by these surveys goes beyond what traditional processing and analysis methods can offer. To face this situation, machine learning (ML) approaches have gained momentum over the last few years offering a suite of alternatives depending on the proposed case.

We present the search for ultracool dwarfs (UCDs, spectral types later than M7) performed across the entire J-PLUS DR2 data set. For this purpose, we apply a methodology driven by the use of multiple VO tools and services that combines J-PLUS data with astrometric information from Gaia EDR3. Furthermore, we explore the ability to reproduce this search with a purely ML-based methodology that relies solely on J-PLUS optical photometry, with a two-step ML method based on Principal Component Analysis (PCA) and Support Vector Machine (SVM) algorithms.

Our methodology starts with a pre-screening process in which we use three different approaches to obtain a shortlist of candidates. Two of these approaches rely on astrometric constraints to preselect the candidates, while the third uses only J-PLUS photometry. Finally, we use VOSA, a Virtual Observatory tool which fits observational data to different collections of theoretical models, to estimate the effective temperature of the candidates and keep only those with $T_{eff}3\,000$ K.

After this process, we ended up with 9\,811 candidate UCDs across the entire sky coverage of J-PLUS DR2. We conducted an in-depth analysis of the kinematics and binarity of these candidates. Also, we developed a Python algorithm to detect flares on Hα and Ca II H and K emission lines, using only J-PLUS photometry, detecting 8 objects with relevant emission peaks in these lines.

When reproducing this search with the ML-based methodology, we were able to remove the hottest objects ($T_{eff} > 4\,100$ K) in the PCA step, using as variables multiple J-PLUS colours. Then, we trained a grid of classification SVMs using as labels the candidate UCDs obtained with the previous methodology. The best recall score (98%) was obtained with a radial basis function (RBF) kernel and hyperparameters $C = 1000$ and $\gamma = 0.001$. Using this model to predict on unseen data, we were able to recover 96% of the candidate UCDs. In contrast with the VO methodology, we deduced that the ML methodology is more efficient in the sense that it allows a greater number of true negatives to be discarded prior to analysis with VOSA, although it is a more restrictive method as it requires objects with good photometry in all the J-PLUS filters used to build the variables.

Poster Session Day 1 / 57

Are large training samples always necessary for machine learning classification models?

Speaker: Stavros Akras (National Observatory of Athens)

Over the last 2 decades machine learning (ML) algorithms have become increasingly popular in astronomy. Several photometric sky-surveys have been conducted and even more are planned for the near future covering a large spectral range. To explore this large amount of data is necessary to apply automatic techniques.
In this talk, I will present the results of the application of ML algorithms to identify new symbiotic stars and planetary nebulae candidates. Despite the small training samples, the performance of the models turns out to be sufficient. For specific problems in astronomy, the combination of data from different spectral wavelengths(different surveys) may be more crucial than the size of the training samples. The detection rate of symbiotic stars spectroscopic discoveries has increased by a factor of three compared to previous attempts.

Poster Session Day 1 / 158

Machine learning applied to X-ray spectra: separating stars from active galactic nuclei

Speaker: Pavan Hebbar (University of Alberta)

Modern X-ray telescopes have detected hundreds of thousands of X-ray sources in the universe. However, current methods to classify these X-ray sources using the X-ray data themselves suffer problems — detailed X-ray spectroscopy of individual sources is too tedious, while hardness ratios often lack accuracy, and can be difficult to use effectively. These methods fail to use the power of X-ray CCD detectors to identify X-ray emission lines and distinguish line-dominated spectra (from chromospherically active stars, supernova remnants, etc.) from continuum-dominated ones (e.g. compact objects or active galactic nuclei [AGN]). In this paper, we probe the use of artificial neural networks (ANN) in differentiating Chandra spectra of young stars in the Chandra Orion Ultradeep Project (COUP) survey from AGN in the Chandra Deep Field South (CDFS) survey. We use these surveys to generate 100,000 artificial spectra of stars and AGN, and train our ANN models to separate the two kinds of spectra. We find that our methods reach an accuracy of 92% in classifying simulated spectra of moderate-brightness objects in typical exposures, but their performance decreases on the observed COUP and CDFS spectra (85–90%), due in large part to the relatively high background of these long-exposure datasets. We also investigate the performance of our methods with changing properties of the spectra such as the net source counts, the contribution of background, the absorption column of the sources, the thermal temperatures of stars, the redshift, and power-law index of AGN, etc. We conclude that these methods have substantial promise for application to large X-ray surveys.

Poster Session Day 1 / 156

Using Machine Learning to Identify Young Stars with TESS Data

Speaker: Carlos Santiago (Universidad de California)

Observations of young stars are essential to developing our understanding of planet formation. Many young stars are found together in well-studied groups, clusters and associations. We aim to identify new candidate young stars, not necessarily in well-studied associations, which may otherwise be missed. While most studies use stellar colors to find young stars, we measure the variability characteristics of known young stars, and compare them to other stars using an algorithm we constructed to group similar objects together based on their lightcurves. In the algorithm we use the Lomb-Scargle periodogram to convert the data into the frequency domain. We then apply machine learning techniques from the scikit-learn library - Principal Component Analysis (PCA) and T-distributed Stochastic Neighbor Embedding (TSNE) - to create a metric to help identify groups of similar stars. We make use of two multi-sector datasets: the Cluster Difference Imaging Photometric Survey (CDIPS) and the Transiting Exoplanet Survey Satellite (TESS) main survey. With the help of UCI's High Performance Community Computing Cluster (HPC), we present a preliminary analysis of several sectors of data from TESS.

Poster Session Day 1 / 137

Obtaining a classification of A-F stars through clustering

analyzing the morphology of the light curves

Speaker: Jose Ramón Rodón (IAA)

Asteroseismology is experiencing a revolution thanks to high-precision asteroseismic space missions (Kepler, K2 and TESS) and their large ground-based monitoring programs. Those instruments have provided an unprecedented wealth of information which allows us to study statistical properties and search for hidden relationships between pulsation and/or physical observables.

Obtaining a large database with well-defined parameters can help to the interpretation of the data. Based on such a DB, this work will focus on the automatic analysis of the morphology of stellar light curves and its relationship with physical parameters.

Previous works have already related morphology with stellar parameters (e.g. metallicity, Teff, luminosity or log g) classifying the observed frequency spectra according to their position in the HR diagram. The novelty of this work lies in the automation of the process and the search for groups with similar morphologies around time and space.

Subsequently, once the morphology information is obtained, unsupervised clustering machine learning techniques are applied, specifically the K-means algorithm and decision trees. Finally, we will see the common characteristics of the groups that we find.

This approach could be particularly useful for stars whose pulsation content is difficult to interpret. This is the case for classical intermediate-mass pulsating stars (i.e., Dor, Scuti, hybrids) for which current theories do not adequately predict the observed oscillation spectra.

Here we use the light curves of stars that have been already studied, taking advantage of the most recent precise stellar characterizations carried out with Asteroseismology. Thus, we obtain a complete set of empirical relationships between morphological characteristics of the stellar light curves and the estimated values of temperature, metallicity, luminosity and surface gravities.

29

Recent trends in machine learning and opportunities for astrophysics

Invited Speaker: Concetto Spampinato (University of Catania)

The talk consists of two parts: an overview of the deep learning paradigm from hand-crafted features to learned ones as well as the recent models for classification, regression, object detection semantic segmentation, generation, and their applications including also astrophysics will be given; the second part will instead discuss the recent techniques to understand why deep models make specific predictions diving into the explainable-AI world.

Deep Learning / 8

A Convolutional Neural Network to characterise the internal structure of stars

Speaker: Juan Carlos Suarez (Universidad de Granada)

In this work we use a convolutional neural network (CNN) to confidently retrieve the mean density and surface gravity of stars from 1.5 to 3 times more massive than the Sun. These physical quantities are key to better understand the structure of these stars as well as the physical processes occurring during their evolution.

Most of A-F stars exhibit variations in their luminosity with time due to stellar pulsations, that is pressure and gravity waves propagating in their interior. As seismology does for the Earth, asteroseismology is nowadays the most powerful tool to probe the interiors of stars, allowing us to test theories on stellar structure and evolution. However, this technique relies on the identification of hundreds of oscillation modes, a very challenging and uncertain task in A-F stars, mainly because of model degeneracy. In the last years, patterns found in the oscillation spectra of these stars have been used to constrain their asteroseismic modelling.

Relying on asteroseismic models we deployed and trained a CNN to detect patterns in the oscillation frequencies and infer their regular separation. The mean density and surface gravity estimates we obtained accurately matched our set of benchmark observations. This method will allow us to analyse massively thousands of A-F stars observed by past, present and future space missions such as CoRoT, Kepler/K2, TESS, and the upcoming PLATO mission (ESA, launch foreseen in 2026).

Deep Learning / 7

Deep learning searching for cluster galaxies from multi-band imaging and extensive spectroscopy

Speaker: Giuseppe Angora (Istituto Nazionale di Astrofisica)

With the upcoming of next-generation large and data-intensive surveys, the development of methods able to automatically extract information from the vast amount of data has exponentially grown up in the last decade. In this work, we explore the classification capabilities of Convolutional Neural Networks to identify galaxy Cluster Members, by disentangling them from foreground and background sources directly from Hubble Space Telescope images by exploiting extensive spectroscopic surveys (CLASH-VLT and MUSE), without any additional photometric information. We train the neural network with squared multi-band thumbnails extracted from HST ACS and WFC3 imaging by combining 15 clusters at redshift 0.2< <0.6, with 3800 spectroscopic redshifts in total. We find that a typical purity and completeness of 90% in identifying Cluster Members can be achieved by feeding the networks only with HST image cut-outs, avoiding the complexity of photometric measurements in cluster fields.

Deep Learning / 9

Characterization of Convolutional Neural Networks for the identification of Galaxy-Galaxy Strong Lensing events

Speaker: Laura Leuzzi (University of Bologna)

Galaxy-galaxy strong lensing events occur when the light emitted by a background source is highly deflected by a foreground galaxy's gravitational potential. Studying these systems allows to tackle several problems, that include the reconstruction of the mass distribution of the lens galaxies and the estimation of the Hubble constant. Thousands of new events are expected to be detected in upcoming imaging surveys, such as the one that will be carried out by the Euclid space telescope, but the potential candidates will have to be identified among the billions of sources that will be observed. In this context, the development of automated and reliable techniques for the inspection of large volumes of data is of crucial importance. Machine Learning and Deep Learning methods have already proven their effectiveness in this field, thus it is expected that they will also play a key role in the future of astronomical data analysis methodologies. In particular, Convolutional Neural Networks are a Deep Learning technique that, in the recent years, has proven to be a powerful tool for the analysis of large datasets of images, because of its speed of execution and ability of generalization. In this work, we implement three Network architectures: a VGG-like Network (Simonyan & Zisserman, 2015), an Inception Network (Szegedy et al. 2015; 2016) and a Residual Network (He et al. 2016; Xie et al. 2017) and we apply them to the problem of identifying galaxy-scale strong lenses in survey images. For this purpose, we train and test our models on a dataset of 40000 Euclid-like mock images simulated by the Bologna Lens Factory. We divide the data into four portions, that progressively include larger fractions of faint lenses. In this way, we evaluate how the inclusion of borderline lenses in the training set impacts the classification of both the clear and the faint events. Initially, the classification is solely based on the morphological features of the systems, i.e. the distortion of the background source in wide arcs and rings around the lens galaxy, since we consider single-band simulations. Afterwards, we also evaluate the importance of adding information about the colour difference between the lens and source galaxies by repeating the same training on multi-band images. Our analysis confirms the potential of the application of this method for the identification of clear lenses, since our models find samples of these systems with >90% precision and completeness. On the other hand, we suggest that specific training for different classes of lenses might be needed for finding the faint lenses as well, since not even the addition of the colour information yields a relevant improvement in this sense.

Deep Learning / 10

Unravelling galaxy merger histories with deep learning

Speaker: Connor Bottrell (Kavli IPMU)

Mergers between galaxies can be drivers of morphological transformation and various physical phenomena, including star-formation, black-hole accretion, and chemical redistribution. These effects are seen clearly among galaxies that are currently interacting (pairs) – which can be selected with high purity spectroscopically with correctable completeness. Galaxies in the merger remnant phase (post-mergers) exhibit some of the strongest changes, but are more elusive because identification must rely on the remnant properties alone. I will present results from my recent paper combining images and stellar kinematics to identify merger remnants using deep learning (arXiv:2201.03579). I show that kinematics are not the smoking-gun for improving remnant classification purity and that high posterior purity remains a significant challenge for remnant identification in the local Universe. However, an alternative approach which treats all galaxies as merger remnants and reframes the problem as an image-based deep regression yields exciting results.

Deep Learning / 151

Multi-band photometry and photo-z estimation from narrow-band images

Speaker: Laura Cabayol (IFAE)

In recent years, the amount and quality of galaxy survey data have increased, requiring more precise and efficient methods for data analysis. Imaging galaxy surveys require precise photometry and photometric redshifts measurements, which are crucial for an extensive set of science applications.

In previous work, we developed Lumos, a deep learning-based algorithm to predict the flux probability distribution of single exposure galaxy images. This method has been tested on the Physics of the Accelerating Universe (PAUS) data, an imaging survey observing with a 40 narrow-band filters camera covering a wavelength range from 4500A to 8500A.
On PAUS observations in the COSMOS field, Lumos increased the SNR of the observations by a factor of 2 compared to an aperture photometry algorithm and has proven more robust towards distorting effects as e.g. blended galaxies, cosmic rays, and scattered light.

We have extended Lumos to predict multi-band photometry and the photometric redshift directly from the astronomical images. In contrast to Lumos, this network uses information from all the narrow-band images observed for one galaxy to predict the photometry in all of the narrow bands. This enables the network to exploit the correlations between independent images of the same galaxy (e.g. the underlying SED or the fact that the galaxy has the same morphology in all images). On PAUS image simulations, this increases the precision of the photometry measurements by a factor of ~4.

Furthermore, the network also uses the information from all the images of one galaxy to predict its photometric redshift. In this way, the network has access to additional information beyond the photometry, e.g. the galaxy morphology and nearby sources or artefacts that may be affecting the image. Moreover, as the network directly works on the galaxy images, it has the potential of detecting spurious effects that could affect the photo-z predictions.

The network also internally co-adds several exposures of the same galaxy in the same band. This enables a more flexible method that can detect problematic observations already at the image level, enabling a more optimal combination of exposures. The network's architecture allows for a different number of exposure images per galaxy and band using an adaptive-pooling layer.

On PAUS simulations, we obtain a photo-z scatter of sigma(z) / (1+z) < 0.01 for a test sample to i_AB <22.5. It also increases the photometry precision by a factor of ~4 at the bright end (i_AB < 20) and ~2 at the fainter end (i_AB > 22) with respect to single-band photometry.

30

Anomaly Detection in Astronomical Data using Machine Learning

Invited Speaker: Michelle Lochner (University of the Western Cape/SARAO)

The next generation of telescopes such as the SKA and the Vera C. Rubin Observatory will produce enormous data sets, far too large for traditional analysis techniques. Machine learning has proven invaluable in handling large data volumes and automating many tasks traditionally done by human scientists. In this talk, I will discuss how machine learning for anomaly detection can help automate the process of locating unusual astronomical objects in large datasets thus enabling new cosmic discoveries. I will introduce Astronomaly, a general purpose framework for anomaly detection in astronomical data using active learning and overview some recent results.

Unsupervised Learning and Pattern Discovery / 26

Unsupervised classification reveals new evolutionary pathways

Speaker: Malgorzata Siudek (IFAE)

While we already seem to have a general scenario of the evolution of different types of galaxies, a complete and satisfactory understanding of the processes that led to the formation of all the variety of today's galaxy types is still beyond our reach. To solve this problem, we need both large datasets reaching high redshifts and novel methodologies of dealing with them.

The statistical power of the VIPERS survey which observed ~90,000 galaxies at z>0.5 and the application of an unsupervised FEM clustering algorithm allowed us to select 12 galaxy classes at z~1: 3 passive, 3 intermediate, 5 star-forming, and a class of broad-line AGNs. Physical properties - in particular, those which were not used for classification purposes - of all these subtypes differ from each other, and the transition between different subtypes is not smooth.

Studies of environmental dependence indicate that the FEM classification may actually reflect different evolutionary paths of different subclasses of passive, star-forming, and intermediate subtypes of galaxies. For instance, the most passive class of red galaxies, residing in dense environments is the most compact and ~20% smaller than other red galaxies of a similar stellar mass. This indicates that unsupervised machine-learning techniques were able to automatically distinguish a rare population of red nuggets, a population of red compact galaxies that avoid merger processes and give us a unique opportunity to study the formation and evolution of red galaxies. In my talk, I discuss the clustering methodology and emerging scenarios of galaxy evolution.

Unsupervised Learning and Pattern Discovery / 13

Patterns in the chaos? An unsupervised view of Galactic SNRs

Speaker: Cristobal Bordiu (Istituto Nazionale di Astrofisica)

Supernova Remnants (SNRs) are the remains of supernova explosions, the cataclysmic deaths of certain types of stars. These rapidly expanding structures have a notorious impact into the surroundings, releasing vast amounts of energy and processed matter to the interstellar medium. The study of SNRs is then crucial to understand the structural, dynamical and chemical evolution of the Galaxy as a whole. However, SNRs constitute a remarkably heterogeneous population of roughly 300 known objects, that exhibit a wide variety of observational properties: their morphologies include shells, bubbles, knots and filaments, often showing strong asymmetries; some of them are bright at infrared wavelengths owing to dust emission; others only exhibit thermal and non-thermal radio emission, etc. These differences likely arise from the progenitor star's nature, existing inhomogeneities in the pre-SN circumstellar material, explosion mechanism and other factors that are quite difficult to constrain.

Is there any underlying order in this apparent chaos? With the goal of answering such a question and systematically look for patterns in the SNR population, we resorted to unsupervised deep learning methods. This approach allows for unbiasedly clustering sources in a multidimensional space that takes into account most of the observational properties of SNRs. In this talk we present a multi-stage unsupervised pipeline, consisting of two complementary steps: (i) a feature extraction phase, able to achieve a compact representation of multiwavelength images by means of convolutional autoencoders and manifold representation; and (ii) a clustering stage that groups together objects that display similar multiwavelength features in the resulting latent space. By applying this pipeline to infrared and radio continuum imagery (22 m, 70 m and 30 cm), we have been able to identify physically meaningful subsets within a representative sample of the Galactic SNR population. This work, developed in the frame of the H2020 NEANIAS project, underlines the potential of unsupervised methods for Galactic Science in the era of SKA precursors.

Unsupervised Learning and Pattern Discovery / 173

An unsupervised approach to galaxy spectra classification

Speaker: Julien Dubois (University of Grenoble)

In a data-intensive era, turning to automated statistical methods has become necessary in most scientific fields, including astronomy and astrophysics. Such methods make possible the analysis of large quantities of data of various forms, ranging from images to spectra, time series, and much more. They are capable of tackling lots of challenging tasks: inference problems, clustering, pattern recognition, and so on. In this presentation, we focus on the topic of clustering and its application to galaxy spectra.

Deep-learning algorithms have gained a lot of popularity in the past few years. However, their supervised nature makes them fully dependent on the quality and completeness of the data samples used in the training processes. And when it comes to galaxy spectra, there are simply no suited training samples available yet. We thus have adopted a data-driven unsupervised approach using the discriminant latent mixture-model based algorithm Fisher-EM (Bouveyron and Brunet, 2012).

In this talk, we will present our results described in Fraix-Burnet et al. 2021 and Dubois et al. 2022 (submitted). We investigated the discriminative capacity of the method based on the analysis of a sample of galaxy spectra simulated with the code CIGALE, and we have successfully applied Fisher-EM to observed data of nearby galaxies from the SDSS survey. We are currently extending this work to redshifts up to 1.2 using the VIPERS survey, and are studying the physical specificities of the classes using the galaxy SED modeling code PEGASE. Finally, we will illustrate how those results can be useful for supervised methods.

Unsupervised Learning and Pattern Discovery / 12

Modelling Galactic Microwave emission with ML techniques

Speaker: Giuseppe Puglisi (University of Roma)

One of the major challenges in the context of the Cosmic Microwave Background (CMB) radiation is to detect a polarization pattern, the so called B-modes of CMB polarization, that are thought to be directly linked to the space-time metric fluctuations present in the Universe at the very first instants of life. To date, several challenges have prevented to detect the B-modes partly because of the lower sensitivity of the detectors partly because of the polarized emission coming from our own Galaxy acting as a contaminant. In this talk, I will show how novel techniques involving unsupervised learning (e.g. clustering methods) can improve the quality of the recovered CMB polarization maps once Galactic emission is removed.
This work has been recently published online in Puglisi et al. 2022.
Moreover, I will show recent developments (Puglisi&Bai 2020 and Krachmalnicoff&Puglisi 2021) in improving modeling of the Galactic polarized emission at sub-millimetric wavelengths by means of Deep Neural Networks like Generative Adversarial Networks and Auto-Encoder. This is particularly relevant in the context of future CMB experiments (e.g. SO, LiteBIRD, CMB-S4) where high sensitivity measurements are expected to be achieved and a better characterization of the foreground contamination is thus required.

Anomaly Detection / 16

In Search of the Peculiar: An Unsupervised Approach to Anomaly Detection in the Transient Universe

Speaker: Dennis Crake (University of Edinburgh)

The era of big data time-domain Astronomy is here, and with planned projects such as the Vera-Rubin Telescope, the scale of the data available is escalating at an astonishing pace. Perhaps, the most scientifically promising aspect of these surveys is their potential for discovery across the transient universe. Nonetheless, current methods are restricting the potential for discovery due to their inability to handle databases of extreme size and multidimensional nature. Currently, rapid analysis to systematically identify scientifically compelling objects in time for relevant spectroscopic follow up using traditional methods is near impossible. We present an approach that tackles the challenges presented by the onslaught of data to identify the most anomalous light curves in current and future time-domain surveys. Using observations from the first two years of TESS, we deploy an Unsupervised Random Forest method using a combination of normalized light curve points and their spectral power spectrum as features to systematically identify anomalous objects. Our method identifies a wide range of variability patterns and successfully pinpoints several documented scientifically compelling targets discovered by the community to date. This approach complements the TESS Objects of Interest (TOI's) list by expanding the discoveries beyond the primary Exoplanet focused mission. We increase the census of rare variable classes such as pulsating stars and eclipsing binaries by publishing our entire list of anomaly scores along with notes from systematic inspection of over 10,000 anomalies, many of which are previously unidentified. We combine our results with Gaia photometry to establish a relation between our "Weirdness metric" and the evolutionary stage of anomalies, revealing fascinating candidates within the instability strip, young stellar objects, white dwarf stars amongst many others. Furthermore, we discover a link between the anomaly score and physical properties, such as the orbital parameters of eclipsing binaries.

Anomaly Detection / 15

Searching for changing-state AGNs in massive datasets with anomaly detection

Speaker: Paula Sanchez (ESO)

The classic classification scheme for Active Galactic Nuclei (AGNs) was recently challenged by the discovery of the so-called changing-state (changing-look) AGNs (CSAGNs). The physical mechanism behind this phenomenon is still a matter of open debate and the samples are too small and of serendipitous nature to provide robust answers. In order to tackle this problem, we need to design methods that are able to detect AGN right in the act of changing–state.
In this talk I will present an anomaly detection (AD) technique designed to identify AGN light curves with anomalous behaviors in massive datasets, in preparation for the upcoming Vera Rubin Observatory. The main aim of this technique is to identify CSAGN at different stages of the transition, but it can also be used for more general purposes, such as cleaning massive datasets for AGN variability analyses. To test this algorithm, we used light curves from the Zwicky Transient Facility data release 5 (ZTF DR5), containing a sample of 230,458 AGNs of different classes. The ZTF DR5 light curves were modeled with a Variational Recurrent Autoencoder (VRAE) architecture, that allowed us to obtain a set of attributes from the VRAE latent space that describes the general behaviour of our sample. These attributes were then used as features for an Isolation Forest (IF) algorithm. We used the VRAE reconstruction errors and the IF anomaly score to select a sample of 8810 anomalies. These anomalies are dominated by bogus candidates, but we were able to identify promising CSAGN candidates.

Anomaly Detection / 163

Predicting and detecting very-high-energy blazar flares with deep learning

Speaker: Hermann Stolte (Humboldt University Berlin)

Blazars are a subclass of active galactic nuclei (AGNs) with relativistic jets pointing toward the observer. They are notable for their flux variability at all observed wavelengths and time scales. The very-high-energy (VHE) emission observed during blazar flares may be used to probe the population of accelerated particles, together with simultaneous measurements at lower energies. However, optimally triggering observations of blazar high states can be challenging. Notable examples include: identifying a flaring episode in real time; predicting VHE flaring activity based on lower energy observables; and devising an optimal follow-up strategy in the context of other observations.

For this purpose, we have developed a new deep learning analysis approach, based on data-driven anomaly detection techniques. We utilise a recurrent neural network architecture for the forecasting of multi-wavelength light curves, and for learning their condensed representations. Based on non-flaring training data, we derive a background model from the learned representations of the model. This is accomplished by integrating the output of our neural network within a Bayesian clustering framework. Finally, we identify future variability in the source as anomalous configurations in the cluster-space of the background model.

We demonstrate our approach using simulations of blazar light curves in two energy bands, corresponding to sources observable with the Fermi Large Area Telescope, and the upcoming Cherenkov Telescope Array (CTA).

Anomaly Detection / 14

Galaxy Zoo: Practical Methods for Large-Scale Learning

Speaker: Michael Walmsley (University of Manchester)

Deep learning is fundamental to creating Galaxy Zoo's latest catalogs. In this talk, we explore the methods we've developed to best exploit large-scale human labels and how other researchers can benefit from them.

We open by presenting Galaxy Zoo LegS - new deep-learning-powered detailed morphology measurements for 8 million galaxies imaged by the DESI Legacy Surveys. Our models are trained on human labels collected over 8 years, during which time different volunteers answered different questions and followed different instructions. We describe how we overcome the resulting label distribution shift to learn from more human responses than any previous astronomical model.

We next show how answering every Galaxy Zoo question simultaneously forces the resulting models to learn meaningful semantic representations of galaxies. These representations can then be directly used for similarity search and to outperform a recent approach at personalized anomaly-finding. Further, and crucially for other researchers, because the models are trained on a diversity of tasks (answering every GZ question), the trained models make excellent base models to finetune to new tasks. We demonstrate this by finetuning to find ringed galaxies. Models pretrained on all GZ questions are better able to find rings than models pretrained on a single GZ question or on ImageNet. We go on to exploit this to create the largest ringed galaxy catalog to date by an order of magnitude. Our trained models are available for the community to finetune for their own tasks at www.github.com/mwalmsley/zoobot (in both TensorFlow and PyTorch) .

Finally, we describe our very latest work combining self-supervised approaches with broad supervised pre-training on Galaxy Zoo to classify galaxies better than with either alone. We believe such approaches are ideally suited to Euclid and Rubin because they allow us to leverage both the millions of human labels collected over the last decade and the raw scale of unlabelled images these new surveys will produce.

Poster Session Day 2 / 139

Star- and planet formation caught-in-the-act

Speaker: Gabor Marton (Konkoly Observatory)

Since the beginning of the last decade the analysis of huge amounts of data is a new challenge that researchers in general have to cope with. This is also true for astronomers and luckily the number of infrared facilities and the amount of data collected by them increased several order of magnitudes, leading us to new discoveries through data mining and knowledge discovery in databases using modern statistical methods, supervised and unsupervised machine learning. On-going surveys in other domains of the electromagnetic spectrum are providing us with a data avalanche at the moment and allow us to catch phenomena that we have never seen before. I present the project NEMESIS (Novel Evolutionary Model for the Early stages of Stars with Intelligent Systems) that aims to build the largest panchromatic dataset of Young Stellar Objects (YSOs) and our methods that are efficiently used in YSO discoveries in large catalogues based on data from IR space telescopes like AKARI, WISE and Herschel and that help to identify eruptive young stars in present and future alert systems like the Gaia Photometric Science Alerts System. NEMESIS has also the aim to revisit the YSO evolutionary timescales and identify evolutionary stages that needs to be analysed in great details. YSOs in these rare stages and alerts can be potential targets for the JWST and future high spatial and/or temporal resolution facilities and can provide details about star and planet formation that we had no chance to capture before.

Poster Session Day 2 / 88

Classification of system variability using a convolutional neural network.

Speaker: Jozef Magdolen (Slovak University of Technology in Bratislava)

It is common that a binary system can report during observation some kind of variability. It may be caused by changing the actual luminosity of the observed object or due to changes in light that can reach the detector by blocking the source. Such variations have been observed in nova outbursts. Novae occur in interacting binaries where matter is accumulated from a companion star to the surface of a white dwarf. After a critical amount is reached, the accumulated hydrogen-rich envelope expands. Both, intrinsic changes in luminosity and/or occultations are present. By creating a light curve and then a periodogram from the detected signal, it is possible to determine the significance of certain frequencies as well as their character. A very decisive factor is the nature of the detected variability, e.g. whether the frequency is variable or stable. We performed multiple simulations of light curves with different types of variability based on XMM-Newton observations of binaries Cal 83 and KT Eri. Both systems were discussed as probably variable in frequency. Next, we created their dynamic power spectra (DPS). Using the AlexNet convolutional neural network (CNN) algorithm and training it on the created DPSs we try to find out, whether it is possible to distinguish between variability in amplitude and variability in frequency. Trained CNN based on Cal 83 reports 92% accuracy and the system was classified as variable in both, frequency and amplitude with a probability of 99.97%. On the other hand, CNN trained on KT Eri reports only 47% accuracy, and the system was classified as variable in frequency with also 48.91% variability in amplitude. The cause of such a difference in accuracy is discussed.

Poster Session Day 2 / 95

Time series anomaly detection in a cataclysmic variable using Support Vector Machine

Speaker: Denis Benka (Slovak University of Technology)

Cataclysmic variable stars (CVs) often show multiple frequency components with a quasi-periodic occurrence. Those can be very subtle and their confidence using standard statistical methods is often of a less significance, e.g., falls under the 1- interval. In our study we aim to use a Support Vector Machine (SVM) to train a model to detect those components with a plausible confidence. We used the lightcurve of MV Lyrae, which is a very bright member of the CVs family. Those are known to have a specific pattern in the variability of their brightness. We used a 272 day long light curve obtained at a cadence of   59 s from the Kepler satellite archive. Using Lomb-Scargle (1982) algorithm we created periodograms and averaged all its values within the bins to get the power density spectrum (PDS). The searched characteristic frequency component is located near log(f/Hz)   -3.4. We simulated the data using the Timmer & König (1995) method based on the PDS with the frequency component. The quasi-periodic frequency in the simulated data had a low confidence level (1-) which was based on the same confidence level as in the observed PDS. The simulated data had a quasi-periodic peak near log(f/Hz)   -3.4. The dataset was divided in two categories, with the presence of the quasi-periodic oscillation and without it (removing it from the simulated PDS with the oscillation). These data, as a time series, were used to train a supervised SVM model. We used a Gaussian kernel to find the support vectors and a hyperplane. Optimal parameters of the regularization parameter C as well as the kernel coefficient ( ) were found using a cross-validated grid-search. We trained several models using different sizes of the training dataset. The model was tested using simulated data as well as the PDS from the observation of our studied CV. The classification accuracy reflects the confidence of the manifesting quasi-periodic frequency.

Poster Session Day 2 / 87

Background Estimation in Fermi Gamma-ray Burst Monitor lightcurves through a Neural Network

Speaker: Riccardo Crupi (University of Udine)

The aim of this work is to provide a data-driven approach to estimate a background model for the Gamma-Ray Burst Monitor (GBM) of Fermi satellite. We employ a Neural Network (NN) to estimate each detector background signal given the information of the satellite: position, velocity, direction of the detectors, etc.

The estimated background can be employed into a triggering algorithm to discover significant long/weak events that are and previously not detected by other approaches.

We show the potentiality of the model by estimating the background on GBM data for Gamma-Ray Bursts (GRBs) present in GBM cataloge, the long GRB 190320 and ultra-long GRB 091024.

The proposed approach is straightforwardly generalizable to estimate the background model of other satellites.

Poster Session Day 2 / 134

Scope: the Zwicky Transient Facility Variable Source Classification Project

Speaker: Joannes Van roestel (California Institute of Technology/University of Amsterdam)

The Zwicky Transient Facility (ZTF) is an optical survey telescope that observes the northern sky every night. Lightcurves in g,r, and i have been obtained of than a billion stars down to magnitude 20.5. Identification and classification of all variables in this huge dataset is required for multiple science cases. ZTF's volume of data, multicolour lightcurves, and (highly) irregular cadence contains a lot of information. These attributes also make classification a challenging problem, but also an excellent testing ground to learn how to deal with these challenges.
I will present our efforts to classify ZTF variable stars, how we dealt with these challenges and the design choices we made in the process. This includes the technical challenge of processing billions of lightcurves, the design of a comprehensive classification scheme and use of many one-vs-all classifiers, and our active learning approach to continuously correct and improve the classifiers.

Poster Session Day 2 / 62

Reconstructing blended galaxies with Machine Learning

Speaker: Lavanya Nemani (Istituto Nazionale di Astrofisica)

Galaxy blending is a confusion effect created by the projection of photons from galaxies on the same line of sight to the sky 2D plane (Dawson & Schneider 2014). The upcoming deep extragalactic surveys like LSST and EUCLID expect to see a blending fraction of up to 50% in the densest regions (Reiman & Göhre 2018). For standard aperture photometry and for more complex techniques such as psf-fitting and template-fitting algorithms, de-blending, the process of reconstructing the individual light profiles from blended sources, becomes crucial. The current standard de-blending algorithms (e.g. SExtractor, Bertin & Arnouts 1996) are based on threshold methods that simply assign each pixel to a single object, often failing to correctly take into account the real properties of the blended galaxies. With the advent of Machine Learning and Computer Vision in Astronomy we want to explore an unbiased and more accurate method of reconstructing individual light profiles using generative models.

Variational Auto-Encoders (Kingma & Welling 2013), VAE, are a type of probabilistic generative models that consist of two parts: In the encoder part the model learns to reduce the high-dimensional input to an encoded representation, and in the decoder part it learns how to reconstruct the input from the lower dimensional representation (called the bottleneck). Two distinct networks are needed to deblend galaxies using VAEs: One which learns how to reconstruct galaxy light profiles in isolation, and another one which uses the trained part of the first network to actually deblend overlapping pairs reconstructing their individual light profiles (e.g. Arcelin et al 2020).

In our work we simulate stamps of galaxy images as expected from the EUCLID survey in the VIS band, to test the results of using this ML technique for deblending as opposed to standard deblending methods like SExtractor. The galaxy images are built starting from a mock input catalogue created using EGG, a code that can generate mock galaxy catalogs with realistic positions, morphologies and fluxes; the catalogue either feeds the image simulation toolkit GalSim to generate analytical double-Sersic profiles, or it is used in parallel with autoencoders on HST images to obtain "euclidized" realistic profiles (Euclid Morphology Challenge, Bretonniere et al. in preparation).

The main focus of our work is to obtain accurate flux and morphology estimates for blended objects and clean light profiles to be used as priors for template fitting (e.g. T-PHOT, Merlin et al 2015). We use several metrics to test our performance of the VAE algorithm comparing the reconstructed light profile of each simulated galaxy to the true input one, such as mean-square error and cosine distance. We also compare fluxes obtained on predicted images with the input catalogue fluxes. With our current best approach we are able to retrieve the original flux within 10% for 1 sigma (whereas SExtractor is within 16% for 1 sigma). In the future, we want to optimize the network architecture used for training an ML algorithm to perform deblending by hyper-parameter training and including state-of-the-art ML architecture blocks like VGG, Inception-Residual, ResNet.

Poster Session Day 2 / 144

Estimating Bayesian Posteriors for Galaxy Morphological Parameters using Machine Learning

Speaker: Aritra Ghosh (Yale University)

Galaxy morphology is connected to various fundamental properties of a galaxy and its environment, such as galaxy mass, star formation rate, stellar kinematics, merger history, etc. Thus, studying the morphology of large samples of galaxies can be a crucial clue to understanding galaxy formation and evolution.

In the past few years, although machine learning has been increasingly used to determine the morphology of galaxies, most previous works have provided only broad morphological classifications (without parameter estimation or uncertainty quantification). They have also required large sets of pre-classified training data.

We have developed Galaxy Morphology Posterior Estimation Network (GaMPEN), a machine learning framework that can estimate the Bayesian posteriors for a galaxy's bulge-to-total light ratio, effective radius, and flux. The computation of full Bayesian posteriors is crucial for drawing scientific inferences that properly account for uncertainty (e.g., deriving robust scaling relations, performing tests of theoretical models using morphology).

To predict posteriors, GaMPEN uses the Monte Carlo Dropout technique and incorporates the full covariance matrix in its loss function. The latter novel step empowers GaMPEN to simultaneously predict accurate posteriors for all three variables by incorporating the structured relationships between the output variables into its predictions.

GaMPEN also contains a novel Spatial Transformer Network (STN) that automatically crops input galaxy frames to an optimal size before determining their morphology. The STN trains along with the rest of the framework, with no additional supervision, and will be crucial in applying GaMPEN to new survey data with no radius measurements.

We have used GaMPEN to determine robust morphological parameter posteriors for ~ 1 million $z < 0.6$ galaxies in the Hyper Suprime-Cam Wide survey. In order to not require a large pre-classified training set of real galaxies, we first trained GaMPEN on realistic simulations of galaxies and then performed transfer-learning/domain adaptation using a small amount of real data.

We have demonstrated that GaMPEN's predicted posteriors are well-calibrated and accurate ($\leq \pm 4\%$ of corresponding confidence levels). Testing has shown that GaMPEN predictions become less precise for especially small or faint galaxies, where the algorithm correctly predicts correspondingly larger uncertainties. We have also demonstrated that by predicting qualitative labels (instead of quantitative values) in certain small regions of the parameter space where GaMPEN's residuals are higher, we can achieve accuracies of $\geq 95\%$.

GaMPEN is the first machine learning framework for determining joint posterior distributions of multiple morphological parameters, does not require large amounts of real training data, and is also the first application of an STN in astronomy.

Poster Session Day 2 / 93

Radio Galaxy detection prediction with ensemble Machine Learning.

Speaker: Rodrigo Carvajal (University of Lisbon)

The study of Active Galactic Nuclei is fundamental to comprehend the processes regarding the birth and evolution of Super-Massive Black Holes (SMBHs) and its connection with star-formation history and general galaxy evolution.

Up to this moment, only ~300 AGN have been identified in the EoR (z>6) of which a small fraction have radio detections, making it difficult to thoroughly study their properties at these early times. Simulations and models predict though that this is just but the tip of the iceberg and that future observatories might increase these numbers by at least an order of magnitude.

In view of the development and operation large-scale radio observatories (e.g., SKA), with their large data volumes, the use of regular AGN detection and z determination techniques will become inefficient. Critical attention has been drawn, then, to the development of Machine Learning (ML) methods to predict the detection of AGN and some of their properties (redshift being one of the most relevant).

We have developed, thus, a series of ML models that, using multi-band photometry, can produce a list of Radio Galaxy candidates, along with their predicted z values. The training stage of these models can then be extended to reach sources at redshift values close to the early epochs of galaxy formation.

We will present the results of training these models with data of NIR-selected sources in the HETDEX Spring Field, covered by the LoTSS Survey, and applying them to photometric data on the Stripe 82 Field.

Poster Session Day 2 / 80

Radio Image Segmentation with Variational Autoencoders

Speaker: Hattie Stewart (University of Bristol)

The Square Kilometre Array (SKA) will be the world's largest radio telescope, producing data at a rate of about 1Tb per second. Even after conversion to images, traditional methods of source detection and classification will not be sufficient. The pre-construction phase of the SKA project saw the launch of SKA Data Challenge 1 (SDC1), a model dataset released for analysis by the community. This work develops a machine-learning approach to detecting and classifying the full radio source population. A Variational Autoencoder (VAE) is presented as a method of image segmentation. The trained network reconstructs the raw image data as a binary segmentation map. The segmentation map describes the angular size, eccentricity, position angle and location of the source. The classes of the source population can be represented by the latent space of the network if appropriate latent vectors are chosen. This work serves as a proof of concept that a VAE can detect and classify radio source populations from SKA-like data.

31

Time Domain Astroinformatics

Invited Speaker: Massimo Brescia (Istituto Nazionale di Astrofisica)

Astronomy has entered the multi-messenger data era and Data Science has found widespread use in a large variety of applications. The exploitation of synoptic (multi-band and multi-epoch) surveys, like Rubin-LSST, requires an extensive use of automatic methods for data processing and interpretation. With data volumes in the petabyte domain, the discrimination of time-critical information has already exceeded the capabilities of human operators and crowds of scientists have extreme difficulty to manage such amounts of data in multi-dimensional domains. I will introduce the machine/deep learning paradigms suitable to explore Time Domain Astronomy in an efficient and semi-automatic way.

Transients and Time Series / **20**

Forecasting solar cycle 25 and next solar minimum using a multi-step Bayesian deep neural network model

Speaker: Ilaria Bizzari (University of Torino)

The latest advances in deep learning techniques have provided new effective prediction models that allow forecasting in detail the evolution of cosmogeophysical time series such as the solar activity, which is also crucial to anticipate potentially adverse space weather effects on the Earth's environment.
Because of the underlying complexity of the quasi-periodic solar dynamo mechanism, the predictions offered by state-of-the-art machine learning algorithms represent valuable supplementary tools for our understanding of the solar cycle progression. As a plus, Bayesian deep learning is particularly compelling due to recent advances in the field that provide improvements in both accuracy and uncertainty quantification compared to standard training. In this work, a deep learning long short-term memory model is applied to the monthly sunspot number series to predict the Solar Cycle 25. Moreover, an accurate uncertainty estimation of the predicted sunspot number is obtained by applying a unique Bayesian approach.
The performance of the applied algorithm was assessed through two different validation techniques, namely the Train-Test split and the time-series k-fold Cross-Validation, which gave compatible results, thus demonstrating the robustness of the method as the length of training and test data changes.
We show the forecasted complete profile of Solar Cycle 25 and discuss the obtained values of phase and amplitude in the frame of the actual estimates available in literature.

Transients and Time Series / **152**

Upgrading PlaNET: A Deep Neural Network for Searching for Transiting Exoplanets with the Next Generation Transit Survey

Speaker: Alexander Chaushev (University of California Irvine)

Exoplanet transit surveys produce flux time-series for
hundreds of thousands of stars to search for the tell-tale signs of a transiting planet. In the process, they provide a rich dataset for the application of machine learning (ML) methods. One focus so far has been the classification of exoplanet signals as genuine or instrumental false positives, particularly by using deep neural networks. I will discuss one such network, PlaNET, and its ongoing upgrade as part of the Next

Generation Transit Survey (NGTS) pipeline. PlaNET has helped vet thousands of candidate signals over the past two years, and in the process we have learnt important lessons about how the network operates in a real-world setting. In particular, interpretability of the results is key, and we show how the application of existing 'explainable AI' methods can greatly illuminate the inner workings of PlaNET. As a result, we have changed the network structure and dataset greatly improving performance. Finally, I will discuss the prospect of other applications of ML to NGTS data such as: identifying unusual variability, searching for clusters of similar stars, and improving the sensitivity of transit searches.

Transients and Time Series / 17

Finding stellar flares with recurrent deep neural networks

Speaker: Attila Bodi (Konkoly Observatory)

At present, low-mass, cool M dwarfs are the prime targets of planet searches, since the habitable zone is much closer to the central object in cool stars than in the case of a solar-like star; thus, detecting a possibly habitable Earth-like planet is easier. However, the late spectral type of these stars and the magnetic activity associated with these could pose a threat to habitability. To study this threat, we have to find stellar flares that are an important tracers of magnetic activity but automatically and accurately finding them is still a challenge to researchers in the Big Data era of astronomy. In this presentation, we present an experiment to detect flares in space-borne photometric data using deep neural networks. Using a set of artificial data and real photometric data we trained a set of neural networks, and found that the best performing architectures were the recurrent neural networks (RNNs) using Long Short-Term Memory (LSTM) layers. The aim for the trained network is not just detect flares but also be able to distinguish typical false signals (e.g. maxima of RR Lyr stars) from real flares. The best trained network detected flares over 5 with 80% recall and precision. Testing the network –trained on Kepler space telescope observations– on Transiting Exoplanet Survey Satellite (TESS) light curves showed that the neural net is able to generalize and find flares –with similar effectiveness– in completely new data with different sampling and characteristics from those of the training set.

Transients and Time Series / 19

Classifying Supernova Time-Series Photometry with Convolutional Neural Networks

Speaker: Helen Qu (University of Pennsylvania)

The use of type Ia supernovae (SNe Ia) as standardizable candles led to the Nobel Prize-winning discovery of the accelerating expansion of the universe and cemented their role in the quest to understand the nature of dark energy. Accurate cosmological parameter estimation requires a sample of pure SNe Ia with minimal non-Ia contamination, but spectroscopic confirmation of supernova type is logistically infeasible in most cases. Thus, a reliable algorithm for photometric supernova classification is vital to expanding the set of cosmologically useful SNe Ia beyond the spectroscopic sample. In this talk, I will present SCONE (Supernova Classification with a Convolutional Neural Network), a novel deep learning-based method for photometric supernova classification. While traditional photometric classification algorithms relied on extracting handcrafted features from supernova photometry, deep learning methods bypass this requirement by identifying and using features optimized for the classification task. SCONE is a convolutional neural network (CNN), an architecture prized in the deep learning community for its state-of-the-art image recognition capabilities. Supernova time-series photometry is preprocessed into 2D "images" using Gaussian processes in both wavelength and time dimensions to generate the input to the model. This alleviates the issue of irregular sampling between filters and, along with the choice of an asymmetric convolutional kernel covering the full wavelength range, allows the CNN to learn from information in all filters simultaneously. In addition, our model requires raw photometric data only, precluding the necessity for accurate redshift approximations. SCONE has achieved 99.73±0.26% test set accuracy differentiating SNe Ia from non-Ia, as well as 98.18±0.3% test accuracy performing 6-way classification of supernovae by type. SCONE also exhibits impressive performance classifying supernovae by type as early as the second detection with very few epochs of photometric observations. SCONE achieved 60% average accuracy across 6 supernova types at the date of trigger and 70% accuracy 5 days after trigger. The incorporation of redshift information improves these results significantly to 75% accuracy at the date of trigger and 80% accuracy 5 days after trigger. The model also has relatively low computational and dataset size requirements without compromising on performance – the above results are from models trained on a 10^4 sample dataset in around 15 minutes on a GPU. SCONE's ability to produce impressive early-time classification results as well as perform highly accurate SNe Ia vs. non-Ia classification makes it an excellent choice for spectroscopic targeting and the development of photometric SNe Ia samples for cosmology.

Transients and Time Series / 164

An Unsupervised Dive Into Gamma-ray Burst Afterglow Classification

Speaker: Eliot Ayache (Stockholm University)

The Neil Gehrels "Swift" Observatory has been detecting and measuring emission from gamma-ray bursts (GRBs) and their associated afterglow for the last 17 years. Today, over 1500 bursts have been observed, with light curves displaying different morphologies in the succession of decay regimes with time. We explore prospects for acquiring physical inference from machine-learning models by investigating the presence of intrinsic classes (or lack thereof) in the morphology of GRB afterglow X-ray light curves. Ignoring the well-known divide between long and short GRBs, we carry out unsupervised classification of Swift-XRT time-series data using a convolutional variational autoencoder. The generative aspect of the model can provide physical insight by highlighting the discriminative features in the light curves.

We compare the classification results obtained with the traditional functional-form-based classification, and investigate the resulting level of segregation in the dataset. We evaluate our model's ability to identify different morphological classes by carrying out training on synthetic data. We find that the data creates over-densities in the latent-space. However, the observed gradual transition in between unifies the prevalent classification of GRBs based on their X-ray data into a single continuum, supporting the idea that light curves of different types should be unified under a single model.

In a deeper investigation of this afterglow population, we make use of variational deep embedding, where the level of clustering can be more easily quantified, for which I will present our latest results.

32

Exploring the Universe with the world's largest radio telescope

Invited Speaker: Grazia Umana (Istituto Nazionale di Astrofisica)

The Square Kilometre Array (SKA) is an ambitious global science and engineering project aimed at building one of the most impressive astronomical infrastructures ever built. Currently in the construction phase, the SKA will be a network of radio telescopes, distributed over two continents, characterized by high sensitivity and resolving power and by the application of innovative technologies to the development of receivers, to the transport and processing of signal and calculation.

Italy, through the National Institute of Astrophysics (INAF), has been one of the major players in the pre-construction phase, participating in the first line in the design, and actively contributes to the definition of scientific cases through the SKA Science Working Groups.

Even before the SKA comes online, a series of demonstrator telescopes and systems, known as pathfinders and precursors, are already operational or under development across the world, paving the way for the kinds of technology which the SKA will need to pioneer to make the huge data available to scientists.

In this talk information about the project and on its status will be provided. In the framework of the preparatory activity for the SKA, I will also present some scientific highligths from SKA precursors, with particular regards to ASKAP and MeerKAT in the field of Galactic Radioastronomy.

33

Machine Learning developments for Radio Astronomy in the SKA era

Invited Speaker: David Cornu (LERMA, Observatoire de Paris)

In the context of the preparation of the first SKA observations in a few years from now, more and more public surveys based on precursor and pathfinder instruments are getting released. The analysis of the resulting datasets can already be very challenging as they are getting closer to real upcoming SKA observations. Even if the type of task to perform on such datasets is often rather classical (detection, classification, denoising, etc.), they have become heavily demanding for classical approaches due to datasets size and dimensionality. It is not a surprise then, that many astronomers started to focus their work on Machine Learning approaches that demonstrated their efficiency in similar applications. However, radio-astronomical images are very different from images used to train state-of-the-art pattern recognition algorithms. Moreover, astronomers have specific expectations for the predicted results, be it in terms of robustness, reproducibility, or explainability. As a direct consequence, these methods do not always perform as well as expected when directly applied to astronomical datasets. For this reason, astronomers have started to propose in-depth modifications of widely adopted approaches and also have initiated the development of new dedicated methods.

In this talk, I will present an overview of several Machine Learning approaches that have successfully been employed for HI analysis. I will describe methods for a variety of tasks including image denoising, foreground removal, model inversion, etc. but I will put the emphasis on galaxy detection, classification, and characterization techniques, which is a necessary preliminary task for the vast majority of studies. For this, I will present a technical overview of different Machine Learning methods that have been developed by various teams that participated in the SKAO Science Data Challenge 2, which consisted of a 3D detection and characterization task inside a 1TB simulated cube of HI emission. I will discuss the main difficulties identified that are specific to this type of dataset and the various tweaks used by the different teams to mitigate them. I will also discuss what could be future technical developments for these methods in order to overcome the remaining difficulties. Finally, I will present some of the efforts being made in the application of these approaches on pathfinder and precursor instruments, along with the additional difficulties that arise, like the proper definition of learning samples.

SKA and Precursors / 37

Deep Learning Processing and Analysis of Mock Astrophysical Observations

Speaker: Claudio Gheller (Istituto Nazionale di Astrofisica)

The challenge facing astronomers in the upcoming decade is not only scientific, but also technological.
A flurry of complex data will be delivered by new telescopes such as SKA ant its precursors and pathfinders, Athena ecc. This data will be difficult to manage with traditional approaches. Data will have to be stored in dedicated facilities, providing the necessary capacity at the highest performance. Corresponding data processing will have to be performed local to the data, exploiting available high performance computing resources. Data reduction and imaging software tools will have to be adapted, if not completely re-designed, in order to efficiently run at scale. Fully automated pipelines will
be a compelling requirement for effective software stacks as the richness and complexity of incoming data will inhibit human interaction and supervision.

We have explored the potential of Machine Learning to process and analyse astrophysical data, addressing challenges like the identification and segmentation of low brightness extended sources in noisy images coming from simulated radio and X-ray observations and their denoising in order to extract interesting features with a minimum impact on their intrinsic brightness. Segmentation has been accomplished using a both a Convolutional Neural Network and a U-net based approaches, while denoising exploited a Convolutional Autoencoder methodology.

We present the usage of the different approaches and their effectivness in the repective tasks, highlighting advantages but also drawbacks and limitations. We also discussthe computational performance, addressing, in particular the usage of modern, hybrid HPC systems, in order to reduce the training time and to support larger datasets.

SKA and Precursors / 38

Deep neural networks for source detection in radio astronomical maps

Speakers: Daniel Magro (University of Malta), Renato Sortino (Istituto Nazionale di Astrofisica)

Source finding is one of the most challenging tasks in upcoming radio continuum surveys with SKA precursors, such as the Evolutionary Map of the Universe (EMU) survey of the Australian SKA Pathfinder (ASKAP) telescope. The resolution, sensitivity, and sky coverage of such surveys is unprecedented, requiring new features and improvements to be made in existing source finders. Among them, reducing the false detection rate, particularly in the Galactic plane, and the ability to associate multiple detected islands into physical objects. To bridge this gap, we developed a new source finder, based on the deep learning Mask R-CNN framework, capable of both detecting, classifying, and segmenting/masking compact sources, radio galaxies, or imaging sidelobes in radio images. The model was trained using ASKAP data, taken during the Early Science phase, and previous radio survey data. The final model achieves Reliability (Precision) above 66% and Completeness (Recall) above 86% on sources and galaxies. This results in an F1 Score of 0.75 across all object classes.

SKA and Precursors / 40

A machine learning classifier for LOFAR radio galaxy cross-matching techniques

Speaker: Lara Alegre (University of Edinburgh)

New-generation radio telescopes like LOFAR are conducting extensive sky surveys, detecting tens of millions of radio sources. To maximise the scientific value of these surveys, radio source components must be properly associated into physical sources before being cross-matched with their optical/infrared counterparts. We use machine learning to identify those radio sources for which either source association is required or statistical cross-matching to optical/infrared catalogues is unreliable. We train a binary classifier using manual annotations from the LOFAR Two-metre Sky Survey (LoTSS) first data release. We find that, compared to a classification model based on just the radio source parameters, the addition of features of the nearest-neighbour radio sources, the potential optical host galaxy, and the radio source composition in terms of Gaussian components, all improve model performance. Our best model, a gradient boosting classifier, achieves an accuracy of 95 per cent on a balanced dataset and 96 per cent on the whole (unbalanced) sample after optimising the classification threshold. Unsurprisingly, the classifier performs best on small, unresolved radio sources, reaching almost 99 per cent accuracy for sources smaller than 15 arcseconds, but still achieves 70 per cent accuracy on resolved sources. It sends 68 per cent more sources than required to visual inspection, but this is still fewer than the manually-developed decision tree used in LoTSS, while also having a lower rate of wrongly accepted sources for statistical analysis. The results have an immediate practical application for the cross-matching of LoTSS second data release and can be generalised to other radio surveys.

SKA and Precursors / 39

Analysis of Instance Segmentation Networks for Dispersed FRB Searches

Speaker: Anastasia Seifert (University of Malta)

Next generation telescopes such as the SKA will be collecting a substantial amount of radio data in future surveys. The number of Fast Radio Burst (FRB) events are expected to increase as a result of the SKA having a higher sensitivity and resolution; thus increasing the volume of data collected which will consequently need to be analyzed and processed. Therefore, the previous methods of FRB searches will no longer be efficient as these processes will need to rival the incoming data rates. Recently, neural networks have been applied to classify or detect radio objects such as FRBs and pulsars. These methods were also found to be faster than previous search methods with high precision and accuracy, particularly with the use of GPUs rather than CPUs. The dedispersion step of the FRB search process, which is necessary to correct the propagation effects from the Interstellar Medium, is the most computationally expensive stage and the use of neural networks may be beneficial in the identification of FRBs without needing to dedisperse. However in order to attain the correct dispersion measure (DM), dedispersion should still occur at a later stage. The aim of our investigation will explore the application of state of the art instance segmentation and object detection algorithms with regards to simulated dispersed FRBs. For this study, a simulated FRB dataset was generated using the Injectfrb software with observed parameters selected from the FRB Catalogue and observation parameters specific to the Green Bank Telescope (GBT). Using the pulsar analysis program Sigproc, Gaussian noise was simulated in filterbank format and was used as background noise for the FRB to be injected into. The generated dataset is used to train and evaluate various instance segmentation models and the performance of these models are evaluated based on community standard metrics. The results of our models are compared to existing methods of searching for FRBs in noisy radio telescope streams.

SKA and Precursors / 41

Deep learning 21cm light-cones in 3D

Speaker: Caroline Heneka (UHH)

Interferometric measurements of the 21cm signal with the Square Kilometre Array are a prime example of the data-driven era in astronomy and astrophysics we are entering with current and upcoming experiments. To optimally learn the Universe from low to high redshift I advocate for the use of multiple lines (multi-line intensity mapping) and complementary galaxy survey data, as well as the development of well-tailored modern machine learning techniques to increase information content inferred and its robustness. Tomography of 21cm intensity maps targeted by SKA-LOW will teach about source properties, IGM state and cosmology during the epoch of reionisation, while imaging with SKA-MID can tell about HI galaxy properties at lower redshifts. In this talk I firstly showcase the use of deep networks that are tailored for the structure of tomographic 21cm light-cones of reionisation and cosmic dawn to directly infer e.g. dark matter and astrophysical properties jointly without an underlying Gaussian assumption. I compare different architectures and highlight how a comparably simple 3D network architecture (the 3D-21cmPIE-Net) that mirrors the data structure as the best-performing model. I present well-interpretable gradient-based saliency maps and discuss robustness against foregrounds and systematics via transfer learning. I discuss first findings on reliable error calibration on the way to a 3D Bayesian network. I complement these findings with a discussion of lower redshift results for the recent SKA Science Data Challenge 2, where hydrogen 21cm sources where to be detected and characterised in a large (TB), again 3D, cube. I will highlight my team's lessons-learned on the use of machine learning methods for such data, where our networks performed especially well when asked to characterise flux and size of sources bright in 21cm.

Poster Session Day 3.1 / 85

Effects of incompleteness in the training sample for photoz estimation by DNF algorithm

Speaker: Laura Toribio (CIEMAT)

One of the crucial keys in the cosmological studies is the estimation of an accuracy redshift of a large number of galaxies. Sometimes, the spectroscopic sample used as training sample for ML approaches doesn't cover the same magnitude and color space as the target sample. This issue raises doubts about the confidence of the photometric redshift provided by the algorithms.

In this talk, we present the effect of using complete or incomplete spectroscopic training samples to determine the photo-zs by DNF algorithm. We compare the photo-zs estimate for the validation sample using both training samples. We provide a new method for determining the level of confidence in the photo-z values and the incompleteness assessment of the results. Finally, we compare the DNF photo-zs with templates methods.

Poster Session Day 3.1 / 121

Machine learning investigations for LSST: Strong Lens Mass Modeling and Photometric Redshift estimation

Speaker: Stefan Schuldt (MPA/TUM)

Strong lensing analyses and photometric redshifts are both integral for cosmological studies with the Rubin Observatory Legacy Survey of Space and Time (LSST). With convolutional neural networks, we can obtain for both significant gain in the performance and speed. In my talk, I will present the new achievements of the HOLISMOKES collaboration in modeling galaxy-scale lenses using a residual neural network that was trained, validated and tested on very realistic mocks. This network enables us to predict the mass model parameters values with uncertainties, allowing us to analyze the huge amount of lenses soon to be discovered by LSST and predicting in advance the next appearing image and corresponding time delays in case of a lensed transient such as a supernova. I will further present our dedicated model comparison of 32 real lenses, once obtained with the network and once with traditional techniques, for which we developed an automation procedure to minimize the user input time drastically. Furthermore, I will briefly highlight our newly developed method NetZ to predict the photo-z based only on the pure galaxy images. Both networks are able to estimate the parameter values in fractions of a second on a single CPU while the lens modeling with non-automated traditional techniques takes typically weeks to month. With both networks and also our automated traditional pipeline, we are ready to process the huge amount of images obtained with LSST in the near future.

Poster Session Day 3.1 / 65

Detection of point sources in the CMB intensity using machine learning techniques

Speaker: Patricia Diego-Palazuelos (Instituto de Fisica de Cantabria, CSIC-UC)

To obtain a clean measurement of the Cosmic Microwave Background (CMB), we must first separate its signal from all the other galactic and extragalactic components of the sky. In this way, extragalactic sources emitting in the microwave range, like radio-loud active galactic nuclei and dusty galaxies, constitute a contaminant that appears in the form of unresolved point-like objects in CMB maps due to the limited resolution of most CMB experiments.

As an alternative to traditional methods, we address the blind detection of point sources on maps of the CMB intensity from a machine learning standpoint. We approach source detection as an image segmentation problem and design a convolutional neural network (CNN) able to successfully solve it while maintaining a simple architecture. By treating source detection as a binary segmentation operation, we manage to decouple the process of localization from flux estimation. This allows the CNN to outperform conventional detection algorithms based on optimal flux estimators like the matched filter. Preliminary results confirm that our CNN yields higher completeness at all fluxes than the matched filter when applied to CMB and instrumental noise simulations.

Recognizing its potential, we prepare the CNN for the application to data by including Galactic foregrounds in the training process. Galactic foreground emission presents an anisotropic and complex structure that vastly dominates over extragalactic sources and the CMB in most of the sky. We address this additional level of complexity by dividing the sky into separate regions of progressively increasing foreground intensity and independently training specialized CNNs for each of them. However, such partition of the sky greatly reduces the volume of data available for training in each region, which severely hinders the CNN generalization ability and eventually leads to overfitting. Preliminary results show that this problem can be solved through data augmentation and dropout techniques.

With overfitting under control, we are ready to extend the training to different frequency bands and eventually apply the CNN to data to produce deeper and more complete extragalactic source catalogs.

Poster Session Day 3.1 / 168

Discovering factors that determine dark matter halo abundance with interpretable deep learning

Speaker: Ningyuan Guo (UCL)

The halo mass function describes the abundance of dark matter halos as a function of halo mass and depends sensitively on the cosmological model. Accurately modelling the halo mass function for a range of cosmological models will enable forthcoming surveys such as Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST) to place tight constraints on cosmological parameters. Due to the highly non-linear nature of halo formation, understanding which quantities determine the halo mass function for different cosmological models is difficult. We present an interpretable deep learning framework that allows us to find, with minimal prior assumptions, a compressed representation of the information required to predict the halo mass function. We apply this framework to investigate whether in addition to peak height, information on growth history is required to accurately model the halo mass function. We use neural network models that consist of an encoder-decoder architecture: the encoder compresses the input linear matter power spectrum and growth function into a low-dimensional representation, and the decoder uses this representation to predict halo abundance given a halo mass and redshift. We interpret the representation by quantifying mutual information between the representation and the ground truth halo number densities. This can enable us to gain new insights on what physics is involved in the process of halo formation, and a better understanding of how to accurately model the halo mass function for different cosmological models.

Poster Session Day 3.1 / 75

New applications of Graph Neural Networks in Cosmology

Speaker: Farida Farsian (University of Bolonia)

The Golden age of Cosmology with high precision surveys is highly entangled with the so-called "Big-Data" era, as the future astrophysical and cosmological experiments will produce massive amount of data. Therefore, new methodologies, such as Artificial Intelligence and Deep Learn-ing, should come to play to handle the computational expensive operations, automation, and to extract non explored data features and statistics. On the other hand, there are plenty of cosmo-logical models, specifically in the Dark Energy field, which have to be tested and explored. The latter would be essential in order to analyse the data provided by the ESA Euclid satellite which will be launched in 2023.

Up to now, standard cosmological analyses based on abundances, two-point and higher-order statistics have been widely used to investigate the properties of the Cosmic Web. However, these statistics can only exploit a sub-set of the whole information content available. Along these lines, we are studying a new description of cosmic web data in the form of graphs, in which the clustering information of the matter density field would be automatically included. This form of data can be fed to Graph Neural Networks (GNN).

In this talk, we will present a new application of GNN in the large-scale structure field. We will show that by making use of raw dark matter halo catalogues, considering only mass and coor-dinates as features, the GNN can robustly discriminate among different dark energy models. Specifically, we will show the results obtained by applying the GNN on different simulated halo catalogues with various dark energy equation parameters.

Poster Session Day 3.1 / 130

Inferring the Dark Matter halo mass in galaxies from other observables with Machine Learning

Speaker: Carlo Cannarozzo (UNAM)

In the context of the galaxy-halo connection, it is widely known that the dark matter (DM) halo of a galaxy exhibits correlations with other physical properties, like the well-studied stellar-to-halo-mass relation. However, given the complexity of the problem and the high number of galaxy properties that might be related to the DM halo in a galaxy, the study of the galaxy-halo connection can be approached relying on machine learning techniques to shed light on this intricate network of relations. Hence, with the aim of inferring the DM halo mass and then finding a unique functional form able to link the halo mass to other observables in real galaxies, in this talk I will present the results of this project obtained relying on the state-of-the-art Explainable Boosting Machine (EBM) algorithm, a novel method with very high accuracy and intelligibility that exploits some machine learning techniques like boosting or bagging in the field of the generalized additive models with pairwise interactions (GA^{2}M). Unlike a simple GAM, EBM is an additive model that makes final predictions as a summation of *shape functions* of each individual feature, considering also any possible pairwise interaction between two features. I will illustrate an analysis performed on a sample of galaxies at different redshifts extracted from the IllustrisTNG cosmological simulation. This method has been proving to be very promising, finding, at all redshifts, a scatter of < 0.06 dex between the actual value of $M_{\rm DM}$ from the simulation and the value predicted by the model.

This project is supported by the *Marie Skłodkowska-Curie grant* received in 2019, and financed by the *Horizon2020-MSCA-RISE-2017 Project 777822 GHAIA "Geometric and Harmonic Analysis with Interdisciplinary Applications"*.

Poster Session Day 3.1 / 49

Mimicking the halo-galaxy connection using machine learning

Speaker: Natali Soler Matubaro de Santi (University of Sao Paulo)

As far as we know, galaxies form inside dark matter halos and elucidating this connection is a key element in theories of galaxy formation and evolution. In this work, we propose a suite of machine learning (ML) tools to analyze these intricate relations in the IllustrisTNG300 magnetohydrodynamical simulation. We apply four individual algorithms: extremely randomized trees (ERT), K-nearest neighbors (kNN), light gradient boosting machine (LGBM), and neural networks (NN). Moreover, we combine the results of the different methods in a stacked model. In addition, we apply all these methods in an augumented dataset using the synthetic minority over-sampling technique for regression with Gaussian noise (SMOGN), in order to alleviate the problem of unbalanced datasets, and show that it improves the shape of the predicted distributions. Overall, the all the ML algorithms produce consistent results in terms of predicting central galaxy properties from a set of input halo properties that include halo mass, concentration, spin, and halo overdensity. For stellar mass, the (predicted vs. real) Pearson correlation coefficient is 0.98, dropping down to 0.7-0.8 for specific star formation rate, colour, and size. We also demonstrate that our predictions are good enough to reproduce the power spectra of multiple galaxy populations, defined in terms of stellar mass, sSFR, colour, and size with high accuracy. Our analysis adds evidence to previous works indicating that certain galaxy properties cannot be reproduced using halo features alone.

Poster Session Day 3.1 / 128

Inference of galaxy cluster mass profile using deep learning

Speaker: Alessia Sbriglio (Università degli studi di Roma)

Clusters of galaxies are the largest gravitationally bound systems in the Universe resulting from the natural evolution of cosmic structures. They are crucial tracers of the structure formation history and their mass function at different epochs is of key importance to constrain cosmological parameters. Therefore, it is essential to infer the mass of the observed clusters, which unfortunately is not a direct observable and is affected by different biases related to the applied observational estimates.To overcome these obstacles we exploit a modern method, provided by machine learning algorithms, that turn out to outperform conventional statistical methods. In a previous work, Convolutional Neural Networks (CNNs) were applied to Compton-y parameter maps from the Planck satellite, to estimate the masses of clusters defined at a fixed aperture radius corresponding to 500 times critical overdensity . We now extend this study to estimate the radial profiles of the cluster total mass in order to compare them with real observation.In our case, we make use a deep learning architecture based on Autoencoders to find the most efficient compact representation of the input data. The training of the architecture is performed on mock images of the Sunyaev-Zel'dovich signal generated by a large set of hydrodynamically simulated galaxy clusters from the "THE THREE HUNDRED" project.

34

HPC-cloud convergence is the missing link between scientific computing and applied-AI

Invited Speaker: Marco Aldinucci (Università di Torino)

First, HPC infrastructures are embracing GPUs for their superior performance-per-watt ratio against general-purpose multicores. Second, the next-generation scientific workflows are integrating AI-based steps for their accuracy in approximating and analyzing complex phenomena. Third, AI and specifically Machine Learning (ML), is a perfect workload for GPUs in terms of performance and development time. Today, we cannot still close the circle seamlessly running AI-enabled scientific workloads into HPC infrastructures because their system software and development tools are not designed for modern workloads, such as ML frameworks designed for the cloud. HPC-cloud convergence is likely to bridge the gap. In the talk, we will present Streamflow and CAPIO, two development tools for HPC-cloud convergence.

Deep Learning for Astroparticle Physics / 35

ML Event Reconstructions for Neutrino Telescopes

Speaker: Philipp Eller (TU Munich)

Neutrino telescopes, such as IceCube, KM3NeT or GVD, consist of thousands of photo sensors distributed over cubic kilometer volumes. The Cherenkov light from particle interactions within create signals of varying shapes and sparsity. For the reconstruction of such physics events, one aims to infer quantities like the interaction vertex, the deposited energy, the angles, and the topology of the interaction. This reconstruction step is of central relevance to many physics analyses and searches, and improvements in both accuracy and speed have a direct, positive impact on the science. This talk will present two novel event reconstruction algorithms based on machine learning (ML). One of our algorithms is based on graph neural networks, that are capable of accurate reconstruction at very high speeds, even applicable to online event processing. Our second approach is a hybrid likelihood-ML method using a likelihood decomposition and a "ratio estimator trick" to arrive at a fast and precise approximation of the true reconstruction likelihood. This allows to employ frequentist and Bayesian inference techniques alike. We will round off the talk with performance figures and comparisons.

Deep Learning for Astroparticle Physics / 36

Gamma/hadron separation for the MAGIC Imaging Atmospheric Cherenkov Telescopes through Convolutional Neural Networks trained with real data samples

Speaker: Stefano Truzzi (University of Siena/INFN Pisa)

Imaging Atmospheric Cherenkov Telescopes (IACT) are able to indirectly detect gamma rays from the ground with energies beyond several tens of GeV emitted by the most energetic objects known, including Pulsar Wind Nebulae, Active Galactic Nuclei, and Gamma-Ray Bursts. Gamma rays and cosmic rays are detected by IACTs through imaging, the Cherenkov light produced by the charged superluminal leptons in the extended air shower of secondary particles originated when the primary particles interact with the atmosphere. These Cerenkov flashes dominate the light of the night sky for short integration times at the nanosecond scale. From the image topology and other observables, gamma rays can be separated from the more numerous cosmic rays, and thereafter incoming direction and energy of the primary gamma rays can be reconstructed. The standard algorithm in MAGIC data analysis for the gamma/hadron separation is a Random Forest, working on a parametrization of the stereo events based on the Hillas parameters. Until a few years ago, such a treatment was also restricted by the limited computational resources. Modern IT resources, such as GPUs make it possible to work directly on the pixel-wise information. Most CNN applications in the field perform the training on Monte Carlo simulated data for the gamma-ray sample. This choice is prone to systematics arising from discrepancies between observational data and simulations. Instead, we trained a well-known Deep CNN scheme, the Inception ResNet V2, with real data from a giant flare of the bright TeV blazar Mrk 421 observed by MAGIC in 2013. We show the preliminary results of this method for the gamma/hadron separation, which we found able to rival the results of the standard MAGIC analysis based on Random Forest, but showing potential for further improvement as well.

Outreach and Citizen Science / 42

GWitchHunters - Machine Learning and Citizen Science to support Gravitational Wave detection

Speaker: Massimiliano Razzano (University of Pisa and INFN-Pisa)

Gravitational waves opened a new window on the Universe and paved the way to a new era of multimessenger observations. Ground-based detectors such as Advanced LIGO and Virgo have been extremely successful in detecting gravitational wave signals from the coalescences of black holes and/or neutron stars. In order to improve over the actual sensitivities, the background noise must be investigated and removed. Transient noise events called "glitches" can affect data quality and mimic real astrophysical signals, and it is therefore of paramount importance to characterize them and find their origin, a task that will support the activities of detector characterization of Virgo and other interferometers. Machine learning offer a promising approach to characterize and remove noise glitches in real time, thus improving the sensitivity of interferometers. A key input to the preparation of a training datasets for these machine learning algorithms can originate from citizen science initiatives, where volunteers contribute to classify and analyze signals collected by detectors. We will present GWitchHunters, a new citizen science project focused on the study of gravitational wave noise, that has been developed within the REINFORCE project (a "Science With And For Society" project funded under the EU's H2020 program). We will present the project, its development and the key tasks that citizens are participating in, as well as its impact on the study and characterization of noise in the Advanced Virgo detector.

Outreach and Citizen Science / 43

Citizen Science and Machine Learning: Towards a Robust Large-Scale Automatic Classification in Astronomy

Speaker: Manuel Jimenez (Instituto de Astrofísica de Andalucía, IAA-CSIC)

Citizen science, traditionally known as the engagement of amateur participants in research, is demonstrating a great potential for large-scale processing of data. Using the power of the web, virtual communities of volunteers have been able to coordinate the classification of hundreds of thousands of images in a reasonable amount of time. In areas such as astronomy or geo-sciences, where emerging technologies generate huge volumes of data, this approach entails image classification at a rate not possible to accomplish by experts alone, although at the expense of worse quality in the classifications made by amateur participants. Despite its success in astronomy, as evindenced by the numerous editions of the Galaxy Zoo project, the current and upcoming massive surveys highlight its limitations, and the inclusion of machine learning methods towards a more robust automatic classification is considered mandatory. However, current efforts attempting the exploitation of citizen science outcomes with machine learning tools have ignored their inherent uncertainty as well as the potential of expert classifications to ameliorate this issue. Their ultimate goal has mainly been to replicate the amateur performance, thus propagating their biases and limitations and disregarding the fact that, apart from the data labelled by amateurs, there is also available (limited) expert knowledge of the problem along with vast amounts of unlabelled data that have not been exploited yet within a unified learning framework.

Our research delves into the development of automated approaches for astronomical classification problems that have been aided by citizen science projects on the web, aiming to leverage the inherent uncertainty in their results and all levels of knowledge available about the problem. We introduce an innovative learning paradigm for citizen science projects in astronomy capable of taking advantage of expert- and amateur-labelled images, and unlabelled images. As an implementation of this learning framework, we present the Citizen Science Learning (CzSL), an algorithm that first learns from unlabelled data with a convolutional autoencoder, and then exploits amateur and expert labels via the pre-training and fine-tuning of a convolutional neural network, respectively. As a case study, we focus on the classification of galaxy images from the first edition of the Galaxy Zoo project, from which we test binary, multi-class, and imbalanced classification scenarios, although the methodology is not limited to any classification problem in particular. Our results demonstrate an improved classification performance in comparison to a representative set of baseline approaches, showing a more comprehensive use of these resources that are available to the astronomical research community.

Poster Session Day 3.2 / 81

matryoshka: A suite of neural network based emulators for the power spectrum.

Speaker: Jamie Donald-McCann (University of Portsmouth)

Cosmological inference can be a computationally expensive task. Calculation of posterior distributions often requires sampling form a high dimensional parameter space, with a large number of nuisance parameters. Typically analyses require hundreds of thousands or even millions of model evaluations. Therefore, even analyses that use the most efficient perturbative models for predicting the power spectrum still require significant resources. Emulators offer a solution since they can be trained to accurately reproduce expensive model outputs whilst greatly reducing computational cost. In this talk I will present a publicly available, Python implemented, suite of neural network based emulators: matryoshka. I will present example analyses using some of the emulators included in matryoshka, and demonstrate that when using these emulators cosmological inference can be done in a coffee break rather than days or even weeks.

Poster Session Day 3.2 / 119

Integration and deployment of Model Serving Framework to serve machine learning models at production scale in easy way.

Speaker: Francesco Caronte (Altecspace)

With the help of machine learning systems, we can examine data, learn from that data and make decisions. Now machine learning projects have become more relevant to various use case, but too many models are difficult to manage. For this reason several MLOps tools were born. These tools are the main platforms, hosting the full machine learning process lifecycle, starting with data management and ending with model versioning and deployment.
NEANIAS (Novel EOSC Services for Emerging Atmosphere, Underwater & Space Challenges) is an ambitious project that comprehensively addresses the challenges set out in the 'Roadmap for EOSC' foreseen actions.
NEANIAS drives the co-design and delivery of innovative thematic services, derived from state-of-the-art research assets and practices in three major sectors: Underwater research, Atmospheric research and Space research.
The machine learning core services identified in the NEANIAS Project allow the scientists to define machine learning models and manage their lifecycle.
The Machine learning core services is composed by a JupyterHub instance (C3.1) that enable different profile server:
(C3.2) Serve machine learning model in production enviroment.
(C3.3) Do deep learning training framework using a Horovod cluster.
(C3.4) Perform distributed calculation using Apache Spark.
All these services combined and integrated are used for astrophysics use cases. In particular, two environments including two deep learning models tailored to object detection (based on maskRCNN) and semantic segmentation (based on Tiramisu) and trained on radio dataset from the Square Kilometre Array (SKA) precursors were integrated in C3.1 to classify radio astronomical sources, galaxies and image artefacts.
The "C3.2 - Model Serving" is a set of applications, part of Neanias ML core services, that is employed to simplify machine-learning model deployment and to run high-performance model serving at scale. This makes it easy to turn ML models into prediction services that are easily deployable on container platform. BentoML is the software identified to satisfy these goals. In the project described in this paper, It has been improved in the frame of the NEANIAS project to increase the interoperability of the applications.

"C3.2 - Model Serving" works on top of Kubernetes platform, provide REST API documented with Swagger and use GRPC proxy to guarantee authentication and authorization at the model saved and deployed.

Poster Session Day 3.2 / 116

Predictive Maintenance for a Cherenkov Telescope array

Speaker: Salvatore Gambadoro (Unict-INAF)

Cherenkov telescope arrays are equipped with a multitude of sensors spread all over the instrumentation and collect a large volume of housekeeping and auxiliary data coming from telescopes, weather stations and other devices in the array site.
In this poster we will present how we intend to exploit the sensor's information, together with the most advanced artificial intelligence algorithms, to perform predictive maintenance (PdM). This technique will be useful to detect in advance the remaining useful life of the array components, and to estimate the correct timing for performing their maintenance. The application of PdM will allow to minimize the array downtime, to increase the telescopes sub-components longevity, and to reduce the costs due to unforeseen maintenance. Our model will be trained and tested with time series data coming from a number of different sensors (temperature, current, torque, etc.) dedicated to monitor several mechanical components of the telescopes (engines, cameras, encoders, etc.). The adopted supervised machine learning approach will allow us to perform the correct trade-off between preventive and corrective maintenance.

Poster Session Day 3.2 / 77

Reconstruction and Particle Identification with CYGNO detector

Speaker: Atul Prajapati (Gran Sasso Science Institute-INFN)

CYGNO is developing a gaseous Time Projection Chamber (TPC) for directional dark matter searches, to be hosted at Laboratori Nazionali del Gran Sasso (LNGS), Italy. CYGNO uses He:CF4 gas mixture at atmospheric pressure and relies on Gas Electron Multipliers (GEMs) stack for the charge amplification. Light is produced by the electrons avalanche thanks to the CF4 scintillation properties and is then optically read out by a high-resolution scientific CMOS Camera (sCMOS) and Photo-Multiplier Tubes (PMT). sCMOS are designed for low readout noise, uniformity, and linearity and are therefore capable to track particles down to O(keV) energies. These high-resolution images (2D event projection) are combined with the PMT signal (relative z coordinate information) to obtain 3D reconstruction, with the aim of particle identification and to determine track direction of arrival.
sCMOS images are well suited to be analyzed with Deep Learning techniques (using Convolutional Neural Networks (CNNs), Deep Neural Networks (DNNs)) because of their high granularity and low noise. We will present the CYGNO features and achieved experimental performance, and then focus on the MonteCarlo sCMOS images simulation, reconstruction performed with density-based algorithms (DBSCAN) and Geodesic Active Contour (GAC) clustering algorithm to identify and track particles. Using the morphological features in the track from the reconstructed data, Deep Leaning Models like Gradient Boosted Decision Trees (GBT), Random Forest Classifier (RFC), and Deep Neural Network (DNN) are employed to classify the tracks into different energy classes of Electron recoils and Nuclear Recoils. Future

work focuses on the use of a CNN-based model for track reconstruction (instance segmentation) and classification.

Poster Session Day 3.2 / 102

Deep neural networks for single-line event direction reconstruction in ANTARES

Speaker: Juan Garcia Mendez (Universidad Politécnica de Valencia)

ANTARES, the first large undersea neutrino telescope, has recently stopped taking data after nearly 16 years of operation. ANTARES consisted of 12 vertical lines forming a 3D array of photo-sensors, instrumenting about 10 megatons of Mediterranean seawater to detect Cherenkov light induced by secondary particles from neutrino interactions.

The event reconstruction and background discrimination is challenging. Here, we present a method based on deep learning aimed to improve the direction reconstruction of low-energy single-line events, for which the reconstruction of the azimuth angle of the incoming neutrino is particularly difficult. Our results show a promising improvement in resolution over former reconstruction techniques, at least doubling our sensitivity in the low energy range, which is highly relevant for dark matter searches and other physics studies.

The network was trained with Monte Carlo simulations of track-like events: (anti)muon-neutrino interactions. The dataset was randomly divided in three parts: training, validation (to avoid overfitting through early stopping), and test (to evaluate the generality of the results). The architecture of the neural network combined the properties of Deep Convolutional Networks (DCNs) with those of Mixture Density Networks (MDNs): the convolutional layers made the initial processing of the inputs, whereas using MDNs in the output layer allowed predicting not only the direction angles but also their uncertainty. Results of the deep neural network on the test set showed a mean deviation upon the true value of 7.4° for the zenith angle reconstruction, in contrast to 15.5° using former methods. We also computed the estimation of the azimuth angle, which was missing previously for single-line events due to low precision. The results of (i) randomized, and (ii) date-sorted, K-fold cross-validation (with k=5) confirmed the validity and robustness of the approach.

To confirm the viability of the tool in real data analyses, we ran additional tests with simulated data of different sorts. First, we checked the behavior against background noise. As input, we used information from a random single line other than that of the single-line event. As expected, predictions showed a large deviation from the true value in the Monte Carlo simulation. More importantly, this test confirmed the generality of the uncertainty output reflecting the precision of predictions. Second, different types of simulated events, atmospheric muons and shower events, were tested. We confirmed accurate predictions for their direction and uncertainty, despite these events were not explicitly trained. Moreover, these results showed the power of the method in discriminating muon-related events into two subtypes based on direction criteria: muons originated from up-going neutrinos crossing the Earth vs. down-going muons directly detected after they cross the atmosphere. Finally, we checked that zenith and azimuth distributions, and their uncertainties, were in good agreement with those from real data events, especially when cutoffs based on quality parameters were applied. Thus, we conclude that this method is ready to be applied in physics analyses.

Poster Session Day 3.2 / 56

Machine-learning based detector optimization of the future P-ONE neutrino telescope

Speaker: Christian Haack (TU Munich)

P-ONE is a planned cubic-kilometer-scale neutrino detector in the Pacific ocean. Similar to the successful IceCube Neutrino Observatory, P-ONE will measure high-energy astrophysical neutrinos to help characterize the nature of astrophysical accelerators. Using existing deep-sea infrastructure provided by Ocean Networks Canada (ONC), P-ONE will instrument the ocean with optical modules - which host PMTs as well as readout electronics - deployed on several vertical cables of about 1km length. While the hardware design of a first prototype cable is currently being finalized, the detector geometry of the final instrument (up to 70 cables) is not yet fixed. Traditionally, the detector geometry would be optimized using a large-scale simulation campaign, which results in detector resolutions only for discrete points in the geometry phase space.
Recently, a new approach for detector optimization is emerging: Using ML-based surrogate models, a differentiable parameterization of the expected detector response is obtained. This model is then used to optimize the detector design with respect to the expected detector resolution and acceptance.
In this talk, I will discuss the prospects and current state of applying an ML-based optimization approach to P-ONE.

Poster Session Day 3.2 / 68

Cats and Dogs vs Gamma and Hadrons

Speaker: Francesco Visconti (Istituto Nazionale di Astrofisica)

A simple Convolutional Neural Network architecture suitable for cats and dogs classification can discriminate gammas from hadrons on Montacarlo data for a single ASTRI telescope, much better than classic methods based on Hillas parameters.

Poster Session Day 3.2 / 141

Background discrimination for DSNB detection using high-resolution convolutional neural networks

Speaker: David Maksimovic (Johannes Gutenberg-University Mainz)

The Diffuse Supernova Neutrino Background (DSNB) is the faint signal of all core-collapse supernovae explosions on cosmic scales. A prime method for detecting the DSNB is finding its inverse beta decay (IBD) signatures in Gadolinium-loaded large water Cherenkov detectors like Super-Kamiokande (SK-GD).Here, we report on a novel machine learning method based on Convolutional Neural Networks (CNNs) that offer the possibility for a direct classification of the PMT hit patterns of the prompt events.

While the enhanced neutron tagging capability of Gadolinium greatly reduces single-event backgrounds, correlated events mimicking the IBD coincidence signature remain a potentially harmful background. Especially in the low-energy range of the observation window, Neutral-Current (NC) interactions of atmospheric neutrinos dominate the DSNB signal, which leads to an initial signal-to-background (S:B) ratio inside the observation window of about 1:10.

Based on the events generated in a simplified SK-GD-like detector setup, we find that a trained CNN can maintain a signal efficiency of 96 % while reducing the residual NC background to 2 % of the original rate, corresponding to a final signal-to-background ratio of about 4:1. This provides excellent conditions for a DSNB discovery.