Experimental Astronomy (2022) 53:1–43
https://doi.org/10.1007/s10686-021-09827-4

REVIEW ARTICLE

Astronomical big data processing using machine learning: A comprehensive review

Snigdha Sen[1,2] · Sonali Agarwal[1] · Pavan Chakraborty[1] · Krishna Pratap Singh[1]

Received: 15 July 2021 / Accepted: 27 December 2021 / Published online: 14 January 2022

Abstract

Astronomy, being one of the oldest observational sciences, has collected a lot of data over the ages. In recent times, it is experiencing a huge data surge due to advancements in telescopic technologies with automated digital outputs. The main driver behind this article is to present various relevant Machine Learning (ML) algorithms and big data frameworks or tools being applied and can be employed in large astronomical data-set analysis to assist astronomers in solving multiple vital intriguing problems. Throughout this survey, we attempt to review, evaluate and summarize diverse astronomical data sources, gain knowledge of structure, the complexity of the data, and challenges in the data processing. Additionally, we discuss ample technologies being developed to handle and process this voluminous data. We also look at numerous activities being carried out all over the world enriching this domain. While going through existing literature, we perceived a limited number of comprehensive studies reported so far analyzing astronomy data-sets from the viewpoint of parallel processing and machine learning collectively. This motivated us to pursue this extensive literature review task by outlining up-to-date contributions and opportunities available in this area. Besides, this article also discusses briefly a cloud-based machine learning approach to estimate the extra-galactic object redshifts considering photometric data as input features. As the intersection of big data, machine learning and astronomy is a quite new paradigm, this article will create a strong awareness among interested young scientists for future research and provide an appropriate insight on how these algorithms and tools are becoming inevitable to the astronomy community day by day.

Keywords Astronomical big data · Distributed processing · Machine learning · Cloud computing

Sonali Agarwal, Pavan Chakraborty and Krishna Pratap Singh contributed equally to this work.

✉ Snigdha Sen
rwi2019003@iiita.ac.in

Extended author information available on the last page of the article.

1 Introduction

Astronomy and Astrophysics have been data-rich and with the advent of numerous large aperture ground-based telescopes, survey telescopes such as upcoming large synoptic survey telescope (LSST) and space-based telescopes [1], deeper and fainter objects are being observed. The high-resolution cameras and back- end instruments have further improved and automated the data acquisitions and it is expected to generate approximately a data of 15TB/night [2]. All this has led to an exponential rise in data collection. The storage of the data has been systematized and large data archives have been made for better accessibility. Therefore there is a strong demand for better, faster, and efficient data analysis and complex data analysis for appropriate decision-making. Therefore, a significant requirement of new algorithms and novel techniques to analyze [3] and understand that data-set arises.

The capabilities of the current traditional centralized [1] system of data processing are not adequate to handle, manage, and work with this large amassing data further. Moreover, identifying patterns hidden behind the data set is incredibly difficult for individual humans. Additionally, astronomical operations are very much compute-intensive and costly as well. Hence computer engineers and astronomers require to utilize the essence of parallel processing using big data techniques, embrace ML methods for quality prediction. These certainly help to characterize patterns and comprehend the universe better by extracting potential and useful information. This data analysis undeniably accelerates discoveries by leveraging advanced data mining tools, machine learning algorithms, and distributed frameworks.

ML algorithms not only automate the tasks by reducing human intervention but help to find a correlation between hidden patterns among complex astrophysical datasets. To perform better prediction, visualization, and build quick decision, ML is exceptionally important in this domain and have substantial potential to introduce inspiring developments and success. In this regard, data-driven astronomy ought to be explored, studied, and analyzed accurately to resolve various unanswered queries related to universe expansion, black holes, distance estimation of various celestial bodies, and many more issues. It can in turn benefit astronomers significantly and help in obtaining an unprecedented, unseen look of the universe. The role of these technologies needs to be proliferated further in this field which is still in a nascent stage in most places.

Mining vast data and obtaining valuable information is undoubtedly a difficult task for astronomers alone. Therefore, a collaboration of astronomy, information science, and machine learning experts is the need of the hour in the new interdisciplinary branch of Astro-informatics [4]. Plenty of researches are being carried out on distinct topics such as searching for an exoplanet, redshift estimation, classification of galaxies, and habitability score finding, etc across the world. In recent times, Wang et al. [5] explained the application of computational intelligence in astronomy mostly focusing on astronomical issues from the perspective of machine learning and fuzzy logic.

In addition to several ML application to astronomy, our article will explore and highlight the way how distributed data processing and cloud platform has huge potential to offer faster, scalable and novel solutions to astronomers for their future

research. In the end, we will be illustrating briefly the redshift estimation task using the cloud-based AWS SageMaker platform for bulk data processing. From the literature we believe, this is the first work discussed using AWS SageMaker for redshift analysis. Although there is quite a large number of the enriched research area in astronomy where ML, Big data concepts are being experimented, we explain some of them here to save space.

Further organization of our article is structured as follows. Section 2 explains advances in various machine learning methods, statistical approaches, and forward modeling techniques, etc. used in astronomy. In Section 3 we elaborate on big data concepts, astronomical big data challenges, data sources, and data formats. In Section 4, various data analysis tools have been illustrated. We discuss major activities and collaboration carried out in this field in Section 5. Section 6 presents precisely about redshift prediction task using cloud-based framework AWS Sagemaker. Lastly, we conclude our paper with prospective future scope and enhancement in this domain.

2 Machine learning approaches towards astronomical problems

Artificial intelligence (AI) techniques are helpful to mimic and incorporate human intelligence inside a computer system. The AI-assisted machine can think like a human and perform more efficiently than a human. Generally, ML is considered a subsection of AI. Fluke et al. [6] surveyed the importance and growing application of AI and ML in astronomy. Their research focuses on various ML algorithms used in classification as well as regression tasks for petabyte data analysis. Apart from working on the existing field, they investigated and identified some new emerging fields such as planetary studies, the non-stellar component of the milky way, and the stellar cluster that requires huge support from AI and ML. For deep space exploration and object detection AI-based approach has been described by Bird et al. [7]. Similarly, Ntampaka et al. [8] also discussed how new opportunities and challenges are waiting to use ML for the next decade cosmological application.

To deal with and analyze a huge dataset, a large number of diverse research have been conducted to propose many tools and techniques in this field. For example, a thorough survey and analytical study on the intersection of machine learning and physical sciences is presented by Carleo et al. [9]. They discussed several generative modeling and Generative Adversarial Networks (GAN) techniques to create new data instances similar to observed training data. GAN is a type of generative modeling that learns the pattern and irregularities from input data and generates new samples through the generator. Generator and discriminator are two important concepts in GAN. On the contrary, the discriminator differentiates between actual samples and generated samples. Another interesting method, the Approximate Bayesian computation (ABC) approach is discussed to estimate the posterior distribution of model parameters. Working on the principles of Bayesian statistics, ABC does not require calculating the likelihood function. This made ABC a popular choice in a wider range of applications for solving complex problems.

The authors also summarised that application areas such as particle physics, quantum physics, and quantum computing can be largely benefited by ML. In order to understand the different challenges involved in the implementation of likelihood-free interference and free energy surfaces, the authors suggested necessary methods. A different approach Quantum machine learning differs from traditional ML in the way that it uses quantum and qubits to improve the efficiency of computation speed and data storage. Both ABC and quantum machine learning help to understand milky way evolution better as per the authors.

Another concept of Deep Boltzmann machine [10] was discussed in which model consists of many directionless connections in the hidden layers and includes Markov random field to deal with unlabelled data. Furthermore, the authors here explained the application of various other unsupervised ML models like the generative model and variational autoencoders in the area of quantum physics, biological physics, and electronic structure calculation. All these methods are good at handling irregularities in data. In a broader sense, Autoencoders are of type artificial neural network (ANN) that is used in data compression and visualization. On the contrary variational autoencoder follows a probabilistic approach to generate high-quality images.

Apart from these, a few other areas such as star galaxy separation, galaxy classification, supernova classification, and more have been benefited by ML methods extensively. Researchers working in the field of exoplanet searching, habitability score calculation, and galaxy identification have experimented with plenty of ML approaches and obtained notable performance. Applying ML and AI to research studies like next-generation cosmic microwave background (CMB), observations of the epoch of reionization, identification of strong lensing curve, and pipeline optimization shows a huge success. Undoubtedly a rich and extensive contribution is being made from eminent researchers in this area, we summarized a few significant articles above.

2.1 Fundamentals of machine learning

For the review, we briefly describe some basic concepts of ML. Broadly we can categorize machine learning into four major divisions. a) Supervised; b) Unsupervised; c) Semi-supervised, and d) Reinforcement learning [11]. The term supervised implies data is labeled [3] indicating the target variable is known. A mapping function between input and output is formulated and consequently, the output is predicted for new sets of inputs. Unlike the traditional model fitting technique, these are not predefined. Model has to learn from the existing labeled dataset here. Classification and regression are the two most prevalent methods used in this category. The target variable in a regression task is continuous whereas classification requires discrete values.

Contrarily, unsupervised algorithms learn from the pattern of the complex unlabelled dataset and predict accordingly by grouping and clustering similar data. No prior information about the target variable is available in this method. It is mainly used to find a complex relationship and hidden features from the dataset. In astronomy dimensionality reduction, cluster analysis, as well as an outlier and anomaly detection, are the major tasks that require unsupervised learning. Baron had rightly

mentioned the need for machine learning algorithms in astronomy particularly focusing on unsupervised methods [3] such as K means, PCA (Principal component analysis), and hierarchical clustering. Such tools help to create a group of clusters based on distance measures of data points. We discuss details of these methods in their corresponding subsection of astronomical application.

In semi-supervised learning, as labeled data is small, it uses a large set of unlabelled datasets. Extrapolation of known samples onto unknown samples is done through clustering. The last category reinforcement learning [12] works in the concept of earning an award or punishment based on situation and observations in a dynamic environment. Given an input and output, the agent has to learn moves going towards the right direction to reach the output.

Although supervised and unsupervised techniques are being adopted majorly in astronomy, semi-supervised and reinforcement learning have not been explored yet much, therefore generating opportunities for further research. Figure 1 discusses the traditional architecture of the ML pipeline.

Before training of any ML model, during data splitting, a major percentage of data (80-90%) is generally used for training purposes, remaining (10-20%) is kept for validation. The model learns hyperparameters and complex relations from training data and subsequently, the performance of the model will be checked against validation data which is a subset of the input dataset. During training, various combinations of hyperparameters like epoch, iterations, and configuration variables are altered to observe the performance of the model. Finally, model selection is done based on the best performing one and is checked against completely unseen test data. The optimization process then starts based on the calculation of the error or loss function to make a more accurate model.

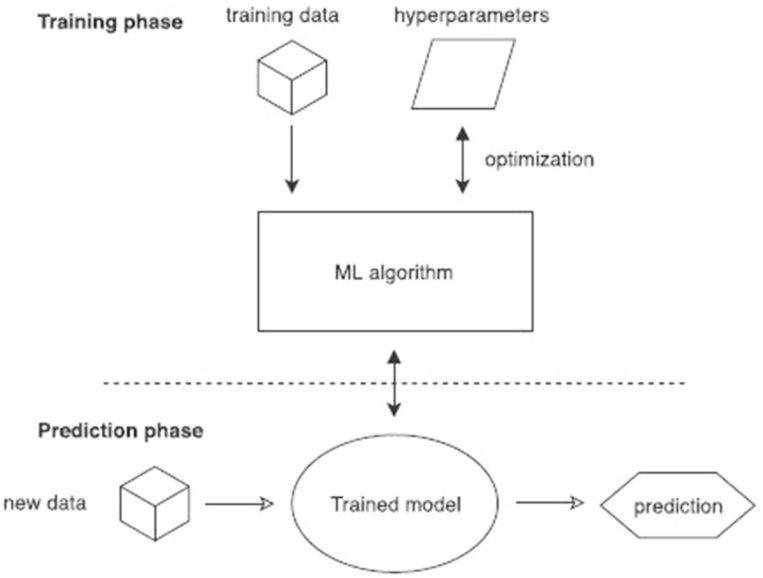

Fig. 1 Typical architecture of ML pipeline [11]

Several metrics exist to evaluate and assess the performance of a model. Precisely, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R2score are for regression task whereas Receiver operating characteristic (ROC) curve, precision, recall, and confusion matrix are frequently used in classification. MAE and MSE used in regression problem can be calculated as [3] $MAE = 1/n \sum_{j=1}^{n} |y - y1|$ where y is actual output and y1 is predicted output. Similarly $MSE = 1/n \sum_{j=1}^{n} (y - y1)^2$ whereas ROC curve is a visualization tool for binary classifier and confusion matrix presents classification results in the form of a matrix. Similarly precision and recall measure the effectiveness of the model to identify relevant data.

Overfitting and underfitting are two crucial aspects of machine learning. Overfitting occurs when the model performs better for training data but shows poor performance during testing. On the contrary, underfitting arises when model performance is bad for training as well as testing. The review by Mehta et al. [13] focuses on machine learning usage for physicists specifying areas of overfitting, underfitting, and bias-variance tradeoff. These all techniques are used in machine learning for better generalization of the model and tuning hyperparameters for powerful predictive analysis. Since regularization, batch normalization, and dropout are precisely regularization methods to treat overfitting, it builds a more feasible and sustainable model by reducing overfitting. Their work also reveals that the use of a python jupyter notebook for many physics-related applications is emerging.

2.1.1 Supervised techniques

This section presents a few well-known supervised techniques applied on particularly astronomical data sets. Out of multiple algorithms, we cite significant few approaches here. The concept to mine astronomy data using ML was introduced by Ball et al. [14] and later they showed the necessity and effectiveness for data mining and machine learning algorithms for astronomers. The data mining approach tries to discover useful patterns from a massive dataset and works well for both simulated and real data samples. Their study elaborately explained how data mining can be applied to select active galactic nuclei (AGN) in a better way. The term AGN refers to the center region of a galaxy. Real-time data analysis and time-domain analysis will be more rigorous once LSST starts operation. Indeed, it requires special attention from data mining. Effective handling of multiple observations also can be done using data mining.

One of the prominent and widely used algorithm Support Vector Machine (SVM) is widely used as a classifier as well as a regressor. It solves classification problems through a hyperplane having maximum margin between data points and plane. Using the SVM kernel trick, nonlinear input can be mapped to a linear relationship by projecting data points into a higher dimension. More precisely, the kernel trick helps to classify non-linearly separable data efficiently and inexpensively. Radial basis function (RBF), Gaussian, and polynomials are mostly used kernel tricks for SVM. Besides classification, SVM is used for regression and termed as Support vector regression (SVR). SVM [15–23] has been tested on many applications of astronomy.

Other ML algorithm K-nearest neighbor (KNN) shows [24–28] improved performance as it extracts all information from an input samples through distance measurement. Although KNN is applied to regression and classification, its usage is more evident as a classifier. Specifically, this algorithm assumes that two data points nearby to each other falls in the same class. K implies the number of neighboring data points to be considered. Challenge lies here to choose the value of K and the algorithm is computationally expensive. On the other hand, a Logistic Regression algorithm was applied for deriving probabilities [29] for star candidates in stellar classification.

Ensemble methods generally combine multiple algorithms into a single one for improved prediction power. Robust algorithms such as Random Forest [30, 31], Decision tree, and Extratreeregressor are few examples of ensemble approaches that demonstrated enhanced results on astronomical data without spending much time and effort. The random forest is primarily a combination of multiple decision trees. The decision tree presents attributes of the dataset as a tree-like structure having the root node as the starting point and the leaf node as a path label. Following the top-down approach, it first identifies the root node, presents the condition of one feature through each node, and finally constructs the tree. Both random forest and decision tree algorithms are used in classification as well as in regression tasks. Another ensemble method consisting of Random Forest, multilayer perceptron and GAN was experimented successfully on nanosatellite mission [32].

2.1.2 Unsupervised techniques

On the other hand, the unsupervised algorithm is not much far behind in its contribution towards astronomy. K means, Hierarchical clustering, Gaussian mixture (GM), and Agglomerative Hierarchical clustering are frequently used methods on astronomy datasets. K means clustering iteratively calculates centroid and groups data points into clusters based on shortest distance measurement. In K means [33, 34] cluster centroid is calculated using Euclidean distance, Manhattan distance, or some other distance measurement. This algorithm reassigns data points every time based on a new centroid and finally converges when the cluster centroid becomes stable. Cluster analysis, being a major task in astronomy, is used mainly in finding stellar spectra, x-ray spectra, galaxy spectra, and more.

Hierarchical clustering builds a tree-like structure of clusters whereas Agglomerative Hierarchical clustering is known as the bottom-up approach dealing with similarity measures to group data points. Here similar clusters are merged one after another. Besides, hierarchical clustering is easily implementable. The Gaussian mixture is also a well-known probabilistic approach for cluster analysis. It works on the principles of the Gaussian distribution of data points having unknown parameters. Data points belonging to different Gaussian distributions are put in separate clusters. Ward [35] applied hierarchical clustering on x-ray spectra and galaxy images, etc to show the superiority of this method over K- means and GM method. Experiments from previous researches show that hierarchical clustering is more powerful than K means and GM in detecting clusters. Moreover, it is less sensitive to outliers.

Similarly, Kernel density estimation (KDE) [36–42] is a non parametric approach for finding probability density function (PDF). This algorithm has been used widely for clustering and is capable of handling a higher dimensional dataset such as astronomy. But very often it is observed that for small fluctuation in data, density estimation is much higher which leads to degradation of performance. Kernel width and dimension are being used as two important parameters.

A self-organizing map (SOM) is a type of Artificial Neural Network (ANN) that uses an unsupervised approach for training the dataset and generating a low-dimensional map for input samples. Instead of using backpropagation in ANN, this algorithm applies competitive learning methods for representing data points. Korhonen successfully applied the neural network model (SOM) [43, 44] for unsupervised learning. In the galaxy morphology problem, SOM has been used by Galvin et al. [45] to visualize the dataset in a higher dimension. Also, SOM has been tested on the photometric redshifts problem by Masters et al. Recently in 2020, Wilson [46] experimented with a self-organizing map (SOM) on mid-range (z between 0 to 2) spectroscopic redshifts problem for training and testing to predict photometric redshift. Gomes et al. have shown accuracy [47] improvement after adding angular features and near the infrared magnitude of galaxies using the gaussian process.

As many algorithms can not manage a large number of features in a dataset, Principal Component Analysis (PCA) is the best solution to compress and reduce the dimensionality of a large dataset [48, 49]. PCA finds out physical parameters from spectra and estimates multivariate correlation. Using PCA entire data sample can be presented through its subset because it uses the largest eigenvectors to indicate the maximum variance of data. In this approach, unrelated fields are eliminated and only relevant features are considered to reduce complexity. Govada et al. [50] proposed distributed load balancing principal component analysis (DLPCA) as an extension of PCA that is distributed in nature. Furthermore, DLPCA reduces transmission and download costs for the user respectively. Working of PCA is demonstrated as follows:

1. Input data need to be centerized by subtracting the mean $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$ from each data point
2. Need to calculate covariance matrix S using the centered data as $S = 1/NXX^T$
3. Find out eigen value and eigen vectors from covariance matrix S
4. Choose first K eigenvectors u_k with eigen values λ_k
5. The final reduced K dimension data can be found by $Z = U^T X$ where U = [u1 .... uK] is (D x K) and embedding matrix Z is (K x N)

2.2 Neural network based techniques

On contrary to various traditional ML algorithms, the neural network (NN) based approach is heavily inspired by the human nervous system where through a series of neurons information propagates. NN is known to be a sub-component of ML widely. Artificial neural network (ANN), a basic multilayer perceptron model outperforms other ML methods in predictive power with its layered structure [51] and hundred of tunable parameters. ANN can understand excessively complex non-linear relations from a large dataset. As ANN consists of hidden layers associated with weighted

connections and activation functions, it confirms better results compared to its traditional counterpart. Sigmoid, rectified linear unit (RELU), SoftMax, and LeakyRelu are being frequently applied activation functions in hidden layers and output layers based on the nature of the task either classification or regression.

Data sets are trained multiple times to minimize the loss or error through back propagation [52]. During backpropagation, weight adjustment is done in every layer starting from the output layer to the input layer. A cost or loss function is primarily defined as the deviation between the predicted and actual value. The idea of a neural network has been applied first in astronomical problems by Angel et al. [53] and it was examined in the context of star galaxy separation and galaxy morphological classification [54]. A genetic algorithm-based ANN approach [55] has been applied by David et al. for preprocessed light curve-related classification model measurement. In the field of molecular astronomy, Alejandro et al. [56] have applied a neural network for predicting many molecule parameters from emission lines. ANN can be used as a deep architecture model by increasing the number of hidden layers and the associated neurons in each layer.

Figure 2 is an illustration of the structure of ANN.

2.2.1 Deep learning techniques

Considered as a subpart of the neural network, Deep learning (DL) shows excellent accomplishment in solving critical astronomical issues through its automatic feature extraction capabilities. Predominantly neural networks act as a backbone of deep learning algorithms. Although concepts of deep learning exist since 1980, due to the

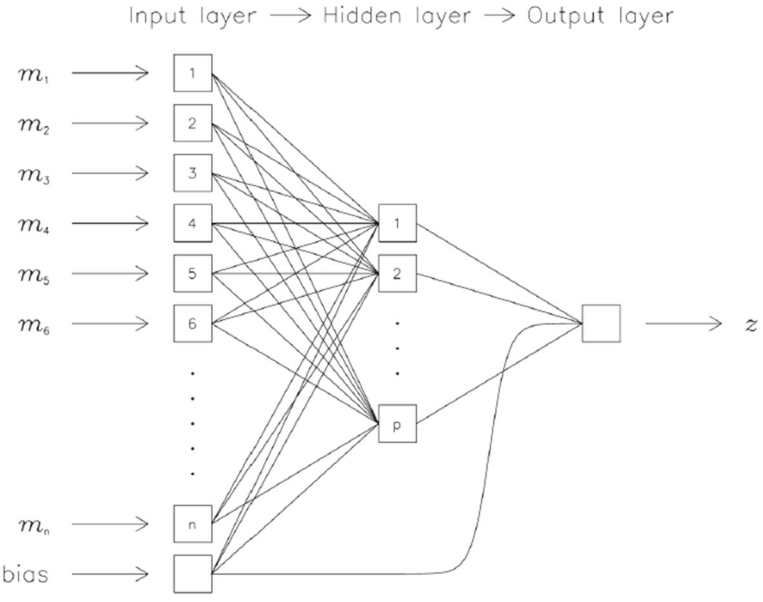

Fig. 2 General architecture of ANN [51]

lack of high-end systems and computational power it was not being experimented on vastly at that time. Increased storage capacity and availability of more computation power made deep learning an essential tool in modern data analysis. For example, Hinners et al. [57] discussed the effectiveness of long short term memory (LSTM) and recurrent neural network (RNN), two representation learning techniques in stellar curve classification. Both these methods are efficient enough to process sequential data and time series data with their input, output, and forget gates. In a recent study, Bassi et al. used LSTM for classifying variable star light curves. To detect anomalies in real-time transient light curves LSTM has been used by Muthukrishna et al. [58] Very recently Barchi et al. [59] have summarized a comparative study between ML techniques and deep learning techniques to galaxy morphological classification. Then [60] ML and DL are being experimented on galaxy detection and identification by Roberto et al. Searching for distant galaxies using DL has been reported by Cacho et al. [61]. Deep learning provides automatic feature selection [62] and powerful prediction of various issues. The automatic feature selection capability has made it far more superior than its counterpart which is beneficial to analyze data with high dimensions.

Iten et al. [63] applied neural network as a general-purpose tool to work as a representational learning method to retrieve physical parameters from experimental data and gain deeper insights into data. Their neural network structure is based on an encoder and decoder and used as a feed-forward connection. Sedaghat et al. [64] also used an encoder-decoder based deep convolutional neural network to extract useful and hidden patterns from large astronomical data itself without any human intervention and labels. Their proposed methods help to understand learned features and physical properties of stellar spectra.

Moreover, advancement in GPU boosts deep learning performance unexpectedly. Convolutional neural network (CNN), a DL algorithm shows impressive outcomes in the context of astronomical image processing. Mainly CNN consists of multiple layers such as convolution layer, pooling layer, and fully connected layers, etc. CNN learns features from input images, and through several layers and activation functions, it identifies the class of the images at the output. Huan et al. reported better accuracy [65] while applying SOM and CNN together. Other deep learning approaches, Deep belief network (DBN) is good at learning features directly from labeled and unlabeled images whereas LSTM has feedback connections and performs better on many applications including the space weather forecast. Usage of auto encoders [66] also has proven an effective solution towards data reduction, compression, and visualization. Autoencoder reconstructs output same as input through its two components encoder and decoder.

From the astronomical point of view, Gravitational lensing is very much useful in the study of galaxy dark matter distribution. CNN has been an attractive choice in weak gravitational lensing [67] in recent times because of its ability to extract required and essential data from high dimension astronomical sample datasets. CNN has the power of extracting a varied range of astrophysical parameters compared to other approaches which use convergence maps by simulation. CNN is proven to be more powerful in producing relatively small contours than the power spectrum. Moreover, CNN has been used in parameter fitting of cometary dust by Yue et al. [68]

Compared to traditional CNN, deep CNN uses more convolutional layers and pooling layers which handles automatic feature extraction efficiently and improves accuracy. Yashar et al. [69] suggested that utilization of deep CNN provides faster results in finding several lensing parameters from a massive set of data. It has shown remarkable achievement in spotting image distortions that can happen due to the deformation of distant image sources by strong gravitational lensing. The use of CNN in the automatic detection of the strong gravitational lens has been proposed by Pearson et al. [70]. Deep CNN has also been applied as a strong gravitational lensing detector for ground and spaceborne surveys by Schaefer et al. [71]. Furthermore, CNN helps in the automatic removal of lense light too. Lanusse et al. [72] proposed a deep learning-based approach CMU DeepLens that aids in an automatic finding of the galaxy-galaxy lens and is useful in many cosmological constraints. Based on different image simulations, this ML-based method provides reliable success.

As the research in time-domain astronomy is rapidly increasing, ML models are being applied to transient detection of late. Searching for astrophysical transients is a prominent research area in cosmology. Researchers [73] have applied to encode and decoding along with CNN for real-time transient detection. This approach further facilitates the registration of images, removal of noise, and background subtraction. Then Sadeah et al. [74] proposed a deep learning-based robust algorithm to search gamma-ray transients. Another work by Mong et al. [75] discusses a DL-based approach for creating a training dataset containing multiple different images for Gravitational-wave Optical Transient Observer (GOTO) prototype.

2.3 Optimization techniques

Optimization is an analytical technique aimed to find out the optimal solution to a problem and is a favored choice in astronomy. In recent years, searching for exoplanets through habitability scores is an emerging and interesting area of research . A significant contribution is being made by a group of researchers while applying optimization techniques to work with exoplanets.

For instance, Saha et al. has contributed mostly to exoplanet search. They [76] carried out a comparative study of different classification algorithms towards finding habitability scores of exoplanets. Categorizing exoplanets into various classes using ML has been demonstrated by [77]. The analysis on quasar-star classification using machine learning [78] and derivation of novel activation function [79, 80] for exoplanet classification suggest notable achievements. The activation function proposed by them requires less hyperparameter tuning effort compared to the traditional activation function. Moreover, it is proved to be an effective solution to the first-order differential equation. Continuing research in a similar direction, Saha et al. experimented with ANN in habitability classification problems seeking good results. Thereafter a novel elastic KNN model has been developed for the classification of exoplanets by Bora et al. [81]. As per the authors, this model can input a large parameter set while ensuring global optima.

Particle swarm optimization (PSO) is a widely used heuristic optimization method that works on the principles of iteratively improving candidate solutions known as

swarm intelligence. PSO is highly capable of solving complex mathematical functions. PSO is easy to implement and effectively manages memory and speed. This method [82] is experimented to handle early convergence and constraint issues of a model to estimate accurate habitability scores. The efficacy of a boosted decision tree for the newly invented exoplanet has been illustrated by Saha et al. [83]. Authors have classified exoplanets into different classes based on their habitability score using the XGBOOST tree. In order to create a strong classifier, multiple weak learners are combined through boosting. Similarly, a gradient boosting method XGBOOST offers a fast and robust solution for many ML tasks.

Another critical work [84] applied concepts of multi-objective optimization towards exoplanet habitability. Based on Cobb–Douglas habitability (CDHS) production function, a new metric has been proposed for calculating habitability score using convex optimization by Basak et al. [85]. Authors have introduced a novel habitability metric, the Constant Elasticity Earth Similarity Approach (CEESA) that uses a derivative-free optimization method and eccentricity as a feature. Their research showed the applicability of the metaheuristic method while minimizing complexity and curvature problem. CESSA's superiority over existing CDHS lies in handling missing or zero input effectively.

Although the usage of the surrogate model is not completely new, it acts as a black box to produce the output of some other complex model. In its broader sense, a surrogate model is a data-driven approach that accelerates simulation-based analysis and helps to optimize the system without much analytical expression. It successively reduces computation time through an inexpensive and limited number of complex simulations. Problems like uncertainty propagation and deterministic design where computationally expensive simulation is required, surrogate modeling shows an overwhelming response. Heitmann et al. [86] first built a trial simulator by applying two surrogate models namely estimated mass function and linear power spectrum and later constructed a highly effective precision emulator using 26 models and new dimensions. They proposed Mira–Titan universe which deals with a simulation-based approach for many astronomical observations.

Varma et al. [87] experimented with a surrogate model of expensive Numerical relativity (NR) waveform to provide reliable and quicker output. Their study shows that outside the parameter space of the training region, the surrogate model performs quite well. On contrary, the Bayesian-based surrogate model has been applied by Ford et al. [88] on real and simulated exoplanet data which offers computation efficiency. Khan et al. [89] proposed an ANN-based surrogate model for gravitational waves using GPU-based parallel processing. Recently, Angulo et al. [90] had proposed a 3D mass distribution framework taking cosmological parameters as input. Combining two approaches such as rescaling and baryonification, their approach can find out degeneracies in the non-linear power spectrum.

Among different space missions, the Euclid space mission (ESA), dedicated completely to dark matter and dark energy-related research looks very promising and worth noting. ESA's work towards gravitational lensing and wide coverage (15000 square degrees) galaxy clustering has also gained a lot of attention quite often. Euclid collaboration [91] is set to launch in 2020 to investigate the reason behind universe expansion and the source of expansion. This mission also searches the root cause

of how gravitational lensing affects the modification of shapes of galaxies. In this regard, a surrogate model has been developed as a part of Euclid-II [92] that can compute desired cosmological parameters within a fraction of a second and can emulate non-linear correction also when compared to the desktop system.

2.4 Simulation based techniques and statistical approaches

In terms of computational efficiency, the nested sampling-based approach has become a popular choice for Bayesian-related computations as it outperforms a standard sampling method to handle bulk data. It is a powerful bayesian simulation method introduced by Skilling [93]. Precisely, nested sampling helps in posterior sample calculation and marginal likelihood approximation. Improved computational power and resources with ever-increasing large datasets are pushing astronomy to solve more complex operations by applying Bayesian methods in the calculation of model parameter distribution.

In several applications of astronomy, such as light curve fitting for transient sources and stellar mass constraint calculation, nested sampling provides reasonable success. It efficiently helps in cosmological parameter selection. As likelihood calculation requires more computation, using less number of the likelihood function, Bayesian computation can provide posterior inferences. Model selection and parameter estimation [94] are two crucial areas of Markov Chain Monte Carlo (MCMC) method. MCMC is a simulation method aiming to find the posterior distribution of samples. This also helps in estimating uncertainty and detecting Bayesian objects. More specifically, MCMC is proposed as an improvement on Monte Carlo to deal with high-dimensional data. It combines two methods Markov Chain and Monte Carlo to calculate statistical inferences and probability distribution.

A generalized version of nested sampling, known as dynamic nested sampling chooses samples dynamically to boost accuracy and provides better computational efficiency. To overcome a few drawbacks of nested sampling such as sampling strategy and posterior integration, dynamic nested sampling is introduced and is more preferred than nested sampling in many use cases. For instance, based on dynamic nested sampling Speagle et al. [95] proposed a python package named dynesty for marginal likelihood and the Bayesian posteriors. While useful for multimodal distribution, authors claimed its superiority over the popular MCMC method in various astronomical applications such as analysis of gravitational waves, 3D dust mapping, and searching of exoplanets, etc. Graff et al. [96] discussed the usefulness of Bayesian algorithm and nested sampling strategies to estimate likelihood function. Then Higson et al. [97] introduced dynamic nested sampling methods with improved accuracy as allocation of live samples are done more efficiently and dynamically for posterior samples. Using this approach, both parameter estimation and evidence calculation can be enhanced compared to the general nested sampling approach. Existing work says that the dynamic sampling approach offers better computational efficiency and more robust results.

Another important term likelihood-free method has been a sudden interest among the astronomy community as it eliminates the need for computing likelihood function for many statistical inferences. It is a kind of Approximate Bayesian computation

(ABC) method. More broadly, likelihood implies data model probability distribution, and the likelihood-free approach deal with model parameter estimation. The usage of this method is more dominant in the simulation-based task. Mikelson et al. showed likelihood-free estimation is unbiased for other likelihood estimators and likelihood-free nested sampling is useful for parameter inference of biochemical reaction networks. A lot of astronomical problem uses diffusive nested sampling which was proposed by Brewer et al. [98] that uses a series of MCMC for the nested probability distribution. Dnest software was implemented based on this concept to solve several astronomy issues.

Gravitational-wave data analysis requires to use of the error estimation method and nested sampling. From data to encoding signal, Bayesian sparse reconstruction is very much useful. Akeret et al. [99] applied the ABC method where they claimed that in absence of likelihood function, forward modeling simulation of the mock dataset can be advantageous in image calibration of large range cosmological applications. Another area in astronomy like the modeling of complex type and posterior samplers resulted in a good performance after employing the Bayesian technique. Very recently to calculate cosmological interference, a likelihood-free parallel method using density estimation has been proposed by Taylor et al. [100].

One of the popular and powerful methods of Bayesian analysis uses the Bayes theorem to calculate the probability of a hypothesis based on prior events. Using observed data and prior probability, it calculates the posterior probability of parameters and finally forms statistical inferences. In astronomy application of this technique is vast because of its simplicity. For instance, Savage et al. [101] proposed a Bayesian-based astronomical source extraction method for identifying image background from a point source. For measuring flux, local background and point spread function is being used by them. This method is effective to distinguish between point and extended source. Upcoming Akari Far-Infrared Survey or all-sky survey plans to use this method. Rogers et al. [102] reported that in high dimensional space likelihood computation is very much essential. The requirement of emulators lies in mapping between observable and parameters that help to reduce computational complexity by considering sample points for mapping. Emulation uses training sets with a Bayesian procedure. Here authors claimed that their work on Bayesian optimization applied to large data set is completely novel.

Ishida et al. [103] reported that if the actual likelihood is not available, the ABC method helps to understand the complex system by interfering parameters. They used variations of ABC and estimated posterior probability distribution in identifying galaxy cluster parameters without calculating the likelihood function. Next, Cameron et al. [104] discussed the usefulness of the ABC method in understanding transformation in the morphology of a huge set of galaxies in the redshift range 1.5 to 3. The Sequential Monte Carlo method is the most efficient statistical algorithm to find the age and mass of star clusters based on their Spectral Energy Distribution (SED), according to the authors.

On a similar note, to tackle the problem of likelihood-free estimation, Bayesian interference is essential suggested by Leclercq [105]. Their method combines both the Gaussian process and Bayesian optimization to obtain proper training data. With the limited number of simulations and budget, the authors achieved reliable posterior

approximation. As interpreting cosmological parameters is becoming expensive day by day, the authors [106] applied a gaussian emulator of the likelihood function for parameter estimation that provides the accurate distribution of posterior. This method uses less likelihood compared to MCMC methods.

Most of the time complex astronomy data contain large missing values, hence analyzing those data should account for some extent of uncertainty. Apart from clustering, Gaussian mixture model (GMM) and the hierarchical Bayesian model (HBM) are often used to calculate error and uncertainty. GMM [107] directly learns from noisy data considering uncertainty as input whereas HBM offers a probabilistic framework in understanding uncertainties from multiple sources. HBM frees developers from extensive computations.

Kristiadi et al. [108] reported that along with the Bayesian neural network approach, ensemble class of neural network with mixture models is gaining attention in recent times. This compound density network can be trained with maximum likelihood along with the Bayesian interference and is proved to be better than other approaches in uncertainty estimates. Author Long too described numerous statistical methods useful for astronomy. Censoring and truncation are often used to handle the issue of biased samples caused by brighter objects. Usage of HBM in characterizing galaxies and supernova light curve modeling have been experimented by many authors. Generalized linear models (GLMs) that allow response variable as non-normal distribution was applied in many applications such as finding the distance of galaxies as a function of their colors. To deal with predictive uncertainty, Zhu et al. [109] introduced the Bayesian approach based deep encoder and decoder technique that is very effective compared to the ensemble and Gaussian methods.

2.5 Forward modeling techniques

In recent years, forward modeling started gaining popularity in the astronomical context as all major associations and partnerships are going in that direction. Broadly forward modeling is evolving as a data analysis tool as it aids in finding a problem's solution where no direct technique for inversion is available. In an astronomical context, forward modeling is mostly used in solving an inverse problem. Forward modeling and measurement of uncertainty are both related to the likelihood function. MCMC method combined with forward modeling has great potential in estimating maximum likelihood. Researchers have used this method for understanding the exoplanet atmosphere condition. The forward model uses specific parameters and produces data that can then be compared with the actual observations.

Chen et al. proposed a fast and accurate forward modeling technique of gravity field using prismatic grids. To understand transmission spectra of exoplanet goyal et al. [110] investigated a forward model which is supposed to be scalable up to H2/He dominated atmospheres. Recently Schmidt [111] proposed EFT based forward model for large scale redshift surveys of galaxies. Astronomical data being high dimension requires forward modeling technique to estimate stellar astronomical parameters collected from data of GAIA survey [112] whereas Halotools (v0.2) is proposed by Hearin et al. for testing connection towards galaxy halo and respective halo models. Lately, Sartori et al. [113] applied a forward modeling based approach for finding

out variability in active galactic nuclei. Based on PTF survey data, the authors implemented GPU based simulation setup for forward modeling of longer light curves. This framework is believed to help upcoming large surveys. Furthermore, Forward photometric modeling can be used in stellar mass prediction which in turn helps in galaxy evolution understanding.

2.6 Evolutionary computational techniques

To tackle huge data floods in astronomy, evolutionary computation has a major role to play. Fuzzy set theory and Genetic algorithm (GA) are two widely used search heuristic approaches in this category. Generally, fuzzy set theory is proven to be a useful tool to deal with ambiguous and vague data. Starting from data reduction to automatic classification, fuzzy set theory has been tested by many researchers throughout the years. For instance, it was first applied by Hu et al. [114] for solar activity prediction. In finding the motion of stars, the evolutionary method was used first by Metcalfe [115]. A few years later, Ordonez et al. [116] proposed GA based algorithm using a neural network for classifying stellar objects. To classify complex stellar spectra, faint galaxy and to remove cosmic ray and solar image segmentation, fuzzy set theory has been used widely by eminent researchers [117–121]. In the task of asteroids classification, Freistetter [122] applied fuzzy logic also to achieve better performance. A survey on various existing soft computing techniques illustrating how fuzzy logic-based algorithms can be beneficial for automating astronomical image analysis is presented by Shmair et al. [123]. The fuzzy logic controller has been implemented for position tracking of a telescope by Attia in 2009 [124]

On the other hand, GA (Genetic Algorithm) shows success in both constraint and unconstrained optimization and search problems. It has the potential to find a near-optimal solution in a short duration. Based on concepts of natural selection, the most suitable individual is selected for the next phase. Concepts of GA [125] have been employed on many vital issues of astronomy such as anomaly observation in distant bodies, determination of galaxy's rotation curve, review of stellar dynamics, and study of supermassive bodies, etc. Another noteworthy application of GA along with an annealing approach for parameter estimation has been reported by Han [126].

Although we have listed here a limited number of important papers, a lot of research effort has been put in this direction and these techniques are certainly becoming a potential tool for astronomers for their wide range of problems. As discussed before, multiple areas of astronomy are being enriched through the usage of ML approaches, we discussed a few here.

3 Big data concepts and challenges

As our main focus is handling astronomical big data, this section includes a few fundamental concepts of big data, data format, challenges, and techniques.

The term big data is relative and varies from person to person. Yesterday's big data can be small today based on storage and computing power, etc. In general, big data denotes massive data coming from heterogeneous sources [127] in different formats

that are beyond the processing capacity of traditional computing. Therefore, a new storage facility and real-time data analysis are very much important. There are four different V's that generally describe big data characteristics [128].

- Volume: It implies the quantity and size of data that is becoming a worry among data scientists. In recent days volume measurement goes up to tens of zettabytes.
- Variety: It indicates certain formats of data such as image, audio, video, and email, etc. It can be either structured, unstructured, and semi-structured format which needs to be processed in distinct ways. Data complexity introduces a very big challenge when data comprises numerous data formats like audio and video and multimedia content, etc. All possible types of data collected online or offline are a threat to storage and need to be handled carefully.
- Velocity: It signifies the speed of data on how it gets generated or processed. It is represented in terms of the batch, real-time and streaming data, etc. High-velocity real-time data need to be processed at a quicker rate using advanced tools.
- Veracity: It implies data purity, noise-free and trustworthy data, and filters out unimportant data for a particular task.

Along with the above major four characteristics, data validity checks and finds out relationships or correlation among data items because valid input is a necessity for proper data analysis. Another characteristic of big data that is data volatility tells us how long data is valid and needs to be stored. Most of the time non frequently used data can be moved to an archive. Besides, the confidentiality of data needs to be handled properly. Lastly, if this massive data doesn't add any value in decision making it becomes useless. In recent times, venue [129] is also considered to be added as part of big data characteristics. The venue describes different platforms from where data is coming.

3.1 Different sources of astronomical data

As data collection remains the first and important task in any kind of analysis, an extensive data source is very essential. We illustrate a few significant data sources here.

SDSS survey [130] (http://www.sdss.org), being the largest, provides an in-depth wide-range(approximately one third coverage of sky) [131] catalogs of spectroscopic and photometric data. Starting from DR1 (Data release 1), currently serving DR16 it stores images, spectra, and redshifts, etc. It provides a SQL-like interface named CasJob server through which we can write the query to download data from the server. This huge data set can be stored in the local system or directly in the cloud using the necessary command.

Another survey VIPERS provides a catalog of around 90000 redshift [132] samples using an 8m large telescope. This survey mainly aims to obtain measurements when the universe was half its current age. PDR2 is the latest second and final data release which includes photometric information, full spectroscopic data, and survey masks. Through a web interface, the tar file of the dataset can be accessed by the user. VIPERS covers sky up to median approx. redshift z=0.7 and generates volume for

galaxy mass, magnitude, and spectral data which [133] can be comparable to other sky surveys like 2dFGRS and SDSS with similar galaxy populations. VIPERS [134] has been used recently for studying galaxy's color-magnitude diagram.

The 2-MASS astronomical survey conducted at near-infrared wavelength [135] generates a catalog of over 300 million objects containing low mass stars and galaxies in a span of 4 years from 1997-2001. NASA and Caltech support this survey and data can be accessed from their online repository.

Another in-depth survey DEEP2 which is currently with DEEP2 data release 4, surveys distant galaxies using the Keck telescope. It contains image spectra, redshifts in the range during its 2002-2008 years of operation covering 3 square degrees sky. With its wide depth coverage and high precision redshift (approx z=1) [136] it can be comparable with VLT deep survey.

Future survey LSST uses a Large Synoptic Survey Telescope (LSST) comprising of 3.2 billion-pixel camera. Through this, it can capture the night sky [137] of billions of objects rapidly. Currently, under construction in Chile, it can detect rare and fainter objects beyond the capacity of the human eye. The main objective of the LSST is to cover as much area as possible such as 40 square meters. This powerful instrument will help to discover the changed brightness of objects from seconds to months and makes us understand the milky way galaxy in depth.

SKA observatory is very powerful and the world's largest telescope [138] governed by jointly Australia and South Africa. This is set to release in 2021 and explore a wide depth survey to find dark matter, dark energy, and early universe and exoplanet. All together 13 other countries are involved in this project.

LIGO survey funded by NSF, Caltech, and MIT, focuses on gravitational-wave [139] data. As of 2019 Ligo has collected 50 observations of gravitational waves. Caltech Ligo astrophysics group focuses on the study of the properties of gravitational waves. Gravitational-Wave open science center enables access to LIGO data, tutorials, and tools to users. LIGO-India is a collaboration among three Indian institutes and the LIGO laboratory is the initiative by the Indian government.

Apart from above mentioned sky surveys, there are multiple other surveys such as VVDS [140], UKIRT Infrared Deep Sky Survey [141], Alhambra survey [142], GOTO, DPOSS [143], GALAXY ZOO [144], All-Wise [145], Pan-STARRS1, KiDS [146], ZTF, PRIMUS and GAIA, etc offer reliable source of astronomical data. Figure 3 and Table 1 show and explain data volumes generated by multiple sky surveys respectively.

3.2 Astronomical data format/types

Knowing the core of the data structure and their formats are necessary for better data analysis. Generally, astronomical data is available mainly in FITS and Comma Separated Value (CSV) format. We discuss a few relevant data formats and types in this section.

The calibrated data in numerical form are stored in CSV format and can be downloaded for further use. Researchers work with IMAGES frequently for many applications. Astronomical data mostly deals with spectral images like optical spectra and infrared spectra, etc. Through different navigator tools, we can access captured

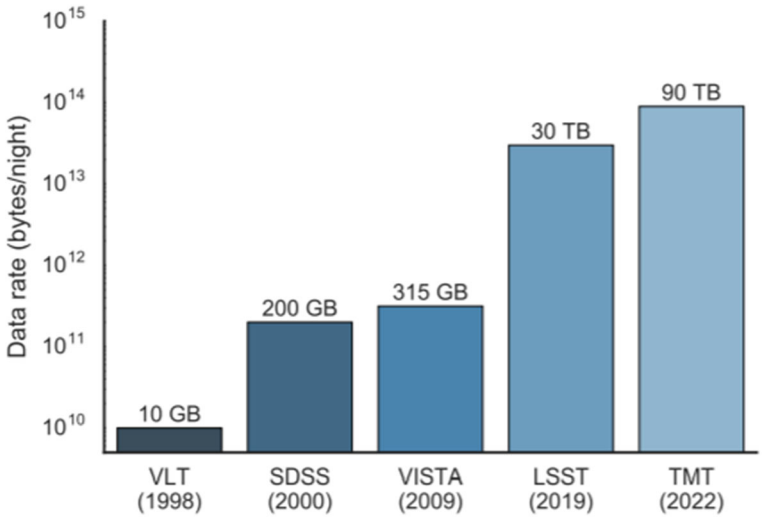

Fig. 3 Sky surveys [1]

images of the night sky. Compressed images are sometimes stored as.gif or .jpeg format.

Another very popular and recognized format is FITS which stands for flexible image transport system [148]. It is mainly used for the delivery and transportation of observatory data between various environments. The majority of astronomical data including images are stored in FITS (extension .fits/.fit/.fts) format digitally. It is a preferred data format as it is well documented, tested, and overcomes operating systems incompatibility. FITS popularity rises because of the ease of universal human and machine readability and extensibility. Currently, in version 4, FITS includes metadata and ASCII or binary data in the same file and also describes photometric and spatial calibration information. Astronomers use this format to transport,

Table 1 Volume of Data [147]

Different astronomical survey	Generated data
GOTO	50GB image/night
DPOSS	3 TB
2MASS	10TB
GALEX	30TB
SDSS	40TB
SMSS	500 TB
GBT	20PB
PanSTARRS	40 PB expecting
LSST	200 PB expecting
SKA	4.6 EB expecting

analyze, and archive scientific data files. FITS is capable to store image data in two-dimensional and three-dimensional forms. Besides imaging data, non-image data such as spectra, photon lists, or even structured data in form of multi-table databases can be stored in FITS. SAOImage DS9 is an open-source FITS viewer software for Windows that opens and analyses FITS files. It is a cross-platform software as well and helps to convert FITS files to other formats like JPEG, PNG, and TIFF, etc.

On the other hand, HDF5 is basically version 5 of HDF (Hierarchical data format) and is considered a strong alternative to FITS. In many applications mainly for LOFAR radio data encapsulation [149], the use of this format is prominent. It also helps in the custom design of data encapsulation. In order to read data, h5py and paytables are used. LSST and SKA also explored the feasibility to use HDF5 for their data storage. In addition to supporting multiple storage options, it supports variable-length arrays and structures. Compared to HDF, HDF5 is much more stable.

There are still other significant formats like CDF (Common and self describing data format) [150], NDF (N-dimensional data format) [151], VOLTABLE [152], FITSML [153] and ASDF [154] have been used in many extensive research applications.

3.3 How to handle big data

Improvement in storage and computing power effectively handles big data problems without incurring extra cost. The major advantages of big data technologies are the usage of cheaper hardware to reduce the data processing cost. As the data volume is substantial, efficient and reliable techniques are beneficial in this regard.

3.3.1 Distributed computing

In a broader sense, through distributed computing data and computing resources are shared across multiple machines in different locations for quicker processing. One such instance, cloud technology has proved itself as a domain solution for solving big data storage problems. Cloud provides all computer resources as a service in a pay-per-use model. So instead of spending time and money upfront in upgrading the local system(s) with higher configuration, huge data can be stored and processed in the cloud without any scalability issue. Precisely, Infrastructure as a service (IAAS), platform as a service (PAAS), and software as a service (SAAS) [155] are three major models offered by the cloud to facilitate users with computing services. These offer unlimited computing instances, storage, and networking services on a demand basis.

Some leading cloud providers available in the market presently like Amazon Web Service (AWS), Microsoft Azure, and Google offer cost-effective solutions to business and industrial problems. Using cloud platforms hardware costs and maintenance costs can be reduced tremendously, instead, we can thoroughly focus on data analysis and algorithm design. Amazon EMR, Redshift, Kinesis, Dynamo and S3, and Glacier, etc. are a few such big data solutions from AWS. Berriman et al. [156] experimented with AWS cloud technology for better cost-effective solutions for the NASA Kepler mission. Recently, Araya et al. [157] proposed a cloud-based jupyter notebook facility for astronomical data analysis.

Another distributed architecture Grid computing facilitates multiple machines and researchers across all over the world to connect and collaborate on a specific task for efficient resource sharing and utilization. Here data loads are stored across multiple machines in a distributed fashion facilitating data access faster with high throughput. Unlike cloud computing, users don't have to pay once the setup is functional. UK's GridPP collaboration [158] uses distributed computing to address dark matter related issues.

3.3.2 Distributed processing framework

To analyze a big dataset, Apache Hadoop is one of the widely accepted open-source scalable and distributed processing frameworks. Primarily Hadoop consists of two sub-components namely Hadoop Distributed File System (HDFS) [127] for storage and Map-reduce for processing. HDFS can support hundreds of nodes in a cluster to offer reliable storage and operate as a master-slave architecture. Nodes in the cluster are designated as a name node (Master Node) or data node (Slave Node). Through name nodes and data nodes, HDFS manages and coordinates data load distribution. HDFS offers a low-cost storage solution as it uses commodity hardware. Map-reduce, a programming framework relies on two operations- Map and Reduce for parallel processing. Dividing the input dataset into multiple partitions is done by Map operation and aggregating and combining results from multiple nodes is carried out by Reduce function.

As an alternative, Apache Spark offers in-memory [159] real-time distributed data processing which is 100 times faster than Map-reduce. Spark's main component is Resilient Distributed Data Set (RDD) which is highly fault-tolerant and stored across multiple machines. Additionally, Spark's machine learning library MLLIb is a very effective tool for executing iterative machine learning algorithms. It works efficiently with java, scala, python, and R. Parallel processing in Spark happens through driver nodes and multiple worker nodes. In order to deal with large FITS files, Spark-fits works as a Scala-based spark connector. Apache Mahout, an open-source software run on the Hadoop environment is widely used for implementing machine learning algorithms and linear algebra operations. Similarly, Caffe and TensorFlow both are mostly used frameworks for training deep learning models and computer vision tasks. Internally, they execute based on tensors which is very much beneficial for deep learning.

Nonetheless, Apache has introduced several other tools such as Flume that collects streaming [160] data and fed input to HDFS. Indeed, Flume is capable of handling massively distributed data sources. Real-time scalable software Flink efficiently works with multiple machine learning algorithms and supports java, python, R, and Scala. Although similar to the Hadoop cluster, Storm supports fast real-time data processing and hides the complexity of distributed processing efficiently. It is capable to process millions of data per second. However, for analyzing real-time machine-generated data Splunk can be used.

Pig is an open-source framework that helps [161] in the reduction of a map-reduce complexity. Pig reduces lines of code as well as development time. Scoop helps to import data from RDBMS to Hadoop and export from Hadoop to RDBMS using

SQL. Scoop offers faster performance and fault tolerance. Workflow manager Oozie is an Apache product that uses directed acyclic graphs (DAGs) [162] of actions and helps in scheduling and executing Hadoop jobs in a distributed environment. Developed by Microsoft as a counterpart of Hadoop, Dryad is applied to upscale a small-scale setup to a large scale with a parallel and distributed programming model. Dryad is mainly for graph processing framework.

Open-source NoSQL distributed database Cassandra is mainly applied in online transactions that help in real-time processing with no single point of failure. It is highly scalable and reliable software with proper support for structured, semi-structured, and unstructured data. Here data gets stored in key-value pairs using the gossip protocol. Similarly, Statwings offers easy to use statistical tool [163] to process a massive amount of data.

To handle schema-less Not only SQL (NoSQL) database and Google Big Query, Apache Drill is developed. It offers an interactive analysis of distributed data processing. In order to manage multidimensional astronomical data management, SciDB database [164] is introduced as an excellent robust database software from Paradigm4. It uses cheap hardware and follows shared-nothing architecture. Lately, this software is co-created by Turing Award winner Michael Stonebraker. Apache Tez [165] framework also works on DAG as spark whereas Apache Hive, a SQL-like interface generally uses TEZ or map-reduce for data processing.

Previously, astronomers used to rely on high-performance supercomputing for their data-intensive scientific calculations but optimization of arithmetic efficiency and usage of complex architecture requires careful tuning. The benefit of high-performance cluster computing (HPCC) is achieved through a data analytics supercomputer. It provides open-source parallel processing for huge data using multi-core systems. Julia computing helps in parallel supercomputing [166] that analyzes 188 million images in 15 minutes. Same way Map-Reduce merge [167] is considered as an improved and efficient version of the map-reduce model for bulk heterogeneous and relational data processing and parallel programming. It supports various relational algebra operations as well.

Although we are seeing usage of spark and other tools in the astronomy community recently, to some extent it is still limited. Using these tools, standard astronomy data analysis can be done quite easily without the requirement of many programming skills by astronomers.

3.4 Challenges in astronomical big data

While astronomical data analysis looks promising, a significant amount of challenges are also introduced dealing with this data set. The main challenges in astronomical data lie in its noisy [168] and sparse nature. Thermal noise, as well as interference from WI-FI, mobile phone, and TV, increases SNR (Signal to Noise ratio) for most of the observation. In astronomy, input as well as output both are non-gaussian and data are often of type multi-temporal, multi-wavelength, and multi-scale. At the same time, it is not clean and incorrect [168]. Enormous volume data sets collected in a different format from various large digital sky surveys lead to complexity as well which compels usage of expensive astronomical operations [169]. One such job that

astronomers and sometimes data analysts do is, cross verification [170] and cross-matching of data samples collected from various observations to prepare the catalog of the dataset. Moreover, scarcity of labeled data, lack of standard query language [170] while accessing data are also major issues. Very often, detection of faint objects is a vital concern and most of the time improper and uneven density dataset causes highly skewed structure. As many astronomical catalogs use Parquet's non-linear columnar design format, it imposes challenges sometimes. In parquet format, data is stored contiguously in a single column.

In the astronomical community, there is no proper standard for data collection, and the labeling of data makes data processing even more difficult. Also, different telescopes are having structural differences and astronomical images are of 2-D or 3-D, or one-dimensional spectra. Converting high-dimensional raw data to calibrated data too imposes challenges. Uncertainties in data too are an integral and inherent part of astronomy as it is observational, not experimental. Unstructured, incomplete, and imprecise information introduces a threat to data mining and ML techniques.

While downloading data from archives, limitation in network bandwidth becomes a critical issue. It in turn aids in making data transfer, and unified access more tedious. As LSST aims to generate petabyte data in the future and detect sky events, the astronomy community is burdened with classifying these events such as class discovery of galaxies based on elliptical, spiral, and spherical shape, dimension reduction of astrophysical data, outlier (Anomaly / Deviation / Novelty) detection and Link Analysis (Association Analysis – Network Analysis).

Furthermore, astronomy is developing problems regarding maintaining the huge infrastructure of on-demand data management for executing several projects. If this problem is addressed by cloud providers, scientists and astronomers can contribute more towards designing a new algorithm rather than spending effort on data management. Computational power and storage are the two crucial factors astronomers can rely on cloud vendors.

3.5 Big data approaches towards astronomy

In this section, we chose to highlight some relevant literature papers that particularly deal with significant big data approaches for astronomical applications. It started with a Caltech project GRIST, developed to facilitate astronomers with the essence of grid-based distributed computing for image processing. The images are collected from multiple observations [171] for better catalog generation. Ivanova et al. [172] examined feasibility in porting SDSS to MonetDB database software. Then Juric [173] implemented a large survey database framework built on python 2.7 for cross-matching catalogs and distributed querying. Continuing the same way, Wiley et al. [174] applied Hadoop MapReduce to work with SDSS images.

Later, [170] Brahem et al. emphasized how astronomers will be benefited from data processing using Astroide framework. Indeed, this framework offers an in-memory distributed data server focusing on astronomical query processing and optimization. To speed up the performance, query processing is done with ADQL and

data partitioning technique HEALPix pixilationcis. Brahem's work [175, 176] has ushered the way how spark outperforms other state of art solutions in this field. Lately [177] Zhang et al. implemented an ec2 instance-based astronomy image processing framework, Kira. More precisely, Kira is efficient enough to provide distributed processing, scalability, and flexibility. Furthermore, while comparing the execution of the C program, this framework is almost 2 times faster than a NERSC Edison supercomputer. In order to reuse existing python libraries and leverage distributed processing, spark streaming concepts have been applied.

To query and analyze large astronomical catalogs [178], a Spark-based open-source scalable astronomical data analysis framework AXS is proposed by Zecevic et al. It was built using popular AstroPy packages and SQL statements. Data partitioning in AXS takes place in a distributed zone. While focusing on astrophysical big data challenges, [179] Garofalo et al. indicated ML and big data collectively are being used to process huge data whereas in [180] Ball proposed cloud-based astronomical data mining tool CANFAR+Skytree with improved performance. Users can access CANFAR through virtual machines and SKYTREE is installed in their machine as regular software. Thus, to access petabyte range memory, it offers high-end computing power and storage up to 32GB RAM. According to the literature, this is the first astronomy project implemented on a cloud platform. Recent work on the topological structure of gravitational clustering [181] also has been developed using spark. VeljkoVujčić et al. [182] outlined numerous real-time and heterogeneous astronomy stream data processing methods. Figure 4 describes how astronomical data can be processed using distributed Hadoop architecture.

Table 2 describes few useful big data methodologies applied so far on astronomical applications.

Based on our examination of existing research, it is visible that less work has been done in the area of big data application in astronomy. Hence it is generating a lot of possibilities for future research.

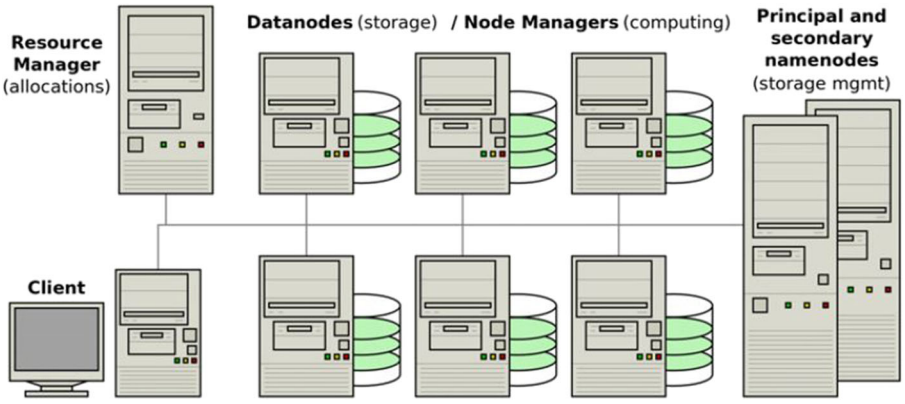

Fig. 4 Distributed processing framework [184]

Table 2 Big data approaches in astronomy

Citation	Approaches	Advantages	Limitations
Zhao et al. [185]	Parallel approach for cross matching for large data set	Virtual spatial indexing approach speeds up data access	NA
Wiley et al. [174]	Combines multiple brief image on cloud using mapreduce	Prefiltering showed performance improvement, faster processing for massive dataset	Irrelevant FITS file need to be removed for better result
Sciacca et al. [183]	Distributed framework to remotely process data, SOA architecture	Low memory usage	NA
Mesmoudi et al. [186]	Usage of hadoopDB, hive and mapreduce for large LSST data and query management	Loading time and execution time is considered.	Can be tested on hybrid in memory and column oriented system
Fillatre et al. [184]	Different solution to astronomical bigdata handling	Distributed processing benifits	NA
Zhang et al. [177]	Spark based toolkit Kira deployed on EC2 cloud	Efficient feature selection improves regression output	No image reprojection and co-addition
Peloton et al. [187]	Scala based spark connector to handle fits file	Scala API gives slightly better performance	Comparison with HPC system not done
Govada et al. [50]	Distributed load balancing reduces cost	Scalable, better accuracy	Latency, memory efficiency, processing speed has not been considered
Xie et al. [188]	SIMBA, as a extension of spark sql	Offers spatial query execution for bigdata	Need to be tested for very high dimensional data
Wei et al. [189]	Python based high performance flexible architecture OpenCluster	Distributed, lightweight, fault tolerant and scalable and cost effective	Does not support resource isolation
Berriman et al. [190]	Montage image mosaic engine, visualization tool	Scalable toolkit, automation and integration into python	NA
Brahem et al. [170]	Spark based distributed data server and HEALPIX scheme to partition data	Faster than other tools	Scalability in terms of increased cluster size and data size need to be explored
Corizzo et al. [191]	Spark based autoencoder	Better performance than python	Need to be tested on other neural network
Sen et al. [192]	Spark based model on cloud for redshift analysis	Provides scalability and lesser training time	Need to be tested on multiple cluster

4 Different astronomical data analysis tools

Python being a wonderful interpreted, open source and platform-independent language supports multithreading, multiprogramming, and portable programs. It is a well-accepted scripting language for data analysis and machine learning. The Global Interpreter Lock (GIL) of python restricts only one thread usage at a time which guarantees exclusive access of a single thread. Although the application of python modules and packages are very much prevalent in astronomical data analysis, we shortly discuss some of them here.

Starting with AstroML [193] which is an exclusive and extensively used community repository to store all kinds of python modules and packages to analyze astronomical data. This library is built on NumPy, Scipy, Scikit-learn and Matplotlib that are being used frequently in data analysis and visualization. It comprises several machine learning algorithms useful for astronomy data. ASTROMLSKIT has been developed as an integrated platform for data-driven analysis and computation. A suite of algorithms was built to handle several types of datasets and open problems [194] under ASTROMLKIT. On the other hand, the AstroPy module developed by community developers is a collection of software packages written in the Python language exclusively for astronomy [195] data operations.

The Astromining package supports data analysis and knowledge discovery software that is written in MatLab. It consists of several statistical and visualization tools to evaluate the performance of neural network models. AstroCV [196] is a computer vision library that offers several python and C++ functions to process and analyze the large astronomical dataset. This library supports classification, segmentation and image recognition, etc. Experiment on galaxy classification and object detection has been explored using this package. Another important package AstroWeka evolved [197] as an extended version of Weka which is an open-source data mining software mainly used to analyze astronomical datasets. It supports voltable data format as well. To load data from observatory and cone search data services, this toolkit uses the Astro runtime library. Similarly, the Astro-Toyz module assists the user with the web-based python framework to work with a large FITS image data set. It also provides a customized environment for users and facilitates remote server access. FITSWebQL supports viewing very large FITS files (approx. 100 GB) in a web browser with minimum RAM. Likewise, AstroLab software [187] helps in developing a solution for astronomical big data. More Precisely, Astrolab provides advanced solutions for challenging and critical research areas.

Apart from these exclusive packages, many other python packages are also being utilized. For example, PYFITS [198] is applied mainly for handling the FITS file. Another AstroLib library, PyWCS enables world coordinate system (WCS) transformation such as conversion between WCS and pixel. But the problem lies with different formats without any coordination between them. One solution can be thought of as the usage of Astropy. Astropy is a community effort to develop one core package for astronomy. PyCS and PyRAF are python-based modules used by astronomers for parallel processing [199]. PyRAF [200] is a command language being applied to process IRAF (Image Reduction and Analysis Facility) tasks. IRAF,

a portable data analysis software is quite popular among astronomers as it offers easy script debugging errors and exceptions.

COLITEC [201] software helps in real-time complex astronomical big data processing effectively. A python library Vaex [202], similar to pandas helps to visualize and work with tabular datasets such as astronomical catalogs like Gaia catalog and other row and column-based datasets. The main advantages are faster processing, less memory usage, and a user-friendly API. Besides, it supports 2D and 3D visualization. Vaex-jupyter, built on Jupyter notebook, helps in visualization and vaex-ml supports machine learning algorithms. Using CPU or GPU, NumPy based python library Theano is being tested in astronomy recently and offers computation for mathematical expression and multidimensional array.

To enable the processing of large astrophysical datasets in a distributed fashion, DAMEWARE is introduced as a data mining tool [203]. It allows users to access web-based interfaces and perform different ML algorithms like regression and classification etc. Many applications of astronomy such as globular cluster classification and photometric redshift evaluation, etc, have used this tool.

Another popular toolkit Data to Knowledge developed by ALG supports multiple java modules and several ML algorithms including SVM, KNN, decision tree, and ANN, etc. It supports parallelism and distributed environment as well [204]. CANFAR+Skytree is the first cloud-based machine [180] developed at a Canadian astronomy data center to work with seven popular machine learning methods. Astronomers use TOPCAT software to process large and sparse tabular data and various object catalogs. Most of the astronomical data formats are supported by it.

Ipython is developed as an improved version of the default python interpreter [205] that provides powerful interactive and parallel computing. It helps in data visualization as well. The current version (IPython 7.12.0), released on Jan 31st, 2020 offers an interactive interface to the python language. Similarly, Vizic assists [206] in flexible interactive visualization tools. Customization and analysis of user-generated scripts can be done by this. Moreover, support for JavaScript and python has made it unique.

Online statistical software StatCodes [207] links to public domain software focusing on statistical problems for astronomy. Keeping in mind, the versatility of R, recently Vostat came up as a statistical web service to perform various data analysis methods like plotting, regression, filtering, etc., and 3D graphics using R. It's main goal is to encourage astronomers to use R language. The montage toolkit proposed by Jacob et al. is used for mosaicing [208] multiple images collected from the telescope. It was already tested on SDSS, DPOSS, and 2MASS survey data whereas Autonlab [147] offers data mining for statistical analysis and pattern detection.

5 Major contribution and activities

As the objective of our article is to provide insight into this area meticulously, we discuss a lot of activity in terms of workshops and conferences as well as talks, blogs, and courses. We especially present various forums and groups being formed all over the world to promote this interdisciplinary research. Additionally, we add

the description of noteworthy exclusive books useful for researchers working in this direction.

A long program on Mathematical and Computational Challenges in the Era of Gravitational Wave Astronomy was planned from September 2021 to December 2021 in Los Angeles. 2 days Workshop hosted by NOAO on Tools for astronomical bigdata [209] was organized in March 2015 in Arizona. Talk on Data science and Visualization was delivered in Feb 2016 at NASA, UC Riverside. Big data and astronomy [210] workshop was held in August 2017 at Mauritius. A workshop on Big data tools in physics and Astronomy was held in June 2017 in Amsterdam. Furthermore, the SKA bigdata workshop [211] was held in April 2018 at ZTE, Shanghai. Exclusive workshop on Artificial Intelligence in astronomy [212] had been conducted in July 2019 at ESO Garching. 3 days conference on Machine Learning Tools for Research in Astronomy [213] was organized in December 2019 at Ringberg castle focusing on the area of ML methods for simulation and observations. Collaborators from Switzerland and South Africa conducted a bilateral workshop on bigdata [214] for 2 days in Switzerland. Workshop on Innovation in data-driven astronomy [215] was held in May 2011 in Green Bank, West Virginia.

Data science for physics [216] and astronomy workshop was conducted in December 2019 at Alan Turing Institute. PES University, Bangalore, India hosted the first international conference on modeling machine learning [217] and astronomy in November 2019. In order to discuss majorly radio astronomy, a workshop on Big data and Digital Technology [218] was held in September 2019 at Chiang Mai. 3 days workshop on Applications of Data Science in Astrophysics [219] was organized in November 2019 by IIIT Allahabad, India. Moreover, Machine Learning in Astronomical Data Analysis [220] workshop was there in January 2019 at Washington. A discussion on Big Data Challenge in Astronomy [221] was held in June 2019 in Shanghai whereas AstroInformatics [222] conference takes place each year in US. 2 days Astronomical Data Science [223] workshop was held in Feb 2020 at Texas AM. Their discussion was mainly on new opportunities generated by huge data collection. Recently IISER, Tirupati had organized a Workshop in SKA in India [138] in February 2020. IAU Symposia hosted a meeting to discuss Big Data Era in Astronomy [224] in December 2020 in Argentina. Recently CAASATRO, CSIRO, and General Assembly planned for some class on big data in astronomy [225] in 2020 at Dallas. The yearly conference on Astronomical Data Analysis Software and Systems (ADASS) [226] is organized by a different institution.

Researchers across the world have written useful blogs in this area to elevate awareness. For example, G. Bruce Berriman [227] had written on GPU-based or cloud-based infrastructure for handling large-scale data in 2011 whereas Megan Ray nicholas [228] has focused on powerful telescopic cameras and sensors in 2020. Muqbir Ahamal [229] talks about Big data tools for the astronomical problem in 2016. Eileen Meyer [230] says about LSST generated data processing in 2018 in a blog. Lyndon Henry [231] states that Data mining and ML tools for space programs in 2017. Discussion on High-end telescopes and data explosion is summarized by Ross Andersen [232] in 2012. James Urton [233] explained Zwicky Transient Facility in 2017. Similar way, Ray Norris [234] focused on huge data generation and scrutinization in 2017 in a blog. Alison Mcguiremay [235] described revamping the

astronomy landscape with big data analytics in 2018. Michael Gordon [236] had written on how machine learning helps in classifying galaxy morphology in 2019. Then Anil Ananthaswamy [237] wrote on handling data crunch by machine learning for multi-messenger astronomy in 2019. Snehanshu Saha and Rahul Yedida maintain [238] maintains a blog site that discusses Different ML algorithms for beginners.

Several online courses and summer schools have been arranged to train interested researchers working in this field. For example, Coursera, University of Sydney is offering a 6-week online course on Data-driven Astronomy [239]. Astrostatistics schools [147] were conducted in Baltimore in 2011. Data Mining and Machine learning in astronomy [240] course were introduced in Spring, 2020 by Arizona university. M.Sc program on Astronomy and Data Science with 2 years full-time duration was introduced by Leiden University, Leiden, Netherlands [241]. During November-December 2018, the University of Turku conducted training school [242] on big data simulation.

In addition to the above-mentioned topics, various forums and groups have been formed by renowned universities to conduct and support multiple activities. To cite a few, Astrostatistics and Astroinformatics Portal(ASAIP) [243] assist in interdisciplinary research for astronomers and data scientists with the help of advanced methodologies and statistical tools for the greater benefit of the astronomy community.

Astroinformatics Research Group [244] work on the intersection of data analytic methods, astronomical issues, and Machine learning supported by IEEE Computer Society. International Astroinformatics Association (IAIA) provides a wide platform to share new ideas and tools and facilitates new collaborations by organizing topical astroinformatics conferences and workshops. Pittsburgh Computational Astrostatistics Group (PiCA) has formed a multidisciplinary group of researchers from astronomy and computer science backgrounds who can work towards data-driven astronomy approaches. Inter-University Institute for Data-Intensive Astronomy established a cloud-based computational facility named IDIA data-intensive astronomy cloud facility [245]. Another group AstroNeural focuses on Data reduction.

Contribution in form of books helps many aspiring authors and researchers. For example, in the book Big Data in Astronomy, Linghe Kong et al. [246] discuss in-depth and up-to-date big data techniques in radio astronomy. It includes project examples from SKA and data processing methods for advanced radio telescopes as well. In the book Astronomy and Big Data, Kieran et al. [247] talk about the Data clustering approach for identifying uncertain galaxy morphology.

A book titled "Knowledge Discovery in Big Data from Astronomy and Earth Observation" written by Skoda et al. [248] widely covers ML and big data tools and methods used in geoscience and astronomy. In the book Multivariate data analysis, Heck et al. [249] discuss PCA, discriminant analysis, cluster analysis, and multivariate methods used in astrophysics.

Wall and Jenkins [250] had written a book on Practical Statistics for Astronomers describing complex data analysis and statistical methods. Concepts on Bayesian and time series analysis also is explained in the book. Connolly et al. [251], in their book "Statistics, Data Mining, and Machine Learning in Astronomy" elucidate a practical

python guide for the analysis of survey data. Stefano Cavuoti [252] discusses concepts related to Data-Rich Astronomy and Mining Synoptic Sky Surveys. We get to know modern Statistical Methods for Astronomy with R Applications from the book written by Feigelson et al. [253]. This book thoroughly covers statistical approaches, Bayesian techniques, and data mining methods. Here authors have shown examples analyzing astronomical data using R.

Statistical challenges in Modern astronomy which mainly focuses on Poisson analysis and maximum likelihood are discussed in the book written by Feigelson, et al.[254]. Podgorski et al. [255] provide useful concepts on Data mining tools and methods for astronomers in their book. In the book "Sparse Image and Signal Processing" Starck et al. [256] describe Sparse images for astronomy data processing. Asis Chattopadhyay et al. [257] completely explain the Statistical method for astronomical data analysis and provide detailed coverage of astronomy data sources and tools to deal with raw data. Saha et al. [258] has written an ebook on Machine learning in astronomy: a workman's manual that illustrates several classification methods for searching exoplanets, calculating habitability score using Cobb Douglas habitability function and supernovae classification and habitability catalog labeling. The contribution towards forums, groups, and other activities is not limited to the above section. As Astronomy is an evolving research area, a lot of activities are being conducted all over the world.

5.1 SciDAC program

Among all other initiatives taken towards the collaboration of cross-disciplinary research, SciDAC [259] program is a notable mention here considering its significant contribution in this field. This well-recognized SciDAC program encourages and promotes cross-disciplinary problem solving and accelerates scientific discovery through advanced computing using a special dedicated software program and supercomputers. Top researchers across the world join hands to solve critical issues through advanced computational techniques. US department of energy (DOE) Office of Science has started this mission in 2001 to develop a proper infrastructure and facilitate the use of supercomputers in advance computation which will lead to unknown discoveries. In 2020, they had announced to fund $60 million to SciDAC program to proliferate growth towards new tools and discoveries. This program was a collaboration involving all 6 offices of the science program (SC) from the US department of energy, laboratories, universities, and the office of nuclear energy.

Collaborators working in the multidisciplinary area from applied science, mathematics, and computer science initiated this project to make use of advanced scientific computing for path-breaking results to enrich and facilitate more promising research. In 2001, SC proposed some scientific challenging problems and issues which need to be addressed using advanced computational techniques and powerful high-end systems. The SciDAC program was re-competed in 2006 and recently in 2011 and 2017 to address the challenges of petabytes data. They support facilitating enhancements and innovations with the improved computational framework for computing extensive operations. Advanced Scientific Computing Research (ASCR) provides funding to the SciDAC program. Recent SciDAC projects [260] such as Accelerating HEP

Science: Inference and Machine Learning at Extreme Scales, Advancing Catalysis Modeling: From Atomistic Chemistry to Whole System Simulation, A New Discrete Element Sea-ice Model for Earth System Modeling and Improving the Numerical Solution of Atmospheric Physics in ACME use advanced techniques while offering great research potential.

The Focus area of the SciDAC program is that using high-end systems and effective simulation code, fast computation methods and research capabilities need to be formulated and enhanced. It aims towards scientific computing and algorithm design as well. Fast-paced computation needs to be carried out for better efficient results. Analysis of large data set need to be rigorous and more collaboration with academia need to be encouraged and fostered. Bridging the gap between advanced research and traditional research with effective algorithms and better infrastructure is the key factor to achieving more accurate results for many challenging critical tasks. SciDAC projects help astronomers to understand the reason behind the supernovae explosion through computational simulation. The simulation was run for 300,000 processor hours by the team of supernovae science center to provide a breakthrough achievement that was otherwise impossible without SciDAC partnership. Moreover, to answer queries related to the SciDAC program, a SciDAC outreach center has been formed.

Milestone achieved by SciDAC:

- Simulation of nuclear combustion for Type 1a supernovae, the exploding star through which we get a better picture of the universe
- Climate controller simulation
- First large scale 3-D flame simulation
- Detailed study of tiny particles
- Development of a flexible and portable Chroma software system that provides a highly effective calculation for high energy physics.

In addition to computations and microprocessor power, parallel computing and simulation have taken a rise in the last few years. To leverage the capabilities of supercomputers, necessary tools and resources need to be developed. 2 SciDAC institutes have done a collaboration with 24 institutions and started research with 12 billion dollars funding. To boost awareness and motivate young scientists, the SciDAC program organizes conferences, talks, workshops, etc frequently. For example, the SciDAC Conference was held in 2007 in Boston. After SciDAC-1 from 2001 – 2006, SciDAC-2 started as the second phase from 2006 – 2011 and mainly focuses on high-end computing and the advancement of high energy physics. They clearly suggested new tools with enhanced power and capabilities to be used for ILC, LHC, Tevatron, and electron cooling and beam dynamics. According to SciDAC, supercomputers equipped with high-performing software are essential in this area. Under SciDAC-2, a new data management tool has been developed to handle petascale data.

SciDAC-3 [261] (2011 – 2016) talks about the Scientific Computation Application Partnership project which dedicates to the accurate calculation of nuclear physics with new software developments. SciDAC-4 (2017-2021) involves enriching research on high-energy physics and biological issues using advanced HPC. To make it more effective, collaboration and partnership with the institutes are encouraged on a large

scale. A notable project Lattice Quantum Chromo-Dynamics (LCQD) [262] is executed as a part of SciDAC-4 whereas SciDAC-5 institutes are aiming to deliver expert and intelligent methods to collaborators. FASTMath which provides frameworks, algorithms, and Scalable Technologies for Mathematics and RAPIDS, a SciDAC institute for computer science are two major SciDAC-5 [263] institutes. The first one focuses on scalable software and algorithm design whereas the second one solves computational challenges. Active collaboration from renowned universities and institutes is accelerating growth to enrich the wider research community as a part of the SciDAC program.

6 Redshift estimation- a crucial application in cosmology

While examining the cosmological application, we noticed measurement of redshift has paramount importance in this domain. By measuring the redshifts of a million galaxies, we can create a three-dimensional picture of our local neighborhood of the universe. Famous Euclid mission [264] also rely on accurate redshift values. Although plenty of ML-based methods has been proposed over the years to estimate redshift accurately and precisely, we find the application of cloud platforms was not explored much. Hence we highlight where researchers can extend their work and explore new opportunities. One such area we briefly discuss here is the usage of a cloud-based distributed framework using AWS sagemaker [265] towards redshift prediction. Sagemaker is Jupyter like notebook interface which facilitates easy code development and deployment. It is mainly a GPU-based integrated effective tool for building machine learning models without much effort and processing cost. This approach will be hugely beneficial to handle a large amount of data generated by upcoming surveys like LSST etc.

As storing huge datasets in local systems is not feasible, the potential and performance of AWS cloud services for handling bulk astronomical data can be analyzed. Therefore we intend to implement an ML model by utilizing Amazon SageMaker that not only provides high storage capability but also offers high speed and low latency to carry out predictions and analytics as quickly as possible.

7 Conclusions

Different sections of our review focus that efficient ML, DL techniques, and big data frameworks will surely automate and facilitate the task of data analysis which was not possible till a few years ago. Modern days astrophysicists are contemplating applying machine learning to create a similar model as AlphaGo which is a computer program that defeats humans in the game. Therefore young computer researchers, on large scale are expected to come forward, engage themselves, and put rigorous effort towards designing a new data-driven architecture to shape a better future for astronomy.

This paper also provides a definite insight that novel innovation needs to be done, a better faster algorithm should be designed to expedite data processing. Implementation of the ML model on the cloud will help us to experiment with more than terabyte range data in the future. Our article fairly introduces an idea on usage of AWS Sagemaker as an example for bulk data processing and consequently contributes to analyzing data-driven astronomy. Additionally, we deeply explained the challenges associated with data analysis. Despite recent important developments, many shortcomings still exist. Thus, as a future enhancement, we try to investigate the feasibility of downloading the entire data set from the sky server to the cloud directly to reduce download cost as well as a burden on a local system. As the domain is data-rich, collaboration, association, and interdisciplinary studies are the key factor to generate new possibilities and opportunities.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Kremer, J., Stensbo-Smidt, K., Gieseke, F., Pedersen, K.S., Igel, C.: Big universe, big data: Machine learning and image analysis for astronomy. IEEE Intell. Syst. https://doi.org/10.1109/mis.2017.40 (2017)
2. Tallada, P., Carretero, J., Casals, J., Acosta-Silva, C., Serrano, S., Caubet, M., Castander, F.J., César, E., Crocce, M., Delfino, M., et al.: Cosmohub: Interactive exploration and distribution of astronomical data on hadoop. Astron. Comput., 100391 (2020)
3. Baron, D.: Machine Learning in Astronomy: a practical overview (2019)
4. Borne, K.D.: Astroinformatics: A 21st century approach to astronomy (2009)
5. Wang, K., Guo, P., Yu, F.: Computational intelligence in astronomy: A survey. Int. J. Comput. Intell. Syst. **11**, 575–590 (2018)
6. Fluke, C.J., Jacobs, C.: Surveying the reach and maturity of machine learning and artificial intelligence in astronomy. Wiley Interdisc. Rev. Data Mining Knowl. Discov **10**(2), 1349 (2020)
7. Bird, J., Petzold, L., Lubin, P., Deacon, J.: Advances in deep space exploration via simulators & deep learning. New Astron. **84**, 101517 (2021)
8. Ntampaka, M., Avestruz, C., Boada, S., Caldeira, J., Cisewski-Kehe, J., Di Stefano, R., Dvorkin, C., Evrard, A.E., Farahi, A., Finkbeiner, D., et al.: The role of machine learning in the next decade of cosmology. arXiv:1902.10159 (2019)
9. Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., Zdeborová, L.: Machine learning and the physical sciences. Rev. Modern Phys. **91**(4), 045002 (2019)
10. Navamani, T.: Efficient deep learning approaches for health informatics. In: Deep Learning and Parallel Computing Environment for Bioengineering Systems, pp. 123–137. Elsevier (2019)
11. Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., Rellermeyer, J.S.: A survey on distributed machine learning. ACM Comput. Surv. (CSUR) **53**(2), 1–33 (2020)
12. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. J. Artif. Intell. Res. **4**, 237–285 (1996)
13. Mehta, P., Bukov, M., Wang, C.-H., Day, A.G., Richardson, C., Fisher, C.K., Schwab, D.J.: A high-bias, low-variance introduction to machine learning for physicists. Phys. Rep. **810**, 1–124 (2019)
14. Ball, N.M., Brunner, R.J.: Data mining and machine learning in astronomy. Int. J. Modern Phys. D **19**(07), 1049–1106 (2010). https://doi.org/10.1142/s0218271810017160
15. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
16. C.J., B.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. **2**, 121–167 (1998)

17. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (2013)
18. Cristianini, N., Shawe-Taylor, J., et al.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge (2000)
19. Kecman, V.: Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. MIT Press, Cambridge (2001)
20. Schölkopf, B., Smola, A.J., Bach, F., et al.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
21. Abe, S.: Support Vector Machines for Pattern Classification, vol. 2. Springer, New York (2005)
22. Lin, C.-F., Wang, S.-D.: Fuzzy support vector machines with automatic membership setting. Support vector machines: Theory and applications, 233–254 (2005)
23. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
24. Fix, E.: Discriminatory Analysis: Nonparametric Discrimination, Consistency Properties. USAF School of Aviation Medicine (1951)
25. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans Inf Theory **13**(1), 21–27 (1967)
26. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Mach. Learn. **6**(1), 37–66 (1991)
27. Dasarathy, B.V.: Nearest neighbor (nn) norms: Nn pattern classification techniques. IEEE Comput. Soc. Tutorial (1991)
28. Shakhnarovich, G., Darrell, T., Indyk, P.: Nearest-neighbor methods in learning and vision. IEEE Trans. Neural Netw. **19**(2), 377 (2008)
29. Beitia-Antero, L., Yáñez, J., de Castro, A.I.G.: On the use of logistic regression for stellar classification. Exp. Astron. **45**(3), 379–395 (2018)
30. Carliles, S., Budavári, T., Heinis, S., Priebe, C., Szalay, A.S.: Random forests for photometric redshifts. Astrophys. J. **712**(1), 511 (2010)
31. Baron, D., Poznanski, D.: The weirdest sdss galaxies: results from an outlier detection algorithm. Mon. Not. R. Astron. Soc. **465**(4), 4530–4555 (2017)
32. Cao, H., Bastieri, D., Rando, R., Urso, G., Luo, G., Paccagnella, A.: Machine learning on compton event identification for a nano-satellite mission. Exp. Astron. **47**(1), 129–144 (2019)
33. Steinhaus, H.: Sur la division des corps materiels en parties. bull. acad. polon. sci., c1. iii vol iv: 801-804 (1956)
34. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297, Oakland, CA, USA (1967)
35. Ward Jr, J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)
36. Parzen, E.: On estimation of a probability density function and mode. Ann Math. Stat. **33**(3), 1065–1076 (1962)
37. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification and Scene Analysis, vol. 3. Wiley, New York (1973)
38. Silverman, B.W.: Density Estimation for Statistics and Data Analysis, vol. 26. CRC Press, Boca Raton (1986)
39. Scott, D.W.: Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, New York (2015)
40. Taylor, C.: Classification and kernel density estimation. Vistas Astron. **41**(3), 411–417 (1997)
41. Wasserman, L.: All of Statistics: a Concise Course in Statistical Inference. Springer, New York (2013)
42. Klemelä, J.S.: Smoothing of Multivariate Data: Density Estimation and Visualization, vol. 737. John Wiley & Sons, New York (2009)
43. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**(1), 59–69 (1982)
44. Kohonen, T.: An overview of som literature. In: Self-Organizing Maps, pp. 347–371. Springer (2001)
45. Galvin, T.J., Huynh, M., Norris, R.P., Wang, X.R., Hopkins, E., Wong, O., Shabala, S., Rudnick, L., Alger, M.J., Polsterer, K.L.: Radio galaxy zoo: Knowledge transfer using rotationally invariant self-organizing maps. Publ. Astron. Soc. Pac. **131**(1004), 108009 (2019)
46. Wilson, D., Nayyeri, H., Cooray, A., Häußler, B.: Photometric redshift estimation with galaxy morphology using self-organizing maps. Astrophys. J. **888**(2), 83 (2020)

47. Gomes, Z., Jarvis, M.J., Almosallam, I.A., Roberts, J.S.: Improving photometric redshift estimation using gpz: size information, post processing, and improved photometry. Mon. Not. R. Astron. Soc. **475**, 331–342 (2018)
48. Boroson, T.A., Green, R.F.: The emission-line properties of low-redshift quasi-stellar objects. Astrophys. J. Suppl. Ser. **80**, 109–135 (1992)
49. Djorgovski, S.: The fundamental plane correlations for globular clusters. Astrophys. J. **438**, 29–32 (1995)
50. Govada, A., Sahay, S.K.: A communication efficient and scalable distributed data mining for the astronomical data. Astron. Comput. **16**, 166–173 (2016)
51. Collister, A.A., Lahav, O.: Annz: estimating photometric redshifts using artificial neural networks. Publ. Astron Soc Pac. **116**(818), 345 (2004)
52. Sadeh, I., Abdalla, F.B., Lahav, O.: Annz2: photometric redshift and probability distribution function estimation using machine learning. Publ. Astron Soc Pac. **128**(968), 104502 (2016)
53. Angel, J.R.P., Wizinowich, P., Lloyd-Hart, M., Sandler, D.: Adaptive optics for array telescopes using neural-network techniques. Nature **348**(6298), 221–224 (1990)
54. Bazell, D., Peng, Y.: A comparison of neural network algorithms and preprocessing methods for star-galaxy discrimination. Astrophys. J. Suppl. Ser. **116**(1), 47 (1998)
55. Andrešič, D., Šaloun, P., Pečíková, B.: Large astronomical time series pre-processing for classification using artificial neural networks. In: Intelligent Astrophysics, pp. 265–293. Springer (2021)
56. Barrientos, A., Holdship, J., Solar, M., Martín, S., Rivilla, V.M., Viti, S., Mangum, J., Harada, N., Sakamoto, K., Muller, S., et al.: Towards the prediction of molecular parameters from astronomical emission lines using neural networks. Exp. Astron., 1–26 (2021)
57. Hinners, T.A., Tat, K., Thorp, R.: Machine learning techniques for stellar light curve classification. Astron. J. **156**(1), 7 (2018)
58. Muthukrishna, D., Lochner, M., Webb, S.: Real-time detection of anomalies in large-scale transient surveys (2019)
59. Barchi, P., de Carvalho, R., Rosa, R., Sautter, R., Soares-Santos, M., Marques, B., Clua, E., Gonçalves, T., de Sá-Freitas, C., Moura, T.: Machine and deep learning applied to galaxy morphology-a comparative study. Astron. Comput. **30**, 100334 (2020)
60. González, R.E., Munoz, R.P., Hernández, C.A.: Galaxy detection and identification using deep learning and data augmentation. Astron. Comput. **25**, 103–109 (2018)
61. Cacho Martínez, R.: Distant galaxies analysis with deep neural networks. http://hdl.handle.net/10609/107807 (2020)
62. Hoyle, B., Rau, M.M., Bonnett, C., Seitz, S., Weller, J.: Data augmentation for machine learning redshifts applied to sloan digital sky survey galaxies. Mon. Not. R. Astron. Soc. **450**(1), 305–316 (2015)
63. Iten, R., Metger, T., Wilming, H., Del Rio, L., Renner, R.: Discovering physical concepts with neural networks. Phys. Rev. Lett. **124**(1), 010508 (2020)
64. Sedaghat, N., Romaniello, M., Carrick, J.E., Pineau, F.-X.: Machines learn to infer stellar parameters just by looking at a large number of spectra. Mon. Not. R. Astron. Soc. **501**(4), 6026–6041 (2021)
65. Mu, Y.-H., Qiu, B., Zhang, J.-N., Ma, J.-C., Fan, X.-D.: Photometric redshift estimation of galaxies with convolutional neural network. Res. Astron. Astrophys. **20**(6), 089 (2020)
66. Schawinski, K., Turp, D., Zhang, C.: Exploring galaxy evolution with latent space walks. AAS **231**, 309–01 (2018)
67. Ribli, D., Pataki, B.Á., Zorrilla Matilla, J.M., Hsu, D., Haiman, Z., Csabai, I.: Weak lensing cosmology with convolutional neural networks on noisy data. Mon. Not. R. Astron. Soc. **490**(2), 1843–1860 (2019)
68. Yue, Y., Cao, Z., Gu, H., Wang, X.: Dynamic simulation and parameter fitting method of cometary dust based on machine learning. Exp. Astro, 1–34 (2021)
69. Hezaveh, Y.D., Levasseur, L.P., Marshall, P.J.: Fast automated analysis of strong gravitational lenses with convolutional neural networks. Nature **548**(7669), 555–557 (2017)
70. Pearson, J., Pennock, C., Robinson, T.: Auto-detection of strong gravitational lenses using convolutional neural networks. Emergent Sci. **2**, 1 (2018)
71. Schaefer, C., Geiger, M., Kuntzer, T., Kneib, J.-P.: Deep convolutional neural networks as strong gravitational lens detectors. Astron. Astrophys. **611**, 2 (2018)

72. Lanusse, F., Ma, Q., Li, N., Collett, T.E., Li, C.-L., Ravanbakhsh, S., Mandelbaum, R., Póczos, B.: Cmu deeplens: deep learning for automatic image-based galaxy–galaxy strong lens finding. Mon. Not. R. Astron. Soc. **473**(3), 3895–3906 (2018)
73. Sedaghat, N., Mahabal, A.: Effective image differencing with convolutional neural networks for real-time transient hunting. Mon. Not. R. Astron. Soc. **476**(4), 5365–5376 (2018)
74. Sadeh, I.: Deep learning detection of transients. arXiv:1902.03620 (2019)
75. Mong, Y.-L., Ackley, K., Galloway, D., Killestein, T., Lyman, J., Steeghs, D., Dhillon, V., O'Brien, P., Ramsay, G., Poshyachinda, S., et al.: Machine learning for transient recognition in difference imaging with minimum sampling effort. Mon. Not. R. Astron. Soc. **499**(4), 6009–6017 (2020)
76. Agrawal, S., Basak, S., Saha, S., Rosario-Franco, M., Routh, S., Bora, K., Theophilus, A.J.: A comparative study in classification methods of exoplanets: Machine learning exploration via mining and automatic labeling of the habitability catalog (2015)
77. Basak, S., Agrawal, S., Saha, S., Theophilus, A.J., Bora, K., Deshpande, G., Murthy, J.: Habitability classification of exoplanets: a machine learning insight. arXiv:1805.08810 (2018)
78. Viquar, M., Basak, S., Dasgupta, A., Agrawal, S., Saha, S.: Machine learning in astronomy: A case study in quasar-star classification. In: Emerging Technologies in Data Mining and Information Security, pp. 827–836. Springer (2019)
79. Saha, S., Nagaraj, N., Mathur, A., Yedida, R.: Evolution of novel activation functions in neural network training with applications to classification of exoplanets. arXiv:1906.01975 (2019)
80. Saha, S., Mathur, A., Bora, K., Agrawal, S., Basak, S.: Sbaf: A new activation function for artificial neural net based habitability classification. arXiv:1806.01844 (2018)
81. Bora, K., Saha, S., Agrawal, S., Safonova, M., Routh, S., Narasimhamurthy, A.: Cd-hpf: New habitability score via data analytic modeling. Astron. Comput. **17**, 129–143 (2016)
82. Theophilus, A., Saha, S., Basak, S., Murthy, J.: A novel exoplanetary habitability score via particle swarm optimization of ces production functions. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 2139–2147. IEEE (2018)
83. Saha, S., Basak, S., Safonova, M., Bora, K., Agrawal, S., Sarkar, P., Murthy, J.: Theoretical validation of potential habitability via analytical and boosted tree methods: An optimistic study on recently discovered exoplanets. Astron. Comput. **23**, 141–150 (2018)
84. Khaidem, L., Saha, S., Kar, S., Saha, S., Basak, S.: Quantifying exoplanet habitability via penalized multi-objective optimization (2019)
85. Basak, S., Saha, S., Mathur, A., Bora, K., Makhija, S., Safonova, M., Agrawal, S.: Ceesa meets machine learning: A constant elasticity earth similarity approach to habitability and classification of exoplanets. Astron. Comput. **30**, 100335 (2020)
86. Heitmann, K., Bingham, D., Lawrence, E., Bergner, S., Habib, S., Higdon, D., Pope, A., Biswas, R., Finkel, H., Frontiere, N., et al.: The mira–titan universe: precision predictions for dark energy surveys. Astrophys. J. **820**(2), 108 (2016)
87. Varma, V., Field, S.E., Scheel, M.A., Blackman, J., Kidder, L.E., Pfeiffer, H.P.: Surrogate model of hybridized numerical relativity binary black hole waveforms. Phys. Rev. D **99**(6), 064045 (2019)
88. Ford, E.B., Moorhead, A.V., Veras, D., et al.: A bayesian surrogate model for rapid time series analysis and application to exoplanet observations. Bayesian Anal. **6**(3), 475–499 (2011)
89. Khan, S., Green, R.: Gravitational-wave surrogate models powered by artificial neural networks: The ann-sur for waveform generation. arXiv:2008.12932 (2020)
90. Aricò, G., Angulo, R.E., Hernández-Monteagudo, C., Contreras, S., Zennaro, M., Pellejero-Ibañez, M., Rosas-Guevara, Y.: Modelling the large-scale mass density field of the universe as a function of cosmology and baryonic physics. Mon. Not. R. Astron. Soc. **495**(4), 4800–4819 (2020)
91. Blanchard, A., Camera, S., Carbone, C., Cardone, V., Casas, S., Clesse, S., Ilić, S., Kilbinger, M., Kitching, T., Kunz, M., et al.: Euclid preparation-vii. forecast validation for euclid cosmological probes. Astron. Astrophys. **642**, 191 (2020)
92. Collaboration, E., Knabenhans, M., Stadel, J., Marelli, S., Potter, D., Teyssier, R., Legrand, L., Schneider, A., Sudret, B., Blot, L., et al.: Euclid preparation: Ii. the euclidemulator–a tool to compute the cosmology dependence of the nonlinear matter power spectrum. Mon. Not. R. Astron. Soc. **484**(4), 5509–5529 (2019)
93. Skilling, J., et al.: Nested sampling for general bayesian computation. Bayesian Anal. **1**(4), 833–859 (2006)

94. Feroz, F., Hobson, M.P.: Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses. Mon. Not. R. Astron. Soc. **384**(2), 449–463 (2008)
95. Speagle, J.S.: dynesty: a dynamic nested sampling package for estimating bayesian posteriors and evidences. Mon. Not. R. Astron. Soc. **493**(3), 3132–3158 (2020)
96. Graff, P., Feroz, F., Hobson, M.P., Lasenby, A.: Neural networks for astronomical data analysis and bayesian inference. In: 2013 IEEE 13th International Conference on Data Mining Workshops, pp. 16–23. IEEE (2013)
97. Higson, E., Handley, W., Hobson, M., Lasenby, A.: Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation. Stat. Comput. **29**(5), 891–913 (2019)
98. Brewer, B.J., Pártay, L.B., Csányi, G.: Diffusive nested sampling. Stat. Comput. **21**(4), 649–656 (2011)
99. Akeret, J., Refregier, A., Amara, A., Seehars, S., Hasner, C.: Approximate bayesian computation for forward modeling in cosmology. J. Cosmol. Astropart. Phys. **2015**(08), 043 (2015)
100. Taylor, P.L., Kitching, T.D., Alsing, J., Wandelt, B.D., Feeney, S.M., McEwen, J.D.: Cosmic shear: Inference from forward models. Phys. Rev. D **100**(2), 023519 (2019)
101. Savage, R.S., Oliver, S.: Bayesian methods of astronomical source extraction. Astrophys. J. **661**(2), 1339 (2007)
102. Rogers, K.K., Peiris, H.V., Pontzen, A., Bird, S., Verde, L., Font-Ribera, A.: Bayesian emulator optimisation for cosmology: application to the lyman-alpha forest. J. Cosmol. Astropart. Phys. **2019**(02), 031 (2019)
103. Ishida, E., Vitenti, S., Penna-Lima, M., Cisewski, J., de Souza, R., Trindade, A., Cameron, E., Busti, V., collaboration, C., et al.: Cosmoabc: likelihood-free inference via population monte carlo approximate bayesian computation. Astron. Comput. **13**, 1–11 (2015)
104. Cameron, E., Pettitt, A.: Approximate bayesian computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift. Mon. Not. R. Astron. Soc. **425**(1), 44–65 (2012)
105. Leclercq, F.: Bayesian optimization for likelihood-free cosmological inference. Phys. Rev. D **98**(6), 063511 (2018)
106. Pellejero-Ibañez, M., Angulo, R.E., Aricó, G., Zennaro, M., Contreras, S., Stücker, J.: Cosmological parameter estimation via iterative emulation of likelihoods. Mon. Not. R. Astron. Soc. **499**(4), 5257–5268 (2020)
107. Gao, G., Jiang, H., Vink, J.C., Chen, C., El Khamra, Y., Ita, J.J.: Gaussian mixture model fitting method for uncertainty quantification by conditioning to production data. Comput. Geosc, 1–19 (2019)
108. Kristiadi, A., Däubener, S., Fischer, A.: Predictive uncertainty quantification with compound density networks. arXiv:1902.01080 (2019)
109. Zhu, Y., Zabaras, N.: Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification. J. Comput. Phys. **366**, 415–447 (2018)
110. Goyal, J.M., Wakeford, H.R., Mayne, N.J., Lewis, N.K., Drummond, B., Sing, D.K.: Fully scalable forward model grid of exoplanet transmission spectra. Mon. Not. R. Astron. Soc. **482**(4), 4503–4513 (2019)
111. Schmidt, F., Elsner, F., Jasche, J., Nguyen, N.M., Lavaux, G.: A rigorous eft-based forward model for large-scale structure. J. Cosmol. Astropart. Phys. **2019**(01), 042 (2019)
112. Bailer-Jones, C.A.: The ilium forward modelling algorithm for multivariate parameter estimation and its application to derive stellar parameters from gaia spectrophotometry. Mon. Not. R. Astron. Soc. **403**(1), 96–116 (2010)
113. Sartori, L.F., Trakhtenbrot, B., Schawinski, K., Caplar, N., Treister, E., Zhang, C.: A forward modeling approach to agn variability–method description and early applications. Astrophys. J. **883**(2), 139 (2019)
114. Hu, F.-M., Jiang, M.-H.: The fuzzy classification of the solar cycle and the prediction for the 22nd solar cycle. ChJSS **5**, 237–244 (1985)
115. Metcalfe, T.S.: Genetic-algorithm-based light-curve optimization applied to observations of the w ursae majoris star bh cassiopeiae. Astron. J **117**(5), 2503 (1999)
116. Ordóñez, D., Dafonte, C., Manteiga, M., Arcay, B.: Parameterization of rvs synthetic stellar spectra for the esa gaia mission: Study of the optimal domain for ann training. Expert Syst. Appl. **37**(2), 1719–1727 (2010)

117. Spiekermann, G.: Automated morphological classification of faint galaxies. In: Digitised Optical Sky Surveys, pp. 209–213. Springer (1992)
118. Dumitrescu, A., Pop, A., Dumitrescu, D.: Structural properties of pulsating star light curves through fuzzy divisive hierarchical clustering. Astrophys. Space Sci. **250**(2), 205–226 (1997)
119. Rodrıéguez, A., Arcay, B., Dafonte, C., Manteiga, M., Carricajo, I.: Automated knowledge-based analysis and classification of stellar spectra using fuzzy reasoning. Expert Syst. Appl. **27**(2), 237–244 (2004)
120. Liu, Z.-B., Gao, Y.-Y., Wang, J.-Z., et al.: Automatic classification method of star spectra data based on manifold fuzzy twin support vector machine. Spectrosc. Spectr. Anal. **35**(1), 263–266 (2015)
121. Revathy, K., Lekshmi, S., Nayar, S.P.: Fractal-based fuzzy technique for detection of active regions from solar images. Solar Phys. **228**(1-2), 43–53 (2005)
122. Freistetter, F.: Fuzzy characterization of near-earth-asteroids. Celest. Mech. Dyn. Astron. **104**(1-2), 93–102 (2009)
123. Shamir, L., Nemiroff, R.J.: Astronomical pipeline processing using fuzzy logic. Appl. Soft Comput. **8**(1), 79–87 (2008)
124. Attia, A.-F.: Hierarchical fuzzy controllers for an astronomical telescope tracking. Appl. Soft Comput. **9**(1), 135–141 (2009)
125. Charbonneau, P.: Genetic algorithms in astronomy and astrophysics. Astrophys. J. Suppl. Ser. **101**, 309 (1995)
126. Jin-shu, H.: Parameter estimation of stellar population synthesis using a combined genetic algorithm. Chin. Astron. Astrophys. **39**(4), 454–465 (2015)
127. Oussous, A., Benjelloun, F.-Z., Lahcen, A.A., Belfkih, S.: Big data technologies: A survey. J. King Saud Univ.-Comput Inf. Sci. **30**(4), 431–448 (2018)
128. Furht, B., Villanustre, F.: Introduction to big data. In: Big Data Technologies and Applications, pp. 3–11. Springer (2016)
129. Kapil, G., Agrawal, A., Khan, R.: A study of big data characteristics. In: 2016 International Conference on Communication and Electronics Systems (ICCES), pp. 1–4. IEEE (2016)
130. York, D.G., Adelman, J., Anderson Jr, J.E., Anderson, S.F., Annis, J., Bahcall, N.A., Bakken, J., Barkhouser, R., Bastian, S., Berman, E., et al.: The sloan digital sky survey: Technical summary. Astron. J. **120**(3), 1579 (2000)
131. Alam, S., Albareti, F.D., Prieto, C.A., Anders, F., Anderson, S.F., Anderton, T., Andrews, B.H., Armengaud, E., Aubourg, É., Bailey, S., et al.: The eleventh and twelfth data releases of the sloan digital sky survey: final data from sdss-iii. Astrophys. J. Suppl. Ser. **219**, 12 (2015)
132. Vipers:the vimos public extragalactic redshift survey. http://vipers.inaf.it (2020)
133. Guzzo, L., Scodeggio, M., Garilli, B., Granett, B., Fritz, A., Abbas, U., Adami, C., Arnouts, S., Bel, J., Bolzonella, M., et al.: The vimos public extragalactic redshift survey (vipers)-an unprecedented view of galaxies and large-scale structure at 0.5< z< 1.2. Astron. Astrophys. **566**, 108 (2014)
134. Manzoni, G., Scodeggio, M., Baugh, C., Norberg, P., De Lucia, G., Fritz, A., Haines, C., Zamorani, G., Gargiulo, A., Guzzo, L., et al.: Modelling the quenching of star formation activity from the evolution of the colour-magnitude relation in vipers. New Astron. **84**, 101515 (2021)
135. The Two Micron All Sky Survey at IPAC. https://old.ipac.caltech.edu/2mass/ (As on June, 2020)
136. Conselice, C., Bundy, K., Trujillo, I., Coil, A., Eisenhardt, P., Ellis, R., Georgakakis, A., Huang, J., Lotz, J., Nandra, K., et al.: The properties and evolution of a k-band selected sample of massive galaxies at z 0.4–2 in the palomar/deep2 survey. Mon. Not. R. Astron. Soc. **381**(3), 962–986 (2007)
137. The large synoptic survey telescope. https://www.lsst.org/lsst (2020)
138. SKA in India: science with big data. https://asi2020.astron-soc.in/workshops/workshop3/ (As on July, 2020)
139. Ligolaser interferometer gravitational-wave observatory. https://www.ligo.caltech.edu (As on June, 2020)
140. Fevre, O.L., Cassata, P., Cucciati, O., Garilli, B., Ilbert, O., Brun, V.L., Maccagni, D., Moreau, C., Scodeggio, M., Tresse, L., et al.: The vimos vlt deep survey final data release: a spectroscopic sample of 35016 galaxies and agn out to z˜ 6.7 selected with 17.5<= i_ {AB} <= 24.7. arXiv:1307.0545 (2013)
141. Lawrence, A., Warren, S., Almaini, O., Edge, A., Hambly, N., Jameson, R., Lucas, P., Casali, M., Adamson, A., Dye, S., et al.: The ukirt infrared deep sky survey (ukidss). Mon. Not. R. Astron. Soc. **379**, 1599–1617 (2007)

142. Pović, M., Huertas-Company, M., Aguerri, J., Márquez, I., Masegosa, J., Husillos, C., Molino, A., Cristóbal-Hornillos, D., Perea, J., Benítez, N., et al.: The alhambra survey: reliable morphological catalogue of 22 051 early-and late-type galaxies. Mon. Not. R. Astron. Soc. **435**(4), 3444–3461 (2013)
143. Djorgovski, S., Gal, R., Odewahn, S., De Carvalho, R., Brunner, R., Longo, G., Scaramella, R.: The palomar digital sky survey (dposs). arXiv:astro-ph/9809187 (1998)
144. Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., et al.: Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. Mon. Not. R. Astron. Soc. **389**(3), 1179–1189 (2008)
145. Cutri, R.e., Wright, E., Conrow, T., Fowler, J., Eisenhardt, P., Grillmair, C., Kirkpatrick, J., Masci, F., McCallon, H., Wheelock, S., et al.: Vizier online data catalog: Allwise data release (cutri+ 2013). VizieR Online Data Catalog, 328 (2021)
146. de Jong, J.T., Kleijn, G.A.V., Kuijken, K.H., Valentijn, E.A., et al.: The kilo-degree survey. Exp. Astron. **35**(1-2), 25–44 (2013)
147. Zhang, Y., Zhao, Y.: Astronomy in the big data era. Data Sci. J. **14** (2015)
148. Wells, D.C., Greisen, E.W.: Fits-a flexible image transport system. In: Image Processing in Astronomy, p. 445 (1979)
149. Anderson, K., Alexov, A., Baehren, L., Grießmeier, J.-M., Wise, M., Renting, A.: Lofar and hdf5: Toward a new radio data standard. arXiv:1012.2266 (2010)
150. Goucher, G., Love, J., Leckner, H.: A discipline independent scientific data management package-the national space science common data format (cdf). step, 691 (1994)
151. Warren-Smith, R., Lawden, M., McIlwrath, B., Jenness, T., Draper, P.: Hds heirarchical data system: Programmer's manual. Technical report, Technical Report. Council for the Central Laboratory of the Research … (2008)
152. Williams, R., Ochsenbein, F., Davenhall, C., Durand, D., Fernique, P., Giaretta, D., Hanisch, R., McGlynn, T., Szalay, A., Wicenec, A.: Votable: A proposed xml format for astronomical tables. CDS: Strasbourg **28** (2002)
153. Thomas, B., Shaya, E., Gass, J., Blackwell, J., Cheung, C.: An xml representation of fits-introducing fitsml. AAS **197**, 116–03 (2000)
154. Greenfield, P., Droettboom, M., Bray, E.: Asdf: A new data format for astronomy. Astron. Comput. **12**, 240–251 (2015)
155. Patidar, S., Rane, D., Jain, P.: A survey paper on cloud computing. In: 2012 Second International Conference on Advanced Computing & Communication Technologies, pp. 394–398. IEEE (2012)
156. Berriman, G.B., Juve, G., Deelman, E., Regelson, M., Plavchan, P.: The application of cloud computing to astronomy: A study of cost and performance. In: 2010 Sixth IEEE International Conference on e-Science Workshops, pp. 1–7. IEEE (2010)
157. Araya, M., Osorio, M., Díaz, M., Ponce, C., Villanueva, M., Valenzuela, C., Solar, M.: Jovial: Notebook-based astronomical data analysis in the cloud. Astron. Comput. **25**, 110–117 (2018)
158. Grid computing to tackle the mystery of the dark universe. https://astronomynow.com/2016/11/26/grid-computing-to-tackle-the-mystery-of-the-dark-universe/ (As on December, 2020)
159. Spark. http://spark.apache.org/ (As on June, 2020)
160. Flume. https://flume.apache.org/ (As on June, 2020)
161. Apache Pig. https://pig.apache.org/ (As on June, 2020)
162. Apache Oozie. https://oozie.apache.org/ (As on June, 2020)
163. Statwing. https://www.statwing.com/ (As on June, 2020)
164. Stonebraker, M., Brown, P., Zhang, D., Becla, J.: Scidb: A database management system for applications with complex analytics. Comput. Sci. Eng. **15**(3), 54–62 (2013)
165. Saha, B., Shah, H., Seth, S., Vijayaraghavan, G., Murthy, A., Curino, C.: Apache tez: A unifying framework for modeling and building data processing applications. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1357–1369 (2015)
166. Parallel supercomputing for astronomy
167. Liu, L., Liu, D., Lü, S., Zhang, P.: An abstract description method of map-reduce-merge using haskell. Math. Probl. Eng. **2013** (2013)
168. Zhou, L., Huang, M.: Challenges of software testing for astronomical big data. In: 2017 IEEE International Congress on Big Data (BigData Congress), pp. 529–532. IEEE (2017)
169. Szalay, A.S., Gray, J., Kunszt, P., Thakar, A., Slutz, D.: Large databases in astronomy. In: Mining the Sky, pp. 99–116. Springer (2001)

170. Brahem, M., Zeitouni, K., Yeh, L.: Astroide: a unified astronomical big data processing engine over spark. IEEE Trans. Big Data (2018)
171. Jacob, J.C., Katz, D.S., Miller, C.D., et al.: Grist: Grid-based Data Mining for Astronomy, Astronomical Data Analysis Software and Systems XIV, ASP Conference Series, Vol. XXX (2005)
172. Ivanova, M., Nes, N., Goncalves, R., Kersten, M.: Monetdb/sql meets skyserver: the challenges of a scientific database. In: 19th International Conference on Scientific and Statistical Database Management (SSDBM 2007), pp. 13–13. IEEE (2007)
173. Juric, M.: Large survey database: A distributed framework for storage and analysis of large datasets. AAS **217**, 433–19 (2011)
174. Wiley, K., Connolly, A., Gardner, J., Krughoff, S., Balazinska, M., Howe, B., Kwon, Y., Bu, Y.: Astronomy in the cloud: using mapreduce for image co-addition. Publ. Astron. Soc. Pac. **123**(901), 366 (2011)
175. Brahem, M., Zeitouni, K., Yeh, L.: Hx-match: In-memory cross-matching algorithm for astronomical big data. In: International Symposium on Spatial and Temporal Databases, pp. 411–415. Springer (2017)
176. Brahem, M., Lopes, S., Yeh, L., Zeitouni, K.: Astrospark: towards a distributed data server for big data in astronomy. In: Proceedings of the 3rd ACM SIGSPATIAL PhD Symposium, pp. 1–4 (2016)
177. Zhang, Z., Barbary, K., Nothaft, F.A., Sparks, E.R., Zahn, O., Franklin, M.J., Patterson, D.A., Perlmutter, S.: Kira: Processing astronomy imagery using big data technology. IEEE Trans. Big Data (2016)
178. Zečević, P., Slater, C.T., Jurić, M., Connolly, A.J., Lončarić, S., Bellm, E.C., Golkhou, V.Z., Suberlak, K.: Axs: A framework for fast astronomical data processing based on apache spark. Astron. J. **158**(1), 37 (2019)
179. Garofalo, M., Botta, A., Ventre, G.: Astrophysics and big data: Challenges, methods, and tools. Proc. Int. Astron. Union **12**(S325), 345–348 (2016)
180. Ball, N.M.: Canfar+ skytree: A cloud computing and data mining system for astronomy. arXiv:1312.3996 (2013)
181. Hong, S., Jeong, D., Hwang, H.S., Kim, J., Hong, S.E., Park, C., Dey, A., Milosavljevic, M., Gebhardt, K., Lee, K.-S.: Constraining cosmology with big data statistics of cosmological graphs. Mon. Not. R. Astron. Soc. **493**(4), 5972–5986 (2020)
182. Vujčić, V., Jevremović, D.: Real-time stream processing in astronomy. In: Knowledge Discovery in Big Data from Astronomy and Earth Observation, pp. 173–182. Elsevier (2020)
183. Sciacca, E., Pistagna, C., Becciani, U., Costa, A., Massimino, P., Riggi, S., Vitello, F., Bandieramonte, M., Krokos, M.: Towards a big data exploration framework for astronomical archives. In: 2014 International Conference on High Performance Computing & Simulation (HPCS), pp. 351–357 (2014). IEEE
184. Fillatre, L., Lepiller, D.: Processing solutions for big data in astronomy. EAS Publ. Ser. **78**, 179–208 (2016)
185. Zhao, Q., Sun, J., Yu, C., Cui, C., Lv, L., Xiao, J.: A paralleled large-scale astronomical cross-matching function. In: International Conference on Algorithms and Architectures for Parallel Processing, pp. 604–614. Springer (2009)
186. Mesmoudi, A., Hacid, M.-S., Toumani, F.: Benchmarking sql on mapreduce systems using large astronomy databases. Distrib. Parallel Databases **34**(3), 347–378 (2016)
187. Peloton, J., Arnault, C., Plaszczynski, S.: Analyzing astronomical data with apache spark. arXiv:1804.07501 (2018)
188. Xie, D., Li, F., Yao, B., Li, G., Zhou, L., Guo, M.: Simba: Efficient in-memory spatial analytics. In: Proceedings of the 2016 International Conference on Management of Data, pp. 1071–1085 (2016)
189. Wei, S., Wang, F., Deng, H., Liu, C., Dai, W., Liang, B., Mei, Y., Shi, C., Liu, Y., Wu, J.: Opencluster: a flexible distributed computing framework for astronomical data processing. Publ. Astron Soc Pac. **129**(972), 024001 (2016)
190. Berriman, G.B., Good, J.: The application of the montage image mosaic engine to the visualization of astronomical images. Publ. Astron. Soc. Pac. **129**(975), 058006 (2017)
191. Corizzo, R., Ceci, M., Zdravevski, E., Japkowicz, N.: Scalable auto-encoders for gravitational waves detection from time series data. Expert Syst. Appl., 113378 (2020)
192. Sen, S., Saha, S., Chakraborty, P., Singh, K.P.: Implementation of neural network regression model for faster redshift analysis on cloud-based spark platform. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 591–602. Springer (2021)

193. Vanderplas, J.T., Connolly, Ž., Ivezić, A.J., Gray, A.: Introduction to astroml: Machine learning for astrophysics, pp. 47–54. https://doi.org/10.1109/CIDU.2012.6382200 (2012)
194. Saha, S., Agrawal, S., R, M., Bora, K., Routh, S., Narasimhamurthy, A.: ASTROMLSKIT: a new statistical machine learning toolkit: a platform for data analytics in astronomy (2015)
195. Astropy. https://www.astropy.org/ (As on June, 2020)
196. González, R.E., Muñoz, R.P., Hernández, C.A.: Astrocv: Astronomy computer vision library. ASCL, 1804 (2018)
197. http://astroweka.sourceforge.net/: Astroweka (Collected on June,2020)
198. pyfits 3.3. https://pypi.org/project/pyfits/3.3/ (As on June, 2020)
199. Singh, N., Browne, L.-M., Butler, R.: Parallel astronomical data processing with python: Recipes for multicore machines. Astron. Comput. **2**, 1–10 (2013)
200. pyraf. http://astro.if.ufrgs.br/ (As on June, 2020)
201. Khlamov, S., Savanevych, V., Briukhovetskyi, O., Pohorelov, A., Vlasenko, V., Dikov, E.: Colitec software for the astronomical data sets processing. In: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), pp. 227–230. IEEE (2018)
202. Breddels, M.A., Veljanoski, J.: Vaex: big data exploration in the era of gaia. Astron. Astrophys. **618**, 13 (2018)
203. https://authors.library.caltech.edu/50265//: Dameware, a web cyberinfrastructure for astrophysical data mining (As on June, 2020)
204. Welge, M., Hsu, W., Auvil, L., Redman, T., Tcheng, D.: High-performance knowledge discovery and data mining systems using workstation clusters. In: 12th National Conference on High Performance Networking and Computing (SC99) (1999)
205. https://ipython.org/: Ipython interactive computing (As on June, 2020)
206. Yu, W., Kind, M.C., Brunner, R.J.: Vizic: A jupyter-based interactive visualization tool for astronomical catalogs. Astron. Comput. **20**, 128–139 (2017)
207. https://astrostatistics.psu.edu/statcodes/: Online statistical software for astronomy and related physical sciences (As on June, 2020)
208. Jacob, J.C., Katz, D.S., Berriman, G.B., Good, J.C., Laity, A., Deelman, E., Kesselman, C., Singh, G., Su, M.-H., Prince, T., et al.: Montage: a grid portal and software toolkit for science-grade astronomical image mosaicking. Int. J. Comput. Sci. Eng. **4**(2), 73–87 (2009)
209. Tools for astronomical big data. https://www.noao.edu/meetings/bigdata/ (As on July, 2020)
210. Big data and astronomy. http://www.astro4dev.org/jan-mar-2017/ (As on July, 2020)
211. 2nd Australia-China SKA big data workshop. https://eridanus.net.au/?p=269 (As on July, 2020)
212. Artificial intelligence in astronomy. https://www.eso.org/sci/meetings/2019/AIA2019.html (As on July, 2020)
213. Machine learning tools for research in astronomy. https://www2.mpia-hd.mpg.de/ml2019/ (As on July, 2020)
214. Swiss-SA Astronomy. https://astro.ukzn.ac.za/swiss-sa-astronomy-big-data-workshop/ (As on July, 2020)
215. Innovation in data driven astronomy. https://www.nrao.edu/meetings/bigdata/index.shtml (As on June, 2011)
216. Data science for physics. https://www.turing.ac.uk/events/data-science-physics-and-astronomy-scopin g-workshop/ (As on July, 2020)
217. International conference on modeling, machine learning and astronomy. http://mmla.pes.edu/ (As on June, 2019)
218. Bigdata and digital technoloy. https://indico.narit.or.th/ (As on July, 2020)
219. Data science in astrophysics. https://dsap.iiita.ac.in/ (As on June, 2020)
220. Machine learning in astronomical data analysis. http://hea-www.harvard.edu/AstroStat/aas233/special.html/ (As on July, 2020)
221. Bigdata challenge. dca2019.csp.escience.cn (As on July, 2020)
222. AstroInformatics virtual conference. https://www.astroinformatics2020.org/ (As on July, 2020)
223. Workshop: astronomical data science. https://tamids.tamu.edu/2020 (As on July, 2020)
224. IAU symposia. https://www.iau.org/science/meetings/future/symposia/2528/ (As on July, 2020)
225. Astronomy in the big data era. https://generalassemb.ly/education/astronomy-in-the-big-data-era/dallas/ (As on July, 2020)
226. ADASS. http://adass2018.astro.umd.edu/ (As on July, 2020)

227. Berriman, G.B., Groom, S.L.: How will astronomy archives survive the data tsunami? Communications of the ACM (2011)
228. Nichols, M.R.: The fast and the curious: How's big data changing astronomy?. https://schooledbyscience.com/big-datas-changing-astronomy/ (2016)
229. How Big Data Analytics is shaping Astronomy. https://runyourbusiness.deskera.in/big-data-analytics-shaping-astronomy/ (As on July, 2020)
230. Big data is transforming. https://www.smithsonianmag.com/science-nature/ (As on July, 2020)
231. Henry, L.: Data's big bang: Applying analytics to astronomy. https://www.informationweek.com/datas-big-bang-applying-analytics-to-astronomy/a/d-id/282405 (2017)
232. Andersen, R.: How big data is changing astronomy (again). https://www.theatlantic.com/technology/archive/2012/04/how-big-data-is-changing-astronomy-again/255917/ (2012)
233. Urton, J.: With launch of new night sky survey, uw researchers ready for era of 'big data' astronomy. https://www.washington.edu/news/2017/11/14/with-launch-of-new-night-sky-survey-uw-researchers-ready-for-era-of-big-data-astronomy/ (2017)
234. Norris, R.: Expect the unexpected from the big-data boom in radio astronomy. https://phys.org/news/2017-09-unexpected-big-data-boom-radioastronomy.html (2017)
235. Mcguire, A.: It's the turn of the celestial world now, big data transforming astronomy!. https://irishtechnews.ie/its-the-turn-of-the-celestial-world-now-big-data-transforming-astronomy/ (As on June, 2020)
236. Data science in astronomy. https://medium.com/trends-in-data-science/data-science-in-astronomy-f0e9b499273/ (As on July, 2020)
237. Ananthaswamy, A.: Faced with a data deluge, astronomers turn to automation. https://irishtechnews.ie/its-the-turn-of-the-celestial-world-now-big-data-transforming-astronomy/ (As on June, 2020)
238. Beginning with ML. https://beginningwithml.wordpress.com/ (2020)
239. Murphy, T.: Data-driven astronomy. https://www.coursera.org/learn/data-driven-astronomy (As on June, 2020)
240. DataMining and machine learning in astronomy. https://www.as.arizona.edu/ (As on July, 2020)
241. University, L.: Astronomy and data science. https://www.mastersportal.com/studies/188902/astronomy-and-data-science.html, (As on June, 2020)
242. BigSkyEarth. https://bigskyearth.eu/ (As on July, 2020)
243. Astrostatistics and astroinformatics portal. https://asaip.psu.edu// (As on June, 2020)
244. Astroinformatics research group. http://astrirg.org/ (As on June, 2020)
245. IDIA data intensive astronomy cloud. http://www.researchsupport.uct.ac.za/idia-data-intensive-astronomy-cloud (As on June, 2020)
246. Linghe Kong, Y.Z., Tian Huang, Y., S.: Big data in astronomy (16th June 2020)
247. Edwards, K.J., Gaber, M.M.: Astronomy and big data. Studies in Big Data. Springer (2014)
248. Skoda, P., Adam, F.: Knowledge discovery in big data from astronomy and earth observation 1st edition (2020)
249. Murtagh, F., Heck, A.: Multivariate data analysis. **131** (2012)
250. Wall, J.V., Jenkins, C.R.: Practical statistics for astronomers (2012)
251. Ivezić, Ž., Connolly, A.J., VanderPlas, J.T., Gray, A.: Statistics, data mining, and machine learning in astronomy: a practical python guide for the analysis of survey data. **1** (2014)
252. Cavuoti, S.: Data-rich astronomy: mining synoptic sky surveys. arXiv:1304.6615 (2013)
253. Babu, G.J., Feigelson, E.D.: Statistical challenges in modern astronomy ii (2012)
254. Feigelson, E.D., Jogesh, B.G.: Statistical challenges in modern astronomy ii (1997)
255. Podgorski, K.: Advances in machine learning and data mining for astronomy. Int. Stat. Rev **82**(1), 153–154 (2014)
256. Kumar, M.: Sparse image and signal processing: Wavelets, curvelets, morphological diversity, by jean-luc starck, fionn murtagh, and jalal m. fadili. J. Electron. Imaging **19**(4), 049901 (2007)
257. Chattopadhyay, A.K., Chattopadhyay, T.: Statistical Methods for Astronomical Data Analysis, vol. 3. Springer, New York (2014)
258. et al., S.S.: Machine learning in astronomy: A workman's manual (2017)
259. Scientific discovery through advanced computing. https://www.scidac.gov/ (As on January, 2021)
260. Project. https://www.scidac.org/tags/projects.html (As on January, 2021)
261. SciDAC-3 scientific computation application partnership project. https://www.bnl.gov/physics/scidac/ (As on January, 2021)

262. LQCD SciDAC-4 project. https://lqcdscidac4.github.io/ (As on January, 2021)
263. Frameworks, algorithms and scalable technologies for mathematics (FASTMath) SciDAC-5 Institute. https://scidac5-fastmath.lbl.gov/home (As on January, 2021)
264. Laureijs, R., Amiaux, J., Arduini, S., Augueres, J.-L., Brinchmann, J., Cole, R., Cropper, M., Dabin, C., Duvet, L., Ealet, A., et al.: Euclid definition study report. arXiv:1110.3193 (2011)
265. AWS SageMaker. https://aws.amazon.com/sagemaker/ (As on July, 2020)

Affiliations

Snigdha Sen[1,2] · Sonali Agarwal[1] · Pavan Chakraborty[1] · Krishna Pratap Singh[1]

Sonali Agarwal
sonali@iiita.ac.in

Pavan Chakraborty
pavan@iiita.ac.in

Krishna Pratap Singh
kpsingh@iiita.ac.in

[1] Department of Information Technology, Indian Institute of Information Technology, Jhalwa, Prayagraj, 211015, Uttar Pradesh, India

[2] Department of CSE, Global Academy of Technology, RR Nagar, Bangalore, 560098, Karnataka, India