International Journal of Mathematics, Statistics, and Computer Science
Vol. 2, 2024
ISSN: 2704-1077 eISSN 2704-1069, DOI: 10.59543/ijmscs.v2i.8561

Overview of Big Data Analytics in Modern Astronomy

Muhammad Faaique*
Department of Computer Science, Mehran University of Engineering and Technology (MUET), Jamshoro, Sindh, Pakistan; memon.faique@gmail.com

Received 17.10.2023, Revised 28.11.2023, Accepted 09.12.2023, Published 19.12.2023

ABSTRACT: Astronomers are increasingly compelled to chart the universe with ever greater precision. Projects like the Sloan Digital Sky Survey (SDSS), Pan-STARRS, and the Large Synoptic Survey Telescope (LSST) generate approximately 100-200 Petabytes of data annually, presenting significant big data challenges in terms of storage, processing, and data transfer. The Square Kilometer Array (SKA), an ambitious project involving 130,000 antennas and 200 dishes spanning two continents, is scheduled to become operational in 2028. It will collect 160 terabytes of data per second, translating to 1 petabyte of data daily. Coping with this immense volume of data necessitates real-time processing and analysis, driving the need for efficient machine learning and image analysis algorithms. Astronomy stands as an ideal domain for big data analytics, pushing the boundaries of data analysis. This review paper will present intriguing applications for data scientists, exploring the challenges and recent technological advancements in big data analytics concerning astronomy. The paper will also critically assess the strengths and weaknesses of various approaches, methodologies, or tools used in big data analytics within the context of astronomy, supported by relevant case studies.

Keywords: Big Data Analytics, Astronomy, Machine learning.

1. INTRODUCTION

Astronomy, the scientific investigation of celestial objects and phenomena, has entered a transformative era defined by the deluge of Big Data [1]. This evolution, characterized by an exponential surge in data volume, has revolutionized our comprehension of the universe, presenting us with thrilling prospects and intricate challenges. In this exploration, we will dive into the profound influence of astronomy on our understanding of the cosmos, with a particular focus on how the influx of data has played a pivotal role in reshaping this field.

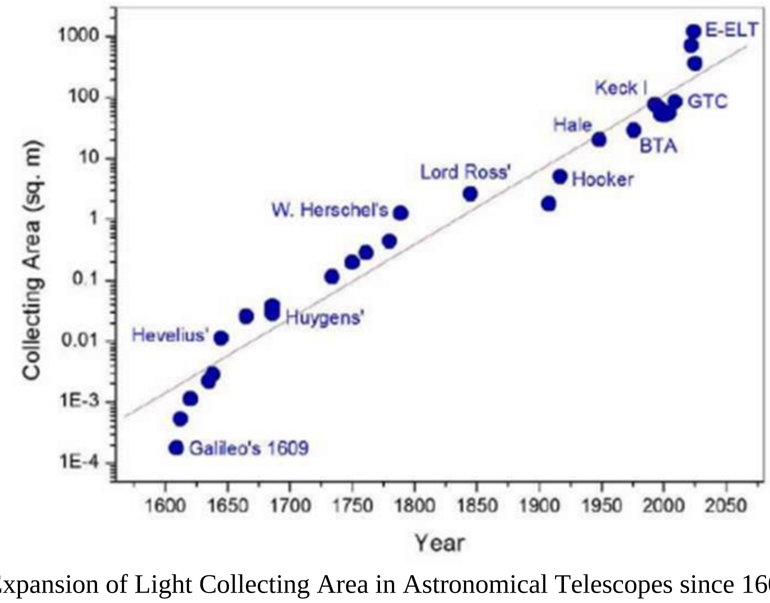

Figure 1: The Expansion of Light Collecting Area in Astronomical Telescopes since 1609. This information is sourced from [11].

Astronomy has traditionally been closely associated with managing vast amounts of data. Nonetheless, there has been a profound paradigm shift in recent decades, with the field embracing what we now recognize as Big Data [2]. This transformation has expanded the horizons of astronomical knowledge, allowing it to transcend the limitations of Earth's physical conditions and explore the diversity of celestial objects. A pivotal factor driving this transformation has been the rise of multiwavelength (MW) observations, encompassing the electromagnetic spectrum from gamma rays to radio wavelengths. These observations have unlocked new horizons in our quest to understand the universe, offering insights that transcend the constraints of traditional astronomy [3].

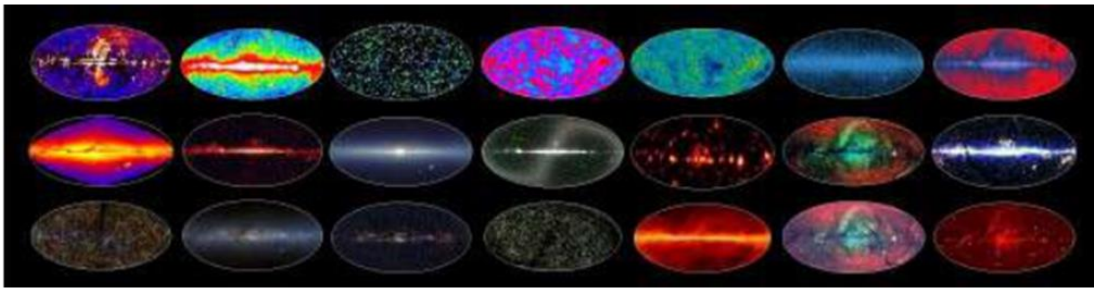

Figure 2: Various Perspectives of the Sky in Diverse Wavelength Ranges: Highlighting the Significance of Multiwavelength (MW) Studies for Gaining Comprehensive Insights into Cosmic Objects and the Universe. This reference is drawn from [11].

Astronomy has undergone a remarkable revolution in data collection, storage, and distribution in the era of Big Data. Astrophysical Virtual Observatories (VOs) have emerged as central players in this landscape, providing standardized access to a wide array of datasets and facilitating sophisticated research endeavors that seamlessly amalgamate data from ground-based and space telescopes. This harmonization across diverse observing methods, temporal domains, and wavelengths has ushered in a new era of possibilities for astronomers [4]. An in-depth examination of the celestial realm is imperative because celestial objects emit energy over a broad spectrum of wavelengths, from radio and infrared to optical, ultraviolet, X-rays, and gamma rays. Each wavelength conveys distinct information about these objects, and the same object can present different appearances at different wavelengths. For instance, a youthful spiral galaxy might manifest as compacted clusters in the ultraviolet spectrum yet exhibit well-defined spiral arms in the optical spectrum. Likewise, X-ray observations can unveil hot and dispersed gases between galaxies within a cluster. To comprehensively comprehend the physical phenomena transpiring within these celestial entities, astronomers must amalgamate observations from a multitude of wavelengths. There is substantial coverage across ten distinct spectral regions, and we anticipate additional data becoming available in at least five more spectral bands soon. Nevertheless, managing and integrating this wealth of data, dispersed across diverse archives, presents a formidable and intricate challenge [4].

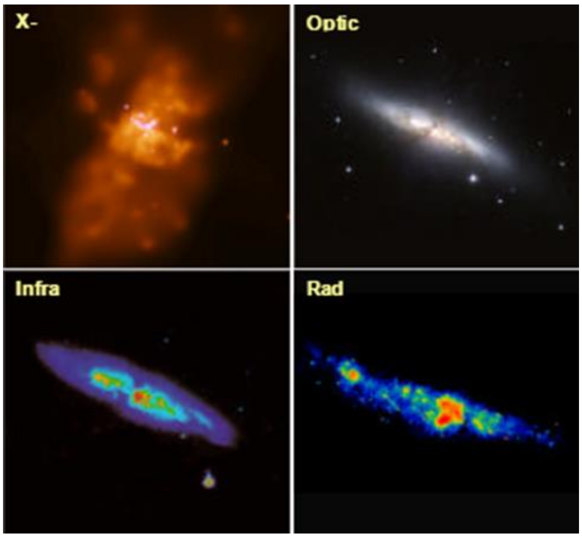

Figure 3: The images showcased offer diverse viewpoints of the star-forming activities happening within M82, each captured at different wavelengths. The sources for these images are as follows: X-ray image -

NASA/CXC/SAO/PSU/CMU, Optical image - AURA/NOAO/NSF, Infrared image - SAO, Radio image - MERLIN/VLA

The four Vs define Big Data in astronomy: Volume, Variety, Velocity, and Veracity. The volume of data generated and stored is enormous, incorporating vast amounts of celestial objects spanning galaxies and beyond, all housed in extensive databases. This data abundance arrives in many formats, including text, images, audio, and video, posing a challenge in effectively extracting meaningful insights from this diverse and heterogeneous collection [5].

Real-time data availability, called Velocity, has evolved into a pivotal element of contemporary astronomical research, propelling us into the era of data-driven discovery. Yet, amidst this data deluge, maintaining data quality, indicated as Veracity, remains equally vital for precise analysis and interpretation [6].

Astronomical surveys, especially all-sky and large-area surveys, are fundamental in amassing observational data and unveiling novel celestial objects and phenomena. These surveys, covering the complete electromagnetic spectrum, have brought about a transformative shift in our comprehension of the universe. From gamma-ray observations carried out by missions such as CGRO [7] and Fermi-GLAST [8] to optical surveys like SDSS [9] and infrared missions like 2MASS [10], a holistic perspective of the cosmos has arisen, fundamentally reshaping our insights into the universe.

The quest for a deeper understanding has driven astronomers to embark on ambitious initiatives like the Hubble Deep Field (HDF) and the Cosmic Evolution Survey (COSMOS). These deep field projects have been pivotal in unveiling the most mysterious and remote objects concealed within the extensive datasets. To handle this vast amount of data, astronomers have leveraged the capabilities of advanced computing technologies, including clusters and grids [11].

Additionally, the global recognition of the necessity for data preservation and sharing has prompted the establishment of the World Data System (WDS). WDS is a unifying platform for data across a spectrum of scientific domains, including astronomy, promoting collaborative research initiatives [11].

In conclusion, the transition of astronomy into the era of Big Data has expanded its reach far beyond the confines of our planet. Thanks to the copious data generated by all-sky and large-area surveys, the field's capacity to contribute to a wide array of scientific disciplines has never been more significant. With the emergence of astrophysical virtual observatories and powerful computing technologies, astronomers are well-prepared to unveil the universe's concealed mysteries hidden within this vast sea of data. As astronomy continues to evolve, its intersection with other scientific realms and its impact on our comprehension of the universe is poised to intensify. In this era of big data, astronomy illuminates the cosmos and the path of scientific discovery.

2. DISCUSSION

2.1 BIG DATA CONCEPT AND CHALLENGES

Big data is a term employed to characterize huge volumes of data that exceed the processing capacity of conventional computing systems. In essence, big data encompasses extensive datasets originating from various sources, frequently in diverse formats, that go beyond the processing capabilities of traditional computing systems [12]. Consequently, there is an increasing need to develop contemporary storage facilities and real-time data analysis mechanisms to handle this data deluge effectively.

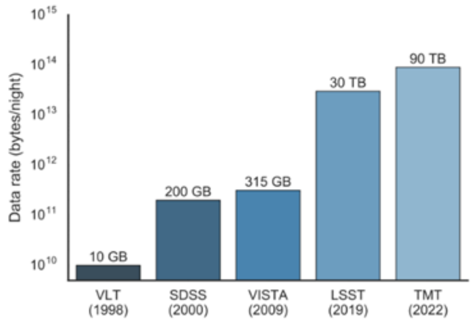

Figure 4: Escalating Data Volumes from Current and Upcoming Telescopes: Notable examples include the Very Large Telescope (VLT), Sloan Digital Sky Survey (SDSS), Visible and Infrared Telescope for Astronomy (VISTA), Large Synoptic Survey Telescope (LSST), and Thirty Meter Telescope (TMT). This information is sourced from [28].

The concept of Big Data is often characterized by the "Four Vs" [13]:

2.1.1. Volume:

This dimension concerns data's sheer quantity and size, a growing challenge for data scientists. In the modern era, data volumes can extend into tens of zettabytes, representing an exponential increase in data generation and storage.

2.1.2. Variety:

Variety encompasses the diverse formats in which data exists, including images, audio, video, emails, and more. Data can be structured, unstructured, or semi-structured, requiring distinct processing methods. Complexity arises when data includes multiple formats, such as multimedia content, which can pose significant storage and processing challenges.

2.1.3. Velocity:

Velocity relates to the speed at which data is generated and processed. It can take the form of batch processing, real-time streams, and streaming data. High-velocity real-time data necessitates rapid processing using advanced tools and techniques.

2.1.4. Veracity:

Veracity emphasizes data quality, focusing on its purity, freedom from noise, trustworthiness, and relevance to a specific task. It involves filtering out irrelevant or unimportant data to extract meaningful insights.

In addition to the primary characteristics (Volume, Variety, Velocity, and Veracity) of big data, several other factors contribute to its complexity [13]:

2.1.5. Validity Checks:

Ensuring data validity involves examining data for accuracy and reliability and identifying relationships or correlations among data items. Valid input is crucial for accurate data analysis.

2.1.6. Data Volatility:

Data volatility concerns how long data remains valid and needs to be stored. Frequently used data typically stays readily accessible, while less frequently used data can be archived to optimize storage resources.

2.1.7. Data Confidentiality:

Handling data confidentiality is essential to safeguard sensitive information and protect against unauthorized access.

2.1.8. Value Addition:

Ultimately, the value of massive data lies in its capacity to inform decision-making. Data that doesn't contribute to meaningful insights and informed choices becomes redundant.

2.2. DIFFERENT SOURCES OF ASTRONOMICAL DATA

Data collection is a crucial initial phase in any analysis, and in astronomy, this step is particularly pivotal. Astronomers depend heavily on extensive and dependable data sources to conduct meaningful research and acquire insights into the universe. Below, we spotlight some notable data sources in astronomy:

2.2.1 SDSS Survey:

The Sloan Digital Sky Survey (SDSS) is one of the most extensive data sources, offering comprehensive details about celestial objects, including stars and galaxies, across approximately one-third of the sky. This data is accessible through their website, and users can store it on their computers or in cloud storage [14].

2.2.2 VIPERS Survey:

The VIMOS Public Extragalactic Redshift Survey (VIPERS) leverages a large telescope to collect data on galaxies, aiming to enhance our understanding of the universe during a period when it was half its current age. Data from VIPERS, which encompasses information about galaxies' mass, brightness, and more, is accessible through their website [15].

2.2.3. 2-MASS Survey:

The Two Micron All-Sky Survey (2MASS) focuses on observing objects within the near-infrared segment of the spectrum. It maintains a vast catalog encompassing over 300 million objects, including stars and galaxies. This data is available online for access [16].

2.2.3 DEEP2 Survey:

The DEEP2 survey is dedicated to examining distant galaxies using the Keck telescope. It encompasses a substantial portion of the sky and offers comprehensive data regarding the positions and properties of galaxies [17].

2.2.5. LSST Survey:

The Large Synoptic Survey Telescope (LSST) is presently in the construction phase in Chile. Equipped with an immense camera, it can capture the night sky with unparalleled detail. The primary design objective of the LSST is to detect rare and faint objects in the celestial expanse [18].

2.2.6. SKA Observatory:

The Square Kilometer Array (SKA) is positioned to become the world's largest radio telescope and is slated for operation in 2028. Its scope encompasses exploring a diverse array of astronomical phenomena, ranging from dark matter and dark energy to studying exoplanets [19].

2.2.7. LIGO Survey:

The Laser Interferometer Gravitational-Wave Observatory (LIGO) is dedicated to detecting gravitational waves resulting from ripples in space-time generated by significant events, such as the collision of black holes. LIGO has accumulated data from multiple gravitational wave events [20].

The data volumes generated by different sky surveys are shown in **Table 1.**

Sky Survey Projects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	40 TB
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	~ 40 PB expected
LSST (The Large Synoptic Survey Telescope)	~ 200 PB expected
SKA (The Square Kilometer Array)	~ 4.6 EB expected

Table 1: Data Volumes for Various Sky Survey Projects. This information is sourced from [5].

Aside from the sky surveys mentioned earlier, numerous other surveys, including VVDS [21], UKIRT Infrared Deep Sky Survey [22], Alhambra survey [23], GOTO, DPOSS [24], GALAXY ZOO [25], All-Wise [26], Pan-STARRS1, KiDS [27], ZTF, PRIMUS, and GAIA, serve as valuable sources of astronomical data. Figure 3 and Table 1 provide additional information regarding the data volumes generated by these diverse sky surveys.

These surveys offer valuable datasets enabling astronomers to explore and understand the universe comprehensively. Encompassing a broad spectrum of celestial objects and phenomena, they contribute significantly to our ongoing quest to gain deeper insights into the cosmos.

2.3. TYPES OF OBSERVATIONS

These surveys employ two primary methods for observing the sky: **spectroscopy** and **photometry**. Spectroscopy involves measuring light in various colors, aiding in identifying the chemical composition of celestial objects. For instance, it can reveal the elements present in stars and galaxies. On the other hand, photometry entails capturing images using a few different filters. However, it may not provide the same level of detail as spectroscopy, but it can still capture fainter objects [28]. While spectroscopy offers precise measurements, it does have its limitations.

While spectroscopy is a powerful tool for studying the chemical composition of celestial objects, it has limitations. It may not detect very faint objects effectively, and it is not the most efficient method when observing many objects simultaneously. Table 3 presents some frequently used feature selection/extraction methods, while other literature sources, such as those mentioned by Zheng and Zhang (2008), provide further insights into these techniques. Numerous books have been published that focus on data mining in astronomy, covering topics like multivariate data analysis, practical statistics, machine learning, and data analysis of astronomical images. These resources contribute to the ongoing advancement of data mining techniques within astronomy, assisting astronomers in extracting meaningful insights from the wealth of available data.

In contrast, photometry shines when faint objects need to be observed. It can capture objects that may be ten times fainter than what can be measured with spectroscopy. For example, consider a faint galaxy so distant that its light has traveled billions of years to reach us. Photometric observations are invaluable to cosmologists as they help understand the early Universe.

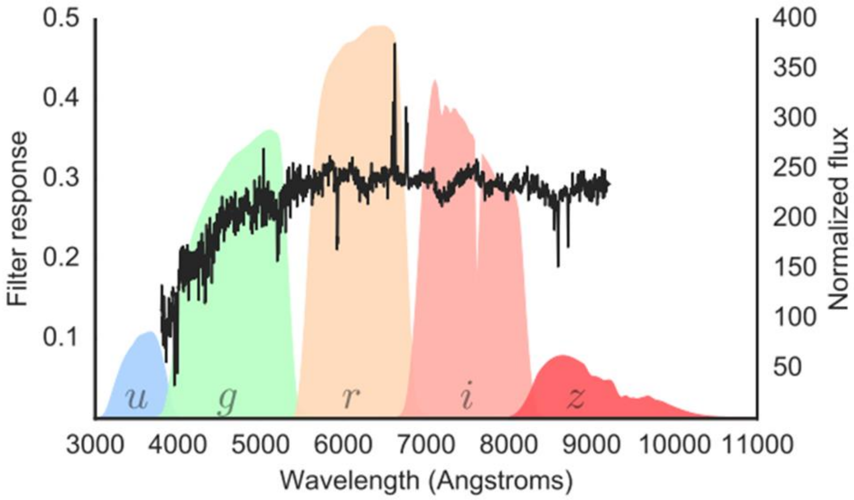

Figure 5: The spectrum of galaxy NGC 5750 is depicted by the black line, as observed by SDSS, in conjunction with the survey's five photometric broad-band filters: u, g, r, i, and z, covering a range from ultraviolet (u) to near-infrared (z). The galaxy's brightness is captured in an image for each of these bands. This information is sourced from [28].

2.4. DATA MINNING

With the rapid expansion in the volume of data generated by many sky surveys, data repositories have evolved from gigabytes to terabytes and even petabytes. The advent of astroinformatics is particularly timely, providing solutions to tackle the challenges and harness the opportunities stemming from the enormous scale, speed, and complexity of data produced by next-generation telescopes. This field leverages data mining tools to manage and analyze extensive astronomical datasets efficiently, optimizing data resources and creating innovative tools to address astronomical inquiries [5].

In the age of big data, data mining assumes a pivotal role, empowering researchers to effectively extract valuable information and knowledge from the extensive datasets housed in databases, data warehouses, and various information repositories. Depending on the patterns sought, data mining tasks encompass summarization, classification, regression, clustering, association, time-series analysis, and outlier/anomaly detection. Numerous reviews have delved into the application of data mining in astronomy, encompassing diverse techniques, such as using neural networks and outlier detection methods.

Data Mining Tasks	Applied Approaches	Applications in Astronomy
Classification	Artificial Neural Networks (ANN) Support Vector Machines (SVM) Learning Vector Quantization (LVQ) Decision Trees Random Forest K-Nearest Neighbors Naïve Bayesian Networks Radial Basis Function Network Gaussian Process Decision Table ADTree	Known knowns: – Spectral classification (stars, galaxies, quasars, supernovas) – Photometric classification (stars and galaxies, stars and quasars, supernovas) – Morphological classification of galaxies – Solar activity
Regression	Artificial Neural Networks (ANN) Support Vector Regression (SVR) Decision Trees Random Forest K-Nearest Neighbor Regression Kernel Regression Principal Component Regression (PCR) Gaussian Process Least Squared Regression Random Forest Partial Least Squares	Known unknowns: – Photometric redshifts (galaxies, quasars) – Stellar physical parameter measurement ([Fe/H], Teff, log*g*)
Clustering	Principal Component Analysis (PCA) DBScan K-Means OPTICS Cobweb Self Organizing Map (SOM) Expectation Maximization Hierarchical Clustering AutoClass Gaussian Mixture Modeling (GMM)	Unknown unknowns: – Classification – Special/rare object detection
Outlier Detection or Anomaly Detection	Principal Component Analysis (PCA) K-Means Expectation Maximization Hierarchical Clustering One-Class SVM	Unknown unknowns: – Special/rare object detection
Time-Series Analysis	Artificial Neural Networks (ANN) Support Vector Machines (SVM) Random Forest	Known unknowns: – Novel detection – Trend prediction

Table 2: Applied approaches and their applications for the principal data mining tasks in astronomy are detailed in the source [5].

In brief, **Table 2** provides an overview of the primary data mining tasks commonly utilized in astronomy. The growth in scale, depth, multi-wavelength capabilities, and time-domain surveys has resulted in a substantial expansion of astronomical data, which, in turn, has introduced challenges related to high dimensionality and algorithm efficiency. Feature selection and extraction become vital components of the data preparation process to improve the efficiency and effectiveness of data mining approaches. Feature selection is often preferred over feature extraction as it preserves the physical attributes of objects, enhancing the interpretability of the results.

Table 3 offers an overview of commonly employed feature selection and extraction methods, and additional insights into these techniques can be found in literature sources, including those referenced by Zheng and Zhang in 2008. Moreover, several books have been published explicitly focusing on data mining in astronomy, encompassing topics such as multivariate data analysis, practical statistics, machine learning, and data analysis of astronomical images. These valuable resources play a significant role in advancing data mining techniques within astronomy, assisting astronomers in extracting meaningful insights from the vast pool of available data.

Feature selection/extraction	Applied approaches	Applications in astronomy
Feature Selection	Best First Exhaustive Search Greedy Stepwise Random Search Rank Search Race Search Genetic Search Random Forest ReliefF Fisher Filtering Other wrapper methods	–Reducing dimension –Choose effective features
Feature Extraction	Principal Component Analysis (PCA) Independent Component Analysis (ICA) Linear discriminant analysis (LDA) Latent semantic index (LSI) Singular Value Decomposition (SVD) Multidimensional Scaling (MDS) Partial Least Squares (PLS) Locally Linear Embedding (LLE) ISOMAP Factor analysis Kernel LDA Kernel PCA Kernel Partial Least Squares (KPLS)	–Noise reduction/removal –Reducing dimension

Table 3: Feature selection/extraction methods. Taken from [5]

2.5. DATA MINING SOFTWARE AND TOOLS

Across diverse scientific domains, including astronomy, the demand for managing extensive and distributed datasets and conducting intricate knowledge discovery tasks is on the rise. Data mining experts have devised software and tools to confront these challenges, and these tools are increasingly finding applications in business, medicine, science, and engineering [5].

2.5.1. StatCodes:

A website offers links to valuable statistical codes for astronomy and related fields. This website is a resource to promote the application of statistics within the astronomy community and among researchers in related fields [5].

2.5.2. VOStat:

A statistical web service with a user-friendly interface implemented in the R language is available. This service is designed to conduct various statistical analyses, encompassing tasks such as plotting, data smoothing, spatial analysis, time series analysis, and more. Furthermore, it aims to encourage the adoption of R among astronomers and researchers in related fields, promoting the use of statistical tools for their work [5].

2.5.3. Weka:

An open-source data mining tool that implements machine learning algorithms for many tasks, including data preprocessing, classification, regression, clustering, association rules, and visualization, is available. AstroWeka comprises a set of extensions to the Weka data mining software specifically customized for astronomical data mining [5].

2.5.4. AstroML:

A Python module is accessible for machine learning and data mining, constructed using well-known libraries like Numpy, SciPy, and Scikit-learn. This module offers a library of statistical and machine-learning routines, open astronomical datasets, and illustrative examples for the analysis and visualization of astronomical data [5].

2.5.5. DAME (DAta Mining & Exploration):

A web-based, distributed data mining infrastructure is available, specializing in exploring extensive datasets through machine learning techniques. This infrastructure finds applications in astrophysics, such as evaluating photometric redshift and classifying transients [5].

2.5.6. Auton Lab:

There is a dedicated research group that focuses on statistical data mining. This group offers talks, tutorials, and software related to data mining and machine learning, serving as a valuable resource for researchers in this field [5].

2.5.7. Skytree Jump Start:

A machine learning platform is available to assist users in exploring the viability of applying machine learning to specific objectives. This platform is well-suited for handling large datasets and can be employed in astronomy for various tasks [5].

2.5.8 Photometric Redshift Tools:

Astronomers have devised a range of tools and methodologies for photometric redshift estimation, a regression task used to ascertain redshift values for galaxies and quasars. These tools include BPZ, Hyperz, ANNz, and ZEBRA [5].

Collaboration between astronomers, statisticians, computer scientists, and data mining experts is pivotal in advancing the development of data mining tools and techniques customized for astronomy. These collaborative efforts are instrumental in enabling the analysis of astronomical data and the discovery of valuable insights within this field.

3. MACHINE LEARNING APPROACHES TOWARDS ASTRONOMY

Integrating artificial intelligence (AI) techniques, notably machine learning (ML), in astronomy has gained significant prominence. AI empowers computers to emulate human intelligence, enhancing efficiency in performing various tasks. Researchers have identified several key areas where AI and ML are making noteworthy contributions to astronomy [29].

Below are some notable data mining tools and resources that find applications in astronomy:

ML Algorithms in Astronomy: Artificial intelligence (AI) and machine learning (ML) techniques have found valuable applications in classifying and analyzing the immense datasets produced by advanced telescopes. These applications extend to various domains, including planetary research, examining non-stellar constituents in the Milky Way, and investigating stellar clusters [29]. **Deep Space Exploration:** AI-based approaches are employed for deep space exploration and object detection, enhancing our ability to discover and study celestial objects [30]. **Generative Modeling:** Generative Adversarial Networks (GANs) are used to create new data instances that resemble observed training data. GANs consist of generators and discriminators and have applications in various domains, including astronomy [31].
Approximate Bayesian Computation (ABC): ABC is used to estimate the posterior distribution of model parameters without requiring the likelihood function, making it suitable for complex problems in astronomy [31]. **Quantum Machine Learning:** Quantum computing and qubits are leveraged to improve astronomy computational speed and data storage. This approach helps us understand the Milky Way's evolution [31].
Deep Boltzmann Machine: This model features numerous undirected connections in hidden layers and is used for handling unlabeled data in astronomy [32]. **Unsupervised ML Models:** Generative models and variational autoencoders handle data irregularities in quantum physics, biological physics, and electronic structure calculations [32]. **Applications in Astronomy:** ML methods find applications in star-galaxy separation, galaxy classification, supernova classification, exoplanet searches, habitability score calculations, galaxy identification, and more. They also play a crucial role in studying cosmic microwave background, the epoch of reionization, strong lensing curves, and pipeline optimization.

3.1. SUPERVISED TECHNIQUES

In astronomy, supervised techniques are indispensable tools that facilitate tasks such as classification and regression. They empower researchers to extract valuable insights from vast and intricate datasets. The application of machine learning (ML) to analyze astronomical data was pioneered by Ball et al. [33], who underscored the significance of data mining and ML algorithms for astronomers. Their pioneering work showcased the efficacy of these techniques in identifying active galactic nuclei (AGN), the central regions of galaxies. As the Large Synoptic Survey Telescope (LSST) commences, the demand for real-time and time-domain analyses intensifies, underscoring the need for advanced data mining approaches. Furthermore, these techniques are pivotal in effectively managing multiple observations [33].

In astronomy, the Support Vector Machine (SVM) algorithm [34] stands as a prominent and versatile tool. SVM serves a dual role, functioning as both a classifier and a regressor, demonstrating exceptional proficiency in addressing intricate classification challenges [35]. Its approach involves the construction of a hyperplane that maximizes the margin between the data points and the plane. By applying the kernel trick, SVM can adeptly transform non-linear input data into a linear relationship, rendering it well-suited for classifying astronomical data that is not linearly separable. Widely employed kernel tricks for SVM encompass Radial Basis Function (RBF), Gaussian, and polynomial kernels [36]. SVM's versatility extends to regression tasks, known as Support Vector Regression (SVR) [37]. In astronomy, researchers have harnessed SVM extensively for various applications, including classifying stars and identifying Active Galactic Nuclei (AGN) within galaxies.

The K-Nearest Neighbor (KNN) machine learning algorithm is a common choice in astronomy. KNN serves primarily as a classifier, utilizing distance measurements to derive information from input samples [38]. Its fundamental premise is that data points nearby are likely to belong to the same class, and the parameter "K" signifies the number of neighboring data points to be considered [38]. However, determining an optimal value for "K" can pose a challenge, and the algorithm may demand significant computational resources due to its proximity-based computations.

In stellar classification, Logistic Regression is pivotal in estimating probabilities for identifying star candidates [39]. Furthermore, ensemble methods have garnered attention for their ability to enhance prediction accuracy in astronomy. These methods encompass algorithms like Random Forests, Decision Trees, and ExtraTreeRegressor [40,41]. For example, Random Forest leverages multiple decision trees to deliver resilient outcomes when classifying astronomical data. Decision trees, which represent attributes in a tree-like structure, serve as versatile tools employed in astronomy classification and regression tasks.

Additionally, astronomers have ventured into pioneering ensemble methods, including the fusion of Random Forests with multilayer perceptron and Generative Adversarial Networks (GANs), yielding encouraging outcomes, even in intricate contexts like nanosatellite missions [42]. These supervised techniques persist as indispensable tools in astronomer's arsenal, empowering them to unveil the enigmatic facets of the universe concealed within extensive datasets.

3.2. UNSUPERVISED TECHNIQUES

Unsupervised techniques hold equal importance in astronomy, making substantial contributions to diverse data analysis domains. Methods like K-means clustering, Hierarchical clustering, Gaussian Mixture (GM), Agglomerative Hierarchical clustering, and Kernel Density Estimation (KDE) assume pivotal roles in unveiling inherent patterns within astronomical datasets.

The K-means clustering algorithm [43,44] is an iterative process that computes centroids and categorizes data points into clusters according to the shortest distance metric, typically using Euclidean or Manhattan distances. It continually reassigns data points based on newly computed centroids until the algorithm converges. This method finds extensive utility in tasks such as the identification of stellar spectra, x-ray spectra, and galaxy spectra.

Hierarchical clustering is a technique that builds a tree-like structure of clusters, and Agglomerative Hierarchical clustering takes a bottom-up approach by merging similar clusters. This method is especially valuable for clustering data points using similarity measures, like X-ray spectra and galaxy images. Hierarchical clustering has advantages over K-means and Gaussian Mixture (GM) methods in detecting clusters and proves robust against outliers [45].

Gaussian Mixture is a probabilistic approach that assumes data points follow Gaussian distributions with unknown parameters. It divides data points into distinct clusters based on these Gaussian distributions. Ward's research emphasized the effectiveness of hierarchical clustering compared to K-means and Gaussian Mixture (GM) methods, highlighting its robustness and sensitivity to outliers [45].

Kernel Density Estimation (KDE) is a non-parametric approach used to estimate probability density functions (PDFs) [46]. It is commonly applied in clustering high-dimensional astronomical datasets, although it can be sensitive to minor data fluctuations, potentially affecting its performance. Critical parameters in KDE include kernel width and dimension [46].

A self-organizing map (SOM), a type of Artificial Neural Network (ANN), employs an unsupervised approach to create a low-dimensional map for input samples. Unlike traditional ANNs using backpropagation, SOM uses competitive learning methods to represent data points [47, 48]. Researchers have effectively applied SOM to address various astronomy challenges, including galaxy morphology [49] and photometric redshift prediction [50].

Principal Component Analysis (PCA) is a vital tool for reducing the dimensionality of extensive datasets [51,52], and compressing data while preserving its essential characteristics. PCA helps identify physical parameters from spectra and capture multivariate correlations. It accomplishes dimension reduction by employing the largest eigenvectors representing the data's maximum variance. PCA is particularly advantageous when dealing with datasets containing numerous features. Furthermore, a distributed load balancing PCA (DLPCA) extension has been proposed [53] to reduce user transmission and download costs, further enhancing PCA's utility in astronomy data analysis. The steps for implementing PCA include centering input data, calculating the covariance matrix, identifying eigenvalues and eigenvectors, selecting relevant eigenvectors, and generating the final reduced-dimension data.

3.3. NEURAL NETWORK TECHNIQUES

Unlike conventional machine learning algorithms, the neural network (NN) approach draws inspiration from the human nervous system, where information flows through interconnected neurons. NN, a subset of machine learning, encompasses artificial neural networks (ANN) known for their predictive power, thanks to their layered structure and numerous adjustable parameters [54]. ANNs excel at comprehending intricate non-linear connections in vast datasets. They comprise hidden layers connected by weighted links and activation functions, including Sigmoid, Rectified Linear Unit (RELU), SoftMax, and Leaky Relu, chosen according to the specific task, whether it's classification or regression.

Training datasets undergo numerous iterations to minimize loss or error through backpropagation [55]. This process involves adjusting weights in each layer, starting from the output layer and moving back to the input layer. A cost or loss function quantifies the disparity between predicted and actual values. The application of neural networks in astronomy can be traced back to Angel et al. [56], who used them for tasks like star-galaxy separation and galaxy morphological classification [57]. In a different context, David et al. applied a genetic algorithm-based ANN for classifying preprocessed light curves [58]. Alejandro et al. employed a neural network in molecular astronomy to predict various molecule parameters from emission lines [59]. ANNs can also function as deep architecture models by increasing the number of hidden layers and neurons in each layer.

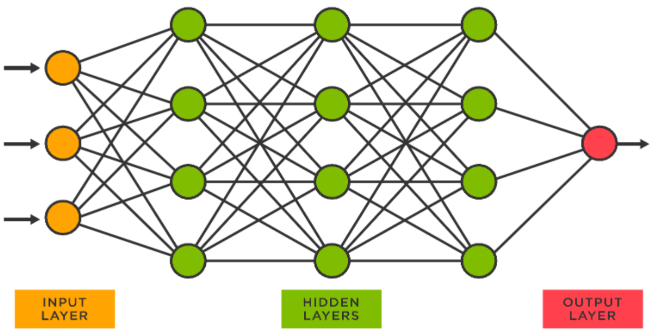

Figure 6: General architecture of ANN

3.4. DEEP LEARNING TECHNIQUES

Deep Learning (DL), a potent subset of neural networks, has contributed substantially to solving crucial astronomy problems, primarily due to its capacity for automatic feature extraction. Despite the existence of DL concepts since the 1980s, their extensive exploration was hampered by limitations in computational power and high-end systems. Nevertheless, the increased storage capacity and computational resources available today have transformed DL into an indispensable tool for modern data analysis, particularly in astronomy.

Deep Learning (DL) excels in various astronomy applications, as mentioned in the references provided. It has demonstrated its efficiency in tasks such as processing sequential and time series data using techniques like Long Short Term Memory (LSTM) and recurrent neural networks (RNN) for variable star light curves and real-time transient light curve anomaly detection [60][61]. Additionally, DL has been effectively applied in galaxy morphological classification [62], galaxy detection [63], and identification, showcasing its versatility in astronomy.

Deep Learning (DL) offers automatic feature selection, making it a powerful tool for high-dimensional data analysis. For example, Iten et al. utilized neural networks to extract physical parameters from experimental data [65], demonstrating the capabilities of DL in handling complex datasets. Additionally, Sedaghat et al. applied encoder-decoder-based deep convolutional neural networks to uncover hidden patterns within extensive astronomical datasets without requiring human intervention, highlighting the potential of DL in data analysis [66].

Incorporating graphics processing units (GPUs) has played a pivotal role in enhancing the performance of deep learning (DL). Convolutional Neural Networks (CNNs) have demonstrated remarkable capabilities in learning features from input images within astronomical image processing. Specifically, CNNs excel in extracting astrophysical parameters and delivering accurate outcomes, surpassing alternative methods [67]. Deep CNNs, characterized by increased convolutional and pooling layers, further enhance feature extraction and accuracy. This enhancement proves valuable in tasks such as parameter fitting for cometary dust and detecting image distortions induced by gravitational solid lensing [68].

Convolutional Neural Networks (CNNs) have found versatile applications in various astronomical domains, including weak gravitational lensing, where they are crucial in extracting essential information from high-dimensional datasets [69]. CNNs efficiently identify lensing parameters, detect gravitational solid lenses, and even autonomously eliminate lens-induced light distortions. Deep learning methods have also significantly contributed to transient detection, especially in searching for astrophysical and gamma-ray transients. Researchers have leveraged encoder-decoder models featuring CNNs for real-time transient detection and background subtraction, showcasing their effectiveness in these applications [70].

In conclusion, Deep Learning (DL) has firmly established itself as a cornerstone of modern astronomy. DL's potent automatic feature extraction capabilities have significantly advanced our comprehension of intricate astronomical phenomena.

4. CASE STUDIES

4.1 MEASUREMENT OF GALAXY MORPHOLOGIES [71]

Machine learning methods have proven their capability to establish relationships between input data, such as galaxy images, and desired outputs, like the physical properties of galaxies, through the analysis of input-output samples. These methods have already demonstrated success in various astrophysical applications. For instance, one notable application involves the measurement of galaxy morphologies, traditionally achieved through visual inspection, where galaxies are assigned classes based on their appearance. Recent advancements, accelerated by projects like Galaxy Zoo involving citizen scientists, have generated over 100 million galaxy classifications. These classifications have enabled astrophysicists to explore connections between the visual aspects (morphology) of galaxies and their internal and external characteristics, leading to numerous discoveries about the processes governing galaxy evolution. However, quantifying a galaxy's morphology concisely remains challenging, and automated machine-learning methods are highly desired. While some efforts have been made to replicate these classifications using machine learning, more sophisticated systems will be essential for handling the data products of next-generation telescopes.

Figure 7: An illustrative comparison of two morphology categories in astronomy. On the left, we observe the captivating spiral galaxy M101, while on the right, the elliptical galaxy NGC 1132 displays a different and distinctive morphology (Image credit: NASA).

Image analysis in astronomy streamlines automated classification [72] and fosters innovative strategies for investigating galaxy morphology. For instance, researchers have delved into utilizing the shape index [73] to predict a key indicator of galaxy evolution, the star formation rate. The shape index evaluates the local structure surrounding each pixel within an image, encompassing diverse structures from dark areas to valleys, saddle points, ridges, and, ultimately, bright regions. This index measures local morphology on a per-pixel basis, as depicted in Figure 8. The investigation unveiled that the shape index captures crucial galaxy information that traditional approaches might overlook. Integration of shape index features led to a 12% reduction in the rootmean-square error (RMSE) when forecasting the star formation rate. This methodology showcases how image analysis can provide fresh insights into galaxy characteristics beyond conventional methods.

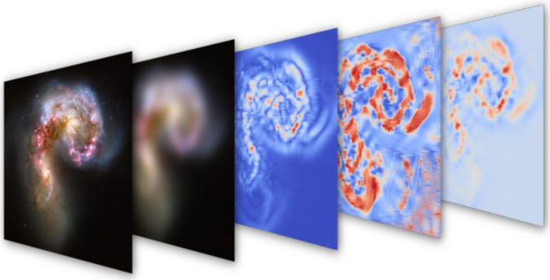

Figure 8: The image displays the Antennae galaxies as observed by the Hubble Space Telescope (credit: NASA)

4.2 CLASSIFICATION FOR ASTRONOMICAL SURVEYS WITH DEEP LEARNING (DETECTRON2) [74]

The pre-trained models examined in Detectron2's Model Zoo adhere to the Generalized RCNN meta-architecture, a versatile framework within the codebase. This architecture serves as a flexible and overarching structure that can accommodate a range of modifications as long as they support three essential components: (1) a backbone for per-image feature extraction, (2) region-proposal generation, and (3) per-region feature extraction and prediction. The schematic representation of this meta-architecture is depicted in Figure 8.

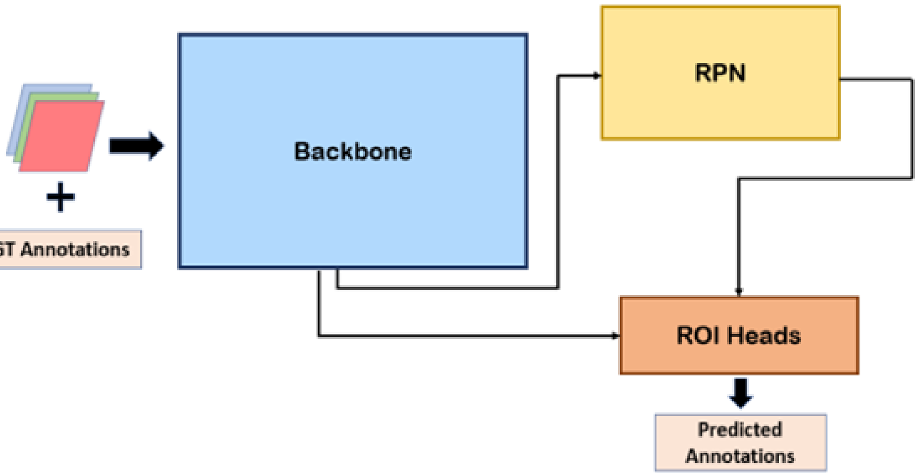

Figure 9: The meta-architecture of the Generalized Region-based Convolutional Neural Network (RCNN). In this depiction, a multi-channel image and provided ground truth object annotations are input to the backbone feature extractor. These extracted features are subsequently directed to the RPN (Region Proposal Network) and ROI (Region of Interest) heads, where they are utilized to predict object locations and annotations. The source of this image is credited to [74].

4.2 SUPERNOVA CLASSIFICATION [75]

A supernova is a powerful stellar explosion that briefly outshines an entire galaxy. These explosions can result from nuclear fusion in a degenerate star or the core collapse of a massive star, generating immense amounts of energy. Supernova shockwaves can lead to the formation of new stars and provide valuable distance indicators for astronomers. Supernovae are typically classified based on certain features in their spectral profiles. According to Rudolph Minkowski, there are two main classes: Type-I and Type-II. Type-I is further divided into Type-Ia, Type-Ib, and Type-Ic, while Type-II is sub-classified as Type IIP and Type IIn. Classifying supernovae can be challenging because they evolve, transitioning between types. Astronomers face even more significant challenges when spectral data is unavailable, relying solely on photometric measurements for classification.

Machine learning methods offer valuable assistance in real-time data analysis. These methods involve constructing models from input data and utilizing learning algorithms to extract knowledge. Machine learning approaches can be supervised, relying on a training set with known target properties, or unsupervised, requiring initial input data without known classes. They empower researchers to efficiently analyze and classify astronomical data, even when spectral information is limited or absent, facilitating their studies and discoveries.

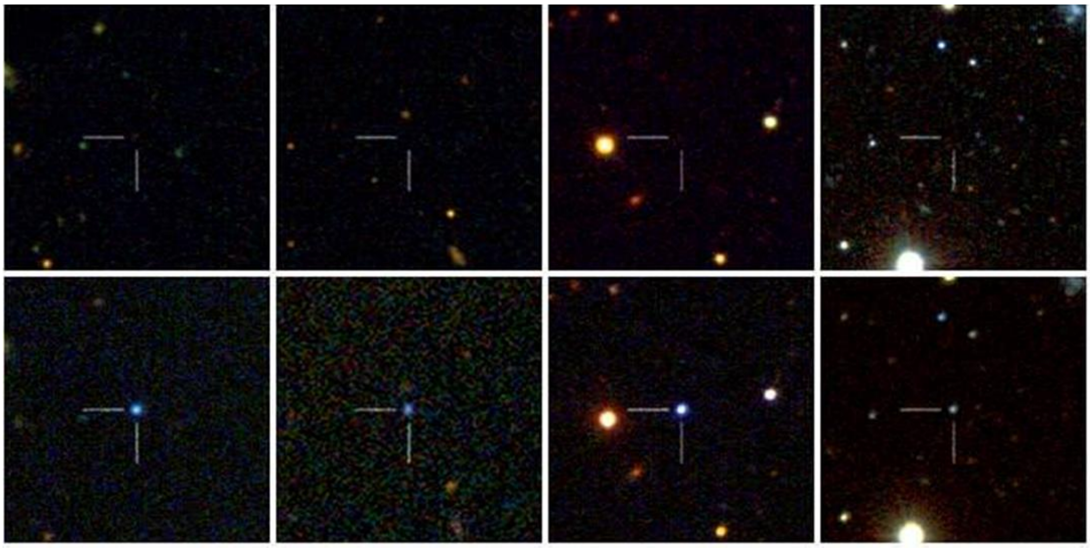

Figure 10: The four supernovae discovered by the Palomar Transient Factory. Left: before explosion. Right: after explosion. From top to bottom, the supernovae are PTF09atu, PTF09cnd, PTF09cwl, and PTF10cwr. [Credit: Caltech/Robert Quimby/Nature].

5. CHALLENGES OF BIG DATA ANALYTICS IN ASTRONOMY

Observational astronomy is undergoing a significant transformation due to the emergence of large digital sky surveys, which have become the primary source of astronomical data. These surveys currently encompass over 100 terabytes of data in major archives [11], and their volume is rapidly increasing. Today's typical sky survey involves approximately 10 terabytes of image data, around 1 billion detected sources, and about 100 measured attributes per source. These surveys cover the full range of wavelengths, from radio to X-ray, providing a panchromatic and less biased view of the universe.

Future advanced astronomical facilities are poised to produce unprecedented volumes of data, and integrating data from various surveys taken at different times and wavelengths presents both challenges and opportunities. Some of the most compelling scientific discoveries may arise from combining diverse datasets, such as studying multi-wavelength characteristics of celestial objects and analyzing time-series data from variable sources. The era of big data in astronomy encompasses every step of the data-to-knowledge process, from data generation and collection to transformation, storage, management, preprocessing, mining, visualization, understanding, evaluation, and explanation.

Various tools and technologies are needed to tackle the challenges posed by massive astronomical data [76]. Cloud storage and cloud computing offer promising solutions, but they are still in the early stages of development. Moving algorithms closer to data is essential to avoid data transformation, which requires significant internet bandwidth. Effective data management is critical, especially for well-characterized archival data and heterogeneous datasets. New database technologies are emerging, and careful consideration is required when choosing the appropriate database type.

The rise of big data has also necessitated the development of efficient data mining algorithms. Speed and efficiency play crucial roles in data exploration, and algorithms must be parallelized, distributed, or optimized for GPU-based or cluster-based processing. Collaboration among astronomers, statisticians, mathematicians, computer scientists, information scientists, and data scientists has become essential to extract meaningful insights from massive datasets. Interdisciplinary and multidisciplinary collaboration is the key to addressing complex challenges and making the most of big data in astronomy [77,78,79]. These studies [80-90] widely exploited different big frameworks for healthcare and transport applications. These studies implement big data analytics and cloud computing services for large applications.

One exemplary project, the Sloan Digital Sky Survey (SDSS), is a testament to the success of collaboration and data sharing in astronomy. Over its 15-year history, SDSS has produced a wealth of scientific results, including the discovery of distant quasars, brown dwarfs, gravitational lenses, Milky Way sub-structure, low surface brightness galaxies, asteroid families, hyper-velocity stars, and baryon acoustic oscillations. SDSS's collaborative approach involving astronomers, computer scientists, and physicists has been instrumental in its success.

Hence, the era of big data in astronomy presents exciting opportunities and significant challenges. Collaboration among diverse fields, support from governments and communities, and the training of the next generation of scientists are crucial for realizing the full potential of big data in advancing our understanding of the universe.

6. FUTURE WORK

In the context of the rapid advancement of image analysis and machine learning systems in astronomy, there are several promising avenues for future work and collaboration between astronomy and computer science:

6.1. Advanced Machine Learning Algorithms:

Develop and refine machine learning algorithms capable of handling even larger datasets with increased accuracy. Focus on creating models that can adapt and learn from the data they process, identifying new object classes and structures without human intervention.

6.2. Automated Object Classification:

Continue to improve automated object classification systems. Create models that classify known object types and discover new categories and sub-classes based on data patterns. This could involve unsupervised learning approaches.

6.3. Data Quality Assurance:

Develop machine-learning systems for data quality assurance. These systems can automatically flag and correct data anomalies, improving the reliability of astronomical databases and ensuring that machine-learning models receive high-quality input data.

6.4. Real-time Processing:

Enhance real-time data processing capabilities to keep up with the enormous data rates generated by instruments like LSST. Create algorithms that can process terabytes of data in near real-time, providing astronomers with timely insights.

6.5. Data Visualization and Interpretability:

Improve data visualization techniques that allow astronomers and data scientists to better understand the outputs of machine learning models. Develop tools for interpreting model decisions, particularly in cases where novel discoveries are made.

6.6. Ethical Considerations:

As machine learning systems play a more prominent role in data analysis, ensure that ethical considerations are integrated into research practices. Address issues related to bias, fairness, and data privacy in the context of astronomical data.

6.7. Scalable Computing Infrastructure:

Invest in a scalable computing infrastructure that can support the growing computational demands of astronomy. Cloud computing and distributed computing resources can be valuable in this regard.

7.CONCLUSION

In conclusion, the fusion of astronomy with the era of Big Data has propelled our understanding of the cosmos to unprecedented heights. This transformative journey has unleashed a torrent of data characterized by its sheer volume, diversity, velocity, and the demand for veracity. Astronomers have harnessed the power of data mining, machine learning, and deep learning techniques to sift through this vast celestial data trove, revealing hidden patterns, classifying celestial objects, and advancing our comprehension of the universe.

Significant data sources, such as the Sloan Digital Sky Survey, the VIMOS Public Extragalactic Redshift Survey, and the Two Micron All-Sky Survey, have provided invaluable insights into the cosmos. Combined with innovative data mining tools, these datasets have allowed astronomers to classify stars, galaxies, and other celestial entities with remarkable precision.

Supervised techniques like Support Vector Machines and K-nearest neighbors have excelled in classification tasks. At the same time, unsupervised methods such as K-means clustering and hierarchical clustering have unveiled hidden structures within astronomical data. Neural networks and deep learning have opened new frontiers, enabling automatic feature extraction and enhancing our understanding of complex astronomical phenomena.

However, challenges persist, ranging from data quality and scalability to ethical considerations and AI explainability. Astronomers and data scientists must collaborate to overcome these hurdles, ensuring the path to scientific discovery remains transparent, ethical, and accessible.

As astronomy continues to evolve in the era of Big Data, its contributions extend far beyond the boundaries of our planet. This exciting journey not only illuminates the cosmos but also illuminates the path to unlocking the universe's most profound mysteries. With ongoing advancements in technology, interdisciplinary collaboration, and a commitment to data-driven discovery, the future of astronomy promises to be as boundless as the universe itself.

Funding: This research received no external funding.

Conflict of interest: The authors declare no conflicts of interest.

REFERENCES

[1] C. Yang, Q. Huang, Z. Li, K. Liu, and F. Hu, "Big Data and cloud computing: innovation opportunities and challenges," International Journal of Digital Earth, vol. 10, no. 1, pp. 13-53, 2017. [Online]. Available: https://doi.org/10.1080/17538947.2016.1239771.

[2] E. D. Feigelson and G. J. Babu, "Big data in astronomy," Significance, vol. 9, no. 4, pp. 22-25, 2012. [Online]. Available: https://doi.org/10.1111/j.1740-9713.2012.00587.x.

[3] A. M. Mickaelian, "Multiwavelength Search and Studies of Active Galaxies and Quasars," Communications of BAO, vol. 1, 2017, pp. 15-38.

[4] A. Szalay and J. Gray, "The World-Wide Telescope," Science (New York, N.Y.), vol. 293, pp. 2037-2040, 2001. doi: 10.1126/science.293.5537.2037.

[5] Y. Zhang and Y. Zhao, "Astronomy in the big data era," Data Science Journal, vol. 14, 2015. [Online]. Available: https://doi.org/10.5334/dsj-2015-011.

[6] "Top 10 Big Data Challenges – A Serious Look at 10 Big Data V's," [Online]. Available: https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/.

[7] R. C. Hartman, D. L. Bertsch, S. D. Bloom, et al., "CGRO," ApJS, vol. 123, pp. 79, 1999.

[8] F. Acero, M. Ackermann, M. Ajello, et al., "Acero F., Ackermann M., Ajello M., et al 2015, ApJS, 218, 23," ApJS, vol. 218, p. 23, 2015.

[9] R. Ahumada, C. Allende Prieto, A. Almeida, et al., "Ahumada R., Allende Prieto C., Almeida A., et al 2020, ApJS, 249, 3," ApJS, vol. 249, p. 3, 2020.

[10] R. M. Cutri, M. F. Skrutskie, S. Van Dyk, et al., "University of Massachusetts and Infrared Processing and Analysis Center (IPAC/California Institute of Technology), 2003."

[11] A. M. Mickaelian, "Big Data in Astronomy: Surveys, Catalogs, Databases and Archives," Communications of BAO, vol. 67, no. 2, pp. 159-180, 2020.

[12] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," J. King Saud Univ.-Comput Inf. Sci., vol. 30, no. 4, pp. 431-448, 2018.

[13] B. Furht and F. Villanustre, "Introduction to big data," in Big Data Technologies and Applications, pp. 3-11, Springer, 2016.

[14] S. Alam et al., "The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III," Astrophys. J. Suppl. Ser., vol. 219, p. 12, 2015.

[15] "VIPERS: The VIMOS Public Extragalactic Redshift Survey," http://vipers.inaf.it, 2020.

[16] "The Two Micron All Sky Survey at IPAC," https://old.ipac.caltech.edu/2mass/, Accessed: June 2020.

[17] C. Conselice 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., vol. 381, no. 3, pp. 962-986, 2007.

[18] "The Large Synoptic Survey Telescope," https://www.lsst.org/, Accessed: 2020.

[19] "SKA in India: Science with Big Data," https://asi2020.astron-soc.in/workshops/workshop3/, Accessed: July 2020.

[20] "LIGO Laser Interferometer Gravitational-Wave Observatory," https://www.ligo.caltech.edu/, Accessed: June 2020.

[21] O. L. Fevre 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.

[22] A. Lawrence et al., "The UKIRT Infrared Deep Sky Survey (UKIDSS)," Mon. Not. R. Astron. Soc., vol. 379, pp. 1599-1617, 2007.

[23] M. Povic et al., "The ALHAMBRA Survey: Reliable Morphological Catalogue of 22 051 Early-and Late-Type Galaxies," Mon. Not. R. Astron. Soc., vol. 435, no. 4, pp. 3444-3461, 2013.

[24] S. Djorgovski et al., "The Palomar Digital Sky Survey (DPSS)," arXiv:astro-ph/9809187, 1998.

[25] C. J. Lintott et al., "Galaxy Zoo: Morphologies Derived from Visual Inspection of Galaxies from the Sloan Digital Sky Survey," Mon. Not. R. Astron. Soc., vol. 389, no. 3, pp. 1179-1189, 2008.

[26] R. E. Cutri et al., "Vizier Online Data Catalog: AllWISE Data Release (Cutri+ 2013)," VizieR Online Data Catalog, vol. 328, 2021.

[27] J. T. de Jong et al., "The Kilo-Degree Survey," Exp. Astron., vol. 35, no. 1-2, pp. 25-44, 2013.

[28] J. Kremer, K. Stensbo-Smidt, F. Gieseke, K. Pedersen, and C. Igel, "Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy," IEEE Intelligent Systems, vol. 32, no. 2, pp. 16-22, 2017. DOI: 10.1109/MIS.2017.40.

[29] C.J. Fluke and C. Jacobs, "Surveying the Reach and Maturity of Machine Learning and Artificial Intelligence in Astronomy," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 2, pp. 1349, 2020.

[30] J. Bird, L. Petzold, P. Lubin, and J. Deacon, "Advances in Deep Space Exploration via Simulators & Deep Learning," New Astron., vol. 84, p. 101517, 2021.C. Donalek et al., "Immersive and collaborative data visualization using virtual reality platforms," 2014 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 2014, pp. 609-614, doi: 10.1109/BigData.2014.7004282.

[31] G. Carleo et al., "Machine Learning and the Physical Sciences," Rev. Modern Phys., vol. 91, no. 4, pp. 045002, 2019.L.: Machine learning and the physical sciences. Rev. Modern Phys. 91(4), 045002 (2019)

[32] T. Navamani, "Efficient Deep Learning Approaches for Health Informatics," in Deep Learning and Parallel Computing Environment for Bioengineering Systems, pp. 123-137, Elsevier, 2019.

[33] N.M. Ball and R.J. Brunner, "Data Mining and Machine Learning in Astronomy," Int. J. Modern Phys. D, vol. 19, no. 07, pp. 1049-1106, 2010. doi: 10.1142/s0218271810017160.

[34] C. Cortes and V. Vapnik, "Support-Vector Networks," Mach. Learn., vol. 20, no. 3, pp. 273-297, 1995.

[35] S. Abe, "Support Vector Machines for Pattern Classification," vol. 2, Springer, New York, 2005.

[36] B. Scholkopf, A.J. Smola, F. Bach, et al., "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," MIT Press, Cambridge, 2002.

[37] I. Steinwart and A. Christmann, "Support Vector Machines," Springer, New York, 2008.

[38] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21-27, 1967.

[39] L. Beitia-Antero, J. Ya´nez, and A.I.G. de Castro, "On the Use of Logistic Regression for Stellar Classification," Exp. Astron., vol. 45, no. 3, pp. 379-395, 2018.

[40] S. Carliles, T. Budavari, S. Heinis, C. Priebe, and A.S. Szalay, "Random Forests for Photometric Redshifts," Astrophys. J., vol. 712, no. 1, p. 511, 2010.

[41] D. Baron and D. Poznanski, "The Weirdest SDSS Galaxies: Results from an Outlier Detection Algorithm," Mon. Not. R. Astron. Soc., vol. 465, no. 4, pp. 4530-4555, 2017.

[42] H. Cao, D. Bastieri, R. Rando, G. Urso, G. Luo, and A. Paccagnella, "Machine Learning on Compton Event Identification for a Nano-Satellite Mission," Exp. Astron., vol. 47, no. 1, pp. 129-144, 2019.

[43] H. Cao, D. Bastieri, R. Rando, G. Urso, G. Luo, and A. Paccagnella, "Machine Learning on Compton Event Identification for a Nano-Satellite Mission," Exp. Astron., vol. 47, no. 1, pp. 129-144, 2019.

[44] H. Steinhaus, "Sur la Division des Corps Materiels en Parties," Bull. Acad. Polon. Sci., C1. III Vol IV, pp. 801-804, 1956.

[45] J.H. Ward Jr., "Hierarchical Grouping to Optimize an Objective Function," J. Am. Stat. Assoc., vol. 58, no. 301, pp. 236-244, 1963.

[46] C. Taylor, "Classification and Kernel Density Estimation," Vistas Astron., vol. 41, no. 3, pp. 411-417, 1997.

[47] T. Kohonen, "Self-Organized Formation of Topologically Correct Feature Maps," Biol. Cybern., vol. 43, no. 1, pp. 59-69, 1982

[48] T. Kohonen, "An Overview of SOM Literature," in Self-Organizing Maps, pp. 347-371, Springer, 2001.

[49] T.J. Galvin, M. Huynh, R.P. Norris, X.R. Wang, E. Hopkins, O. Wong, S. Shabala, L. Rudnick, M.J. Alger, K.L. Polsterer, "Radio Galaxy Zoo: Knowledge Transfer Using Rotationally Invariant Self-Organizing Maps," Publ. Astron. Soc. Pac., vol. 131, no. 1004, p. 108009, 2019.

[50] D. Wilson, H. Nayyeri, A. Cooray, B. Haußler, "Photometric Redshift Estimation with Galaxy Morphology Using Self-Organizing Maps," Astrophys. J., vol. 888, no. 2, p. 83, 2020.

[51] T.A. Boroson and R.F. Green, "The Emission-Line Properties of Low-Redshift Quasi-Stellar Objects," Astrophys. J. Suppl. Ser., vol. 80, pp. 109-135, 1992.

[52] S. Djorgovski, "The Fundamental Plane Correlations for Globular Clusters," Astrophys. J., vol. 438, pp. 29-32, 1995.

[53] A. Govada and S.K. Sahay, "A Communication Efficient and Scalable Distributed Data Mining for the Astronomical Data," Astron. Comput., vol. 16, pp. 166-173, 2016.

[54] A.A. Collister and O. Lahav, "ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks," Publ. Astron Soc Pac., vol. 116, no. 818, p. 345, 2004.

[55] I. Sadeh, F.B. Abdalla, and O. Lahav, "ANNz2: Photometric Redshift and Probability Distribution Function Estimation Using Machine Learning," Publ. Astron Soc Pac., vol. 128, no. 968, p. 104502, 2016.

[56] J.R.P. Angel, P. Wizinowich, M. Lloyd-Hart, and D. Sandler, "Adaptive Optics for Array Telescopes Using Neural-Network Techniques," Nature, vol. 348, no. 6298, pp. 221-224, 1990.

[57] D. Bazell and Y. Peng, "A Comparison of Neural Network Algorithms and Preprocessing Methods for Star-Galaxy Discrimination," Astrophys. J. Suppl. Ser., vol. 116, no. 1, p. 47, 1998.

[58] D. Andresiˇ c, P. ˇ Saloun, and B. Pe ˇ cˇ´ıkova, "Large Astronomical Time Series Pre-processing for Classification Using Artificial Neural Networks," in Intelligent Astrophysics, pp. 265-293, Springer, 2021.

[59] A. Barrientos, J. Holdship, M. Solar, S. Mart´ın, V.M. Rivilla, S. Viti, J. Mangum, N. Harada, K. Sakamoto, S. Muller, et al., "Towards the Prediction of Molecular Parameters from Astronomical Emission Lines Using Neural Networks," Exp. Astron., pp. 1-26, 2021.

[60] T.A. Hinners, K. Tat, and R. Thorp, "Machine Learning Techniques for Stellar Light Curve Classification," Astron. J., vol. 156, no. 1, p. 7, 2018.

[61] D. Muthukrishna, M. Lochner, and S. Webb, "Real-Time Detection of Anomalies in Large-Scale Transient Surveys," 2019.

[62] P. Barchi, R. de Carvalho, R. Rosa, R. Sautter, M. Soares-Santos, B. Marques, E. Clua, T. Gonc¸alves, C. de Sa-Freitas, T. Moura, "Machine and Deep Learning Applied to Galaxy Morphology - A Comparative Study," Astron. Comput., vol. 30, p. 100334, 2020.

[63] R.E. Gonzalez, R.P. Munoz, C.A. Hern ´ andez, "Galaxy Detection and Identification Using Deep Learning and Data Augmentation," Astron. Comput., vol. 25, pp. 103-109, 2018.

[64] B. Hoyle, M.M. Rau, C. Bonnett, S. Seitz, J. Weller, "Data Augmentation for Machine Learning Redshifts Applied to Sloan Digital Sky Survey Galaxies," Mon. Not. R. Astron. Soc., vol. 450, no. 1, pp. 305-316, 2015.

[65] R. Iten, T. Metger, H. Wilming, L. Del Rio, R. Renner, "Discovering Physical Concepts with Neural Networks," Phys. Rev. Lett., vol. 124, no. 1, p. 010508, 2020.

[66] N. Sedaghat, M. Romaniello, J.E. Carrick, F.-X. Pineau, "Machines Learn to Infer Stellar Parameters Just by Looking at a Large Number of Spectra," Mon. Not. R. Astron. Soc., vol. 501, no. 4, pp. 6026-6041, 2021.

[67] Y.-H. Mu, B. Qiu, J.-N. Zhang, J.-C. Ma, X.-D. Fan, "Photometric Redshift Estimation of Galaxies with Convolutional Neural Network," Res. Astron. Astrophys., vol. 20, no. 6, p. 089, 2020.

[68] Y.D. Hezaveh, L.P. Levasseur, P.J. Marshall, "Fast Automated Analysis of Strong Gravitational Lenses with Convolutional Neural Networks," Nature, vol. 548, no. 7669, pp. 555-557, 2017.

[69] D. Ribli, B.A. Pataki, J.M. Zorrilla Matilla, D. Hsu, Z. Haiman, I. Csabai, "Weak Lensing Cosmology with Convolutional Neural Networks on Noisy Data," Mon. Not. R. Astron. Soc., vol. 490, no. 2, pp. 1843-1860, 2019.

[70] N. Sedaghat, A. Mahabal, "Effective Image Differencing with Convolutional Neural Networks for Real-time Transient Hunting," Mon. Not. R. Astron. Soc., vol. 476, no. 4, pp. 5365-5376, 2018.

[71] C.J. Lintott et al., "Galaxy Zoo: Morphologies Derived from Visual Inspection of Galaxies from the Sloan Digital Sky Survey," MNRAS, vol. 389, pp. 1179-1189, 2008.

[72] K.L. Polsterer, F. Gieseke, C. Igel, "Automatic Galaxy Classification via Machine Learning Techniques: Parallelized Rotation/Flipping Invariant Kohonen Maps (PINK)," in Astronomical Data Analysis Software and Systems XXIV (ADASS XXIV), A.R. Taylor and E. Rosolowsky, Eds., vol. 495, Sep. 2015, pp. 81. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2015ASPC..495...81P.

[73] K.S. Pedersen, K. Stensbo-Smidt, A. Zirm, and C. Igel, "Shape Index Descriptors Applied to Texture-Based Galaxy Analysis," University of Copenhagen, Denmark

[74] G. Merz, Y. Liu, C. J. Burke, P. D. Aleo, X. Liu, M. Carrasco Kind, V. Kindratenko, and Y. Liu, "Detection, instance segmentation, and classification for astronomical surveys with deep learning (deepdisc): detectron2 implementation and demonstration with Hyper Suprime-Cam data," Monthly Notices of the Royal Astronomical Society, vol. 526, no. 1, pp. 1122-1137, 2023. doi: 10.1093/mnras/stad2785.

[75] M. M. Kasliwal, S. B. Cenko, S. R. Kulkarni, E. O. Ofek, R. Quimby, and A. Rau, "Discovery of a New Photometric Sub-class of Faint and Fast Classical Novae," The Astrophysical Journal, vol. 735, no. 2, p. 94, Jun. 2011. DOI: 10.1088/0004-637x/735/2/94.

[76] A. A. Ţole, "Big Data Challenges," Database Systems Journal, vol. 4, pp. 31-40, 2013.

[77] C. Yang, Q. Huang, Z. Li, K. Liu, and F. Hu, "Big Data and cloud computing: innovation opportunities and challenges," International Journal of Digital Earth, vol. 10, no. 1, pp. 13-53, 2017.

[78] K. Wadhwani, "Big Data Challenges and Solution."

[79] M. A. Garrett, "Big Data analytics and cognitive computing – future opportunities for astronomical research," in IOP Conference Series: Materials Science and Engineering, vol. 67, no. 1, 2014.

[80] Mohammed, M. A., Lakhan, A., Abdulkareem, K. H., Zebari, D. A., Nedoma, J., Martinek, R., ... & Garcia-Zapirain, B. (2023). Energy-efficient distributed federated learning offloading and scheduling healthcare system in blockchain based networks. Internet of Things, 22, 100815.

[81] Lakhan, A., Mohammed, M. A., Nedoma, J., Martinek, R., Tiwari, P., Vidyarthi, A., ... & Wang, W. (2022). Federated-learning based privacy preservation and fraud-enabled blockchain IoMT system for healthcare. IEEE journal of biomedical and health informatics, 27(2), 664-672.

[82] Lakhan, A., Lateef, A. A. A., Abd Ghani, M. K., Abdulkareem, K. H., Mohammed, M. A., Nedoma, J., ... & Garcia-Zapirain, B. (2023). Secure-fault-tolerant efficient industrial internet of healthcare things framework based on digital twin federated fog-cloud networks. Journal of King Saud University-Computer and Information Sciences, 35(9), 101747.

[83] Lakhan, A., Thinnukool, O., Groenli, T. M., & Khuwuthyakorn, P. (2023). RBEF: Ransomware Efficient Public Blockchain Framework for Digital Healthcare Application. Sensors, 23(11), 5256.

[84] Lakhan, A., Mohammed, M. A., Abdulkareem, K. H., khanapi Abd Ghani, M., Marhoon, H. A., Nedoma, J., ... & Garcia-Zapirain, B. (2023). Secure blockchain assisted Internet of Medical Things architecture for data fusion enabled cancer workflow. Internet of Things, 24, 100928.

[85] Jat, A. S., Grønli, T. M., & Lakhan, A. R. (2023). Towards Next-Generation Healthcare: Architectural Insights into an AI-Driven, Smartwatch-Compatible mHealth Application.

[86] Mohammed, M. A., Lakhan, A., Abdulkareem, K. H., & Garcia-Zapirain, B. (2023). Federated auto-encoder and XGBoost schemes for multi-omics cancer detection in distributed fog computing paradigm. Chemometrics and Intelligent Laboratory Systems, 241, 104932.

[87] Lakhan, A., Mohammed, M. A., Abdulkareem, K. H., Hamouda, H., & Alyahya, S. (2023). Autism Spectrum Disorder detection framework for children based on federated learning integrated CNN-LSTM. Computers in Biology and Medicine, 166, 107539.

[88] Mohammed, M. A., Lakhan, A., Abdulkareem, K. H., & Garcia-Zapirain, B. (2023). A hybrid cancer prediction based on multi-omics data and reinforcement learning state action reward state action (SARSA). Computers in Biology and Medicine, 154, 106617.

[89] Mohammed, M. A., Lakhan, A., Abdulkareem, K. H., Abd Ghani, M. K., Marhoon, H. A., Nedoma, J., & Martinek, R. (2023). Multi-objectives reinforcement federated learning blockchain enabled Internet of things and Fog-Cloud infrastructure for transport data. Heliyon, 9(11).

[90] Lakhan, A., Mohammed, M. A., Abdulkareem, K. H., Jaber, M. M., Kadry, S., Nedoma, J., & Martinek, R. (2023). Fuzzy Decision Based Energy-Evolutionary System for Sustainable Transport in Ubiquitous Fog Network. Human-centric Computing Inform. Sci., 13, 34.