A literature review of supervised and unsupervised machine learning approaches for classification of extended extragalactic objects in astronomical survey data

D. Sergio

Department of Computer Science, Eastern Washington University, Cheney, WA

Abstract

This review summarizes approaches to classification of extended extragalactic astronomical objects using unsupervised and supervised machine learning models, including Random Forest (RF), Probabilistic Random Forest (PRF), Linear Discriminant Analysis (LDA), Equivariant Convolutional Neural Networks (CNN), and deep learning architectures including Mask R-CNN, object detection and segmentation methods, and deblending techniques. Extended extragalactic object classification, e.g. galaxy morphology, present a unique set of challenges due to the inherently complex feature space, noisy data and labels, observational biases, high dimensionality, and model architecture that was not originally designed for such tasks. This review presents approaches to feature engineering, ground-truth labeling systems, dimensionality reduction, data augmentation and model architecture design, highlighting key findings and areas for future research in the fields of astronomical object classification and computer vision more generally.

Keywords: machine learning, supervised learning, unsupervised learning, image classification, equivariant neural networks, random forest, clustering, dimensionality reduction, deep learning

1 Introduction

1.1 What are extended extragalactic objects?

Astronomical objects include a wide variety of entities, ranging from point-like sources such as stars to extended extragalactic structures like galaxies, quasars, blazars, radio galaxies, seyfert galaxies, and other unknown types [26].

Extragalactic spatially extended objects types can be further subdivided into classes of different subtypes of objects. The morphology of a galaxy is the nature of its structure and shape (e.g. spiral, elliptical) [26].

1.1.1 A historical perspective on extended extragalactic object classification

The classification of extragalactic objects in the modern era arguably started with Edwin Hubble's discovery, using Cepheid Variable periodicity as a distance measurement, that the object known as M31 was in fact a separate galaxy, not a nebula [1]. Hubble's morphological classification scheme, introduced in 1926, is based on human visual inspection of galaxies, dividing them into three main categories: elliptical, spiral, and irregular [26]. This scheme laid the groundwork for future studies of galaxy morphology and evolution.

Today, large-scale sky surveys, such as the Sloan Digital Sky Survey (SDSS), produce more data than can be manually classified [9].

However, human classification is still done today. Ferreira et al. uses volunteers for manual human visual classification of 3956 galaxies using a manual decision tree [14]. Human classification of galaxies at the scale of massive astronomical survey datasets, such as the SDSS, have included approaches such as crowdsourcing of manual human visual inspection (e.g. Galaxy Zoo [21] and Galaxy Zoo 2 [38]). This review focuses on machine learning approaches to classification, which in some cases uses human-classified data as training labels [11][13].

1.2 What is machine learning?

Machine learning (ML) is described by Acquaviva as the process of teaching a machine to make informed, data-driven decisions [2]. Supervised learning is the set of algorithms that learn to associate input feature data ground-truth (known) class labels [20]. Unsupervised learning does not require labels, and instead uses only the features, which can make it unbiased, if the labels carry any bias [34].

1.3 Complementary reviews

Complementary reviews include Fotopoulou [15], Baron [4], Zhang [40], Fraix-Burnet [16], and Smith [31].

2 The nature of astronomical data

2.1 Feature engineering

Raw astronomical photometric data, in the case of SDSS, as described by Stoughton et al, is in the form of pixel counts at some position (x, y) measured with a charge-coupled device (CCD) camera in 5 optical wavelength bands u, g, r, i, z [33]. Other wavelength bands include w1, w2, w3, w4 for infrared photometry measured by the Wide-Field Infrared Survey Explorer (WISE) [39]. Logan et al. uses photometric colors (i.e. differences in magnitudes between bands), ranked by their importance, as determined by a Random Forest classifier [23].

High-dimensional features are subject to the curse of dimensionality, i.e. they are sparse, prone to overfitting, and employing clustering on high-dimensional data can be infeasible [15]. For this reason, dimensionality reduction techniques are used to reduce the number of features while preserving the information. Dimensionality reduction is discussed in detail in section 2.5.

2.1.1 Parametric features

Parametric features are raw pixel counts modeled by parametric intensity profiles (not to be confused with ML models), such as the Sérsic profile. Graham et al. describes the Sérsic profile as a function of distance r from its center:

$$I(r) = I_e \exp\left\{-b_n\left[\left(\frac{r}{r_e}\right)^{1/n} - 1\right]\right\}, \qquad (1)$$

where $I(r)$ is the intensity at radius r, I_e is the intensity at the effective radius r_e, enclosing half of the total light, n is the Sérsic index, and b_n is a constant that depends on n [17]. Photometric flux, and associated magnitude, for extended objects, is then derived by fitting the profile model to the observed pixel intensity and summing over the pixels. The de Vaucouleurs profile is a special case of the Sérsic profile with $n = 4$, which attempts to fit elliptical galaxies [12]. Other fit models, such as the exponential disc model is used to describe spiral galaxy discs[33].

The point-spread function (PSF) fit uses optics to profile raw pixels originating from a point-source, such as a star [29]. The PSF along with the de Vaucouleurs profile provides photometric flux and magnitude for point-like and extended objects. Clarke et al. combines the two fits to engineer a $\texttt{resolved}_r$ feature, defined as the difference between the extended fit magnitude and the PSF fit magnitude in the r band [7].

There have been attempts at using multi-component models as well. For example, Peng et. al describes the GALFIT algorithm, a combination of model components, such as bulge and disk [27].

2.1.2 Non-parametric and morphological features

Non-parametric features, on the other hand, do not try to fit pixel intensity to a specific functional model. Instead, they measure statistical properties of light distribution.

One such system defines the non-parametric features Concentration (C), Asymmetry (A), Smoothness (S), Gini coefficient (G), and $M20$ (CASGM) [24], which are computed directly from the pixel intensity values without assuming an underlying model, and quantify the concentration of light is towards the center of the object, symmetry, clumpiness of the light distribution, the inequality in the light distribution, and the second-order moment of the brightest 20% of the light, respectively [24].

Ferrari et al. extends the CASGM system by introducing Entropy (H) and spirality (σ). Entropy measures the randomness in the light distribution, while spirality quantifies non-radial patterns [13]. These features are used this feature space with a Linear Discriminant Analysis (LDA) classifier [3].

2.1.3 Pixel features

Raw pixels can also be used as features, e.g. in a convolutional neural network (CNN). Deileman et al. uses raw pixel data from SDSS as features [10].

2.2 Ground-truth and labeled data

Ground-truth refers to the objective truth about the data, which can be used as a label for training and evaluating ML models. Ground-truth labels are typically derived from detailed analysis, such as spectroscopic fitting, but can also come from crowdsourcing human vision, expert consensus, or model fitting. We will review common approaches to ground-truth labeling in astronomical datasets.

How we assign ground-truth labels to raw data in the form of electromagnetic energy measurements (counts, voltages, fluxes, spectra, polarization, time) originating from astronomical objects which are often billions of light years away, is a non-trivial task [32].

Ground-truth labeling approaches vary based on the type of raw data and the scientific goals of the study. The Galaxy Zoo project [35] is a well-known example of using consensus expert labeling to classify galaxy morphologies from imaging data. This approach can be subjective and prone to human error and bias.

2.2.1 Noisy data

Both features and labels in an astronomical dataset will usually have noise in the form of environmental effects, instrumental effects, and observational biases. Particularly, physical measurements typically have an associated uncertainty value. This is not always factored into the labels or the models for many studies [7]. The performance of the model depends on the signal-to-noise ratio of the data, the uncertainty of the labels, and the observational biases present in the dataset [4]. Reis et al. proposed a Probabilistic Random Forest algorithm that can incorporate measurement uncertainties in both features and labels to improve classification performance [28]. We will examine this model in more detail in 4.3. Song et al. performs an in-depth analysis of the problem of noisy labels in supervised learning [32].

2.2.2 Spectroscopic fitting

More objective methods of labeling a dataset with ground-truth classes include 1D spectroscopy, considered the gold standard for classifying point sources, which can provide definitive classifications based on emission and absorption lines (e.g. hydrogen Balmer lines), continuum shapes (e.g., blackbody radiation for stars, sum of blackbodies for galaxies), redshift measurements (difference between known spectral absorption lines and observed wavelengths), and other spectral features [5]. Clarke et al. used spectroscopically labeling from the Baryon Oscillation Spectroscopic Survey (BOSS), described in detail by Smee et al. [30] to label SDSS objects to learn photometric features of objects without spectra [7].

2.2.3 Spatially resolved spectroscopy

Integral Field Unit (IFU) spectroscopy is the collection of spatially resolved spectra in the form of data cubes (x, y, λ) that enables the identification of different spatial regions within extended objects. For galaxies, this includes star-forming regions, active galactic nuclei (AGN), and velocity gradient maps, which reveal axes and therefore orientation, as described by Weijmans et al. in the SDSS Mapping Nearby Galaxies at APO (MaNGA) Survey. Elliptical galaxies show little rotation and high velocity dispersion, while edge-on disks show velocity gradients along the major axis. Mergers show complex velocity fields [36].

2.3 Rotation

Rotations of extended astronomical objects should not change their predicted morphological classification. Dieleman et al. (2015) notes that while CNNs are translationally invariant (i.e. the classification does not change when the input feature (image) is translated), it is not the case that CNNs are rotationally invariant. To address this, Dieleman et al. augments data by computing rotated and flipped versions of all input images [10]. Model architecture that addresses rotation is discussed in section 4.5 of this paper.

2.4 Orientation viewpoint

Orientation viewpoint, or pose (an extended object's major and minor axes relative to the observer), is an observational bias, and a consequence of Earth's position relative to the object [34]. For example, a spiral galaxy viewed face-on will show its spiral arms clearly, while the same galaxy viewed edge-on will appear as a thin line with a bulge in the center.

Tohill et al. uses auto-encoders for feature extraction to be used in an unsupervised hierarchical clustering (HC) model, and notes that orientation viewpoint is encoded in the latent space [34]. To address this, Tohill et al. rotates all images to a common orientation before training the auto-encoder [34].

While true orientation cannot be determined from most raw pixel samples, 1D spectra, or from photometric multiband fluxes, it can be determined from IFU spectroscopic velocity maps, and in some cases from high-resolution imaging [34].

Liu et al. proposes a general pose estimator trained on data of various poses [22]. It is not known if this method has been used with astronomical data using velocity maps acquired from IFU spectroscopy.

2.5 Dimensionality reduction

Because features may be highly correlated or redundant, there is a need to reduce the number of features while preserving the information by creating a mapping between higher-dimensional feature spaces to lower-dimensional feature space, referred to as the *latent space* [15].

Random Forest (RF) classifiers can be used for dimensionality reduction via feature importance ranking [23].

Principal Component Analysis (PCA) transforms data into a new coordinate system such that the greatest variance by any projection of the data lies on the first coordinate (i.e. the first principal component), meaning the first few principal components capture most of the variance [19].

Logan et al. used a RF classifier to rank photometric colors from a collection of spectroscopic surveys, followed by a PCA for further dimension reduction [23].

Clarke et al. uses Uniform Manifold Approximation and Projection (UMAP) to reduce the number of features from 10 to 2, allowing for visualization of the clustering of different classes of objects in the reduced feature space [7].

The auto-encoder (AE) architecture is a neural network used for unsupervised learning of efficient codings. The AE consists of an encoder and a decoder. The encoder maps the input data to a lower-dimensional latent space, while the decoder reconstructs the original data from the latent representation. By training the AE to minimize the reconstruction error, the model learns to capture the most important features of the data in the latent space [34].

3 Unsupervised learning

3.1 Clustering

Tohill et al. uses auto-encoders (AE) with hierarchical clustering (HC) to identify morphologies of high-redshift galaxies in an unsupervised manner [34]. The AE is trained to learn the latent feature space.

The latent representations are then clustered using hierarchical clustering to group galaxies with similar morphologies. This approach allows for the discovery of distinct morphological classes without relying on labeled training data, enabling the identification of novel or rare galaxy types in large astronomical datasets [34].

4 Supervised learning

4.1 Linear Discriminant Analysis (LDA)

Ferrari et al. uses an LDA classifier to separate extended objects based on non-parametric morphological features described in section 2.1.2, using labels from the Galaxy Zoo project. The goal is to establish a basis for the morphometric space, i.e., the smallest set of independent morphological features that separate different classes of objects [13].

LDA is a supervised learning algorithm that finds a linear combination of features that best separates two or more classes of objects.

4.2 Random Forest

Clarke et al. used a Random Forest (RF) classifier to train photometric samples on a dataset of 3.1 million labeled SDSS objects [7]. The labels were described by Bolton et al. [5]. The model was used to classify a dataset of 111 million unlabeled objects, including approximately 50 million galaxies, 2 million quasars, and 59 million stars [7]. This model, however, does not take into account the uncertainties in the photometric measurements.

Reis et al. proposed a Probabilistic Random Forest (PRF) algorithm that can incorporate these measurement uncertainties in both features and labels to improve classification performance [28].

RF classifiers can also be used for dimensionality reduction via feature importance ranking. For example, Logan et al. used a RF classifier to rank feature importance from a collection of spectroscopic surveys [23].

4.3 Probabilistic Random Forest

Reis et al. modified the RF to account for uncertainties in both features and labels, creating a Probabilistic Random Forest (PRF) as an extension of the traditional RF algorithm that incorporates measurement uncertainties in both features and labels during training and inference [28]. In astronomical datasets, where measurements often come with significant uncertainties due to various observational factors, PRFs can provide more robust and reliable classification results compared to standard Random Forests.

The PRF treats each feature as a probability distribution rather than a single point estimate. During the training phase, the PRF algorithm samples from these distributions to create multiple realizations of the training data, effectively capturing the uncertainty in the measurements. This approach allows the model to learn more robust decision boundaries that account for the inherent noise in the data. The PRF outperforms RF in all cases. It was found to improve classification accuracy by 10% in the case of noisy features, and by 30% in the case of noisy labels [28]. The authors argue that the PRF is particularly well-suited for astronomical applications, where measurement uncertainties are common and so it may outperform RF by an even greater margin.

4.4 Deep neural network architectures

Several deep learning architectures use Mask R-CNN methods, which combines object detection and segmentation to classify and localize objects within an image [18].

Merz et al. uses the Facebook AI `DETECTRON2` library to train a suite of deep neural networks with Mask R-CNN segmentation architecture, ResNets and Transformers to classify galaxy vs. star with Hyper Suprime-Cam Subaru data [25].

Burke et al. uses a Mask Region-based CNN deep learning network to efficiently detect and classify objects with deblending and segmentation [6].

4.5 Rotation-equivariant convolutional neural networks

There is no consistent orientation of objects as observed from Earth (i.e. there is not a fixed reference frame). Traditional CNNs require extensive data augmentation to handle rotated versions of objects, which can be inefficient and may not fully capture the rotational symmetries present in astronomical data. Furthermore, humans are able to learn new concepts and recognize objects from very few examples, while traditional CNNs often require large amounts of labeled data to achieve high performance [8]. This discrepancy highlights the need for more efficient learning algorithms that can generalize better from limited data.

Traditional CNNs are translationally equivariant, i.e. they respond in a predictable way to translations of the input image [8], but not rotationally equivariant. This approach is crucial for effective galaxy classification at scale. Efforts to develop more generalized neural network architecture include E(2)-equivariant convolutions in Steerable CNNs (Weiler et al. [37]).

There have been approaches using convolutional neural networks with layers built in to make models fully or partially equivariant to rotation. Dieleman et al. proposes a set of new CNN layers which perform operations on feature maps that, when used together, make a model partially equivariant to rotation [11].

Traditional CNNs are not inherently rotation invariant, meaning that they may misclassify objects that are rotated versions of those seen during training [10]. This is particularly problematic in astronomy, where celestial objects can appear at arbitrary rotations.

Cohen et al. proposes a Steerable CNN architecture that achieves rotation equivariance by using group convolutions over a rotation group. Steerable CNNs allow filters on both position (i.e. a standard convolution), and pose [8]. Cohen et al. presents a general theory of steerable representations that cover all forms of linear steerability in CNNs [8].

5 Summary

Using ML effectively with astronomical data requires a deep understanding of the nature of the data, as well as a deep understanding of model architectures. Astronomical data often comes with unique challenges, such as limited data, measurement uncertainties, high dimensionality, symmetries, and observational biases, which necessitates very specific modeling approaches.

This encourages an approach that is grounded in first principles, and takes into account measurement uncertainties, observational biases, and symmetries in the data.

References

[1] Cepheids in spiral nebulae. *Pop. Astr.; Vol. 33; Page 252-255*, 33, 1925.

[2] V. Acquaviva. *Machine learning for physics and astronomy.* Princeton University Press, 2023.

[3] S. Balakrishnama and A. Ganapathiraju. Institute for signal and information processing linear discriminant analysis-a brief tutorial.

[4] D. Baron. Machine learning in astronomy: a practical overview, 2019. URL https://arxiv.org/abs/1904.07248.

[5] A. S. Bolton, D. J. Schlegel, Éric Aubourg, S. Bailey, V. Bhardwaj, J. R. Brownstein, S. Burles, Y. M. Chen, K. Dawson, D. J. Eisenstein, J. E. Gunn, G. R. Knapp, C. P. Loomis, R. H. Lupton, C. Maraston, D. Muna, A. D. Myers, M. D. Olmstead, N. Padmanabhan, I. Pâris, W. J. Percival, P. Petitjean, C. M. Rockosi, N. P. Ross, D. P. Schneider, Y. Shu, M. A. Strauss, D. Thomas, C. A. Tremonti, D. A. Wake, B. A. Weaver, and W. M. Wood-Vasey. Spectral classification and redshift measurement for the sdss-iii baryon oscillation spectroscopic survey. *Astronomical Journal*, 144:144, 11 2012. ISSN 00046256. doi: 10.1088/0004-6256/144/5/144. URL https://ui.adsabs.harvard.edu/abs/2012AJ....144..144B/abstract.

[6] C. J. Burke, P. D. Aleo, Y. C. Chen, X. Liu, J. R. Peterson, G. H. Sembroski, and J. Y. Y. Lin. Deblending and classifying astronomical sources with mask r-cnn deep learning. *Monthly Notices of the Royal Astronomical Society*, 490: 3952–3965, 12 2019. ISSN 0035-8711. doi: 10.1093/MNRAS/STZ2845. URL https://dx.doi.org/10.1093/mnras/stz2845.

[7] A. O. Clarke, A. M. M. Scaife, R. Greenhalgh, and V. Griguta. Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million sdss sources without spectra. *Astronomy and Astrophysics*, 639, 5 2020. doi: 10.1051/0004-6361/201936770. URL http://dx.doi.org/10.1051/0004-6361/201936770.

[8] T. S. Cohen and M. Welling. Steerable cnns. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 12 2016. URL https://arxiv.org/pdf/1612.08498.

[9] S. Collaboration, G. A. Pallathadka, M. Aghakhanloo, J. Aird, A. Almeida, S. Amrita, F. Anders, S. F. Anderson, S. Arseneau, C. G. Avila, S. Aviram, C. Aydar, C. Badenes, J. K. Barrera-Ballesteros, F. E. Bauer, A. Behmard, M. Berg, F. Besser, C. M. Bidin, D. Bizyaev, G. Blanc, M. R. Blanton, J. Bovy, W. N. Brandt, J. R. Brownstein, J. Buchner, E. Bulbul, J. N. Burchett, L. Carigi, J. K. Carlberg, A. R. Casey, P. Chakraborty, J. Chanamé, V. Chandra, C. Chiappini, I. Chilingarian, J. Comparat, K. Covey, N. Crumpler, K. Cunha, E. D'Onghia, X. Dai, J. Darling, M. Davis, N. D. Lee, N. Deacon, J. E. M. Delgado, S. Demasi, M. Demianenko, D. Demke, J. Donor, N. Drory, M. A. V. Durango, T. Dwelly, O. Egorov, E. Egorova, K. El-Badry, M. Eracleous, X. Fan, E. Farr, D. P. Finkbeiner, L. Fries, P. Frinchaboy, N. P. G. Fusillo, L. D. S. Félix, B. Gaensicke, E. Galligan, P. García, J. Gelfand, K. Grabowski, E. Grebel, P. J. Green, H. Greve, C. Grier, E. Griffith, P. Guetzoyan, P. Gupta, Z. Hackshaw, P. B. Hall, K. Hawkins, V. Hegedűs, S. Hekker, T. M. Herbst, J. J. Hermes, L. Hernández-García, P. Hiremath, D. W. Hogg, J. Holtzman, K. Horne, D. Horta, Y. Huang, B. Hutchinson, M. Häberle, H. J. Ibarra-Medel, A. P. Ji, P. Jofre, J. W. Johnson, J. Johnson, E. J. Johnston, M. Kaldor, I. Katkov, A. Khalatyan, S. Khoperskov, R. Klessen, M. Kluge, A. M. Koekemoer, J. A. Kollmeier, M. Kounkel, K. Kreckel, D. Krishnarao, M. Krumpe, I. Lacerna, C. Laporte, S. Lepine, J. Li, F.-H. Liang, G. Limberg, X. Liu, S. Loebman, K. Long, Y. Lu, M. Lucey, A. Z. Lugo-Aranda, M. L. M. Martinez-Aldama, K. McKinnon, I. Medan, A. Merloni, S. Morrison, N. Myers, S. Mészáros, J. Müller-Horn, S. Nepal, M. Ness, D. Nidever, C. Nitschelm, A. Oravetz, J. Otto, K. Pan, F. P. Paolino, C. A. N. Peñaloza, M. Pinsonneault, M. T. Popp, A. Price-Whelan, N. Pulatova, A. B. Queiroz, J. Raddick, A. Rankine, H.-W. Rix, C. Román-Zúñiga, D. F. Rosso, J. Runnoe, S. M. Saad, M. Salvato, S. F. Sanchez, N. Sattler, A. Saydjari, C. Sayres, K. Schlaufman, D. P. Schneider, A. Schwope, L. M. Seaton, R. Seeburger, J. Serna, S. Sharma, Y. Shen, A. Sinha, B. Sizemore, M. Sniegowska, Y. Song,

D. Souto, K. Stassun, M. Steinmetz, Z. Stone, A. Stone-Martinez, G. S. Stringfellow, A. M. Sánchez, J. Sánchez-Gallego, J. Tan, J. Tayar, R. Thai, A. Thakar, P. Thibodeaux, Y.-S. Ting, A. Tkachenko, B. Trakhtenbrot, J. G. F. Trincado, N. Troup, J. R. Trump, N. Ulloa, R. P. V. der Marel, P. Vera, S. Villanova, J. Villaseñor, J. Wang, Z. Way, A.-M. Weijmans, A. Wheeler, J. C. Wilson, A. Wofford, T. Wong, Q. Wu, D. Wylezalek, X.-X. Xue, R. Yan, Q. Yang, N. Zakamska, E. Zari, G. Zasowski, G. Zeltyn, Z. Zheng, C. Zucker, and R. de J. Zermeño. The nineteenth data release of the sloan digital sky survey. *18 Szabolcs Mészáros*, 76:48, 7 2025. URL https://arxiv.org/pdf/2507.07093.

[10] S. Dieleman, K. W. Willett, and J. Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450:1441–1459, 4 2015. ISSN 13652966. doi: 10.1093/mnras/stv632. URL https://ui.adsabs.harvard.edu/abs/2015MNRAS.450.1441D/abstract.

[11] S. Dieleman, J. D. Fauw, and K. Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1889–1898, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/dieleman16.html.

[12] F. Ferrari, H. Dottori, N. Caon, A. Nobrega, and D. B. Pavani. The relationship between the sérsic law profiles measured along the major and minor axes of elliptical galaxies. *Monthly Notices of the Royal Astronomical Society*, 347:824–832, 1 2004. ISSN 0035-8711. doi: 10.1111/J.1365-2966.2004.07254.X. URL https://dx.doi.org/10.1111/j.1365-2966.2004.07254.x.

[13] F. Ferrari, R. R. D. Carvalho, and M. Trevisan. Morfometryka – a new way of establishing morphological classification of galaxies. *Astrophysical Journal*, 814, 9 2015. ISSN 15384357. doi: 10.1088/0004-637X/814/1/55. URL https://arxiv.org/pdf/1509.05430.

[14] L. Ferreira, C. J. Conselice, E. Sazonova, F. Ferrari, J. Caruana, C.-B. Tohill, G. Lucatelli, N. Adams, D. Irodotou, M. A. Marshall, W. J. Roper, C. C. Lovell, A. Verma, D. Austin, J. Trussler, and S. M. Wilkins. The jwst hubble sequence: The rest-frame optical evolution of galaxy structure at 1.5 ¡ z ¡ 6.5. *The Astrophysical Journal*, 955:94, 9 2023. ISSN 0004-637X. doi: 10.3847/1538-4357/ACEC76. URL https://iopscience.iop.org/article/10.3847/1538-4357/acec76https://iopscience.iop.org/article/10.3847/1538-4357/acec76/meta.

[15] S. Fotopoulou. A review of unsupervised learning in astronomy. *Astronomy and Computing*, 48:100851, 7 2024. ISSN 2213-1337. doi: 10.1016/j.ascom.2024.100851. URL https://www.sciencedirect.com/science/article/pii/S2213133724000660#b243.

[16] D. Fraix-Burnet, M. Thuillard, and A. K. Chattopadhyay. Multivariate approaches to classification in extragalactic astronomy. *Frontiers in Astronomy and Space Sciences*, 2:3, 8 2015. doi: 10.3389/fspas.2015.00003. URL http://arxiv.org/abs/1508.06756http://dx.doi.org/10.3389/fspas.2015.00003.

[17] A. W. Graham and S. P. Driver. A concise reference to (projected) sérsic r 1/n quantities, including concentration, profile slopes, petrosian indices, and kron magnitudes. *Publications of the Astronomical Society of Australia*, 22:118–127, 2018. doi: 10.1071/AS05001. URL www.publish.csiro.au/journals/pasa.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. URL https://openaccess.thecvf.com/content_ICCV_2017/papers/He_Mask_R-CNN_ICCV_2017_paper.pdf.

[19] S. M. Holland. Principal components analysis (pca). *Department of Geology, University of Georgia, Athens, GA*, 30602:2501, 2008.

[20] Ž. Ivezić, A. Connolly, J. Vanderplas, and A. Gray. *Statistics, Data Mining and Machine Learning in Astronomy.* Princeton University Press, 2014.

[21] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C.

Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389:1179–1189, 9 2008. ISSN 0035-8711. doi: 10.1111/J.1365-2966.2008.13689.X. URL https://dx.doi.org/10.1111/j.1365-2966.2008.13689.x.

[22] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T. Komura, and W. Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. URL https://liuyuan-pal.github.io/Gen6D/.

[23] C. H. Logan and S. Fotopoulou. Unsupervised star, galaxy, qso classification - application of hdbscan. *Astronomy & Astrophysics*, 633:A154, 1 2020. ISSN 0004-6361. doi: 10.1051/0004-6361/201936648. URL https://www.aanda.org/articles/aa/full_html/2020/01/aa36648-19/aa36648-19.htmlhttps://www.aanda.org/articles/aa/abs/2020/01/aa36648-19/aa36648-19.html.

[24] J. M. Lotz, J. Primack, and P. Madau. A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, 128:163, 7 2004. ISSN 1538-3881. doi: 10.1086/421849. URL https://iopscience.iop.org/article/10.1086/421849https://iopscience.iop.org/article/10.1086/421849/meta.

[25] G. Merz, Y. Liu, C. J. Burke, P. D. Aleo, X. Liu, M. C. 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*, 526:1122–1137, 9 2023. ISSN 0035-8711. doi: 10.1093/MNRAS/STAD2785. URL https://dx.doi.org/10.1093/mnras/stad2785.

[26] D. L. Moche. *Astronomy, A Self-Teaching Guide*. Wiley, 2015. ISBN 978-1-62045-990-4.

[27] C. Y. Peng, L. C. Ho, C. D. Impey, and H.-W. Rix. Detailed structural decomposition of galaxyimages*. *The Astronomical Journal*, 124:266, 7 2002. ISSN 1538-3881. doi: 10.1086/340952. URL https://iopscience.iop.org/article/10.1086/340952https://iopscience.iop.org/article/10.1086/340952/meta.

[28] I. Reis, D. Baron, and S. Shahaf. Probabilistic random forest: A machine learning algorithm for noisy data sets. *The Astronomical Journal*, 157:16, 12 2018. ISSN 1538-3881. doi: 10.3847/1538-3881/AAF101. URL https://iopscience.iop.org/article/10.3847/1538-3881/aaf101https://iopscience.iop.org/article/10.3847/1538-3881/aaf101/meta.

[29] V. Sacek. Diffraction image. https://www.telescope-optics.net/diffraction_image.htm, 2006. Accessed: 2025-12-08.

[30] S. A. Smee, J. E. Gunn, A. Uomoto, N. Roe, D. Schlegel, C. M. Rockosi, M. A. Carr, F. Leger, K. S. Dawson, M. D. Olmstead, J. Brinkmann, R. Owen, R. H. Barkhouser, K. Honscheid, P. Harding, D. Long, R. H. Lupton, C. Loomis, L. Anderson, J. Annis, M. Bernardi, V. Bhardwaj, D. Bizyaev, A. S. Bolton, H. Brewington, J. W. Briggs, S. Burles, J. G. Burns, F. J. Castander, A. Connolly, J. R. Davenport, G. Ebelke, H. Epps, P. D. Feldman, S. D. Friedman, J. Frieman, T. Heckman, C. L. Hull, G. R. Knapp, D. M. Lawrence, J. Loveday, E. J. Mannery, E. Malanushenko, V. Malanushenko, A. J. Merrelli, D. Muna, P. R. Newman, R. C. Nichol, D. Oravetz, K. Pan, A. C. Pope, P. G. Ricketts, A. Shelden, D. Sandford, W. Siegmund, A. Simmons, D. S. Smith, S. Snedden, D. P. Schneider, M. S. Rao, C. Tremonti, P. Waddell, and D. G. York. The multi-object, fiber-fed spectrographs for the sloan digital sky survey and the baryon oscillation spectroscopic survey. *The Astronomical Journal*, 146:32, 7 2013. ISSN 1538-3881. doi: 10.1088/0004-6256/146/2/32. URL https://iopscience.iop.org/article/10.1088/0004-6256/146/2/32https://iopscience.iop.org/article/10.1088/0004-6256/146/2/32/meta.

[31] M. J. Smith and J. E. Geach. Astronomia ex machina: a history, primer, and outlook on neural networks in astronomy. *Royal Society Open Science*, 10, 5 2023. doi: 10.1098/rsos.221454. URL http://arxiv.org/abs/2211.03796http://dx.doi.org/10.1098/rsos.221454.

[32] H. Song, M. Kim, D. Park, Y. Shin, and J. G. Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34:8135–8153, 11 2023. ISSN 21622388. doi: 10.1109/TNNLS.2022.3152527. URL `https://ieeexplore.ieee.org/abstract/document/9729424`.

[33] C. Stoughton, R. H. Lupton, M. Bernardi, M. R. Blanton, S. Burles, F. J. Castander, and et al. Sloan digital sky survey: Early data release. *The Astronomical Journal*, 123:485, 1 2002. ISSN 1538-3881. doi: 10.1086/324741. URL `https://iopscience.iop.org/article/10.1086/324741https://iopscience.iop.org/article/10.1086/324741/meta`.

[34] C. Tohill, S. P. Bamford, C. J. Conselice, L. Ferreira, T. Harvey, N. Adams, and D. Austin. A robust study of high-redshift galaxies: Unsupervised machine learning for characterizing morphology with jwst up to z $\tilde{8}$. 2024. doi: 10.3847/1538-4357/ad17b8. URL `https://doi.org/10.3847/1538-4357/ad17b8`.

[35] M. Walmsley, T. Géron, S. Kruk, A. M. Scaife, C. Lintott, K. L. Masters, J. M. Dawson, H. Dickinson, L. Fortson, I. L. Garland, K. Mantha, D. O'Ryan, J. Popp, B. Simmons, E. M. Baeten, and C. Macmillan. Galaxy zoo desi: Detailed morphology measurements for 8.7m galaxies in the desi legacy imaging surveys. *Monthly Notices of the Royal Astronomical Society*, 526: 4768–4786, 10 2023. ISSN 0035-8711. doi: 10.1093/MNRAS/STAD2919. URL `https://dx.doi.org/10.1093/mnras/stad2919`.

[36] A.-M. Weijmans. Manga: Mapping nearby galaxies at apache point observatory. *Proceedings of the International Astronomical Union*, 10:100–103, 8 2015. ISSN 1743-9213. doi: 10.1017/s1743921315003476. URL `https://arxiv.org/pdf/1508.04314`.

[37] M. Weiler and G. Cesa. General e(2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 32, 2019.

[38] K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. Casteels, E. M. Edmondson, L. F. Fortson, S. Kaviraj, W. C. Keel, T. Melvin, R. C. Nichol, M. J. Raddick, K. Schawinski, R. J. Simpson, R. A. Skibba, A. M. Smith, and D. Thomas. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435:2835–2860, 11 2013. ISSN 0035-8711. doi: 10.1093/MNRAS/STT1458. URL `https://dx.doi.org/10.1093/mnras/stt1458`.

[39] E. L. Wright, P. R. Eisenhardt, and et al. The wide-field infrared survey explorer (wise): Mission description and initial on-orbit performance. *The Astronomical Journal*, 140:1868, 11 2010. ISSN 1538-3881. doi: 10.1088/0004-6256/140/6/1868. URL `https://iopscience.iop.org/article/10.1088/0004-6256/140/6/1868https://iopscience.iop.org/article/10.1088/0004-6256/140/6/1868/meta`.

[40] Y. Zhang and Y. Zhao. Astronomy in the big data era. *Data Science Journal*, 14:11–11, 5 2015. ISSN 1683-1470. doi: 10.5334/DSJ-2015-011. URL `https://account.datascience.codata.org/index.php/up-j-dsj/article/view/dsj-2015-011`.