Chapter in **Advancing Astrophysics with the SKA – II**

Source Finding and Characterisation for SKAO Science

Sabyasachi Pal,[1] Souvik Manik,[1] M Carmen Toribio,[2] Geferson Lucatelli,[3,7] Omkar Bait,[4,5] Simone Riggi,[6] Antxon Alberdi,[7] Philippa Hartley,[8] Javier Moldon,[7] Mamta Pandey-Pommier[9] and Rob Beswick[3]

[1]*Department of Pure and Applied Sciences, Midnapore City College, West Bengal, India*
[2]*Department of Space, Earth and Environment, Chalmers University of Technology, Onsala Space Observatory, SE-439 92 Onsala, Sweden*
[3]*Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, The University of Manchester, Manchester M13 9PL, UK*
[4]*National Radio Astronomy Observatory, 520 Edgemont Road, Charlottesville, VA 22903, USA*
[5]*The NSF-Simons AI Institute for Cosmic Origins, USA, 201 E. 24th Street, POB 4.102, Austin, Texas 78712-1229*
[6]*INAF - Osservatorio Astrofisico di Catania, Via Santa Sofia 78, 95123 Catania, Italy*
[7]*Instituto de Astrofísica de Andalucía (IAA-CSIC), Glorieta de la Astronomía s/n, 18008 Granada, Spain*
[8]*SKA Organisation, Jodrell Bank Observatory, Cheshire, SK11 9FT, UK*
[9]*Pole Scientific, University Catholic of Lyon, Campus Saint-Paul, 10 place des Archives 69288, Lyon Cedex 02, France*

E-mail: sabya.pal@gmail.com

The advancements in highly sensitive and powerful radio telescopes, including the Square Kilometre Array Observatory (SKAO) and its precursors, MeerKAT, ASKAP, MWA, and HERA, will enable us to create the deepest radio images of the sky. However, due to the sheer scale of the datasets, manually classifying and analyzing this data is computationally expensive, time-consuming, and laborious. Therefore, the development of automated algorithms to detect and classify complex morphological radio sources from large astronomical surveys is the need of the time. In this chapter, we examine the recent advancements and challenges in source-finding techniques triggered by the analysis of SKAO precursor data and the SKA Data Challenges, both in spectral and continuum modes, along with the growing demand for computational resources and automated source detection methods using machine learning (ML) algorithms for the identification and characterisation of new populations of sources, addressing their complex and diffuse morphologies, as well as transient nature. Additionally, we discuss the critical factors affecting the quality and limitations of automated source-finding techniques, including artefacts from residual continuum emission, sidelobes, radio frequency interference (RFI), technical failures, calibration issues, flagging methods and false positives. With the availability of the full operational phase of the SKAO, continued advancements in source detection algorithms and computational infrastructure will be essential to fully exploit its scientific potential.

1 Introduction

The Square Kilometre Array Observatory (SKAO) is poised to transform radio astronomy, delivering surveys with unprecedented depth, angular resolution, and sensitivity. These surveys will reveal billions of individual radio sources, from compact star-forming galaxies and active galactic nuclei (AGN) to vast diffuse structures such as relics and halos. Turning these data into scientific discovery depends critically on two linked processes: the automated finding of sources in survey images and the subsequent classification of their morphologies. Detection ensures that sources are reliably identified against complex noise backgrounds, while classification assigns physical meaning by distinguishing between structures, evolutionary stages, and emission mechanisms. Together, source finding and morphological classification form the backbone of SKAO data science.

Even this most fundamental task of *source finding* has evolved into a major scientific and computational challenge in the SKAO era. The unprecedented sensitivity, angular resolution, and field of view of the SKAO and its pathfinder facilities, such as the Low-Frequency Array (LOFAR), the MeerKAT telescope, the Australian Square Kilometre Array Pathfinder (ASKAP), the upgraded Giant Metrewave Radio Telescope (uGMRT), the Very Large Array (VLA) and e-MERLIN have ushered in an era of data abundance. These instruments collectively produce petabyte-scale data products, marking a shift from manually curated observations toward fully automated, data-intensive science.

Over the past decade, deep, wide-area continuum surveys such as the LOFAR Two-Metre Sky Survey (LoTSS; Shimwell et al. 2017, 2022), the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE; Jarvis et al. 2016; Hale et al. 2025), and the Evolutionary Map of the Universe (EMU; Norris et al. 2021b) have increased the number of known radio sources by several orders of magnitude compared to earlier surveys like The NRAO VLA Sky Survey (NVSS; Condon et al. 1998) and Faint Images of the Radio Sky at Twenty Centimeters (FIRST; Becker et al. 1995), which each contained about 10^6 sources. SKA Phase I alone is expected to detect tens to hundreds of millions of sources. This growth introduces severe challenges for catalog creation, since every detected source must be characterized in terms of flux density, position, morphology, and spectral properties. Furthermore, the increasing diversity of morphologies, from compact AGN cores and star-forming galaxies to diffuse halos and relics, complicates automated detection even further.

The challenge is not merely one of scale but also of complexity. Radio interferometric imaging involves calibration, deconvolution, and mosaicking steps that can introduce artifacts, correlated noise, and position-dependent sensitivity. These effects can mimic or obscure real signals, making robust extraction difficult. As data volumes continue to grow, manual inspection and cleaning become infeasible. The field has therefore entered the *big data paradigm* of astronomy, where automated, statistically rigorous, and computationally scalable methods are essential (Whiting, 2012; Riggi et al., 2016).

Reliable and complete source catalogues are fundamental to the scientific validity of downstream analyses. Incompleteness or systematic biases can propagate into derived quantities such as source counts, luminosity functions, and clustering statistics, ultimately affecting cosmological and galaxy evolution studies (Hancock et al., 2012; Hale et al., 2019). The ideal source finder must therefore optimize three key metrics: *completeness* (the fraction of real sources detected), *reliability* (the

fraction of detections that are genuine), and *parameter accuracy* (the fidelity of measured fluxes, positions, sizes, etc.). Balancing these simultaneously is a non-trivial problem, particularly for faint or diffuse sources near the noise level.

The evolution of automated source finding has mirrored advances in radio astronomy. Early tools such as SAD and IMSAD provided simple Gaussian fitting for compact sources but were limited by noise and extended emission. Later generations, SExtractor (Bertin and Arnouts, 1996), Aegean (Hancock et al., 2012), PyBDSF (Mohan and Rafferty, 2015), and Selavy (Whiting, 2012), introduced adaptive thresholding, background estimation, and multi-component modeling. These algorithms, now integral to SKA precursor pipelines, incorporate multi-scale decomposition, island-based detection, and local noise estimation. However, they still assume Gaussian noise and simple morphologies, assumptions that often fail in the presence of direction-dependent calibration errors, residual sidelobes, or diffuse emission.

The challenge is even greater for spectral-line data cubes, where the third (frequency or velocity) dimension requires source finders capable of handling spatial-spectral correlations. Tools such as SoFiA (Source Finding Application; Serra et al. (2015)) and DUCHAMP (Whiting and Humphreys, 2012) address these challenges using multidimensional thresholding, smoothing, and reliability analysis. These form the basis of pipelines for HI surveys like WALLABY using ASKAP (Westmeier et al., 2022) and LADUMA using MeerKAT (Blyth et al., 2018), and will be vital for future SKA-MID HI surveys. Nevertheless, they struggle with low signal-to-noise or irregular sources, motivating the integration of more adaptive, learning-based strategies.

A promising frontier is the incorporation of artificial intelligence (AI) and machine learning (ML) into source finding and classification. These approaches can capture complex, non-linear patterns and learn directly from data, bypassing many limitations of rule-based algorithms. Convolutional neural networks (CNNs), U-Net architectures, and generative models have shown success in source segmentation, denoising, and morphological classification (Bonaldi et al., 2021; Magro et al., 2022; Lao et al., 2021). Provided that sufficient representative training data exist, ML approaches can adapt to diverse morphologies, noise conditions, and instrumental effects. Active research continues on training data generation, transfer learning, and model interpretability, especially for the heterogeneous datasets expected from SKA Phase I and II.

These advances must be supported by developments in data management and high-performance computing (HPC). The SKA Science Data Processor (SDP) will be among the largest HPC infrastructures ever built, processing exabytes of raw visibilities into science-ready catalogues. Efficient parallelization, distributed memory handling, and I/O optimization are integral to any SKA-scale source-finding framework. Equally crucial are reproducibility and transparency: pipelines must be automated yet auditable to ensure scientific trust in derived catalogues.

In summary, the SKAO era presents both an extraordinary opportunity and a profound computational challenge. The vast data volumes, complex noise environments, and morphological diversity of radio sources demand innovative, scalable, and intelligent algorithms. The field is evolving from classical detection and fitting toward hybrid methods that combine physical modelling, statistical inference, and AI-driven learning. The success of SKAO science will depend on such next-generation

frameworks that can ensure catalogue integrity, reliability, and computational efficiency.

This chapter reviews the methods developed for source finding and classification, emphasizing lessons from current SKAO precursors and SKAO Data Challenges. It contrasts classical approaches with machine learning methods, assesses their strengths and limitations, and explores their integration into unified pipelines. The chapter concludes with a discussion on scalability, interpretability, and scientific reliability, and how these elements will shape SKA's transformational science.

2 Source Finding Techniques for SKAO Science

In the context of the SKAO and its pathfinder surveys, source finding has evolved from a relatively straightforward image-analysis task into one of the most computationally demanding challenges in modern astronomy. The large data volumes, high dynamic range, and complex source morphologies expected from SKAO observations demand algorithms that are not only statistically robust but also scalable and fully automated. Traditional source finders, developed primarily for smaller, shallower surveys, must now be re-evaluated and adapted to handle petabyte-scale datasets while maintaining high completeness and reliability.

This section outlines the major developments in source-finding methodologies, beginning with classical detection and fitting algorithms, progressing to modern machine-learning approaches, and concluding with a comparative assessment of their performance and applicability to SKA-scale data processing.

2.1 Classical Source Finding Techniques

The rapid growth of modern radio surveys in both sky coverage and data volume has made automated, efficient, and statistically reliable source-finding algorithms indispensable. Contemporary source finders are expected to deliver high completeness and reliability while maintaining computational scalability across petabyte-scale datasets. Over the past two decades, a wide range of source-finding tools have been developed and optimized for different scientific and instrumental contexts. Among the most widely used are Aegean (Hancock et al., 2012, 2018), the Astronomical Point source EXtractor (APEX; Makovoz and Marleau 2005), blobcat (Hales et al., 2012), the Curvature Threshold Extractor (CuTEx; Molinari et al. 2011), Duchamp (Whiting, 2012), and the IFCA Biparametric Adaptive Filter (BAF; López-Caniego and Vielva 2012) and Matched Filter (MF; López-Caniego et al. 2006) algorithms. Additional examples include PyBDSF (Mohan and Rafferty, 2015), the Python Source Extractor (PySE; Spreeuw 2010; Swinbank et al. 2015; Carbone et al. 2018), SAD and its variant HAPPY (Condon et al., 1998; White et al., 1997), Selavy (Whiting and Humphreys, 2012), SExtractor (Bertin and Arnouts, 1996), and the SOURCE_FIND pipeline developed for AMI (AMI Consortium: Franzen et al., 2011). More recent developments include Caesar (Riggi et al., 2016, 2019b) and ProFound (Robotham et al., 2018; Hale et al., 2019). While many of these packages were originally developed for optical imaging (e.g. APEX, CuTEx, ProFound, and SExtractor), they have since been adapted for radio astronomical applications. In addition to two-dimensional (2D) continuum finders, specialized three-dimensional (3D) packages such as SoFiA (Serra et al., 2015; Koribalski et al., 2020) are optimized for spectral-line data cubes and can also operate in 2D mode.

No single source finder performs optimally under all observing conditions or data regimes. Each tool is typically tuned for a particular type of source or image property (Hale et al., 2019; Bonaldi et al., 2021). Broadly, existing algorithms can be categorized into those optimized for compact sources and those tailored for extended or diffuse emission (see Table 1). Some also have specialized applications: for example, BLOBCAT excels at handling linearly polarized data (Hales et al., 2012), DUCHAMP is widely used for H I emission detection (Whiting, 2012), CUTEX performs well in crowded or highly variable background conditions (Molinari et al., 2011), and PYSE is particularly suited for transient searches (Fender et al., 2007; van Haarlem et al., 2013). Several packages, including AEGEAN, SELAVY, and CAESAR, also offer multiprocessing or parallelized implementations to efficiently handle large-scale survey data. In parallel, novel frameworks based on machine learning (Bonaldi et al., 2021; Lao et al., 2021; Magro et al., 2022) and citizen science efforts such as *Radio Galaxy Zoo* (Banfield et al., 2015; Alger et al., 2018) are being explored to complement these traditional approaches.

Most classical source-finding pipelines operate through three fundamental stages: (1) background and noise estimation, (2) island or region detection, and (3) component characterization or model fitting.

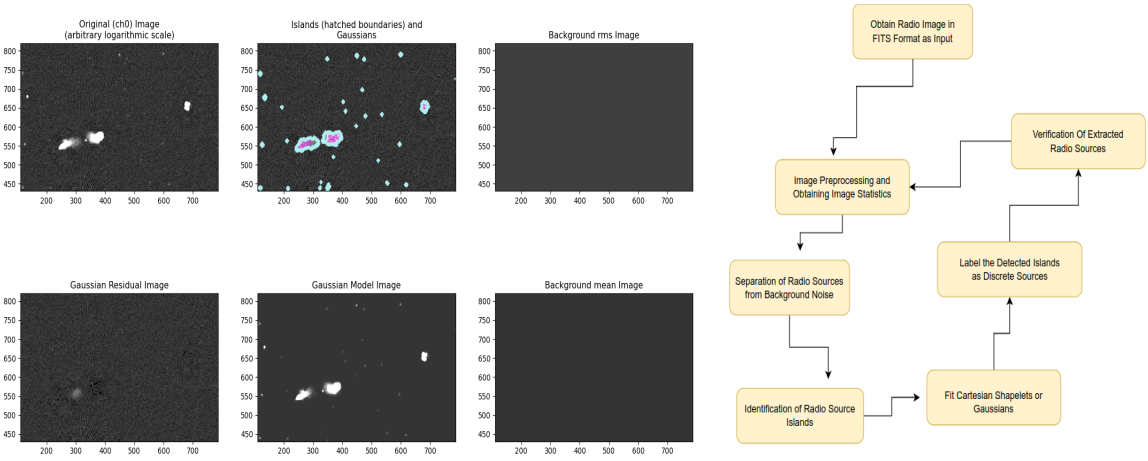

Figure 1: Different stages of the process of characterisation of radio sources with widely used PyBDSF.

2.1.1 Background and noise estimation

A reliable background model is crucial, as it directly affects the sensitivity and false detection rate of the pipeline. Many algorithms, such as AEGEAN, PYBDSF, and SELAVY, use a sliding-box approach in which the local background and noise levels are estimated using neighboring pixels within a defined box size. The choice of box size is critical: too small a region leads to noise overestimation near bright sources, whereas too large a region can smooth over real spatial variations in the background (Huynh et al., 2012). Different statistical estimators are used across finders: for instance, AEGEAN uses the median background with the inter-quartile range (IQR) as a noise measure (Hancock et al., 2012, 2018); PYBDSF relies on mean (μ) and root-mean-square (RMS, σ) noise values (Mohan and Rafferty, 2015); while SELAVY supports both μ/σ and median/MADFM (mean absolute deviation from the median) estimators (Whiting and Humphreys, 2012). Optical-origin

codes such as SExtractor and ProFound apply $\kappa\sigma$-clipping and mode estimation across the image (Bertin and Arnouts, 1996; Robotham et al., 2018). The choice of estimator significantly influences detection completeness and reliability (Huynh et al., 2012).

In practice, the performance of these methods is highly sensitive to the parameter choices, including box size, clipping thresholds, and estimator selection, which often requires domain expertise to optimise for a given dataset. Default parameter settings may not generalise well across surveys with differing resolutions, sensitivities, and noise characteristics, potentially leading to systematic biases in source detection and characterisation. This dependence on expert tuning presents a challenge for fully automated pipelines, highlighting the need for robust, adaptive methods that can minimise manual intervention while maintaining consistent performance across heterogeneous datasets.

2.1.2 Island detection

Once the background is estimated, candidate emission regions, often termed "islands", are identified. Threshold-based detection remains the most common approach, in which contiguous pixels above a specified multiple of the noise level are grouped into islands. Variants of this approach are employed in Duchamp, ProFound, Selavy, and SExtractor (Whiting and Humphreys, 2012; Robotham et al., 2018; Bertin and Arnouts, 1996). Some packages employ more sophisticated region-growing techniques such as flood-fill algorithms (Aegean, PyBDSF, blobcat, and Caesar) or iterative dilation methods inspired by Kron/Petrosian apertures (ProFound; Petrosian 1976; Kron 1980). In certain cases, watershed deblending is used to separate overlapping sources by following the topographic "valleys" in brightness distribution.

2.1.3 Component modeling

After island detection, the emission is decomposed into individual source components. The simplest approach identifies brightness peaks within each island and fits them with 2D elliptical Gaussians, an appropriate model for compact sources that resemble the synthesized beam shape. This strategy is used by Duchamp and Selavy (Whiting and Humphreys, 2012). More complex finders such as PyBDSF and PySE fit multiple Gaussians per island to better represent extended or multi-component sources (Mohan and Rafferty, 2015; Swinbank et al., 2015). Algorithms like Aegean and Caesar incorporate curvature maps or local peak searches to constrain Gaussian fits more robustly (Hancock et al., 2012, 2018; Riggi et al., 2016). Caesar (Riggi et al., 2019a), in particular, combines peak identification with watershed-based segmentation and applies additional filtering (e.g. wavelet, hierarchical clustering, or active contour methods) to extract faint, diffuse, or irregular structures, making it one of the most flexible frameworks for extended emission.

Overall, while classical source finders remain the workhorses of radio survey pipelines, their assumptions about noise, morphology, and source structure limit their applicability in the high-complexity data environment of the SKAO era. This motivates the transition toward hybrid and machine learning based approaches capable of operating at the required sensitivity, speed, and morphological diversity of next-generation surveys.

2.2 3D Source Finding and Spectral Line Surveys

For spectral-line observations, such as HI surveys, the problem of source finding extends into three dimensions: two spatial axes and one spectral (velocity or frequency) axis. In this case, the goal is to detect emission features across multiple frequency channels while accounting for correlated noise and line broadening due to galaxy rotation or turbulence. Classical 2D thresholding is inadequate for this task because real emission often appears as faint, contiguous structures in both spatial and spectral dimensions.

2.2.1 HI emission source finding

A key difference between source-finding in the continuum and in HI emission is that HI emission is relatively faint compared to the noise floor of the given image plane or spectral channel. Fortunately, HI astronomers are able to recover the HI emission through the increase in SNR from integrating across the full frequency range that the HI emission spans. This means that the contiguity of emission across the frequency span provides useful information about the presence of an HI source and is typically crucial in detecting HI sources (and separating true source emission from radio frequency interference (RFI) false positives, for example).

Specialized 3D source finders, such as DUCHAMP (Whiting, 2012) and the Source Finding Application (SoFiA) (Serra et al., 2015), were developed to address this challenge. SoFiA is currently the main tool HI astronomers turn to for automated source finding and characterisation. As evident from SDC2, many of the teams have relied on SoFiA either as the singular tool used to perform source-finding and characterisation, or in combination with newer machine learning methods. DUCHAMP implements a multi-threshold approach with optional smoothing kernels to enhance signal detectability, while SoFiA offers a modular framework combining pre-filtering, detection, reliability analysis, and parametrization. The more recent SoFiA2 (Westmeier et al., 2021) provides better scalability and performance for modern surveys such as WALLABY (ASKAP) and LADUMA (MeerKAT). These tools are critical for the upcoming SKA-MID HI surveys that aim to detect millions of galaxies in emission up to redshift $z \sim 1$. Despite their success, 3D finders also face difficulties in distinguishing faint extended structures from correlated noise, motivating the exploration of machine-learning-based approaches.

The strength of ML-based source finders lies in their ability to model non-Gaussian noise systematics that are inherent in real HI survey datasets. In recent times, several groups have achieved greater performance in source-finding through integrating machine-learning based methods with more traditional tools such as SoFiA (e.g. Barkai et al. 2023; Hartley et al. 2023; Håkansson et al. 2023). Recently, Wang et al. (2025) has also demonstrated the potential to use a machine learning-based method to streamline false positive rejection within the ASKAP WALLABY survey's source finding workflow with excellent efficiency.

2.2.2 HI absorption source finding

The challenge of automated HI absorption source finding is distinct from that of HI emission. This is largely due to the typically narrower linewidths of the narrow component and the relatively weak but broad linewidth of the broadened component. As such, the stability of the bandpass is crucial for accurately identifying an HI absorption line.

The MeerKAT MALS survey employs an automated method based on maximum likelihood optimization and iterative fitting of multiple Gaussian components within an absorption profile, with the final solution determined using Bayesian information criteria (Gupta et al., 2025)..

Similar to SoFiA—which is used to identify HI sources in emission—the ASKAP FLASH survey, a large untargeted HI absorption survey, employs FLASHfinder (Allison et al., 2012), a multi-nested Bayesian Monte Carlo tool that simultaneously detects and fits HI absorption features. The performance of FLASHfinder as a production-ready, automated HI absorption line finder can be found in Yoon et al. (2025). Similar to the ASKAP WALLABY source finding workflow, visual inspection is still required to separate the false positives from the true detections. As such, the team is working to develop more complex machine-learning based solutions to address the non-Gaussian noise statistics that are present in the data products.

Table 1: Summary of source finders design characteristics

Source Finders	**Source Type**			**Multiprocessing**
	Compact	Extended	Diffuse	
PyBDSF	✓	✓		
Aegean	✓			✓
APEX	✓			
ProFound	✓	✓	✓	
blobcat	✓	✓		✓
Caesar	✓	✓	✓	✓
CuTEx	✓			
IFCA BAF/MF		✓		
Selavy	✓			✓
PySE	✓			
SAD	✓			
SExtractor		✓		
SOURCE_FIND		✓		

2.3 Machine Learning in Source Finding

With the advent of large data volumes and increasingly complex source morphologies, machine learning has emerged as a powerful alternative for automated source detection and classification. ML approaches, particularly deep learning (DL), enable the identification of non-linear patterns and subtle features that are often undetectable by traditional algorithms. In radio astronomy, CNNs and U-Net architectures have been successfully applied to tasks such as source segmentation, denoising, and morphological classification (Akeret et al., 2017; Aniyan and Thorat, 2017; Lukic et al., 2019b; Wu et al., 2019a). These models learn directly from labeled datasets, automatically distinguishing real sources from noise or imaging artefacts.

ML-based source finding methods can be broadly divided into supervised and unsupervised approaches, each with distinct advantages and limitations. Supervised methods require labeled training data in which source positions, extents, and morphologies are known, typically derived

from simulations or cross-matched catalogs rather than direct ground truth in real observations. This introduces several challenges: the ground truth for real radio data is rarely known with certainty; simulations may not fully capture the complexity of real images, including correlated noise structures, calibration artefacts, and the morphological diversity of extended sources; and biases present in the training set risk propagating into derived catalogs (Lukic et al., 2019b; Riggi et al., 2023). By contrast, unsupervised and self-supervised approaches learn representations directly from the data distribution without labeled examples, making them potentially more generalisable across heterogeneous surveys, though typically at the cost of less precise source localisation and harder quantitative validation.

The study by Hopkins et al. (2015) demonstrated that while traditional source finders perform well for compact sources, they lack sufficient robustness in extracting extended or diffuse emission. Deep learning architectures such as U-Net offer pixel-level segmentation capabilities, making them effective for detecting extended sources and diffuse structures that classical methods often miss. Similarly, hybrid methods that combine classical pre-processing (e.g., noise estimation or thresholding) with deep learning refinement have shown promising results in improving both completeness and reliability, particularly in crowded or high-noise environments (Vafaei Sadr et al., 2019; Riggi et al., 2023). Recent SKA Data Challenges and precursor surveys have provided valuable benchmarks for evaluating ML-based methods, allowing quantitative comparisons against classical approaches (Bonaldi et al., 2021; Hartley et al., 2023).

AI-based source finders such as COSMODEEP (Gheller et al., 2018), DEEPSOURCE (Vafaei Sadr et al., 2019), and ConvoSource (Lukic et al., 2019b), employing custom CNN architectures, as well as more recent methods including Astro R-CNN (Burke et al., 2019), Tiramisu (Pino et al., 2021), CAESAR-MRCNN/YOLO (Riggi et al., 2023), HeTu (Lao et al., 2023), RadioGalaxyNET (Gupta et al., 2024), YOLO-CIANNA (Cornu et al., 2024), and ContinUNet (Stewart et al., 2024) – adopting either semantic segmentation architectures such as U-Net (Ronneberger et al., 2015) or object-detection frameworks like Mask R-CNN (He et al., 2017) and YOLO (Redmon et al., 2016) – have demonstrated high detection accuracies (often exceeding 90%), particularly in recovering faint and diffuse sources that remain challenging for traditional algorithms.

Notably, Stewart et al. (2024) developed ContinUNet, a U-Net-based framework specifically designed for radio continuum source finding, demonstrating strong performance in recovering faint and extended emission across survey-like conditions.

Taran et al. (2023) presented a novel deep learning framework that bypasses the traditional imaging step entirely by taking as input a low-dimensional vector of sampled visibility (uv-plane) data and directly outputting sky-source positions. They demonstrate with simulations (based on Atacama Large Millimeter/submillimeter Array; ALMA) that their ML-model achieves significantly higher completeness than standard image-plane source-finders (e.g., PyBDSF) in the low-signal-to-noise regime (S/N $\approx 1-10$) and is much faster ($\sim$ 30x) since it avoids the full imaging pipeline. Drozdova et al. (2024) take a different approach: they use simulated ALMA "dirty" radio-interferometric images and apply a conditional denoising diffusion probabilistic model (DDPM) to reconstruct sky models (i.e., clean images) and then perform source localization and flux estimation from those reconstructions. They show that their method delivers high completeness (over 90%) even at S/N

≈ 2, and yields more accurate flux estimates than the traditional CLEAN + PyBDSF approach. Importantly, the stochastic nature of the DDPM also allows uncertainty quantification (in source location and flux density) for detected sources.

Together, these works illustrate two complementary AI/ML-driven approaches for source finding for next-generation radio-interferometric data: one directly from visibilities, the other via improved image reconstruction prior to detection, both promising for the SKAO era where data volumes, faint-sources and imaging pipeline complexity pose major challenges for traditional approaches. Despite their advantages, ML-based methods depend heavily on the quality and diversity of their training datasets. Generating realistic, labeled simulations that capture the full range of SKAO data characteristics—including calibration residuals, instrumental noise, and sky complexity is a non-trivial task. Moreover, the interpretability and generalizability of deep learning models remain open challenges, underscoring the need for transparent and physically motivated architectures.

2.4 Comparative Assessment

Traditional source-finding algorithms rely on deterministic rules such as noise thresholding, Gaussian fitting, and curvature-based detection. While these methods perform well for compact and isolated sources, they often struggle with complex, diffuse, or blended structures, especially in high-noise or crowded regions. Their performance is highly sensitive to parameter tuning, background estimation, and beam characteristics, making them less adaptable to the heterogeneity of modern radio surveys.

Quantitative comparisons of traditional source finders have been carried out in a number of studies. For instance, Hopkins et al. (2015) and Hancock et al. (2018) evaluated multiple algorithms (including Aegean, PyBDSF, and Selavy) using simulated and real radio data, demonstrating that completeness and reliability can vary by ~10–20% depending on source morphology, signal-to-noise ratio, and parameter choices. These studies demonstrate that no single algorithm performs optimally across all source types and survey configurations, particularly for extended and low-surface-brightness emission.

By contrast, ML approaches, particularly those based on deep learning, can learn complex non-linear relationships directly from the data, enabling more robust identification of faint, irregular, or overlapping sources. Recent works (e.g., Lukic et al. 2019b; Wu et al. 2019a) show that ML-based methods can achieve improved completeness at low signal-to-noise levels, often outperforming traditional techniques by ~10% or more in challenging regimes, while maintaining comparable reliability. However, these gains depend on the availability of representative training data and careful validation to avoid biases.

As SKAO and its precursors push toward petabyte-scale imaging, integrating ML within automated pipelines has become increasingly essential for scalable, accurate, and efficient source detection and classification.

3 Morphological Classification of Radio Sources

3.1 The Scientific Motivation

Detecting sources is only the first step in the radio data analysis pipeline; the subsequent morphological classification is crucial for interpreting the underlying astrophysics. Radio morphology encodes key information about the physical mechanisms powering emission, such as relativistic jets from AGN, star formation (SF) processes, or large-scale shocks in galaxy clusters. Distinguishing between compact star-forming galaxies, Fanaroff–Riley type I/II radio galaxies, bent-tailed sources, and hybrid morphology radio sources (HyMoRS) offers insight into feedback processes, jet–environment interactions, and the co-evolution of galaxies and their central black holes.

SKAO and its precursors (ASKAP, MeerKAT, LOFAR, uGMRT), morphological classification is not merely a cataloguing exercise; it underpins several of SKAO's primary science goals, from tracing cosmic magnetism and black hole growth to probing large-scale structure formation. Automated and reliable classification of the billions of sources expected in SKAO continuum surveys is therefore an essential component of future radio astronomy.

3.2 Classical Approaches to Radio Source Classification

Historically, the classification of radio sources has relied heavily on visual inspection and manual interpretation, supported by simple parametric descriptors such as flux ratios, spectral indices, symmetry measures, and Gaussian component fitting (Fanaroff and Riley, 1974; Condon et al., 1998). Early radio surveys such as NVSS and FIRST enabled the construction of large morphological catalogues, but these methods were inherently limited by human subjectivity, scalability, and sensitivity to faint extended emission.

Radio sources exhibit a remarkable diversity of morphologies, reflecting differences in their central engines, environments, and evolutionary stages. The most fundamental classification scheme, introduced by Fanaroff and Riley (1974), divides extended radio galaxies into two broad categories: *Fanaroff–Riley type I* (FR I) and *type II* (FR II). FR I sources display edge-darkened morphologies with brightness peaking near the core, typically associated with lower radio power and turbulent, decelerating jets. In contrast, FR II sources are edge-brightened, with luminous terminal hotspots at the ends of collimated jets, corresponding to higher jet power and often more efficient particle acceleration. This dichotomy remains one of the cornerstones of radio galaxy classification and continues to inform models of jet dynamics and AGN feedback.

Beyond the canonical FR classes, a rich population of atypical and complex morphologies has been identified. These include HyMoRS, which simultaneously exhibits FR I-like structure on one side of the nucleus and FR II morphology on the other, suggesting asymmetric environments or jet instabilities (Gopal-Krishna and Wiita, 2000; Gawroński et al., 2006; Kapińska et al., 2017; Harwood et al., 2020; Kumari and Pal, 2022; Manik et al., 2026). Another prominent subclass comprises *giant radio galaxies* (GRGs), whose projected sizes exceed 0.7 Mpc, making them among the largest single extragalactic structures known in the Universe (e.g. Ishwara-Chandra and Saikia, 1999; Dabhade et al., 2020; Delhaize et al., 2021; Oei et al., 2024; Bhukta et al., 2024; Manik et al., 2025; Charlton et al., 2025). GRGs serve as probes of intergalactic magnetic fields, low-density

environments, and long-term AGN activity cycles.

The *Radio Galaxy Zoo* (RGZ) project (Banfield et al., 2015), a landmark citizen-science initiative, revolutionized morphological classification by engaging tens of thousands of volunteers to visually inspect radio sources from the FIRST (Becker et al., 1995) and Australia Telescope Large Area Survey (ATLAS) (Norris et al., 2006; Mao et al., 2012), cross-matched with Wide-field Infrared Survey Explorer (WISE) (Wright et al., 2010) data. The RGZ effort produced a rich, labeled dataset of complex morphologies and provided invaluable training material for machine-learning models such as CLARAN (Wu et al., 2019b). However, manual or rule-based classification cannot meet the demands of SKAO-scale surveys, which will detect billions of sources spanning orders of magnitude in flux density and angular scale. Traditional approaches also struggle with composite or overlapping systems where multiple jets, lobes, or cores complicate identification.

Apart from the typical double-lobed radio galaxies that define the FR classification, numerous peculiar morphologies have been documented. *Winged radio galaxies* (WRGs), which include *X-shaped* (XRGs) and *Z-shaped* (ZRGs) systems, display secondary lobes or "wings" that may arise from jet reorientation, backflow deflection, or black hole mergers (e.g. Yang et al., 2019; Bera et al., 2020; Bhukta et al., 2022b). *Double–double radio galaxies* (DDRGs) exhibit two distinct pairs of lobes aligned along the same axis, providing evidence for recurrent AGN activity (e.g. Konar and Hardcastle, 2013).

Another notable category is that of *bent-tailed (BT)* radio galaxies, which are subdivided into *wide-angle tail (WAT)* and *narrow-angle tail (NAT)* sources. These are typically located in dense cluster environments, where ram pressure from the intracluster medium bends the relativistic jets (e.g. Missaglia et al., 2019; Bhukta et al., 2022a; Pal and Kumari, 2023). BT galaxies are valuable tracers of cluster dynamics and galaxy motions within the intracluster medium.

Recent surveys have also uncovered entirely new and rare morphologies. One striking example is the class of *odd radio circles* (ORCs), enigmatic ring-like radio structures discovered in deep ASKAP observations (e.g. Norris et al., 2021a; Kumari and Pal, 2023; Lochner et al., 2023; Kumari and Pal, 2024; Kumari et al., 2025). Also see the chapter on the study of circular symmetric diffused radio sources with SKAO in Pal et al. (2026). Their origin remains under debate, with hypotheses ranging from AGN-driven shock fronts to spherical remnants of energetic outflows. Similarly, increasingly sensitive imaging has revealed asymmetric, fragmented, and hybrid structures that challenge existing classification schemes (Harwood et al., 2020; Kapińska et al., 2017; Mingo et al., 2019a; Rudnick, 2021; Kumari et al., 2024, 2026).

These discoveries collectively highlight the morphological richness of the radio sky and the limitations of traditional classification approaches. Manual inspection remains invaluable for discovering rare or peculiar sources, but it is insufficient for the data volumes anticipated in the SKAO era. The next generation of classification frameworks must, therefore, be automated, data-driven, and scalable, capable of recognizing both canonical and unconventional morphologies while maintaining scientific reliability across billions of detections.

Historically, morphology was studied by visual inspection, supported by simple quantitative measures such as flux ratios, axial symmetry, and fitted profiles. Campaigns like *Radio Galaxy Zoo*

demonstrated the power of citizen science in scaling up morphological classification, producing labeled datasets that remain invaluable for training and validation. You may find a chapter on the prospect of citizen science research with SKAO in Hota et al. (2026). However, classical classification cannot meet the demands of SKAO data volumes. Visual inspection does not scale to billions of sources, and rule-based algorithms lack the flexibility to capture the diversity of real morphologies.

3.2.1 Non-Parametric Source Characterisation

Before discussing parametric decomposition approaches, it is important to acknowledge the complementary role of non-parametric morphological metrics. These metrics quantify structural properties without assuming a specific mathematical form for the brightness distribution, making them valuable for characterising irregular, disturbed, or complex morphologies that may not be well-described by smooth analytical functions. In optical astronomy, the CAS system, measuring *Concentration* (C), *Asymmetry* (A), and *Smoothness* or *Clumpiness* (S), has proven effective for distinguishing between galaxy types and systems undergoing merging activity (Conselice et al., 2000; Lotz et al., 2004). This framework has been extended to include the *Gini coefficient* (G), which quantifies the distribution of flux amongst pixels, and the second-order moment of the brightest 20 per cent of the flux (M_{20}), which traces the spatial distribution of the most luminous regions (Lotz et al., 2004, 2008). Collectively, the CASGM system provides a robust, quantitative description of galaxy structure that is sensitive to mergers, interactions, and morphological disturbances.

Although these metrics have been predominantly developed in optical astronomy, their extension to radio wavelengths is straightforward and offers significant potential for characterising the diverse structures observed in radio sources. Concentration indices distinguish centrally dominated AGN cores from diffuse jets and extended star formation, asymmetry measurements quantify disturbances from mergers or AGN feedback, and smoothness parameters reveal clumpiness in star-forming regions and radio lobes. Importantly, non-parametric metrics do not require convergence of complex fitting algorithms, making them computationally efficient and robust even in low signal-to-noise regimes. Their application to radio continuum imaging provides systematic characterisation of source morphology without requiring assumptions about the underlying brightness profile, and can feed knowledge into ML pipelines for automated source classification and characterisation (see Section 3.4).

3.3 Parametric Decomposition and Multi-Component Characterisation

Whilst morphological classification traditionally relies on visual inspection and phenomenological categorisation (e.g. FR I/II, bent-tailed sources), parametric decomposition approaches offer a complementary pathway to source characterisation. These methods model the brightness distribution using mathematical functions, enabling quantitative measurements of structural parameters such as sizes, peak brightness, ellipticity, and concentration indices. Parametric decomposition has been widely employed in optical astronomy (Simard et al., 2002; Peng et al., 2010) and more recently adapted for radio continuum studies (Barcos-Muñoz et al., 2015; Hodge et al., 2019; Song et al., 2021; Lucatelli et al., 2024).

The most commonly used parametric models in radio astronomy include Gaussian profiles (Condon,

1997; Condon et al., 1998), exponential functions, and the Sérsic profile (Sérsic, 1963; Caon et al., 1993). The Sérsic law provides a generalised surface brightness distribution of the form

$$I(R) = I_n \exp\left\{-b_n\left[\left(\frac{R}{R_n}\right)^{1/n} - 1\right]\right\}, \tag{1}$$

where I_n is the intensity at the effective radius R_n (enclosing half the total flux), n is the Sérsic index, and b_n ensures that R_n corresponds to the half-light radius. This law generalises both Gaussian ($n = 0.5$) and exponential ($n = 1$) profiles. Compact structures (e.g. AGN cores) typically follow Gaussian profiles resembling the synthesised beam, whilst extended emission often requires higher Sérsic indices ($n \geq 1$).

3.3.1 Multi-Component Sérsic Fitting

Radio sources frequently exhibit complex, multi-component structures spanning multiple spatial scales, including compact cores (AGN activity or starbursts), nuclear diffuse emission (circumnuclear star formation), and large-scale structures (galactic winds/outflows or extended star formation). Accurately characterising these components requires simultaneous fitting of multiple Sérsic profiles to disentangle their contributions. The general model is a linear superposition

$$I_{\text{total}}(x, y) = \sum_{i=1}^{N} I_i(x, y; \Theta_i), \tag{2}$$

where $I_i(x, y; \Theta_i)$ represents the i-th Sérsic component with parameters $\Theta_i = \{I_{n,i}, R_{n,i}, x_{0,i}, y_{0,i}, q_i, \text{PA}_i, n_i\}$. The model is fitted via non-linear least-squares minimisation using algorithms such as Levenberg-Marquardt (Lourakis, 2004) or Trust Region Reflective methods (Liu et al., 2020; Branch, 1996).

However, multi-component fitting is highly sensitive to initial parameter estimates, and poor starting conditions can lead to non-physical solutions or convergence to local minima rather than the global optimum (Häussler et al., 2007; Andrae et al., 2011). This is particularly problematic for sources with irregular morphologies, where the parameter space is large and degenerate, and small perturbations in initial conditions can produce large variations in best-fit parameters.

3.3.2 Source Extraction as a Constraint for Parametric Fitting

A critical advance in addressing the challenges of multi-component fitting lies in integrating source extraction algorithms to provide physically motivated initial conditions. Rather than relying on arbitrary parameter guesses, the source extraction tools described in Section 2.1 can identify distinct emission regions and compute basic photometric properties: peak brightness position (x_0, y_0), half-light radius R_{50}, axis ratio $q = R_b/R_a$, and position angle (PA). Using these measured properties to initialise and constrain the Sérsic fitting makes the minimisation process more efficient and robust, ensuring that fitted components correspond to actual structures rather than mathematical artefacts.

This method has been successfully applied to decompose U/LIRG radio emission into core-compact, nuclear extended, and large-scale diffuse components, enabling detailed measurements of flux densities, sizes, and brightness temperatures across multiple spatial scales (Lucatelli et al., 2024).

The technique generalises beyond radio astronomy; for instance, applying source extraction to optical or infrared images prior to Sérsic fitting could improve bulge–disc decompositions in disturbed or merging systems.

Beyond classical source extraction algorithms, machine learning approaches (see Section 3.4) offer an alternative pathway for constraining parametric fits. Deep learning models trained to segment radio sources or predict structural parameters can provide data-driven priors that adapt to complex morphologies and varying noise conditions, potentially outperforming traditional methods in challenging regimes. This integration of classical algorithms, parametric fitting, and machine learning represents a promising direction for automated source characterisation in the SKAO era.

3.3.3 Multi-Scale and Multi-Wavelength Decomposition

Beyond multi-component fitting, decomposition techniques across spatial scales and frequency bands provide complementary approaches to characterising radio emission. Interferometric observations with different baseline configurations (e.g. *e*-MERLIN and VLA, or MeerKAT and ASKAP) probe distinct spatial scales, with high-angular resolution images sensitive to compact structures and low-angular resolution images recovering diffuse emission. By combining observations at intermediate angular resolutions and multiple frequencies, it is possible to map the transition between nuclear and extended emission whilst disentangling different emission mechanisms.

The technical framework for multi-scale decomposition involves iterative comparison of images at different angular resolutions. Consider a high-resolution image $I_{\rm hr}$ (restoring beam $\Theta_{\rm hr}$) and a low-resolution image $I_{\rm lr}$ (restoring beam $\Theta_{\rm lr}$). To compare these, the high-resolution image is convolved to match the lower resolution:

$$I_{\rm hr}^{\rm conv} = I_{\rm hr} * \Theta_{\rm conv}, \tag{3}$$

where $\Theta_{\rm conv}$ transforms $\Theta_{\rm hr}$ into $\Theta_{\rm lr}$. The residual emission

$$\mathcal{R}_{\rm lr} = I_{\rm lr} - f_\Theta\, I_{\rm hr}^{\rm conv}, \tag{4}$$

where $f_\Theta = A(\Theta_{\rm lr})/A(\Theta_{\rm hr})$ accounts for beam area changes, isolates extended emission not captured at high resolution. Iterating over intermediate angular resolutions constructs a complete picture of emission distribution across spatial scales.

Extending this approach across multiple frequency bands enables construction of individual spectral energy distributions (SEDs) for each spatial component (Lucatelli et al. in preparation). Radio SEDs encode critical information about emission mechanisms: compact AGN cores often exhibit flat or inverted spectra due to synchrotron self-absorption, whilst extended star-forming regions display steep synchrotron spectra, and free-free thermal emission ($S_\nu \propto \nu^{-0.1}$) from H II can be significant in starbursts at higher frequencies due to recent ongoing SF activity. By performing consistent multi-component decomposition across all frequencies with the SKAO, it is possible to construct component-specific SEDs for compact cores, nuclear extended emission, and large-scale diffuse structures. This spatial-spectral approach yields simpler, more interpretable SEDs than traditional fitting methods that model the entire spectrum without spatial information, enabling

robust estimates of thermal fractions, non-thermal spectral indices, and star formation rates on a component-by-component basis.

These multi-scale and multi-wavelength techniques have proven effective for characterising complex systems such as U/LIRGs. For SKAO, which will deliver data across unprecedented ranges of both angular scales (milliarcsecond VLBI to arcminute single-dish) and fractional bandwidths, such decomposition will be essential for disentangling AGN and star formation contributions and understanding feedback processes, accretion physics, and black hole-galaxy co-evolution.

3.3.4 Synergies and Implications for SKAO Source Characterisation

The integration of non-parametric metrics, source extraction, multi-component and multi-scale source decomposition, with multi-wavelength analysis represents a natural evolution towards more quantitative, robust, and automated source characterisation. These methods are well-suited to the SKAO era, where the volume and complexity of data will necessitate scalable pipelines capable of characterising billions of sources across diverse morphologies, spatial scales, and spectral behaviours.

Non-parametric (CASGM) and parametric approaches are complementary. Non-parametric quantities provide rapid, model-independent characterisation robust in low signal-to-noise regimes, ideal for initial classification and quality assessment. Parametric decomposition provides detailed physical measurements (sizes, flux densities, brightness temperatures) enabling quantitative comparison across sources and their individual components. Combined, these approaches allow SKAO pipelines to flag unusual morphologies for detailed follow-up whilst providing precise measurements for well-behaved sources.

Both metrics types serve as valuable input features for machine learning classifiers. Concentration indices, asymmetry, Gini coefficients, Sérsic indices, and effective radii, combined with multi-frequency spectral information, can train supervised classifiers distinguishing AGN, star-forming galaxies, and hybrid sources. The interpretability of these physically motivated features facilitates scientific understanding and model validation, bridging classical and data-driven approaches.

Multi-wavelength datasets provide a natural pathway for integrating classical methods with machine learning. Spectral indices, thermal fractions, and component-based SEDs carry direct physical information about emission mechanisms, enabling models to recognise patterns associated with specific processes (AGN feedback, starburst-driven winds, post-merger relaxation, etc.). ML can enhance multi-component fitting by predicting optimal initial conditions, whilst parametric fitting results can generate high-fidelity synthetic training datasets, linking the automation between classical and data-driven methods.

Computational scalability is critical for SKAO. Non-parametric metrics are inexpensive and suitable for near-real-time screening of billions of sources. Source extraction and parametric fitting are more demanding but can be parallelised efficiently across distributed computing infrastructures using scalable GPU optimisations such as JAX (Bradbury et al., 2018). Multi-scale and multi-wavelength decomposition can exploit hierarchical pipelines where coarse-scale decomposition informs finer-scale fitting, and low-frequency results constrain high-frequency models.

Polarization information provides an additional and highly diagnostic dimension for source charac-

terisation. Polarized intensity, fractional polarization, depolarization behaviour, and Faraday rotation measures can help distinguish AGN-related synchrotron emission from star-forming processes, trace magnetic-field structures, and identify physical components that may not be evident from total-intensity morphology alone. Incorporating polarization products into future source characterisation pipelines will therefore be of particular value.

The quantitative structural and spectral parameters derived from these integrated methods will be essential for future SKAO scientific cases, including studies of AGN feedback, cosmic star formation history, and black hole-galaxy co-evolution. The SKAO, with its unparalleled sensitivity, angular resolution, and spectral coverage, will fully realise the potential of these characterisation techniques, providing complete catalogues of radio sources and their physical properties.

3.4 Machine Learning in Classification

With the growing availability of large-scale radio survey data, machine learning techniques have evolved beyond simple classifiers toward more structured and scalable frameworks. These developments encompass both model-driven innovations and data-centric strategies that collectively enhance the accuracy, interpretability, and generalisation of automated classification systems. The following sections outline key advances in model architectures, data handling techniques, and learning paradigms that have shaped current approaches to radio galaxy classification.

3.4.1 Model-Based Approaches

In recent years, the rapid progress of artificial intelligence and deep learning has led to substantial innovation in model design and optimization. These developments have naturally extended into radio astronomy, a field increasingly characterized by large, high-dimensional, and data-driven analyses. Beyond incremental improvements to algorithmic performance, advances in model architectures have played a central role in achieving robust classification and morphological interpretation of radio sources.

3.4.1.1 CNN Architectures

CNNs have become essential for automated radio galaxy morphology classification. The network architecture itself, its depth, filter organization, and feature hierarchy, largely govern the model's generalization and discriminative capabilities. Consequently, progressively deeper and more sophisticated architectures have been adapted from computer vision to astronomy, including AlexNet (Krizhevsky et al., 2012; Aniyan and Thorat, 2017), VGG-16 (Ma et al., 2019; Wu et al., 2019a), and DenseNet (Huang et al., 2016; Samudre et al., 2022).

The optimal depth of a CNN varies by application. For example, Lukic et al. (2019a) explored four- and eight-layer networks (CONVNET4 and CONVNET8) for the LoTSS DR1 dataset, while Becker et al. (2021) and Tang et al. (2019) used architectures with up to thirteen layers. Their comparative studies showed that deeper architectures (e.g., CONVNET8) achieved slightly higher precision (94.3%) than shallower networks or capsule models (89.7%), likely due to the increased number of non-linear transformations and feature hierarchies. Overall, deeper models tend to

learn more abstract representations, resulting in improved performance across diverse radio source morphologies.

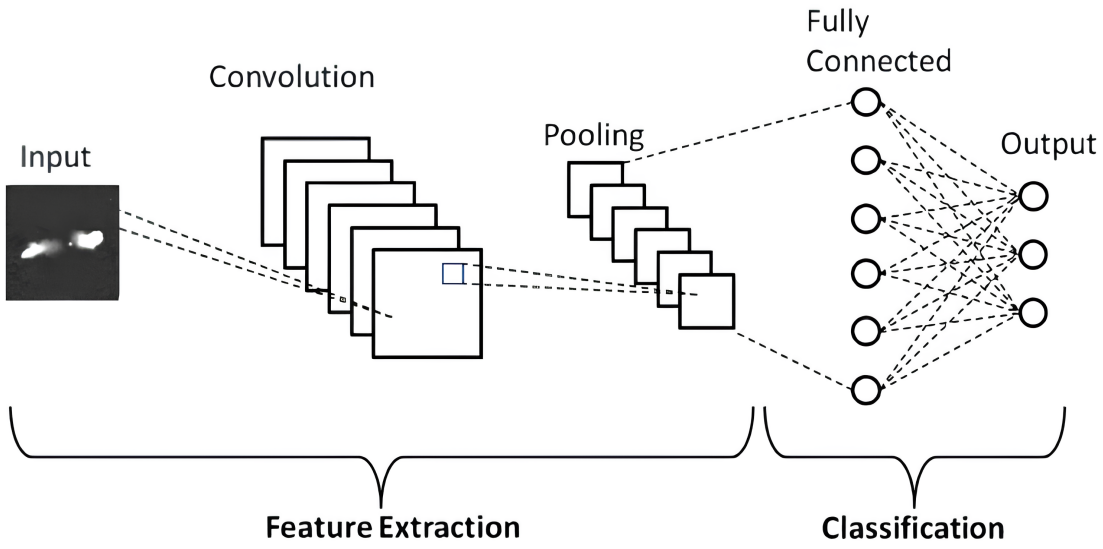

Figure 2: Basic building blocks of a radio galaxy classifier based on a convolutional neural network (CNN).

3.4.1.2 Regularization Techniques

One of the major challenges in deep learning for radio astronomy is overfitting, primarily due to limited labeled data. To address this, various regularization methods are employed to improve model robustness and prevent memorization of the training set.

A widely used approach is *dropout*, which randomly deactivates a subset of neurons during training, thereby reducing co-adaptation of feature detectors (Tang et al., 2019, 2022). Another complementary method is *batch normalization*, which standardizes feature map activations to stabilize and accelerate training by mitigating covariate shifts. These techniques collectively improve convergence speed and enhance generalization across heterogeneous survey data.

3.4.1.3 Specialized Convolutional Blocks

Recent progress has also been driven by the introduction of specialized convolutional blocks and attention mechanisms designed to enhance model interpretability and performance.

Attention gates, inspired by human visual focus, enable models to emphasize salient image regions while suppressing irrelevant background features. Bowles et al. (2021) incorporated attention mechanisms into CNN backbones, demonstrating a 50% reduction in trainable parameters without compromising classification accuracy. The resulting attention maps also offer interpretable insights into model decision-making, advancing the goal of explainable AI in astronomy.

Another architectural innovation is the Group Equivariant CNN (G-CNN), which introduces rotational and reflectional equivariance into convolutional kernels (Cohen and Welling, 2016). This adaptation

allows CNNs to maintain consistent feature representations under geometric transformations, a particularly valuable property given the arbitrary orientations of radio jets. When applied to the MiraBest dataset, G-CNNs demonstrated improved classification stability and accuracy (Scaife and Porter, 2021).

Finally, multi-branch and multi-domain CNNs extend traditional models by incorporating multiple input modalities (e.g., total intensity, polarization maps, or multi-frequency data), enhancing the capacity to capture complex source structures (Tang et al., 2022).

3.4.2 Data-Centric Approaches

While model architecture is critical, the quality and diversity of training data remain equally important. Radio astronomy data must be carefully curated, free from radio-frequency interference (RFI), imaging artifacts, or ambiguous labels to ensure meaningful learning outcomes. Given the scarcity of large, well-annotated datasets, data-centric strategies play a vital role in improving generalization.

3.4.2.1 Data Augmentation

Data augmentation artificially increases dataset size and diversity by applying geometric and photometric transformations such as rotation, flipping, scaling, brightness adjustment, and contrast modification (Aniyan and Thorat, 2017; Alhassan et al., 2018; Lukic et al., 2018; Becker et al., 2021). These techniques help mitigate class imbalance and improve rotational invariance, particularly for underrepresented morphologies like bent or compact sources.

Advanced augmentation strategies use oversampling of minority classes or synthetic data generation with generative adversarial networks (GANs; Rustige et al. 2023). However, care must be taken, as biases in the original data can be propagated or amplified through augmentation. Nonetheless, when properly designed, augmentation remains a cornerstone for robust training in limited-data regimes.

3.4.2.2 Rotation Invariance

Since radio sources can appear at arbitrary orientations, models must exhibit rotational robustness. This can be achieved either through architectural design (e.g., G-CNNs) or through preprocessing techniques that standardize image orientation. Methods such as principal component analysis (PCA) can align source axes prior to model input, effectively normalizing rotation (Polsterer et al., 2019; Brand et al., 2023). Combining augmentation and preprocessing has been shown to significantly enhance classification consistency across surveys.

3.4.2.3 Feature Engineering

In classical machine learning pipelines, handcrafted features, derived from astrophysical or morphological insight, can provide compact, physically meaningful representations of data. Examples include peak brightness, lobe separation, texture measures (e.g., Haralick features; Ntwaetsile and Geach 2021), and rotationally invariant descriptors based on Zernike moments (Sadeghi et al., 2021). Dimensionality reduction methods such as PCA are also used to compress features while retaining key variance (Darya et al., 2023). These engineered features are then input to algorithms

such as random forests, gradient boosting, SVMs, or clustering methods like HDBSCAN, achieving classification accuracies exceeding 95% in some cases. The main limitation, however, lies in the reliance on domain expertise and potential omission of subtle, high-dimensional information.

3.4.3 Weak Supervision Approaches

Given the high cost of manual annotation, weakly supervised paradigms are increasingly adopted to leverage vast unlabeled datasets. Three primary strategies have emerged in radio astronomy: transfer learning, self-supervised/semi-supervised learning, and few-shot learning.

3.4.3.1 Transfer Learning

Transfer learning enables models pre-trained on large datasets to be fine-tuned on smaller, domain-specific samples. This approach effectively transfers learned representations from generic to specialized contexts, drastically reducing the required labeled data. Studies show that CNNs pre-trained on high-resolution surveys (e.g., FIRST; Becker et al. (1995)) retain strong performance when fine-tuned on lower-resolution data such as NVSS (Tang et al., 2019). Architectures like Inception-ResNet-v2 (Szegedy et al., 2017) and DenseNet have achieved accuracies above 90% when applied to FR I/FR II classification (Lukic et al., 2019a).

3.4.3.2 Semi-Supervised and Self-Supervised Learning

Self-supervised and semi-supervised learning (SSL) approaches have recently been explored to mitigate the scarcity of labeled data in radio astronomy. In self-supervised learning, models are first pretrained on unlabeled data using surrogate objectives, such as representation reconstruction or contrastive learning—and subsequently fine-tuned on limited annotated samples. Examples include autoencoder-based pretraining (Ma et al., 2019), as well as contrastive or masked-reconstruction frameworks applied to radio galaxy datasets (Hossain et al., 2023; Slijepcevic et al., 2024; Riggi et al., 2024; Cecconello et al., 2024; Lastufka et al., 2024). In contrast, semi-supervised learning jointly leverages labeled and unlabeled data during training, typically through consistency regularization or pseudo-labeling techniques, as implemented in methods such as FixMatch (i Rabadán et al., 2022). Semi-supervised applications to radio source classification have shown promising results even with scarce annotations (Slijepcevic et al., 2022; Gupta et al., 2023). Both self- and semi-supervised strategies have achieved high precision and recall (often exceeding 90%) while substantially reducing the need for manual labeling, making them particularly relevant for SKAO-scale surveys, where unlabeled data vastly outnumbers labeled examples.

3.4.3.3 N-Shot Learning

Few-shot and N-shot learning frameworks are designed for scenarios with minimal supervision. Siamese or prototypical networks learn similarity metrics rather than direct classification boundaries, enabling them to generalize from a few examples (Koch et al., 2015; Samudre et al., 2022). Though performance remains somewhat lower than fully supervised CNNs, these methods represent a practical route toward scalable, label-efficient classification of radio sources.

3.4.4 Label Ambiguity and Morphological Variability

A significant challenge for supervised ML classification in radio astronomy is the inherent ambiguity of source labels. Galaxy morphologies do not fall into discrete, well-separated categories; instead, they form a continuum of structures that are often difficult to assign unambiguously to a single class. This ambiguity is further compounded by the strong dependence of observed morphology on frequency, angular resolution, and instrumental sensitivity. A source classified as compact or unresolved in one survey may exhibit extended or multi-component structure in another, while diffuse emission visible at low frequencies may be entirely absent at higher frequencies due to spectral index variations (e.g., Miraghaei and Best, 2017; Mingo et al., 2019b).

These frequency- and resolution-dependent effects present a direct challenge for supervised classification models trained on data from a single survey. For example, a model trained on MeerKAT images may not generalise reliably to LOFAR or ASKAP data without retraining or domain adaptation, as differences in resolution, noise properties, and imaging characteristics can significantly alter source appearance. In addition, the use of human-assigned labels, whether from expert annotation or citizen science efforts such as Radio Galaxy Zoo (Banfield et al., 2015), introduces inter-annotator variability, particularly for ambiguous sources. This can propagate label noise into training datasets and degrade classifier performance.

Addressing these challenges requires careful consideration at both the data and model levels. Probabilistic or soft-label approaches, which assign a distribution over classes rather than a single deterministic label, can better capture intrinsic morphological uncertainty. Constructing multi-survey training datasets spanning a range of frequencies, resolutions, and sensitivities can further improve model generalisation, provided that label consistency is maintained across surveys. In addition, domain adaptation and transfer learning techniques offer promising pathways for adapting models trained on one survey to another without requiring complete retraining (Aniyan and Thorat, 2017). As the SKAO begins delivering data across diverse observing regimes, developing classification frameworks that are robust to such observational and labeling ambiguities will be essential for producing reliable, survey-independent source catalogs.

4 Current Limitations and Challenges

Despite significant progress in automated source detection and classification, several limitations continue to hinder the full exploitation of modern radio survey data. Traditional algorithms often struggle with complex, extended, or low-surface-brightness sources, leading to incomplete or fragmented detections. Conversely, ML and DL models, while more flexible, rely heavily on the quality, realism, and diversity of training datasets. The lack of comprehensive, labeled data that fully capture instrumental effects, calibration errors, and the diversity of astrophysical morphologies remains a major bottleneck.

To mitigate this dependency, emerging approaches such as self-supervised and active learning offer promising alternatives, enabling models to learn from partially labeled or even unlabeled data. These strategies can accelerate morphology discovery and adaptively refine models using human feedback where necessary.

Another key challenge lies in the generalization of trained models across different instruments and survey conditions. Models trained on specific datasets (e.g., LOFAR or ASKAP) often underperform when applied to data from other telescopes with differing resolutions, noise characteristics, or beam patterns. Domain adaptation techniques and standardized data representations are thus crucial to ensure robust transferability across observatories.

Most current frameworks are primarily morphology-based and do not yet exploit the spectral or temporal dimensions of radio data, which carry critical physical information about source evolution and emission mechanisms. Integrating spectral indices, variability, and polarization features into detection pipelines will be essential for a more complete astrophysical characterization.

From a computational standpoint, the petabyte-scale data expected from next-generation facilities such as the SKAO demands HPC- and cloud-ready architectures for scalable, real-time analysis. Efficient parallelization, data streaming, and memory optimization are key requirements, along with robust provenance tracking to ensure reproducibility across distributed infrastructures.

Finally, the next leap in automation requires joint source-finding and classification frameworks capable of performing end-to-end inference, from raw visibility data to labeled catalogs, without disjointed intermediate steps. Working directly in the Fourier domain, however, introduces additional challenges. The uv-plane representation of interferometric data is inherently non-uniform in sampling density, and the mapping between visibilities and sky brightness is non-trivial. Moreover, incomplete uv-coverage leads to correlated noise structures in the image plane.

These factors imply that standard ML architectures, typically designed for regular image grids, cannot be directly applied to visibility data without modification. Instead, this requires either specialised network architectures capable of operating on irregular or graph-based representations, or carefully designed gridding and weighting schemes that preserve the statistical properties of the data (Taran et al., 2023; Drozdova et al., 2024). Community-wide benchmarking initiatives, such as the SKAO data challenges, remain vital for developing consistent performance metrics that balance completeness, reliability, and computational efficiency. Addressing these challenges will be essential for building next-generation, interpretable, and physically grounded radio-source analysis systems. Finally, while the current paradigm in radio astronomy has largely involved adapting ML methods developed in other fields, particularly computer vision and natural language processing, to astronomical data, there is growing potential for this relationship to become reciprocal. Radio survey data possess a number of statistically rich and scientifically distinctive properties: high-dimensional, multimodal feature spaces spanning morphology, spectral index, polarization, and variability; complex, correlated noise structures arising from interferometric sampling; and an enormous dynamic range in source properties across cosmic time.

These characteristics present unique challenges that can drive the development of genuinely novel AI architectures and learning strategies, for example, methods for learning from irregular or sparse data representations, scalable uncertainty, aware inference, and self-supervised learning from physically structured data. As next-generation facilities such as the SKAO begin delivering data of unprecedented volume and complexity, the radio astronomy community is well positioned not only to benefit from advances in the broader AI field, but also to contribute to it, with techniques and

insights that may prove broadly applicable beyond astronomy.

5 Future Directions

The upcoming era of large-scale surveys with facilities such as the SKAO, ngVLA, and LOFAR 2.0 will demand a paradigm shift from isolated source detection tasks to fully integrated, intelligent data analysis platforms. These instruments will produce petabytes of data per day, revealing an unprecedented diversity of radio morphologies, spectral behaviors, and transient phenomena. To harness their full scientific potential, the community must advance toward scalable, interpretable, and physically informed machine learning frameworks.

A key direction is the development of multimodal learning systems that jointly analyze morphological, spectral, and temporal information. By fusing continuum maps with polarization, spectral index, and variability data, these models can capture the complete emission behavior of radio sources, enabling more physically meaningful classification. In particular, temporal-domain learning can facilitate the identification of episodic or recurrent AGN activity, while spectral–spatial networks may help disentangle overlapping sources in crowded fields.

In parallel, self-supervised, active, and transfer learning strategies will play a crucial role in reducing the dependence on manually labeled datasets. These approaches can leverage the vast quantities of unlabeled SKAO data, using generative or contrastive objectives to learn representations that generalize across instruments and observing conditions. Human-in-the-loop frameworks, where experts guide the model through targeted feedback, will further accelerate the discovery of new or rare morphological classes.

Future pipelines will likely evolve into end-to-end, cloud-native systems that integrate calibration, source finding, and classification within a unified workflow, deployable on HPC or distributed infrastructures for near-real-time analysis. Explainable AI (XAI) will be equally critical: interpretable models that highlight physically meaningful features—such as jet curvature, lobe symmetry, or spectral steepness, will build scientific trust and enable theory validation. In parallel, adherence to FAIR data principles will ensure interoperability across international facilities.

Ultimately, the synergy between next-generation surveys, physically motivated ML architectures, and collaborative open-science infrastructures will define the next decade of radio astronomy. This integrated approach will not only improve the completeness and reliability of radio-source catalogs but also open new pathways for uncovering the physical processes driving galaxy evolution across cosmic time.

5.1 Integrated Detection and Classification Pipelines

With the advent of large-scale sky surveys, the need for end-to-end automated pipelines that integrate source detection, characterization, and classification has become essential. Such integrated frameworks minimize human intervention and ensure consistent processing of large datasets while maintaining high accuracy and reproducibility. These pipelines combine multiple stages—calibration, source finding, morphological characterization, cross-identification, and classification—into a unified workflow.

Modern implementations exemplify this integration by coupling source detection algorithms with data quality assessment and catalog generation. Notable examples include the ASKAP pipeline Selavy (Whiting, 2012), the LOFAR PyBDSF-based framework (Mohan and Rafferty, 2015), and the MeerKAT Oxkat and CARACal systems (Ramaila et al., 2024; Józsa et al., 2020). These systems are often designed to operate within large-scale data processing environments such as the SKA Regional Centres (SRCs), supporting distributed computation and standardized metadata handling. The SRC network will play a central role in enabling these workflows by providing standardized execution environments, scalable computing resources, and access to common analysis services across the international SKAO community.

In the future, the combination of classical methods with machine-learning techniques is likely to become increasingly common in modern radio surveys. For example, Alegre et al. (2022) used machine-learning techniques to assess the identification of host galaxies for LOFAR radio sources, while Silima et al. (2025) tested whether supervised ML algorithms, including Random Forest and XGBoost, can reproduce source classifications into AGN and star-forming galaxies compared to the traditional diagnostic approach of Whittam et al. (2022). A hybrid approach is also planned for source association within the EMU survey (Vardoulaki et al., 2025). Such hybrid architectures, in which rule-based methods address well-constrained tasks such as noise estimation and Gaussian component fitting while ML modules operate on higher-level inference and morphologically ambiguous cases, provide a practical and scalable template for SKAO-era pipelines, where the volume, complexity, and diversity of the detected source population will fundamentally exceed the capabilities of purely deterministic approaches.

6 Implications for SKAO Science Goals

The integration of advanced machine learning and automated source-finding pipelines has direct and far-reaching implications for the scientific objectives of the SKAO. As the SKAO aims to map billions of radio sources across cosmic time, the ability to detect, classify, and interpret these sources efficiently and accurately is central to its key science drivers, ranging from understanding galaxy evolution and cosmic magnetism to probing the cosmic dawn.

Accurate morphological classification of radio galaxies, including Fanaroff–Riley (FR) types, HyMoRSs, DDRGs, and GRGs, will enable a more complete census of AGN activity and feedback mechanisms. By tracing jet–environment interactions across redshift, SKAO datasets will shed light on how radio-mode feedback regulates star formation and black hole growth in galaxies. Automated pipelines that can identify these diverse morphologies at scale are therefore essential for linking radio structures to their host galaxy and environmental properties.

Furthermore, the improved recovery of diffuse and low-surface-brightness emission will directly benefit studies of cluster physics and large-scale structure formation. Enhanced sensitivity to bent-tailed sources and radio relics can provide new probes of intra-cluster medium dynamics, shock fronts, and magnetization processes in the cosmic web. Deep-learning–based extraction methods, capable of distinguishing faint extended features from noise, are thus crucial for advancing the SKAO's cosmic magnetism and large-scale structure science programs.

The SKAO will also explore transient and time-variable radio phenomena, ranging from episodic AGN activity to explosive events such as fast radio bursts (FRBs) and supernovae. Machine-learning frameworks that incorporate temporal and spectral dimensions will enable real-time detection and classification of these events, facilitating rapid multiwavelength follow-up and cross-correlation with optical, X-ray, and gravitational-wave observatories.

In summary, the advancement of automated, interpretable, and physically informed source detection and classification systems is not merely a technical enhancement but a scientific necessity for achieving the fundamental goals of the SKAO. By bridging data-driven methods with astrophysical insight and embedding these within open and interoperable infrastructures, such frameworks will enable the SKAO to move beyond simple source cataloging toward uncovering the physical mechanisms that govern the radio Universe.

7 Conclusions

The evolution of radio source finding and classification has mirrored the technological transformation of radio astronomy itself—from the early days of manual cataloging to the present era of data-intensive, automated, and intelligent pipelines. Classical algorithms, though foundational, were designed for datasets of limited scale and complexity. As surveys expanded in sensitivity and coverage through instruments such as LOFAR, MeerKAT, ASKAP, and the uGMRT, it became evident that traditional thresholding and Gaussian-fitting approaches could no longer meet the demands of modern data volumes or the diversity of radio source morphologies.

Machine learning and deep learning methods have emerged as effective tools to address these challenges, offering the ability to learn complex, hierarchical features directly from data. Architectures such as convolutional neural networks, U-Nets, and object-detection frameworks have demonstrated exceptional capability in identifying compact, extended, and diffuse emission structures across a wide range of signal-to-noise regimes. When integrated into scalable, HPC environments, these approaches promise to deliver real-time, reproducible, and scientifically reliable catalogues from the vast imaging data expected from the SKAO and its precursors.

The future lies in end-to-end, unified frameworks that jointly perform detection, classification, and characterization of sources, leveraging self-supervised learning, active learning, and transfer learning to adapt to heterogeneous data domains. Such systems must not only optimize completeness and reliability but also incorporate physical interpretability, ensuring that data-driven classifications retain astrophysical meaning.

Ultimately, the development of robust and intelligent source-finding pipelines is not merely a computational challenge but a scientific imperative. Accurate and scalable detection and classification underpin nearly every major SKA) science goal, from tracing the cosmic evolution of AGN and star formation to mapping magnetic fields and large-scale structure. The synergy between AI-driven methods, rigorous physical modeling, and high-performance computing will define the next generation of radio astronomy, unlocking the full scientific potential of the SKAO to probe the radio Universe.

Acknowledgements

We thank Ivy Wong for her valuable insights on H I source-finding techniques. Omkar Bait was supported by the National Science Foundation under Cooperative Agreement 2421782 and the Simons Foundation grant MPS-AI-00010515 awarded to the NSF-Simons AI Institute for Cosmic Origins — CosmicAI, https://www.cosmicai.org/. GL, AA, and JM acknowledge financial support from the Spanish grant PID2023-147883NB-C21, funded by MCIU/AEI/ 10.13039/501100011033, as well as support through ERDF/EU and from the Severo Ochoa grant CEX2021-001131-S funded by MCIN/AEI/ 10.13039/501100011033.

References

J. Akeret, C. Chang, A. Lucchi, and A. Refregier. *Astronomy and Computing*, 18:35–39, Jan. 2017. doi: 10.1016/j.ascom.2017.01.002.

L. Alegre et al. *Monthly Notices of the Royal Astronomical Society*, 516(4):4716–4738, Nov. 2022. doi: 10.1093/mnras/stac1888.

M. J. Alger et al. *Monthly Notices of the Royal Astronomical Society*, 478(4):5547, Aug. 2018. doi: 10.1093/mnras/sty1308.

W. Alhassan, A. R. Taylor, and M. Vaccari. *Monthly Notices of the Royal Astronomical Society*, 480 (2):2085–2093, Oct. 2018. doi: 10.1093/mnras/sty2038.

J. R. Allison, E. M. Sadler, and M. T. Whiting. *Publications of the Astron. Soc. of Australia*, 29(3): 221–228, Oct. 2012. doi: 10.1071/AS11040.

T. M. O. AMI Consortium: Franzen et al. *MNRAS*, 415:2699, 2011. doi: https://doi.org/10.1111/j.1365-2966.2011.18887.x.

R. Andrae, K. Jahnke, and P. Melchior. *Monthly Notices of the Royal Astronomical Society*, 411(1): 385–401, Feb. 2011. doi: 10.1111/j.1365-2966.2010.17690.x.

A. K. Aniyan and K. Thorat. *Astrophysical Journal, Supplement*, 230(2):20, June 2017. doi: 10.3847/1538-4365/aa7333.

J. K. Banfield et al. *Monthly Notices of the Royal Astronomical Society*, 453(3):2326, Nov. 2015. doi: 10.1093/mnras/stv1688.

L. Barcos-Muñoz et al. *Astrophysical Journal*, 799(1):10, Jan. 2015. doi: 10.1088/0004-637X/799/1/10.

J. A. Barkai, M. A. W. Verheijen, E. Talavera, and M. H. F. Wilkinson. *Astronomy and Astrophysics*, 670:A55, Feb. 2023. doi: 10.1051/0004-6361/202244708.

B. Becker, M. Vaccari, M. Prescott, and T. Grobler. *Monthly Notices of the Royal Astronomical Society*, 503(2):1828–1846, May 2021. doi: 10.1093/mnras/stab325.

R. H. Becker, R. L. White, and D. J. Helfand. *Astrophysical Journal*, 450:559, Sept. 1995. doi: 10.1086/176166.

S. Bera, S. Pal, T. K. Sasmal, and S. Mondal. *Astrophysical Journal, Supplement*, 251(1):9, Nov. 2020. doi: 10.3847/1538-4365/abb367.

E. Bertin and S. Arnouts. *Astronomy and Astrophysics Supplement Series*, 117:393–404, June 1996. doi: 10.1051/aas:1996164.

E. Bertin and S. Arnouts. *A&AS*, 117:393, 1996. doi: https://doi.org/10.1051/aas:1996164.

N. Bhukta, S. K. Mondal, and S. Pal. *Monthly Notices of the Royal Astronomical Society*, 516(1): 372–390, Oct. 2022a. doi: 10.1093/mnras/stac2001.

N. Bhukta, S. Pal, and S. K. Mondal. *Monthly Notices of the Royal Astronomical Society*, 512(3): 4308–4323, May 2022b. doi: 10.1093/mnras/stac447.

N. Bhukta, S. Manik, S. Pal, and S. K. Mondal. *ApJS*, 273:2, 2024. doi: 10.3847/1538-4365/ad5184.

S. Blyth et al. In *MeerKAT Science: On the Pathway to the SKA*, page 4, Feb. 2018. doi: 10.22323/1.277.0004.

A. Bonaldi et al. *MNRAS*, 500:3821, 2021. doi: https://doi.org/10.1093/mnras/staa3023.

M. Bowles et al. *Monthly Notices of the Royal Astronomical Society*, 501(3):4579–4595, Mar. 2021. doi: 10.1093/mnras/staa3946.

J. Bradbury et al. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

M. A. Branch. *Inexact Reflective Newton Methods for Large-Scale Optimization Subject to Bound Constraints*. PhD thesis, Cornell University, 1996.

K. Brand et al. *Monthly Notices of the Royal Astronomical Society*, 522(1):292–311, June 2023. doi: 10.1093/mnras/stad989.

C. J. Burke et al. *Monthly Notices of the Royal Astronomical Society*, 490(3):3952–3965, Dec. 2019. doi: 10.1093/mnras/stz2845.

N. Caon, M. Capaccioli, and M. D'Onofrio. *Monthly Notices of the Royal Astronomical Society*, 265:1013–1021, Dec. 1993. doi: 10.1093/mnras/265.4.1013.

D. Carbone et al. *Astronomy and Computing*, 23:92, Apr. 2018. doi: 10.1016/j.ascom.2018.02.003.

T. Cecconello et al. *arXiv e-prints*, art. arXiv:2411.14078, Nov. 2024. doi: 10.48550/arXiv.2411.14078.

K. K. L. Charlton et al. *Monthly Notices of the Royal Astronomical Society*, 537(1):272–284, Feb. 2025. doi: 10.1093/mnras/stae2543.

T. Cohen and M. Welling. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

J. J. Condon. *PASP*, 109(732):166, feb 1997. doi: 10.1086/133871. URL https://dx.doi.org/10.1086/133871.

J. J. Condon et al. *Astronomical Journal*, 115(5):1693–1716, May 1998. doi: 10.1086/300337.

C. J. Conselice, M. A. Bershady, and A. Jangren. *Astrophysical Journal*, 529:886–910, Feb. 2000. doi: 10.1086/308300.

D. Cornu et al. *Astronomy and Astrophysics*, 690:A211, Oct. 2024. doi: 10.1051/0004-6361/202449548.

P. Dabhade et al. *Astronomy and Astrophysics*, 635:A5, Mar. 2020. doi: 10.1051/0004-6361/201935589.

A. M. Darya, I. Fernini, M. Vellasco, and A. Hussain. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023.

J. Delhaize et al. *Monthly Notices of the Royal Astronomical Society*, 501(3):3833–3845, Mar. 2021. doi: 10.1093/mnras/staa3837.

M. Drozdova et al. *Astronomy and Astrophysics*, 683:A105, Mar. 2024. doi: 10.1051/0004-6361/202347948.

B. L. Fanaroff and J. M. Riley. *Monthly Notices of the Royal Astronomical Society*, 167:31P–36P,

May 1974. doi: 10.1093/mnras/167.1.31P.

R. Fender et al. *PoS*, 2007. doi: 10.48550/arXiv.0805.4349.

M. P. Gawroński, A. Marecki, M. Kunert-Bajraszewska, and A. J. Kus. *Astronomy and Astrophysics*, 447(1):63–70, Feb. 2006. doi: 10.1051/0004-6361:20053996.

C. Gheller, F. Vazza, and A. Bonafede. *MNRAS*, 480(3):3749–3761, Nov. 2018. doi: 10.1093/mnras/sty2102.

Gopal-Krishna and P. J. Wiita. *Astronomy and Astrophysics*, 363:507–516, Nov. 2000. doi: 10.48550/arXiv.astro-ph/0009441.

N. Gupta et al. *Publications of the Astron. Soc. of Australia*, 40:e044, Sept. 2023. doi: 10.1017/pasa.2023.46.

N. Gupta et al. *Publications of the Astron. Soc. of Australia*, 41:e001, Jan. 2024. doi: 10.1017/pasa.2023.64.

N. Gupta et al. *Astronomy and Astrophysics*, 698:A120, June 2025. doi: 10.1051/0004-6361/202452407.

H. Håkansson et al. *Astronomy & Astrophysics*, 671:A39, 2023.

C. L. Hale et al. *Monthly Notices of the Royal Astronomical Society*, 487(3):3971–3989, Aug. 2019. doi: 10.1093/mnras/stz1462.

C. L. Hale et al. *Monthly Notices of the Royal Astronomical Society*, 536(3):2187–2211, Jan. 2025. doi: 10.1093/mnras/stae2528.

C. A. Hales et al. *Monthly Notices of the Royal Astronomical Society*, 425(2):979–996, Sept. 2012. doi: 10.1111/j.1365-2966.2012.21373.x.

P. J. Hancock et al. *mnras*, 422(2):1812–1824, May 2012. doi: 10.1111/j.1365-2966.2012.20768.x.

P. J. Hancock, C. M. Trott, and N. Hurley-Walker. *pasa*, 35:e011, Mar. 2018. doi: 10.1017/pasa.2018.3.

P. Hartley et al. *Monthly Notices of the Royal Astronomical Society*, 523(2):1967–1993, Aug. 2023. doi: 10.1093/mnras/stad1375.

J. J. Harwood, T. Vernstrom, and A. Stroe. *Monthly Notices of the Royal Astronomical Society*, 491 (1):803–822, Jan. 2020. doi: 10.1093/mnras/stz3069.

B. Häussler et al. *Astrophysical Journal, Supplement*, 172(2):615–633, Oct. 2007. doi: 10.1086/518836.

K. He, G. Gkioxari, P. Dollár, and R. Girshick. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

J. A. Hodge et al. *Astrophysical Journal*, 876(2):130, May 2019. doi: 10.3847/1538-4357/ab1846.

A. M. Hopkins et al. *pasa*, 32:e037, Oct. 2015. doi: 10.1017/pasa.2015.37.

M. S. Hossain et al. *Procedia Computer Science*, 222:601–612, Jan. 2023. doi: 10.1016/j.procs.2023.08.198.

A. Hota et al. In *Advancing Astrophysics with the SKA – II (AASKAII)*. 2026. arXiv search: Report number AASKAII/Hota01.

G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. *arXiv e-prints*, art. arXiv:1608.06993, Aug. 2016. doi: 10.48550/arXiv.1608.06993.

M. T. Huynh et al. *PASA*, 29:229, 2012. doi: https://doi.org/10.1071/AS11026.

M. M. i Rabadán, A. Pieropan, H. Azizpour, and A. Maki. *ArXiv*, abs/2210.09919, 2022.

C. H. Ishwara-Chandra and D. J. Saikia. *Monthly Notices of the Royal Astronomical Society*, 309(1):

100–112, Oct. 1999. doi: 10.1046/j.1365-8711.1999.02835.x.

M. Jarvis et al. In *MeerKAT Science: On the Pathway to the SKA*, page 6, Jan. 2016. doi: 10.22323/1.277.0006.

G. I. G. Józsa et al. In R. Pizzo et al., editors, *Astronomical Data Analysis Software and Systems XXIX*, volume 527 of *Astronomical Society of the Pacific Conference Series*, page 635, Jan. 2020. doi: 10.48550/arXiv.2006.02955.

A. D. Kapińska et al. *Astronomical Journal*, 154(6):253, Dec. 2017. doi: 10.3847/1538-3881/aa90b7.

G. Koch, R. Zemel, R. Salakhutdinov, et al. In *ICML deep learning workshop*, volume 2, pages 1–30. Lille, 2015.

C. Konar and M. J. Hardcastle. *Monthly Notices of the Royal Astronomical Society*, 436(2):1595–1614, Dec. 2013. doi: 10.1093/mnras/stt1676.

B. S. Koribalski et al. *Astrophysics and Space Science*, 365(7):118, July 2020. doi: 10.1007/s10509-020-03831-4.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. *Advances in neural information processing systems*, 25, 2012.

R. G. Kron. *ApJS*, 43:305, 1980. doi: https://doi.org/10.1086/190669.

S. Kumari and S. Pal. *Monthly Notices of the Royal Astronomical Society*, 514(3):4290–4299, Aug. 2022. doi: 10.1093/mnras/stac1215.

S. Kumari and S. Pal. *Monthly Notices of the Royal Astronomical Society*, 527(4):11233–11239, Dec. 2023. ISSN 1365-2966. doi: 10.1093/mnras/stad3953. URL http://dx.doi.org/10.1093/mnras/stad3953.

S. Kumari and S. Pal. *Astronomy amp; Astrophysics*, 683:A175, Mar. 2024. ISSN 1432-0746. doi: 10.1051/0004-6361/202347816. URL http://dx.doi.org/10.1051/0004-6361/202347816.

S. Kumari, S. Pal, M. J. Hardcastle, and M. A. Horton. *Astronomy amp; Astrophysics*, 689:A301, 2024. ISSN 1432-0746. doi: 10.1051/0004-6361/202347367. URL http://dx.doi.org/10.1051/0004-6361/202347367.

S. Kumari, S. Pal, and S. Manik. *ApJ*, 987(1):10, July 2025. doi: 10.3847/1538-4357/add69d.

S. Kumari, S. Pal, S. Paul, and M. Jamrozy. *Monthly Notices of the Royal Astronomical Society*, 545 (4), Jan. 2026. ISSN 1365-2966. doi: 10.1093/mnras/staf2038. URL http://dx.doi.org/10.1093/mnras/staf2038.

B. Lao et al. *Science Bulletin*, 66(21):2145, Nov. 2021. doi: 10.1016/j.scib.2021.07.015.

B. Lao et al. *Astronomy and Computing*, 44:100728, July 2023. doi: 10.1016/j.ascom.2023.100728.

E. Lastufka et al. *Astronomy and Astrophysics*, 690:A310, Oct. 2024. doi: 10.1051/0004-6361/202449964.

H. Liu et al. *Computers and Geotechnics*, 125:103689, 2020. doi: 10.1016/j.compgeo.2020.103689.

M. Lochner et al. *Monthly Notices of the Royal Astronomical Society*, 520(1):1439–1446, Mar. 2023. doi: 10.1093/mnras/stad074.

J. M. Lotz, J. Primack, and P. Madau. *Astronomical Journal*, 128(1):163–182, July 2004. doi: 10.1086/421849.

J. M. Lotz et al. *Astrophysical Journal*, 672(1):177–197, Jan. 2008. doi: 10.1086/523659.

M. Lourakis. levmar: Levenberg-marquardt nonlinear least squares algorithms in C/C++. http://www.ics.forth.gr/~lourakis/levmar/, 2004. [Accessed on 31 Jan. 2005].

G. Lucatelli et al. *Monthly Notices of the Royal Astronomical Society*, 529(4):4468–4499, Apr. 2024.

doi: 10.1093/mnras/stae744.

V. Lukic et al. *Monthly Notices of the Royal Astronomical Society*, 476(1):246–260, May 2018. doi: 10.1093/mnras/sty163.

V. Lukic et al. *Monthly Notices of the Royal Astronomical Society*, 487(2):1729–1744, Aug. 2019a. doi: 10.1093/mnras/stz1289.

V. Lukic, F. de Gasperin, and M. Brüggen. *Galaxies*, 8(1):3, Dec. 2019b. doi: 10.3390/galaxies8010003.

M. López-Caniego and P. Vielva. *MNRAS*, 421:2139, 2012. doi: https://doi.org/10.1111/j.1365-2966.2012.20444.x.

M. López-Caniego et al. *MNRAS*, 370:2047, 2006. doi: https://doi.org/10.1111/j.1365-2966.2006.10639.x.

Z. Ma et al. *Astrophysical Journal, Supplement*, 240(2):34, Feb. 2019. doi: 10.3847/1538-4365/aaf9a2.

D. Magro et al. Asgard: A deep neural network for the detection of compact sources, extended galaxies, and sidelobes in radio astronomical maps, 2022. In preparation.

D. Makovoz and F. R. Marleau. *PASP*, 117:1113, 2005. doi: https://doi.org/10.1086/432977.

S. Manik, N. Bhukta, S. Pal, and S. K. Mondal. *Astrophysical Journal, Supplement*, 2025. doi: 10.3847/1538-4365/ae0b52.

S. Manik et al. *The Astrophysical Journal*, 997(2):157, Jan. 2026. ISSN 1538-4357. doi: 10.3847/1538-4357/ae1f85. URL http://dx.doi.org/10.3847/1538-4357/ae1f85.

M. Y. Mao et al. *Monthly Notices of the Royal Astronomical Society*, 426(4):3334–3348, 2012. doi: 10.1111/j.1365-2966.2012.21965.x.

B. Mingo et al. *Monthly Notices of the Royal Astronomical Society*, 488(2):2701–2721, Sept. 2019a. doi: 10.1093/mnras/stz1901.

B. Mingo et al. *Monthly Notices of the Royal Astronomical Society*, 488(2):2701–2721, Sept. 2019b. doi: 10.1093/mnras/stz1901.

H. Miraghaei and P. N. Best. *Monthly Notices of the Royal Astronomical Society*, 466(4):4346–4363, Apr. 2017. doi: 10.1093/mnras/stx007.

V. Missaglia et al. *Astronomy and Astrophysics*, 626:A8, June 2019. doi: 10.1051/0004-6361/201935058.

N. Mohan and D. Rafferty. PyBDSF: Python Blob Detection and Source Finder. Astrophysics Source Code Library, record ascl:1502.007, Feb. 2015.

S. Molinari et al. *Astronomy and Astrophysics*, 530:A133, June 2011. doi: 10.1051/0004-6361/201014752.

R. P. Norris et al. *The Astronomical Journal*, 132(6):2409–2423, 2006. doi: 10.1086/508831.

R. P. Norris et al. *Publications of the Astron. Soc. of Australia*, 38:e003, Jan. 2021a. doi: 10.1017/pasa.2020.52.

R. P. Norris et al. *Publications of the Astron. Soc. of Australia*, 38:e046, Sept. 2021b. doi: 10.1017/pasa.2021.42.

K. Ntwaetsile and J. E. Geach. *Monthly Notices of the Royal Astronomical Society*, 502(3):3417–3425, 2021.

M. S. S. L. Oei et al. *Nature*, 633(8030):537–541, Sept. 2024. doi: 10.1038/s41586-024-07879-y.

S. Pal and S. Kumari. *Journal of Astrophysics and Astronomy*, 44(1), Feb. 2023. ISSN 0973-7758. doi:

10.1007/s12036-022-09892-x. URL http://dx.doi.org/10.1007/s12036-022-09892-x.

S. Pal et al. In *Advancing Astrophysics with the SKA – II (AASKAII)*. 2026. arXiv search: Report number AASKAII/SabyasachiPal02.

Y.-j. Peng et al. *Astrophysical Journal*, 721(1):193–221, Sept. 2010. doi: 10.1088/0004-637X/721/1/193.

V. Petrosian. *Astrophys. J. Lett.*, 210:L53, 1976. doi: https://doi.org/10.1086/182301.

C. Pino et al. In *International Workshop on Artificial Intelligence and Pattern Recognition*, pages 393–403. Springer, 2021.

K. L. Polsterer, F. Gieseke, and B. Doser. *Astrophysics Source Code Library*, pages ascl–1910, 2019.

A. Ramaila et al. In *32nd General Assembly International Union (IAUGA 2024)*, page 1735, Aug. 2024.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. doi: 10.1109/CVPR.2016.91.

S. Riggi et al. *Monthly Notices of the Royal Astronomical Society*, 460(2):1486–1499, Aug. 2016. doi: 10.1093/mnras/stw982.

S. Riggi et al. *Publications of the Astron. Soc. of Australia*, 36:e037, Oct. 2019a. doi: 10.1017/pasa.2019.29.

S. Riggi et al. *Publications of the Astron. Soc. of Australia*, 36:e037, Oct. 2019b. doi: 10.1017/pasa.2019.29.

S. Riggi et al. *Astronomy and Computing*, 42:100682, Jan. 2023. doi: 10.1016/j.ascom.2022.100682.

S. Riggi et al. *Publications of the Astron. Soc. of Australia*, 41:e085, Nov. 2024. doi: 10.1017/pasa.2024.84.

A. S. G. Robotham et al. *Monthly Notices of the Royal Astronomical Society*, 476(3):3137–3159, May 2018. doi: 10.1093/mnras/sty440.

O. Ronneberger, P. Fischer, and T. Brox. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

L. Rudnick. *Galaxies*, 9(4):85, Oct. 2021. doi: 10.3390/galaxies9040085.

L. Rustige et al. *RAS Techniques and Instruments*, 2(1):264–277, Jan. 2023. doi: 10.1093/rasti/rzad016.

M. Sadeghi, M. Javaherian, and H. Miraghaei. *The Astronomical Journal*, 161(2):94, 2021.

A. Samudre, L. T. George, M. Bansal, and Y. Wadadekar. *Monthly Notices of the Royal Astronomical Society*, 509(2):2269–2280, Jan. 2022. doi: 10.1093/mnras/stab3144.

A. Samudre, L. T. George, M. Bansal, and Y. Wadadekar. *Monthly Notices of the Royal Astronomical Society*, 509(2):2269–2280, 2022.

A. M. M. Scaife and F. Porter. *Monthly Notices of the Royal Astronomical Society*, 503(2):2369–2379, May 2021. doi: 10.1093/mnras/stab530.

P. Serra et al. *mnras*, 448(2):1922–1929, Apr. 2015. doi: 10.1093/mnras/stv079.

J. L. Sérsic. *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina*, 6:41–43, Feb. 1963.

T. W. Shimwell et al. *Astronomy and Astrophysics*, 598:A104, Feb. 2017. doi: 10.1051/0004-6361/201629313.

T. W. Shimwell et al. *Astronomy and Astrophysics*, 659:A1, Mar. 2022. doi: 10.1051/0004-6361/202142484.

W. Silima et al. *Monthly Notices of the Royal Astronomical Society*, 544(1):799–814, Nov. 2025. doi: 10.1093/mnras/staf1698.

L. Simard et al. *Astrophysical Journal, Supplement*, 142(1):1–33, Sept. 2002. doi: 10.1086/341399.

I. V. Slijepcevic et al. *Monthly Notices of the Royal Astronomical Society*, 514(2):2599–2613, Aug. 2022. doi: 10.1093/mnras/stac1135.

I. V. Slijepcevic et al. *RAS Techniques and Instruments*, 3(1):19–32, Jan. 2024. doi: 10.1093/rasti/rzad055.

Y. Song et al. *Astrophysical Journal*, 916(2):73, Aug. 2021. doi: 10.3847/1538-4357/ac05c2.

J. N. Spreeuw. *Dissertation, Astronomical Institute Anton Pannekoek, University of Amsterdam*, ISBN 978-90-9024055-8, 2010.

H. Stewart et al. *RAS Techniques and Instruments*, 3(1):315–332, Jan. 2024. doi: 10.1093/rasti/rzae019.

J. D. Swinbank et al. *A&C*, 11:25, 2015. doi: https://doi.org/10.1016/j.ascom.2015.03.002.

C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

H. Tang, A. M. M. Scaife, and J. P. Leahy. *Monthly Notices of the Royal Astronomical Society*, 488 (3):3358–3375, Sept. 2019. doi: 10.1093/mnras/stz1883.

H. Tang, A. M. M. Scaife, O. I. Wong, and S. S. Shabala. *Monthly Notices of the Royal Astronomical Society*, 510(3):4504–4524, Mar. 2022. doi: 10.1093/mnras/stab3553.

O. Taran et al. *Astronomy and Astrophysics*, 674:A161, June 2023. doi: 10.1051/0004-6361/202245778.

A. Vafaei Sadr et al. *mnras*, 484(2):2793–2806, Apr. 2019. doi: 10.1093/mnras/stz131.

M. P. van Haarlem et al. *A&A*, 556:A2, 2013. doi: https://doi.org/10.1051/0004-6361/201220873.

E. Vardoulaki et al. *arXiv e-prints*, art. arXiv:2509.19787, Sept. 2025. doi: 10.48550/arXiv.2509.19787.

L. Wang et al. *Publications of the Astron. Soc. of Australia*, 42:e033, Jan. 2025. doi: 10.1017/pasa.2025.14.

T. Westmeier et al. *mnras*, 506(3):3962–3976, Sept. 2021. doi: 10.1093/mnras/stab1881.

T. Westmeier et al. *Publications of the Astronomical Society of Australia*, 39:e058, 2022. doi: 10.1017/pasa.2022.50.

R. L. White, R. H. Becker, D. J. Helfand, and M. D. Gregg. *Astrophysical Journal*, 475(2):479, Feb. 1997. doi: https://doi.org/10.1086/303564.

M. Whiting and B. Humphreys. *PASA*, 29:371, 2012. doi: https://doi.org/10.1071/AS12028.

M. T. Whiting. *Monthly Notices of the Royal Astronomical Society*, 421(4):3242–3256, Apr. 2012. doi: 10.1111/j.1365-2966.2012.20548.x.

I. H. Whittam et al. *Monthly Notices of the Royal Astronomical Society*, 516(1):245–263, Oct. 2022. doi: 10.1093/mnras/stac2140.

E. L. Wright et al. *The Astronomical Journal*, 140(6):1868–1881, 2010. doi: 10.1088/0004-6256/140/6/1868.

C. Wu et al. *Monthly Notices of the Royal Astronomical Society*, 482(1):1211–1230, Jan. 2019a. doi: 10.1093/mnras/sty2646.

C. Wu et al. *Monthly Notices of the Royal Astronomical Society*, 482(1):1211–1230, 2019b. doi: 10.1093/mnras/sty2646.

X. Yang et al. *Astrophysical Journal, Supplement*, 245(1):17, Nov. 2019. doi: 10.3847/1538-4365/ab4811.

H. Yoon et al. *Publications of the Astron. Soc. of Australia*, 42:e088, June 2025. doi: 10.1017/pasa.2025.10046.