

University of Groningen

## Deep learning in high angular-resolution radio interferometry

Rezaei Badafshani, Samira

DOI:  
[10.33612/diss.222496441](https://doi.org/10.33612/diss.222496441)

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*  
Publisher's PDF, also known as Version of record

*Publication date:*  
2022

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*  
Rezaei Badafshani, S. (2022). *Deep learning in high angular-resolution radio interferometry*. University of Groningen. <https://doi.org/10.33612/diss.222496441>

### Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

### Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.



**rijksuniversiteit  
 groningen**

# **Deep learning in high angular-resolution radio interferometry**

## **Proefschrift**

ter verkrijging van de graad van doctor aan de  
 Rijksuniversiteit Groningen  
 op gezag van de  
 rector magnificus prof. dr. C. Wijmenga  
 en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

maandag 27 juni 2022 om 11:00 uur

door

**Samira Rezaei Badafshani**

geboren op 17 juli 1988  
 te Esfahan



**Promotores**

Prof. dr. J. P. McKean

Prof. dr. M. Biehl

**Beoordelingscommissie**

Prof. dr. M. Brüggem

Prof. dr. K. Bunte

Prof. dr. R. Deane

To all the strong women who fight for equality.



**university of  
groningen**

faculty of mathematics  
and natural sciences

kapteyn astronomical  
institute

**Printed by:** Proefschriftspecialist.

**Cover design:** A modified version of a poster by Dan Matutina. The modifications are applied by Mohamadreza Hasani.

The work contained in this thesis is based on research performed within the Data Science and Systems Complexity (DSSC) Doctoral Training Programme, co-funded through a Marie Skłodowska-Curie COFUND (DSSC 754315).

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Radio Interferometry . . . . .	5
1.2.1	Basic principles of radio astronomy . . . . .	5
1.2.2	Very Long Baseline Array . . . . .	7
1.2.3	LOW Frequency ARray (LOFAR) . . . . .	9
1.3	Gravitational Lensing . . . . .	11
1.3.1	Basic formalism of gravitational lensing . . . . .	12
1.3.2	Strong lensing applications . . . . .	14
1.3.3	Strong lens modelling . . . . .	15
1.4	Supervised Learning . . . . .	15
1.4.1	K Nearest Neighbours (KNN) . . . . .	16
1.4.2	Random forest . . . . .	16
1.4.3	Artificial Neural Network (ANN) . . . . .	17
1.5	Deep Learning Techniques . . . . .	19
1.5.1	Convolutional Neural Network (CNN) . . . . .	19
1.5.2	Convolutional autoencoders . . . . .	22
1.6	Visualization techniques . . . . .	22
1.6.1	t-Distributed stochastic neighbour embedding . . . . .	24
1.7	Scope and aim of this thesis . . . . .	24
<b>2</b>	<b>Source counts of VLBI-detected radio sources</b>	<b>27</b>
2.1	Introduction . . . . .	28
2.2	The mJIVE–20 survey final catalogue . . . . .	30
2.2.1	Overview of the mJIVE–20 survey . . . . .	30
2.2.2	Detection rate as a function of radio source surface brightness, compactness and size . . . . .	31
2.3	VLBI-detected radio source counts . . . . .	38
2.3.1	The mJIVE–20 survey sky area . . . . .	38
2.3.2	The mJIVE–20 survey completeness . . . . .	39
2.3.3	A note on the resolution bias of the parent and mJIVE–20 population samples . . . . .	42

2.3.4	Euclidean-normalized and differential source counts . . . . .	43
2.4	Prospects for all-sky VLBI surveys . . . . .	47
2.4.1	Characteristics of the arrays . . . . .	47
2.4.2	Prospects for in-beam calibration . . . . .	49
2.4.3	Expected number of detected sources . . . . .	51
2.5	Conclusions . . . . .	53
<b>3</b>	<b>Point and extended source detection and characterization</b>	<b>55</b>
3.1	Introduction . . . . .	56
3.2	Method . . . . .	59
3.2.1	Simulating a representative training and testing dataset . . . . .	59
3.2.2	Overview of DECORAS . . . . .	64
3.2.3	Preprocessing . . . . .	67
3.2.4	Network structure . . . . .	67
3.2.5	Loss function . . . . .	70
3.2.6	Post processing blob detection . . . . .	71
3.2.7	Surface brightness estimator . . . . .	71
3.3	Source detection . . . . .	75
3.3.1	Defining the true positive and true negative rates . . . . .	75
3.3.2	The performance of BLOBCAT . . . . .	77
3.3.3	Comparing the performance of BCE and MSLE . . . . .	81
3.3.4	DECORAS source detection strategy . . . . .	81
3.3.5	Comparing DECORAS with a traditional source detection algorithm . . . . .	85
3.4	Source characterization . . . . .	85
3.4.1	Recovering the source position . . . . .	88
3.4.2	Recovering the source structure . . . . .	88
3.4.3	Recovering the source peak surface brightness . . . . .	90
3.5	Discussion and Conclusions . . . . .	95
<b>4</b>	<b>Lens detection with the ILT</b>	<b>99</b>
4.1	Introduction . . . . .	100
4.2	Method . . . . .	103
4.2.1	Simulating a training dataset for the ILT . . . . .	103
4.2.2	Network structure . . . . .	106
4.2.3	Preprocessing . . . . .	111
4.2.4	Loss function . . . . .	111
4.2.5	Detection probability uncertainty . . . . .	115
4.2.6	Evaluation criteria . . . . .	116
4.3	Network tests . . . . .	117
4.3.1	Determining the lens probability . . . . .	118
4.3.2	Test on a realistic lensing population . . . . .	118

4.3.3	Test on a lens population of uniformly selected model parameters . . . . .	119
4.3.4	Non-lensed double-lobed radio sources . . . . .	122
4.3.5	Prediction uncertainties . . . . .	126
4.4	Lens detection with the ILT . . . . .	128
4.4.1	Results for the final network test . . . . .	128
4.4.2	Parameter-space of a lens survey with the ILT . . . . .	129
4.5	Conclusions . . . . .	133
<b>5</b>	<b>Conclusions and future prospects</b>	<b>137</b>
5.1	Source counts of VLBI-detected radio sources . . . . .	137
5.1.1	Chapter summary . . . . .	137
5.1.2	General conclusions . . . . .	139
5.1.3	Future prospects . . . . .	140
5.2	Source detection and characterization . . . . .	140
5.2.1	Chapter summary . . . . .	140
5.2.2	General conclusions . . . . .	142
5.2.3	Future prospects . . . . .	142
5.3	Lens detection with the ILT . . . . .	143
5.3.1	Chapter summary . . . . .	144
5.3.2	General conclusions . . . . .	146
5.3.3	Future prospects . . . . .	146
5.4	Preliminary results on lens modelling with deep learning . . . . .	149
5.5	Final remarks . . . . .	154
	<b>Summary</b>	<b>159</b>
	<b>Samenvatting</b>	<b>165</b>
	<b>Acknowledgments</b>	<b>171</b>
	<b>Bibliography</b>	<b>175</b>



# Introduction

## 1.1 Motivation

Artificial intelligence is becoming an inseparable part of doing science in many fields of research, and astrophysics is no exception as it requires the incorporation of smart automated approaches to deal with ever increasing amounts of data. Numerous current and future ground- and space-based instruments make astrophysics a data-rich field, which continually asks new science questions, while presenting progressively more complex data challenges. This thesis is focused on developing solutions to some of the challenges in the area of high angular-resolution radio astronomy with the help of novel intelligent approaches from computing science. It establishes a collaboration between the machines and human experts to deal with these new emerging issues. Here, intelligence can be defined as the ability to process raw data to inform future decisions. In the world of computer science, artificial intelligence can be achieved by building algorithms that mimic certain functions of the human brain, such as image recognition and characterisation. Among the many different types of artificial intelligence that can be applied to real world problems, this thesis is focused on applying data science techniques, such as machine learning (ML) and, in particular, deep learning (DL) algorithms.

ML algorithms are designed to teach a machine how to perform a specific task without being explicitly programmed to do said task. By developing computer programs based on ML algorithms, data is collected, aggregated and learned from by automatically identifying patterns without human assistance. ML algorithms can be used in the context of classification, regression or similar tasks based on the information that can be extracted from the data. Moreover, they allow for obtaining insight to the data and

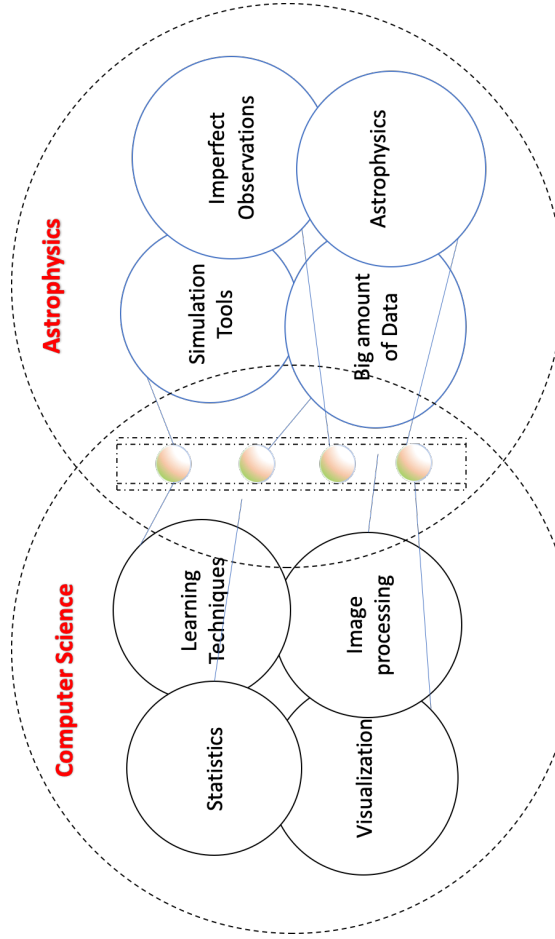


1

the problem at hand. The capability of ML to make fast and reliable decisions or predictions in many applications is an interesting aspect of its use. However, this is only possible when human experts help formulate the problem, acquire and organize data, and design a space for solution possibilities. It is also crucial to select the right learning algorithm and its parameters, apply the algorithm to the data, and validate the resulting solution to determine whether it is useful or not. DL is a subset of ML algorithms that takes the idea of learning from the data a step further by forming a higher level of abstractions. DL algorithms can perform an automatic feature extraction from the raw data, which is how they differ from other ML algorithms. This quality can get us closer to working with larger and more complex datasets.

This thesis is focused on interdisciplinary research that implements a set of ML and DL algorithms to address some of the astrophysical challenges related to complex astronomical observations of the radio sky. Here, radio sources are defined as celestial objects that emit their radiation (photons) at wavelengths of a few centimetres (cm) to metres (m). The primary motivation in linking computing science and radio astronomy is the many millions of celestial objects that are being observed in current wide-field radio surveys, for example, as part of the LOFAR Two Metre Sky Survey (LoTSS; Shimwell et al. 2019). Moreover, upcoming massive astronomical facilities, such as the Square Kilometre Array (SKA), will require reliable and fast solutions to leverage the full potential of the large catalogues and images that are expected to be generated within the next decade. The aim of this thesis is to provide solutions to the rising challenges and opportunities from the exponential growth of data in radio astronomy. In particular, we deal with the complex datasets that come from combining radio telescopes to form an interferometer, with a focus on those that provide the highest possible angular resolution imaging of radio sources. Here, ML and DL algorithms are used to provide a simpler representation of interferometric data that is easier to work with. In other words, by using ML algorithms we aim to provide a simple representation of a complex problem.

Fig. 1.1 shows how the different aspects of astrophysics and computer science are linked together in this thesis. Interferometric data, which is subject to thermal noise fluctuations and the measurement of partial information, require the use of efficient computer science techniques. Simulations are another source of information, which is also generated by using computational tools. Preparing realistic data with existing simulation tools is one of the important aspects of this thesis. Furthermore, image processing, statistics, learning algorithms and data visualization methods are also used to provide novel solutions to various astrophysical questions within the field of high angular-resolution radio astronomy.



**Figure 1.1** – An overview of the different components of the interdisciplinary research in this thesis.

The three main science objectives that are addressed in this thesis are listed below. However, at the end of this introductory chapter, a more descriptive explanation is provided.

- **The source counts of compact radio sources observed with Very Long Baseline Interferometry (VLBI):** In the second chapter of this thesis, real imaging data is analyzed to determine the number of radio sources that exist in the Universe with an extremely high surface brightness at cm-wavelengths. This involves dealing with large catalogues of information and simulating realistic imaging data to understand systematic effects associated with the observing system. The aim of this part of the thesis is to robustly predict the number of radio sources that could potentially be detected from surveys of the entire radio sky with interferometric arrays, like the SKA-VLBI system. These calculations provide an insight to the optimal survey strategies based on the density of such radio sources.
- **Source detection and characterization in imaging data from sparse interferometric arrays:** The third chapter of this thesis investigates the performance of DL algorithms for identifying radio sources in noisy imaging data. This is done to determine whether such algorithms can deliver improvements over traditional image detection techniques. This includes testing methods for deconvolution, noise removal and object detection. In addition, DL algorithms are used to characterize the shape and brightness of the detected radio sources. The aim of this part of the thesis is to determine whether DL algorithms can make wide-field surveys with VLBI instruments more efficient, by improving on the detectability of radio sources, which would result in reducing the required integration time for such surveys.
- **Fast and reliable detection of strong gravitational lens systems:** The fourth chapter of this thesis investigates whether pattern recognition with DL algorithms can be applied to efficiently find radio sources with peculiar shapes, that is, outliers from large samples of objects. In particular, DL algorithms are optimised to find so called *gravitational lenses*, which are caused by an astrophysical phenomena associated with Einstein's General Theory of Relativity. Unlike traditional gravitational lens search techniques, which require laborious visual inspection by human experts, the approach used in this thesis is designed to select genuine gravitational lens candidates from wide-field imaging data taken using the International LOFAR Telescope (ILT), with little or no human interaction. The main aim of this part of the thesis is to develop and test new

methods that can be applied to the millions of radio sources to be observed with the ILT over the next 5 years.

- **Deep learning in high angular-resolution radio interferometry:** The final chapter of this thesis presents a summary of the main results and discusses whether DL algorithms can provide novel solutions to the many challenges within high angular-resolution radio astronomy.

As this thesis brings together a number of concepts within astrophysics and computing science, the next sections of this introductory chapter provide a brief overview of radio interferometry, gravitational lensing, supervised learning, deep learning techniques and visualization techniques. The final section of this chapter gives an outline of the thesis and presents the scientific questions that are to be addressed.

## 1.2 Radio Interferometry

In this section, a brief introduction to radio astronomy, interferometry, and the arrays used in this thesis is given. We refer readers to, for example, Rohlfs & Wilson (2013) for an in-depth discussion of radio astronomy and interferometry.

### 1.2.1 Basic principles of radio astronomy

Radio astronomy focuses on the emission from celestial objects at long wavelengths (typically  $\lambda > 6$  mm), hereafter referred to as radio sources, to investigate and understand their properties. Some of these radio sources emit only in the radio bands, which makes them completely invisible at other wavelengths, such as in the optical bands that we observe with our own eyes. Given the long wavelengths of the radio waves that are emitted, the design and usage of telescopes in this field are different from what is traditionally seen at optical observatories. In fact, radio telescopes are almost identical to the satellite dishes used for television reception and telecommunication around the world. Radio telescopes capture the radio waves and convert them into electrical signals. Then, the signals are amplified, processed and digitised to form the output data. Studying radio waves can provide details about some of the mysteries of the Universe, such as supermassive black holes or the fading jets of plasma that they emitted. Diffuse matter in the large scale structure of the Universe, and the way galaxies interact are further phenomena that can be investigated with radio astronomy (de Gasperin et al. 2021).

The reflection of electro-magnetic waves with frequencies below 10 MHz ( $\lambda > 30$  m) by the Earth's ionosphere is a limiting factor that determines whether radio emission

can be observed from the ground. Those frequencies between 30 MHz and 50 GHz can easily penetrate the Earth's atmosphere (including its troposphere and ionosphere), making it possible to observe the sky at these frequencies from almost any location on Earth. Above these frequencies, sites at higher altitude are needed to avoid the absorption and emission of water in our atmosphere.

The large wavelength corresponds to a low energy of the emitted radio signals. This can affect the efficiency of observations as the faint sources do not produce enough energy to be detected. One solution to overcome this problem is to build larger dishes to collect more signal. On the other hand, larger dishes result in a larger resolving power  $\theta$  for the telescope. The Rayleigh criterion provides  $\theta$  (in radians) as

$$\theta = 1.22 \times \lambda/D \quad (1.1)$$

in which  $D$  is the diameter of the telescope and  $\lambda$  is the wavelength of the observed radiation. This equation shows that at a wavelength of 1 m (300 MHz in frequency), we would need a single dish telescope with a diameter of 200 kilometres (km) to resolve objects that are separated by 1 arcsec, which would be competitive with the resolution of optical telescopes. If it not technologically impossible, this would be very expensive and difficult to achieve. Currently, the largest single dish radio telescope in the world is the Five hundred metre Aperture Spherical Telescope (FAST) in China. However, the solution to this angular resolution problem in radio astronomy is a process called *interferometry*, where multiple radio telescopes (receivers) are connected to work together. With Very Long Baseline Interferometry (VLBI; Broten et al. 1967), we are able to place antennas over large areas such as continents in order to obtain milliarcsec (mas) and sub-mas angular resolution imaging at cm-wavelengths.

In radio interferometry, two or more radio receivers observe the same astronomical object simultaneously at the same frequency. To maximise the signal, the incoming waves are combined in-phase as constructive interference, whilst out-of-phase waves are subjected to destructive interfering measures. Waves that are not entirely in or out of phase exhibit a pattern of intermediate intensity that may be used to detect the relative phase difference. As each wave takes a different route or path to the receiver, the difference in the distance traveled creates a phase difference between them, or in other words, some geometric delay in the arrival time of the waves at the two receivers. The distance between each of the receivers is referred to as the baseline length  $B$ . The angular resolution  $\theta$  for an interferometer is similar to equation 1.1, but now  $D$  is replaced by  $B$ . Larger baselines provide the possibility of observing more compact structures, while the detection of larger structures depends on the shortest baselines.

There are  $N(N - 1)/2$  number of baselines for a system of  $N$  connected antennas.

Radio interferometers do not produce an image of the sky-plane, denoted by coordinates  $(l, m)$ . Instead, the sky brightness distribution,  $I_\nu$  at some frequency  $\nu$ , is measured in the Fourier-plane, denoted by coordinates  $(u, v)$ . The interference pattern of each baseline creates the visibility data, which is a measure of the sky brightness distribution on some angular-scale. The visibility produced by two antennas  $i$  and  $j$  (called a baseline) is given by (for  $w = 0$ ),

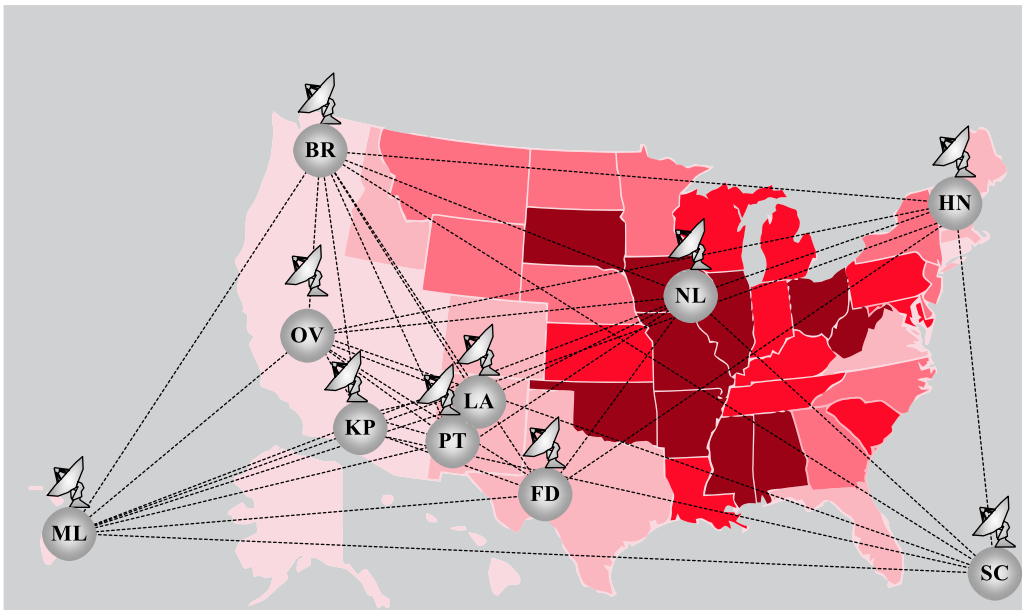
$$V(u, v) = \iint \frac{I_\nu(l, m)}{\sqrt{1 - l^2 - m^2}} \exp[-i2\pi(ul + vm)] dl dm. \quad (1.2)$$

The image of the radio source is then obtained by performing a Fourier transform on the sampled visibility data. It is produced through a sparse sampling of the  $uv$ -plane, based on the antenna locations, the frequency of the observations, and the total integration time used. As the Fourier-plane is only sparsely sampled by the limited number of baselines in the array, there are large gaps in the sampled  $uv$ -plane. This results in a lack of information on certain angular scales for sparse interferometric arrays, which are the focus of this thesis. By using more baselines, aperture synthesis from the Earth's rotation, or a wide frequency bandwidth, the sampling of the  $uv$ -plane can be improved. However, as there will always be gaps in the  $uv$ -plane, the resulting images in the sky-plane will have artifacts. These are typically removed using image deconvolution techniques.

Several interferometers have been built, which can observe the radio sky and provide data to study the Universe. This thesis is based on data collected from two instruments in particular: the Very Long Baseline Array (VLBA) and the Low-Frequency ARray (LOFAR). In the following, these two interferometric arrays are presented.

### 1.2.2 Very Long Baseline Array

The VLBA is a collection of ten identical radio telescopes that are located throughout the continental United States, and in the Virgin Islands and Hawaii (see Figure 1.2 for their locations). The VLBA can observe the radio sky in the frequency range of 0.3 to 83 GHz. Each VLBA telescope has a diameter of 25 m, and the longest baseline is 8600 km, which provides mas angular resolution. The wide range of observing frequencies and the very high angular resolution allows the imaging of very fine structures in radio sources. This includes studies of optically thin relativistic jets from quasars, galactic nuclei, and radio stars. On the other hand, the sparsity of the radio telescopes (45 measurements per time and frequency interval) makes the background noise highly



**Figure 1.2** – VLBA map with marked antenna locations and abbreviated names (credits: NRAO). ML: Mauna Kea – Hawaii, BR: Brewster – Washington, OV: Owens Valley – California, KP: Kitt Peak – Arizona, PT: Pie Town – New Mexico, LA: Los Alamos – New Mexico, FD: Fort Davis – Texas, NL: North Liberty – Iowa, HN: Hancock – New Hampshire, SC: St. Croix – U.S. Virgin Islands.

correlated in the imaging data, and limits the largest recoverable angular-scale to 100–200 mas. This provides a number of imaging challenges that are investigated in this thesis.

VLBA observations at 1.4 GHz are used for the second and third chapters of this thesis. The data come from the mJy Imaging VLBA Exploration at 20 cm (mJIVE–20) survey, which used 408 hours of VLBA filler time (Deller & Middelberg 2014). These data are used to characterize the compact radio source population in Chapter 2, and then to make realistic simulations of VLBI data to test DL algorithms for object detection and characterization in Chapter 3. Further details about the mJIVE–20 survey dataset are given in those two chapters.

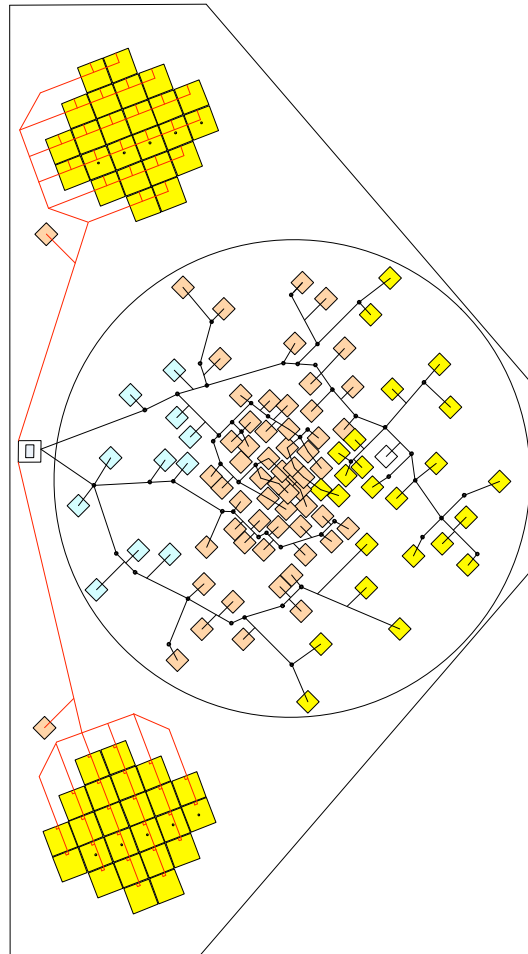
### 1.2.3 LOw Frequency ARray (LOFAR)

LOFAR is an interferometer that is made up of many simple dipole antennas, as opposed to conventional filled aperture dish telescopes. The simple design of the antennas in LOFAR allows astronomers to observe a very wide area of sky at once. It has no moving parts. However, it is still possible to digitally steer the antennas to point at a specific direction using software. This process is referred to as beam forming, where the signals from one particular direction are combined in phase, maximising the response of the array in that direction, whereas the response to signals in other directions is suppressed. LOFAR has also made it possible to observe different parts of the sky at the same time (van Haarlem et al. 2013).

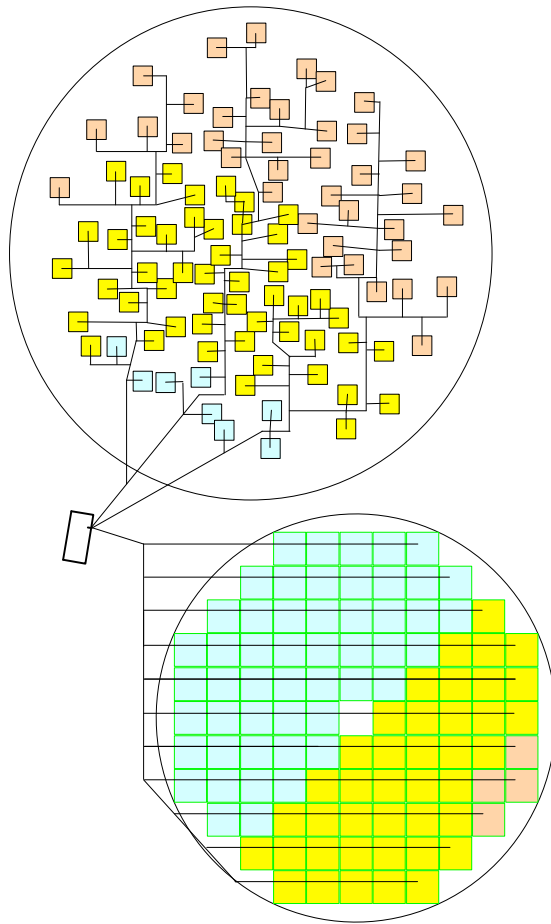
LOFAR is currently the largest radio telescope in the world, and operates at the lowest frequencies observable from the ground. Unlike single-dish telescopes, LOFAR can be considered as a multi-purpose sensor network, with an innovative computer and network infrastructure that can handle extremely large data volumes. Although the core of the array is located in Exloo, with remote stations throughout the Netherlands, currently eight other countries have LOFAR stations. With distinct High Band Antennas (HBA) and Low Band Antennas (LBA), LOFAR can operate almost continuously in the frequency range of 10 to 240 MHz, which corresponds to wavelengths of 30 to 1.2 m. There are 96 LBAs and 48 HBAs in each core and remote station, an example of a core station is presented in Fig. 1.3. The international stations have 96 LBA and HBA antennas (see Fig. 1.4), which when added to LOFAR, form the International LOFAR Telescope (ILT). This expanded array is going through the commissioning phase (Morabito et al. 2021; Jackson et al. 2021; Bonnassieux et al. 2021; Harwood et al. 2021; Sweijen et al. 2021; Timmerman et al. 2021). The ILT currently includes 14 international stations, six in Germany, three in Poland, one in the UK, France, Sweden, Ireland and Latvia. As a result of the distant geographical locations, the ILT can resolve low frequency radio sources that are typically 300 to 500 mas in size. However, an angular resolution of 0.27 arcsec is achievable at 150 MHz, thanks to the longest baseline (1989 km) between Ireland and Poland. Further expansion of the ILT is planned, with the addition of a station in Italy.

The unique angular resolution and wide field-of-view of the ILT should in principle make the array a fantastic instrument for finding gravitationally lensed radio sources. This goal is the focus of Chapter 4 of this thesis, where novel DL algorithms are developed and tested to identify rare lensed events in ILT survey data.





**Figure 1.3** – *The geometric distribution of the LBA and HBA antennas of LOFAR in a core station, reproduced from van Haarlem et al. (2013). The large circles correspond to the LBA antennas, while the small squares are the HBA tiles.*



**Figure 1.4** – *The geometric distribution of LOFAR international station antennas, reproduced from van Haarlem et al. (2013).*

### 1.3 Gravitational Lensing

The purpose of this section is to introduce the fundamental concepts and formalism of gravitational lensing. It explains the applications of gravitational lensing and how it provides a magnified view of distant galaxies. The interested reader is referred to a more detailed description of the phenomenon given by Schneider (1992), Blandford & Narayan (1992), Meylan et al. (2006) and Congdon & Keeton (2018). The following material is relevant to Chapter 4 of this thesis, which is focused on the detection of strong gravitational lens systems and categorizing them.

### 1.3.1 Basic formalism of gravitational lensing

Strong gravitational lensing describes an astrophysical phenomenon in which two galaxies are along the same line-of-sight from the observer. In this thesis, we refer to the most distant galaxy as the "background source", while the closer galaxy is referred to as the "lensing galaxy". The gravitational field of the lensing galaxy bends the light from the background source by an angle  $\hat{\alpha}$ . As a result, the radiation from the background source becomes distorted, and extended arcs and multiple images can be created. The deflection angle is defined as

$$\hat{\alpha} = \frac{4GM}{c^2 b}, \quad (1.3)$$

which is dependent on the enclosed mass of the lensing galaxy ( $M$ ) and the impact parameter (closest approach) to the lensing galaxy ( $b$ ), where  $G$  is the gravitational constant and  $c$  is the velocity of light.

Fig. 1.5 shows the geometry of gravitational lensing in which  $D_s$ ,  $D_l$  and  $D_{ls}$  are the angular diameter distances from observer to source, observer to lens and from lens to source, respectively. From this, the reduced deflection angle,  $\alpha$ , can be shown as

$$\alpha(\theta) = \frac{D_{ls}}{D_s} \hat{\alpha}(D_l \theta) \quad (1.4)$$

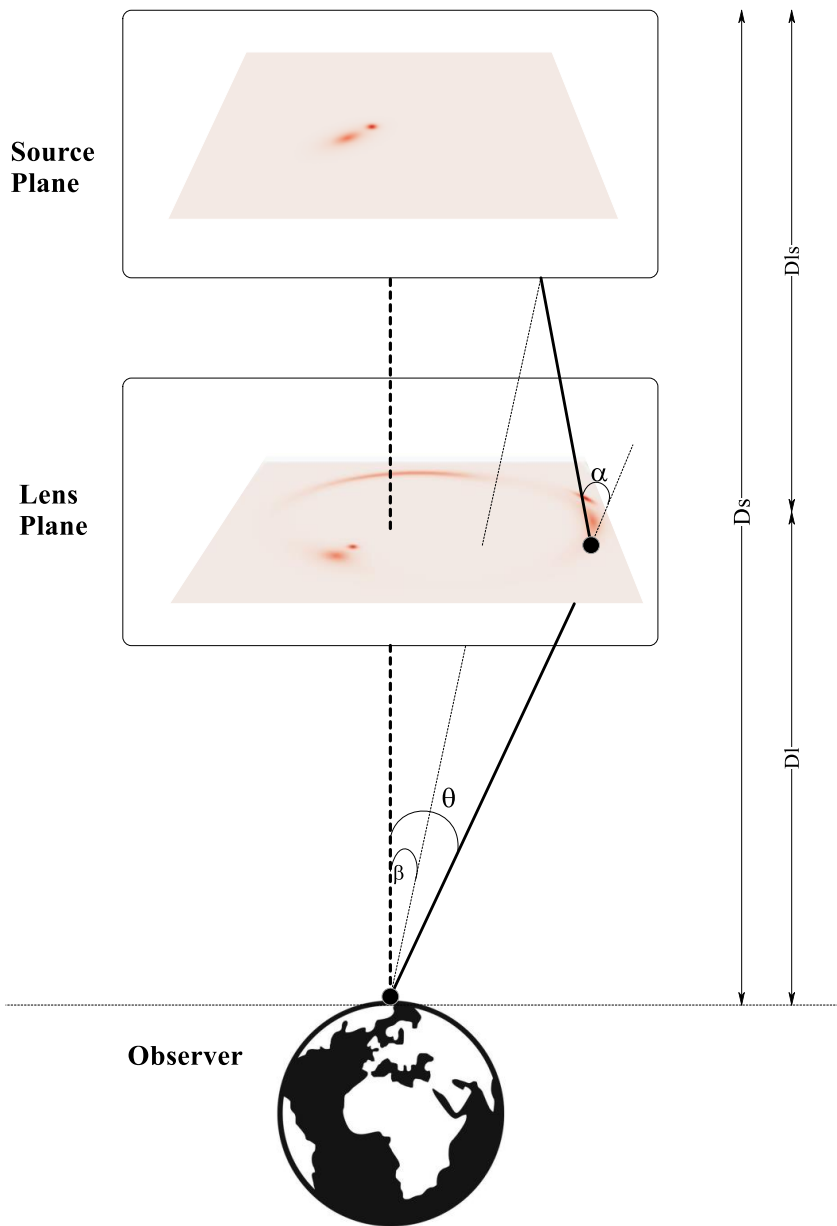
and the final lens equation to be

$$\beta = \theta - \alpha(\theta). \quad (1.5)$$

In the lens equation,  $\beta$  and  $\theta$  are the true and apparent angular position of the source, respectively, as seen by the observer. There are multiple solutions to the lens equation leading to the generation of multiple images at position  $\beta$ . For circularly symmetric lens systems, and when the background source is perfectly aligned with the lensing galaxy, the result is an Einstein ring. The radius of the ring is known as the Einstein radius,  $\theta_E$ , and is given by

$$\theta_E = \sqrt{\frac{4GM}{c^2} \frac{D_s}{D_l D_{ls}}}. \quad (1.6)$$

The radius of an Einstein ring is therefore larger for more massive galaxies. It is also dependent on the relative angular diameter distances between the background source and the lensing galaxy.



**Figure 1.5** – *The geometry of gravitational lensing.*

### 1.3.2 Strong lensing applications

Astronomers are interested in gravitational lenses because they are a major source of information regarding the structure of the Universe. Moreover, gravitational lensing provides the opportunity to study high-redshift galaxies that would be otherwise unobservable (Impellizzeri et al. 2008; Riechers et al. 2011). This phenomenon can be used to determine the expansion history of the Universe, as well as trace galaxy formation and evolution. This can help us to test our current understanding of how the Universe formed over cosmic time (e.g. Stacey et al. 2018). The details of some of these applications are describe below.

Due to the magnification of the background source, the lensing galaxy works in the same way as a telescope. The measured flux of the lensed object is increased with respect to the unlensed background source, as its angular size is increased by lensing. This increased angular size is similar to providing a higher angular resolution image for resolved background galaxies, which results in an opportunity to unveil the structure and dynamic properties of high-redshift sources (Chirivì et al. 2020; Faure et al. 2021; Swinbank et al. 2009).

In gravitational lensing, each individual lensed image of the background galaxy is generated by a different geometric light path. The different gravitational potential depths that the light passes through results in a time-delay between different lensed images (Suyu et al. 2010). Also, the angular diameter distance to the lensed object affects the time delay for each of the lensed images of the background galaxy. Therefore, the observed time delay for each lensed image is inversely proportional to the Hubble constant ( $H_0$ ), which is used to characterize the expansion of the Universe, providing some knowledge of the mass distribution is known (Suyu et al. 2013). In fact, the determination of  $H_0$  is one of the main applications of gravitational lens systems.

Gravitational lensing is a purely geometrical effect that only involves gravity. Therefore, it is independent of the nature of the matter within the lens, that is, whether it is luminous or dark, and also its dynamical state. This means that gravitational lensing provides a promising source of information to study the amount and distribution of mass within the lensing galaxies. This is done through predicting the observed parameters, such as the flux-ratios and/or the relative positions of the lensed images, which allows the astronomer to obtain information on the mass distribution of the lens, such as its position and shape (McKean et al. 2007; Vegetti et al. 2012; Hsueh et al. 2020). Also, how the density of the mass changes as a function of radius within the lens can be determined with gravitational lensing, which can be used to test models for the

formation and evolution of massive galaxies and clusters of galaxies (Koopmans et al. 2006; Nieuwenhuizen, Limousin & Morandi 2021; Bahé 2021).

### 1.3.3 Strong lens modelling

Modelling of gravitational lensing systems is a process that involves reproducing the observed features in the data, namely, the surface brightness distribution of the lensed images using a theoretically motivated model for the lens mass. This process can offer an in-depth description of the mass distribution of the lens, and is required for almost all of the applications of gravitational lensing described above. Modelling is also needed to calculate the lensing magnification, which provides information on the intrinsic physical properties of the background galaxy, such as the size, brightness and luminosity. Lens models are thus required to comprehend the characteristics of the background source.

There are several lens-modeling approaches, which can be roughly classified as parametric and non-parametric. Non-parametric modelling directly retrieves the underlying mass distribution from the generated lensed images (and possibly the background source), while parametric modelling uses assumptions to model both the background and lensing galaxies. It is shown by Meneghetti et al. (2017) that both non-parametric and parametric modelling techniques are good at recovering the averaged mass and density profiles of the lens.

There are around 40 gravitationally lensed radio sources currently known, and it is predicted that with the SKA, this number could grow to as large as  $10^5$  (McKean et al. 2015). This means that fast, efficient and automatic lens modelling techniques are required to study the details of each individual system, or identify classes of lenses to follow-up. Traditionally, a maximum likelihood method has been used for lens modelling (Suyu et al. 2010; Vegetti et al. 2010). However, ML, and in particular DL techniques have become a promising approach that has been widely used in recent years to study a variety of strong lensing applications (Lin et al. 2020; Kim et al. 2020; Marianer, Poznanski & Prochaska 2020).

## 1.4 Supervised Learning

ML algorithms can be divided, in general, into supervised and unsupervised learning. Other learning scenarios, such as reinforcement learning or semi-supervised learning exist, but will not be considered in this thesis. In a supervised learning algorithm, the data acts as an instructor, assisting the model in discovering a relationship between a collection of features and user specified labels. This model is trained to predict the

properties of the unseen data (e.g., Goodfellow, Bengio & Courville 2016). Supervised learning has been used to address various problems in astronomy, for example, the classification of galaxy morphologies in imaging data (Pearson et al. 2019; Nolte et al. 2019), variable star classification using light-curve representations (Becker et al. 2020), and the classification of blazar candidates (Kovačević et al. 2020). This section describes some of the relevant supervised learning algorithms used in this thesis.

### 1.4.1 K Nearest Neighbours (KNN)

K Nearest Neighbours (KNN) is considered as one of the simplest learning techniques (Cover & Hart 1967) that needs no explicit training steps. It is used for classification and regression tasks. To execute the algorithm, the user is asked to specify  $K$  as the number of nearest neighbours. The choice of  $K$  can significantly affect the performance of the algorithm on unseen test data. Although  $K$  is typically defined as the square root of the number of samples (Hassanat et al. 2014), it is necessary to explore over different values of  $K$  and determine which one is the best. For the classification purposes, a voting scheme is applied on  $K$  closest neighbours to specify the membership of the test data. The same methodology can be applied on regression problems, but instead of membership, the output will be the average or the weighted average of the values of  $K$  nearest neighbours. Although KNN is simple and easy to implement, it has a major drawback on the sensitivity of the value of  $K$ . Moreover, the value of  $K$  is considered as a constant for all the test samples, which might cause inaccurate results for some of the test data (Cheng et al. 2014).

### 1.4.2 Random forest

Random forest is an ensemble learning method that works by building a large number of decision trees during training. Similar to KNN, it can be used for both classification and regression tasks with a voting mechanism (Liaw & Wiener 2002). For a classification problem, the most frequently chosen class is selected while the average predicted value of all the individual trees is returned for regression tasks. Although decision trees are easy to use and interpret, they are not accurate enough in classifying unseen data in complex problems. Random forest provides flexibility by creating multiple decision trees using a bootstrap technique. Bootstrap is a sampling technique that allows data points in the original dataset to be selected more than once. It makes the generation of multitude trees in random forest possible by bringing diversity and flexibility to the final model. The following steps explains how random forest is built and how it can generate more generalized results compared to simple decision trees: step 1: creating a bootstrap dataset with the same size as the original dataset;

step 2: creating a decision tree for each bootstrap dataset using a subset of features at each step;

step 3: iteratively executing step 1 and step 2 to generate a number  $N$  decision trees; and

step 4: finally, after generating  $N$  decision trees, a majority voting mechanism is used to classify unseen (test) data.

### 1.4.3 Artificial Neural Network (ANN)

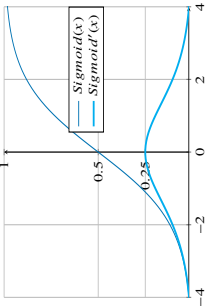
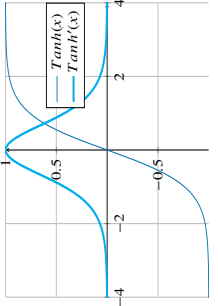
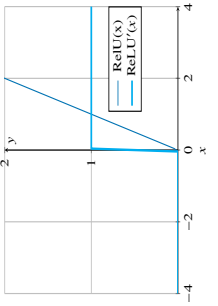
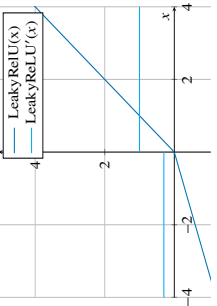
Artificial Neural Network (ANN) is a learning technique that is vaguely inspired from the structure of biological brains (McCulloch & Pitts 1943). In general, it consists of three main types of layers: input, hidden and output layers. Hidden layers are responsible for representing inputs in such a way that the desired output can be achieved. There is usually an activation function following the hidden layers, which introduces nonlinearity to neural networks. Nonlinear activation functions play an important role in the efficiency of neural networks since linear functions are not usually good enough to model complex problems. Moreover, without the use of nonlinear activation functions, several layers of neural networks would still function in the same way as only one layer. In the following, different types of activation functions are reviewed.

#### Activation Functions

Choosing suitable activation functions is one of the main steps in designing neural networks. Usually, activation functions are applied on the weighted sum of all connected neurons along with an optional bias parameter. The activation function determines the level of excitement for each neuron. Three popular activation functions in neural networks are Sigmoid, Tanh and ReLU. Table 1.1 provides the detailed information about several activation functions; (see Apicella et al. 2020; Hayou, Doucet & Rousseau 2019 for an overview of activation functions).

The Sigmoid activation function compresses the input in the range of 0 and 1. Its output can be interpreted as the firing level of a particular neuron: a Sigmoid activation of 0 represents a quiescent neuron (not being active), while the Sigmoid output of 1 represents a fully saturated neuron at maximum activity. The output range between 0 and 1 makes Sigmoid suitable for tasks in which the network predicts the probability of belonging to a particular class of data. The Tanh activation function instead compresses the input to the range of  $-1$  and  $+1$ . ReLU on the other hand simply calculates the  $\max(0, x)$ . This means that the output of this activation function is zero for negative inputs and linear with a slope of 1 for positive inputs. Krizhevsky,



Function	Range	Equation	Derivative	Plot
Sigmoid	$[0, 1]$	$f(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}$	$f'(x) = f(x)(1 - f(x))$	
Tanh	$[-1, 1]$	$f(w^T x + b) = \frac{2}{1 + e^{-2(w^T x + b)}} - 1$	$f'(x) = 1 - f(x)^2$	
ReLU	$[0, \infty)$	$f(w^T x + b) = \begin{cases} w^T x + b < 0 & 0 \\ w^T x + b > 0 & w^T x + b \end{cases}$	$f'(x) = \begin{cases} x < 0 & 0 \\ x > 0 & 1 \end{cases}$	
LeakyReLU	$(-\infty, \infty)$	$f(w^T x + b) = \begin{cases} w^T x + b < 0 & \alpha(w^T x + b) \\ w^T x + b > 0 & w^T x + b \end{cases}$	$f'(x) = \begin{cases} x < 0 & \alpha^* \\ x > 0 & 1 \end{cases}$	

**Table 1.1** – Activation Functions: Sigmoid, Tanh and ReLU.

Sutskever & Hinton (2017) have found that ReLU converges 6 times faster compared to Tanh in the image processing tasks that they considered. It is also perceived as being computationally less expensive, and appears to learn faster compared to networks of Sigmoid or Tanh units. Therefore, ReLU has become a very popular activation function for hidden layers; see Oostwal, Straat & Biehl (2019) for a comparison on the training behavior of ReLU and Sigmoid. Despite all the advantages of ReLU, the constant output of zero for negative inputs might cause some neurons to be permanently inactive during training and never contribute to the learning process. This problem is known as the *dying* ReLU (Lu 2020). The solution to this problem is in using a slightly different activation function, which instead of giving zero for negative values, is linear with a small negative slope. This activation function is known as Leaky ReLU.

## 1.5 Deep Learning Techniques

The term "Deep Neural Network" or "Deep Learning" refers to multi-layered ANN (Goodfellow, Bengio & Courville 2016). It is considered to be one of the most important methods to emerge in the last few decades, and has become extremely common when massive amounts of data need to be handled.

### 1.5.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) have become very popular in astronomical studies. CNNs are one of the most prominent approaches based on ANNs that deal with input data of a topological structure, for example, an astronomical image. Similar to a general ANN, a CNN consists of several layers. Each layer has a unique functionality, which is partly pre-defined. In the following, the most commonly used and basic layers of a convolutional network are described.

#### Convolution Layer

Convolution layers compute the scalar product between a set of weights and the corresponding regions or patches in the input. Kernels, parameterised by the adaptive weights, are typically small, but repeated over the entire input image. Such a layer is trained to activate specific kernels when a certain spatial input pattern occurs in the input (O'Shea & Nash 2015). To demonstrate the functionality of a convolutional layer, an example from the MNIST dataset (LeCun & Cortes 2010) is provided. MNIST is a popular database in the field of image processing and ML. It contains grey-scale images of  $28 \times 28$  pixels, representing handwritten digits. For each image,

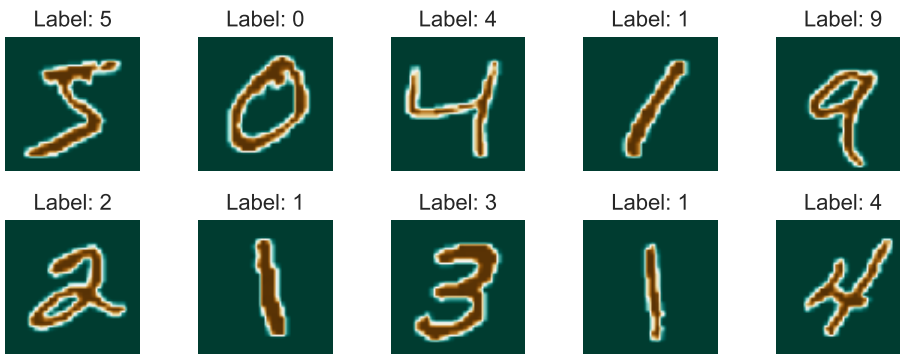


Figure 1.6 – MNIST dataset of handwritten digits.

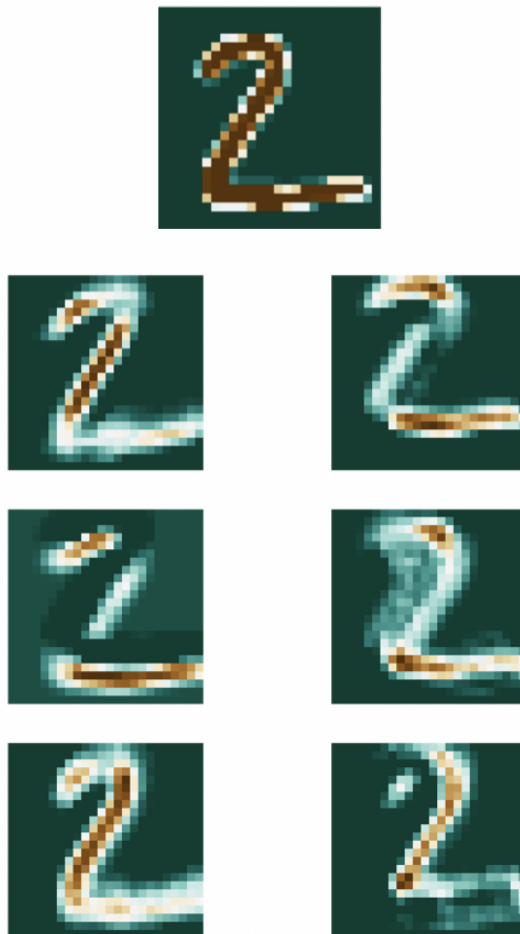
there is a target label as the standard benchmark on the digit in the image. An overview of the MNIST dataset is shown in Fig. 1.6.

The learnt features by the convolution layers are used by densely-connected layers for classification purposes. The convolution layers are responsible for learning a hierarchy of features, those that are highly abstract, as well as those more detailed. Visualizing the trained features can provide valuable information to understand and interpret the performance of a convolutional neural network. Fig. 1.7 shows the output of the convolutional layer; the given input image is shown on the top row, and contains a sample of MNIST dataset associated to label the number ‘2’. Fig. 1.7 shows the lighted up pixels in the (output) feature map, in good agreement with the input image..

### Max Pooling Layer

The purpose of a pooling layer is to combine units (e.g., representing pixels) of a previous layer and generate a version at lower resolution. Considering the ultimate goal of the network is to extract particular features, edges, curves etc., a so-called *max pooling* layer retains only the most activated unit of a given region, while all lower values are disregarded in the following layers of the network. Other types of pooling have been suggested, such as average pooling, which takes the average value of all the pixels within a given block. Max pooling has been observed to outperform average pooling (e.g., Boureau et al. 2010).

Although several arrangements for placing a max pooling layer are possible, it is usually added after a convolutional layer. By grabbing the maximum values of each block, max pooling down-samples the input image. Down-sampling enables the network to look at a larger area of the image, but at a decreased resolution. This



**Figure 1.7** – A representation of the output from a convolutional layer on the MNIST dataset. The input image to the convolution layer is associated with the number '2' (shown on the top row).

reduces the number of adaptive parameters in the network, which results in a lower computational load. The experiments of Scherer, Müller & Behnke (2010) show the significant superiority of CNNs that use max pooling when compared to regular sub-sampling, in which the average over all input values is propagated to the next layer.

### **Dropout Layer**

This layer can work as a regularizer. It is used to overcome overfitting by randomly interrupting data flow through the network. Using dropout, a large model (that can easily be overfitted) will be repeatedly sampled to create smaller sub-models from it. This is done by randomly removing a selection of activation units (and their associated weights). By using dropout, the network cannot be certain on the presence of a particular hidden unit in an individual update. Impaired performance, as a result of using a dropout layer, might be fixed by adding additional layers to the network and making it deeper (Srivastava et al. 2014).

### **Fully connected layers**

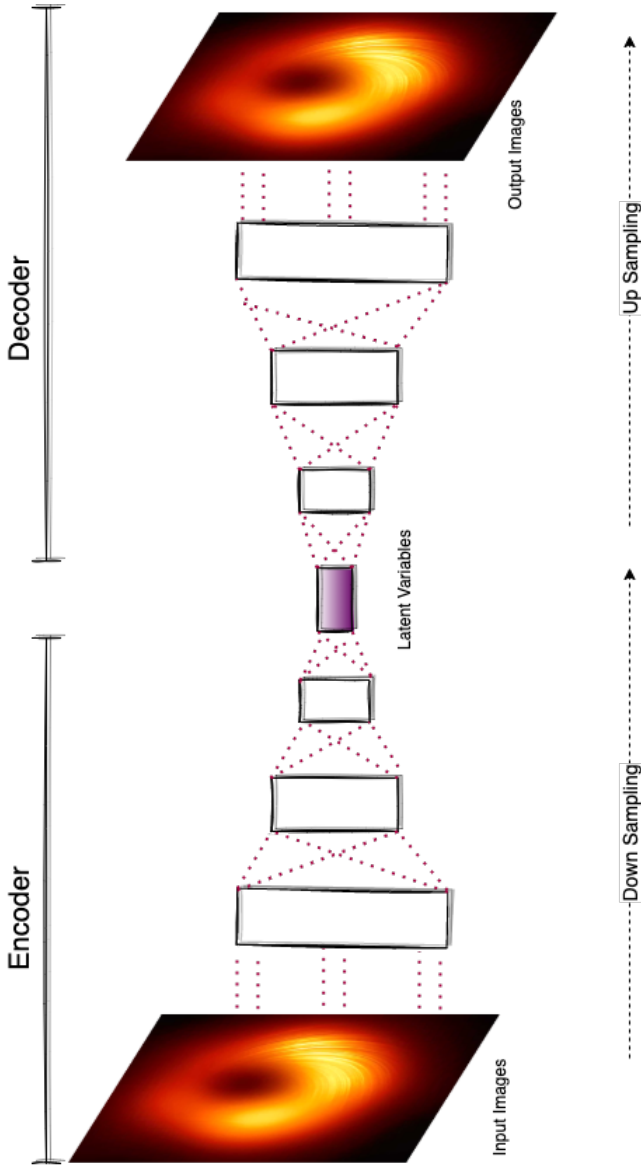
Fully connected layers are usually positioned at the end of a typical CNN. These layers are designed in a way that neurons have full connections to all activations in the previous layer. They have the classic structure of ANNs, as explained in Section 1.4.3.

## **1.5.2 Convolutional autoencoders**

A convolutional autoencoder is a specific type of CNN in which the network learns by encoding the input in an unsupervised approach. It consists of two main components: an encoder and a decoder, which are connected by the bottleneck or the latent variables. The encoder is responsible for compressing the input into a low-dimensional representation (latent variables) that contains the essential features required to reconstruct the output by the decoder. In order to do that, a number of sequentially connected layers, including convolution, max pooling, activation and fully connected layers in both the encoder and decoder are used. Fig. 1.8 shows a symbolic schema of the encoder and decoder sections in autoencoders.

## **1.6 Visualization techniques**

Data visualization is a crucial technique that is used to take data driven decisions when analyzing massive amounts of data. They offer a graphical representation of



**Figure 1.8** – An overview of the existing components in a convolutional autoencoder: Each box represents a selection of layers such as convolution, max pooling and activation. The specific order of layers and number of them can differ based on the application and the defined goal.

1

the data in the form of a map or a graph, to understand data trends, outliers, and patterns. It is one of the key steps in the data science process, which states that data must be visualized after it has been collected, processed, and modelled, before drawing any form of conclusions. In the world of big data, and particularly when using predictive, analytic machine learning algorithms, it is essential to monitor the results in order to make certain models perform as expected. This is because visualizing complex algorithms is usually much easier to understand than numerical outputs. The following visualization technique is used in Chapter 3 when predicting the properties of radio sources.

### 1.6.1 t-Distributed stochastic neighbour embedding

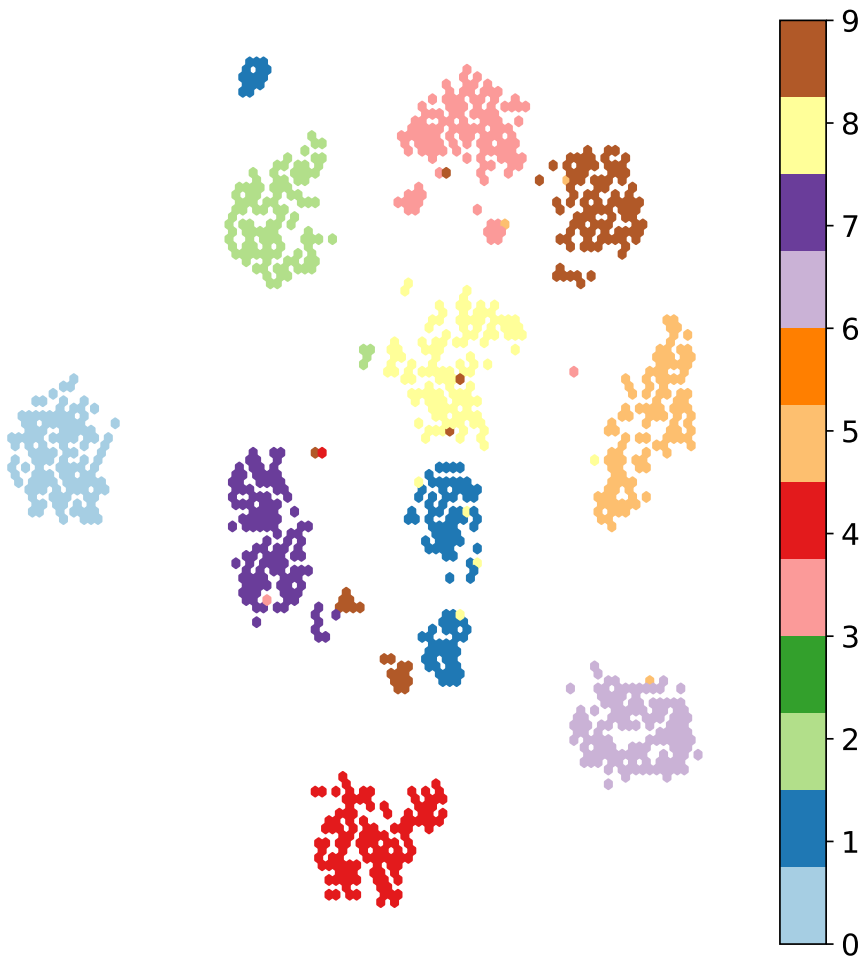
t-Distributed Stochastic Neighbour Embedding (t-SNE) was first presented in 2008 by Van der Maaten & Hinton (2008). It aims to provide a 2-dimensional (2D) representation of data points from a higher-dimensional space. The pairwise similarities of data points in the higher dimensional space are preserved by the generated 2D representation. In other words, similar data points in the original higher dimensional space would also be positioned closely in the lower dimensional space. It is designed to minimize the divergence between the distribution of pairwise similarities in the original space and a distribution of similarities between the corresponding lower-dimensional points. The result of applying t-SNE (with default parameter settings) on 5000 training samples in a MNIST dataset is presented in Fig. 1.9; colours are used to separate different classes of data in the MNIST dataset. Considering the image size of  $28 \times 28 = 784$  pixels in the MNIST dataset, each point in Fig. 1.9 represents an array of size 784. By analyzing the result of applying t-SNE, it is shown that different classes of data are well separated.

## 1.7 Scope and aim of this thesis

This thesis explores several big data challenges in high angular-resolution radio astronomy. The aim is to address several radio astronomical questions (presented in the following) using data science, statistics and machine learning techniques. In particular, there is a focus on novel solutions for object detection and characterization with VLBI observations at cm- and m-wavelengths.

*1. What is the number of VLBI-detected radio sources as a function of flux density? What is the expected number of radio sources with flux densities  $>1$  mJy that can be targeted with SKA-VLBI?*

In recent years, there have been several computational developments that allow objects



**Figure 1.9** – 2-D visualization on a part of MNIST dataset using *t*-SNE. Each datapoint represents an image in the dataset with a size of  $28 \times 28 = 784$  pixels. There are ten class labels in the MNIST dataset corresponding to images of hand written digits.



1 within the full field-of-view of VLBI arrays to be imaged. This can lead to a large number of interesting science use cases that would motivate all-sky surveys with instruments like the VLBA (in the short term) and the SKA-VLBI facility (in the longer term). However, our knowledge of the number of sources that would be found from such surveys, and what makes a source detectable with VLBI is extremely limited. The main aim of Chapter 2 is to measure the number of VLBI-detected radio sources from a  $240 \text{ deg}^2$  pilot survey with the VLBA, and use this measurement to calculate the number of sources that would be detected with SKA-VLBI.

*2. Can data science techniques reliably detect and characterize radio sources from sparse interferometric arrays? Can the efficiency of all-sky VLBI surveys be improved with deep learning?*

Source detection and characterization will always play an important role in making new scientific discoveries, particularly for the all-sky VLBI surveys that are investigated as part of Chapter 2. Reliability, completeness and purity are all important attributes of any survey catalogue. In Chapter 3, a DL algorithm is developed and tested for use during a possible wide-field VLBI survey with the VLBA (or any sparse interferometric array operating at high angular resolution). The main aim of this part of the thesis is to see whether such algorithms offer any improvement over traditional object detection algorithms, and whether they can make all-sky VLBI surveys more efficient by lowering the threshold for detection.

*3. Can deep learning algorithms reliably detect gravitational lenses in interferometric imaging data? What parameter space (in terms of the lens and source properties) is the ILT sensitive to when carrying out a survey for gravitationally lensed radio sources?*

In principle, the ILT should image millions of radio sources at around  $0.3 \text{ arcsec}$  angular resolution, which seems well matched to searching for gravitational lens systems. However, identifying rare lensing events in the imaging data will be challenging. Also, in recent years, a number of DL techniques have been introduced for finding lens systems in optical data. In Chapter 4, such methods are tested for interferometric data, and a new lens detection algorithm for LOFAR is developed. The main aim of this chapter is to provide a novel approach to finding gravitational lenses with the ILT, and to determine what type of lenses the ILT can potentially find, given the sensitivity and angular resolution of the data.

## Source counts of VLBI-detected radio sources

Based on "The source counts of VLBI-detected radio sources and the prospects of all-sky surveys with current and next generation instruments"

S. Rezaei, J. P. McKean, A. Deller and J. F. Radcliffe

To be submitted for publication in MNRAS

### Abstract

We present an analysis of the detection fraction and the number counts of radio sources imaged with Very Long Baseline Interferometry (VLBI) at 1.4 GHz as part of the mJIVE–20 survey. From a sample of 24 903 radio sources identified by FIRST, 4 965 are detected on VLBI-scales, giving an overall detection fraction of  $19.9 \pm 2.9$  per cent. However, we find that the detection fraction falls from around 50 per cent at a peak surface brightness of  $80 \text{ mJy beam}^{-1}$  in FIRST to around 8 per cent at the detection limit, which is likely dominated by the surface brightness sensitivity of the VLBI observations, with some contribution from a change in the radio source population. We also find that compactness at arcsec-scales is the dominant factor in determining whether a radio source is detected with VLBI, and that the median size of the VLBI-detected radio sources is 7.7 mas. After correcting for the survey completeness and effective sky area, we determine the slope of the differential number counts of VLBI-detected radio sources with flux densities  $S_{1.4 \text{ GHz}} > 1 \text{ mJy}$  to be  $\eta_{\text{VLBI}} = -1.74 \pm 0.02$ , which is shallower than in the cases of the FIRST parent population ( $\eta_{\text{FIRST}} = -1.77 \pm 0.02$ ) and for compact radio sources selected at higher frequencies ( $\eta_{\text{JBF}} = -2.06 \pm 0.02$ ). From this, we find that all-sky ( $3\pi \text{ sr}$ ) surveys with the EVN and the VLBA have the potential to detect  $(7.2 \pm 0.9) \times 10^5$  radio sources at

mas-resolution, and that the density of compact radio sources is sufficient ( $5.3 \text{ deg}^{-2}$ ) for in-beam phase referencing with multiple sources (3.9 per primary beam) in the case of a hypothetical SKA-VLBI array.

## 2.1 Introduction

Very Long Baseline Interferometry (VLBI; Broten et al. 1967) is an observational technique where the signals from individual radio telescopes are combined coherently to produce a synthesised unfilled aperture. As the radio telescopes can be widely separated across the Earth (and even located in space), this technique currently provides the highest angular resolution imaging in astronomy (with a synthesized beam size of typically 1 to 10 mas at cm-wavelengths, and reaching 0.025 mas at sub-mm wavelengths; Event Horizon Telescope Collaboration et al. 2019a). VLBI allows a wide range of science goals to be realised, for example, detecting super-luminal motion in radio jets (Cohen et al. 1977), imaging gravitational lenses (Porcas et al. 1979), mapping the accretion disks of supermassive black holes (Miyoshi et al. 1995), tracing the expansion of stellar explosions (O'Brien et al. 2006), imaging outflows of atomic hydrogen from supermassive black holes (Morganti et al. 2013), localising Fast Radio Bursts (Marcote et al. 2020), and making the first images of the shadow of a black hole (Event Horizon Telescope Collaboration et al. 2019b).

It is over 50 years since the first fringes were produced between two unconnected antennas separated by over 3000 km (Broten et al. 1967), yet the unique applications of VLBI at cm-wavelengths have been restricted to studying only  $\sim 25\,000$  radio sources with very high brightness temperatures ( $> 10^5$  K) and over a very small fraction of the observable sky ( $\sim 230 \text{ deg}^2$ ; e.g. Deller & Middelberg 2014; Herrera Ruiz et al. 2017; Petrov 2021). This is significantly lower than the up to 5 million radio sources that have been catalogued from recent all-sky surveys at arcsec-resolution (e.g. Becker, White & Helfand 1995; Condon et al. 1998; Intema et al. 2017; Shimwell et al. 2019; Lacy et al. 2020). This is because the effective field-of-view of a VLBI experiment is typically only a few tens of arcsec in diameter, which is due to historical computational limitations and data-rates that required significant averaging of the visibility data. This means that large numbers of sources can only be observed through many short observations. Coupled with the sparseness of the telescopes forming the synthesised unfilled aperture, this results in only the brightest sources being detectable, that is, those that are sufficiently compact to have a measurable correlated flux on the available baselines of the array.

This limitation of VLBI means that radio sources at relatively high flux densities

(> 1 mJy) tend to make up the target population at cm-wavelengths, which is mainly dominated by radio-loud Active Galactic Nuclei (AGN), or so-called jetted AGN (Padovani 2017). At lower flux densities (< 1 mJy) a higher proportion of radio sources have emission associated with star-formation processes or weak jets (e.g. Condon et al. 2012; Prandoni et al. 2018). Studying such objects at high angular resolution requires much deeper observations, which further limits the number of objects that can be studied (Garrett et al. 2001).

However, due to advances in computing and the development of new techniques for correlating the signals from the different radio telescopes, it is now possible to form multiple phase centres for a given observation (Deller et al. 2011). Using multiple phase centres allows the observer to image small areas around known sources within the same observation. It works by shifting the phase centre to the location of known sources and averaging the obtained visibilities to obtain manageable-sized data sets. For example, the GOODS-N (Radcliffe et al. 2018) and COSMOS fields had 600 and 750 phase centres per observation. However, these are somewhat special cases and routine wide-field VLBI experiments are currently limited to around 100 phase centres from a typical observation. Nevertheless, this technique has revolutionised our view of the mas-scale radio Universe, both through shallow wide-field surveys, like the mJy Imaging VLBI Exploration at 20 cm (mJIVE–20; Deller & Middelberg 2014) survey and through deep narrow-field observations of well-studied deep fields (Middelberg et al. 2013; Herrera Ruiz et al. 2017; Radcliffe et al. 2018).

The results from these studies have demonstrated a proof-of-concept for all-sky VLBI surveys with current instruments, like the European VLBI Network (EVN) or the Very Long Baseline Array (VLBA). Also, in the future, the sensitivity of VLBI arrays are expected to improve dramatically with the construction of the Square Kilometre Array (SKA-VLBI) and the next generation Very Large Array (ngVLA). These instruments will lower the current sensitivity limits, and given their wide fields-of-view and wide frequency bandwidths, they can provide a significant increase in the numbers of sources detected on arcsec- to mas-scales. However, this will require determining the number of detectable sources on VLBI-scales as a function of flux density, so that survey strategies can be properly developed.

In this chapter, we determine the source counts of VLBI-detected radio sources from the wide-area mJIVE–20 survey (Deller & Middelberg 2014). With these data, we estimate the number of radio sources that can be detected by the next generation of wide-field VLBI surveys with the VLBA and the EVN in the short term (< 10 yr) and with a hypothetical SKA-VLBI array in the longer term. This chapter is arranged as follows. In Section 2.2, we present an overview of the final

mJIVE–20 survey catalogue that we use for our study. We determine the source counts of VLBI-detected radio sources in Section 2.3, which includes calculating the mJIVE–20 survey completeness from simulations and determining the effective area of the survey. Using these data and the (expected) capabilities of current and next generation radio interferometers, we also calculate the number of likely detectable radio sources on VLBI-scales for given survey strategies in Section 2.4. Finally, in Section 2.5, we present our conclusions.

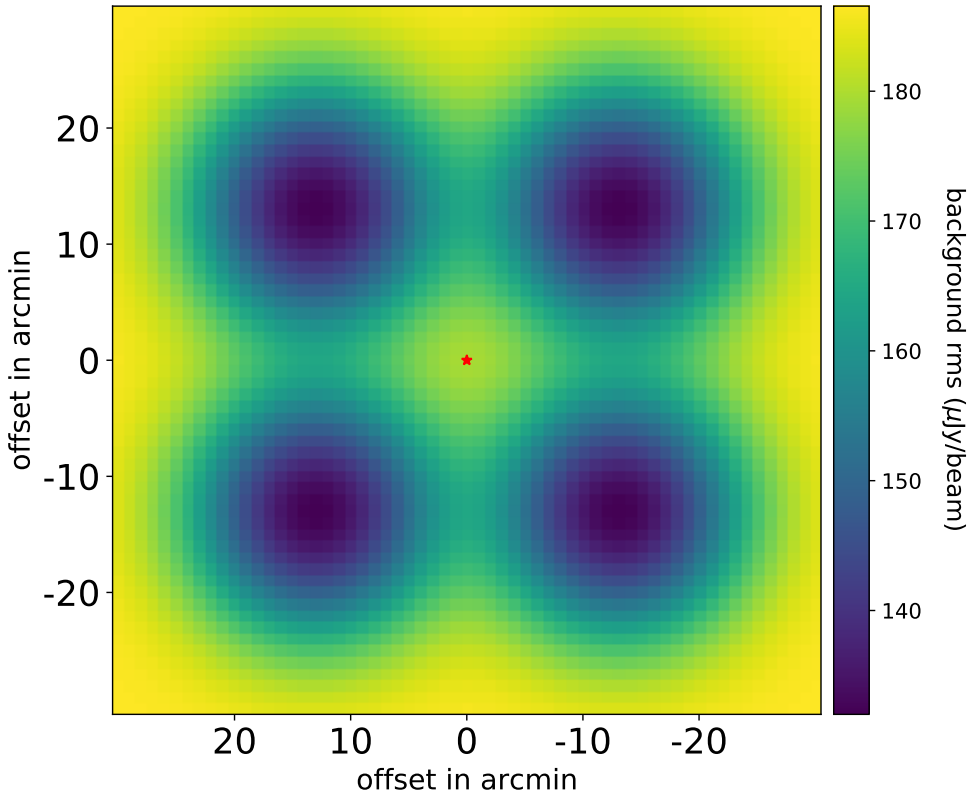
## 2.2 The mJIVE–20 survey final catalogue

In this section, we provide an overview of the mJIVE–20 survey, and discuss the properties and detection rate of the final catalogue that we use to determine the source counts of VLBI-detected radio sources.

### 2.2.1 Overview of the mJIVE–20 survey

The mJIVE–20 survey was carried out with the VLBA at 1.4 GHz with dual polarization and a bandwidth of 64 MHz (recording rate of  $512 \text{ Mbit s}^{-1}$ ). The main aim of the project was to better understand the radio source population at mas resolution by exploiting recent developments in wide-field multi-phase centre correlation techniques (Deller et al. 2011). The survey strategy and initial results were reported by Deller & Middelberg (2014); here we summarise the main properties of the survey and discuss the final catalogue.

Each mJIVE–20 survey observation consisted of four individual pointings of 2 mins duration around a known VLBA calibrator (so that in-beam calibration could be used; see Fig. 2.1) at 7 different hour angles to improve the  $uv$ -coverage. This resulted in a typical synthesised beam-size of  $16 \times 6$  mas for target fields at Declinations between  $-7$  and  $60$  deg. The average rms noise is around  $150 \mu\text{Jy beam}^{-1}$  at each pointing centre. In total, there were 410 separate observations that targeted 24 903 radio sources from the Faint Images of the Radio Sky at Twenty-cm (FIRST) survey (Becker, White & Helfand 1995), making the mJIVE–20 survey the largest targeted sample of radio sources with VLBI to date. From these observations, 4 965 sources were detected above a threshold of  $6.75\sigma$ , where  $\sigma$  is the local rms noise of the imaging data. For our analysis, we use the final source catalogue from 2014 March 31, which was obtained using the BLOBCAT source detection software (Hales et al. 2012).



**Figure 2.1** – An example rms map from combining four separate pointings of the VLBA 25-m antennas at 1.4 GHz. The VLBA calibrator source is denoted by the (red) star in the centre of the image, such that it is always in the primary beam of each antenna for each pointing position. A sky area of about  $1 \times 1 \text{ deg}^2$  is surveyed with this pointing strategy, for each observation of a VLBA calibrator.

### 2.2.2 Detection rate as a function of radio source surface brightness, compactness and size

We now investigate the detectability of radio sources on VLBI-scales, based on their observed properties at lower angular resolution. Note that this comparison is not intended to be a thorough study of the astrophysical characteristics of the radio source population, but instead aims to compare the properties between the parent sample catalogue obtained with the VLA and whether a detection is made with a snapshot observation with the VLBA.

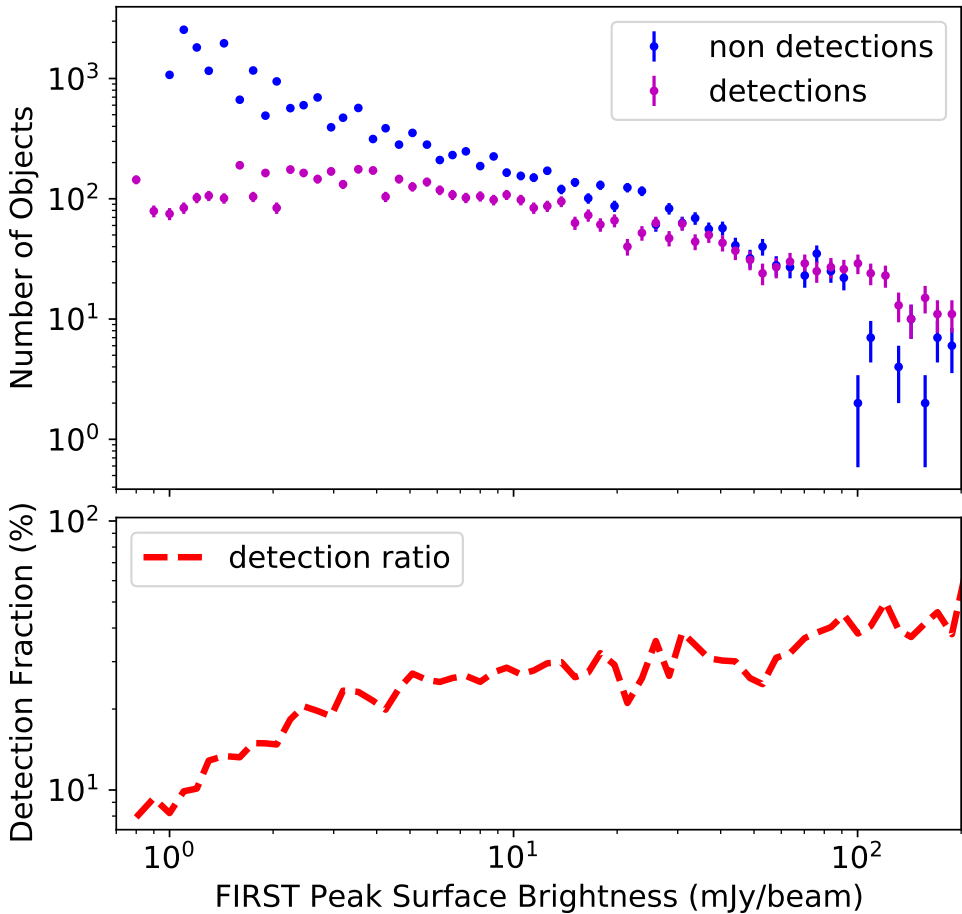
From a straight comparison of the number of sources observed and detected, we see that the mJIVE–20 survey has an overall detection rate of  $19.9 \pm 2.9$  per cent, where the uncertainty is calculated from Poisson statistics. However, to investigate

the detection rate of radio sources on VLBI-scales further, we show in Fig. 2.2 the number of radio sources detected and not detected, and the detection fraction for the mJIVE–20 survey, as a function of the FIRST peak surface brightness. We see that above a peak surface brightness of  $80 \text{ mJy beam}^{-1}$  (for a 5.5 arcsec beam size), the majority of radio sources detected in FIRST also have a VLBI counterpart in the mJIVE–20 survey. Below this surface brightness limit, the detection fraction steadily drops from 50 to around 31 per cent at  $5 \text{ mJy beam}^{-1}$ , before steeply falling to 8 per cent at the FIRST detection limit ( $5\sigma$ ) of  $0.8 \text{ mJy beam}^{-1}$ .

This change in the detection fraction as a function of peak surface brightness in the parent sample could result from a number of factors. First, it could be due to a change in the intrinsic compactness of radio sources, where the brightest objects are relativistically beamed towards the observer, that is, a selection effect. This would explain why the majority of radio sources at the highest surface brightness are also detected on VLBI-scales. Also, even for those cases that are not strongly beamed, the depth of the mJIVE–20 survey data is sufficient to detect a weak radio core in the brightest radio sources in FIRST. This is consistent with the results of Herrera Ruiz et al. (2017), who detected with the VLBA a large number of low flux density radio sources ( $\leq 1 \text{ mJy}$ ) that were previously identified using the VLA. Moreover, for those observations the VLBA recovered between 60 and 80 per cent of the VLA flux density. This suggests that a large portion of sources at and below the mJIVE–20 survey detection limit may still be dominated by AGN activity.

Second, it could be that those sources with a lower surface brightness do have radio emission on VLBI-scales, but the mJIVE–20 survey observations were not deep enough or the  $uv$ -coverage was not sufficient to image them. This would explain the sharp drop in the detection rate towards a lower surface brightness, that is, this is an observational effect. This conclusion is also consistent with extremely deep EVN observations of the GOODS-N field (rms  $9 \mu\text{Jy beam}^{-1}$ ; Radcliffe et al. 2018) when an e-MERLIN/VLA parent sample is used (Muxlow et al. 2020), which finds a similar detection rate of  $25 \pm 6$  per cent (24 out of 94 objects) to the mJIVE–20 survey. Even though the radio source population at this depth is expected to be dominated by star-forming galaxies, as opposed to those harbouring an AGN, it is interesting that providing deep enough observations are carried out, a high fraction of VLBI detections is still found. Therefore, the steep change in the detection fraction towards lower surface brightness for the mJIVE–20 sample is likely a combination of the observing depth and a change in the properties of the underlying source population.

To further investigate the latter, we have looked at several properties of the source population to determine their effect on the mJIVE–20 survey detection fraction. In



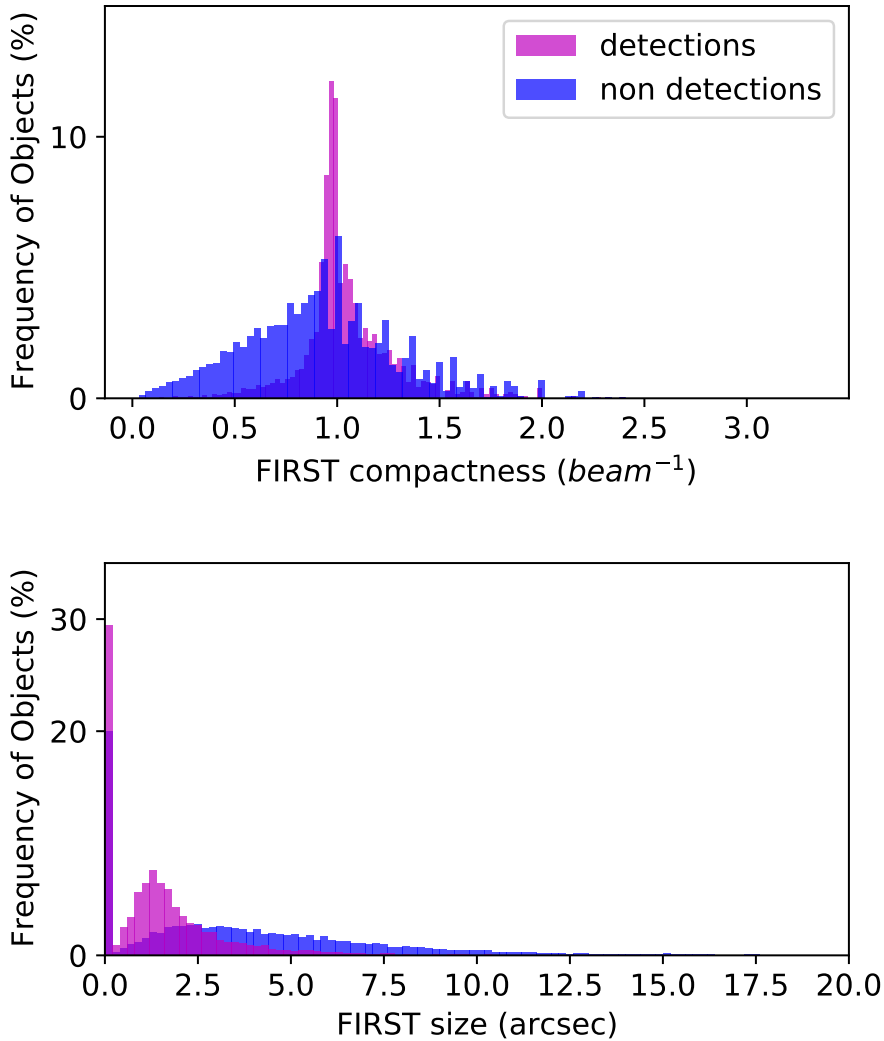
**Figure 2.2** – *Upper panel: The number of objects detected (purple) and not-detected (blue) at VLBI-scales during the mJIVE–20 survey, as a function of peak surface brightness measured from FIRST (5.5 arcsec beam size). Above a peak brightness of around  $90 \text{ mJy beam}^{-1}$ , the number of detections is greater than the number of non-detections. Lower panel: The detection fraction at VLBI-scales during the mJIVE–20 survey, as a function of peak surface brightness measured from FIRST. The detection rate changes from around 50 per cent for the highest surface brightness sources within FIRST, to below 10 per cent at the detection threshold. This is due to the compactness of the radio sources changing as a function of flux density.*



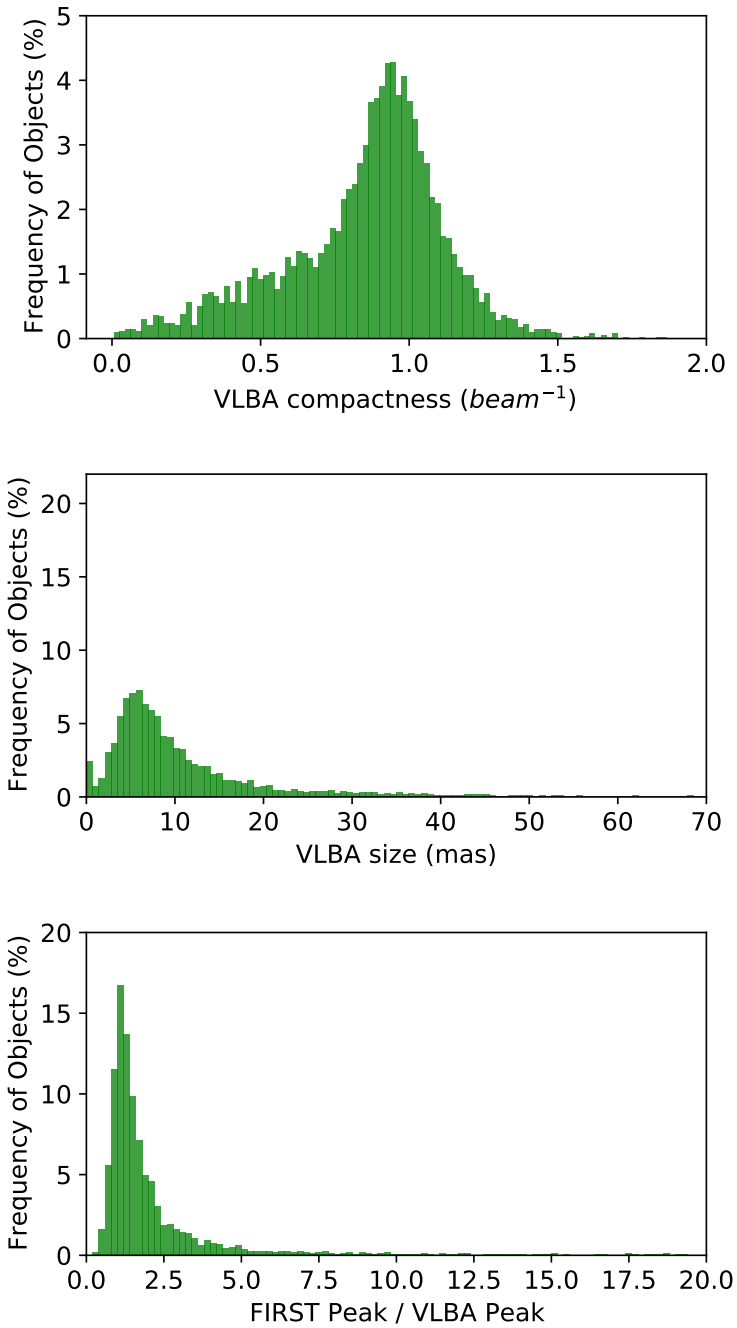
particular, we have analysed the ratio of VLBI peak surface brightness to VLBI integrated flux density (VLBA compactness), the ratio of FIRST peak surface brightness to FIRST integrated flux density (VLA compactness), and the VLBI and FIRST deconvolved major axis (taken from the mJIVE–20 survey catalog obtained using BLOBCAT), which are defined as  $V_{\text{comp}}$ ,  $F_{\text{comp}}$ ,  $V_{\text{size}}$  and  $F_{\text{size}}$ , respectively. Distributions for the VLA compactness and size are presented in Fig. 2.3. The same distributions for the VLBA properties, and also the ratio of peak VLBI surface brightness to the peak FIRST surface brightness, which we refer to as  $P_{F/V}$  (radio source compactness), are presented in Fig. 2.4.

As expected, those objects with a higher compactness and a smaller size on VLA-scales are more likely to also have a detection on VLBI-scales. For the VLA compactness, the mean (with standard deviation) and median of the distributions for the VLBI detections and non-detections are 1.05 ( $\sigma = 0.24$ ) and 1.00 beam<sup>-1</sup>, and 0.92 ( $\sigma = 0.38$ ) and 0.92 beam<sup>-1</sup>, respectively. In the case of the size, the mean (with standard deviation) and median of the distributions for the VLBI detections and non-detections are 1.60 ( $\sigma = 2.01$ ) and 1.24 arcsec, and 4.10 ( $\sigma = 4.10$ ) and 3.27 arcsec, respectively. Also, we find that those objects detected on VLBI-scales by the mJIVE–20 survey tend to have a higher compactness and smaller intrinsic size; the mean (with standard deviation) and median for the compactness and size on VLBI-scales is 0.85 ( $\sigma = 0.27$ ) and 0.9 beam<sup>-1</sup>, and 11.5 ( $\sigma = 13.0$ ) and 7.7 mas. Finally, we find that  $P_{F/V}$  has a mean of 3.7, with a standard deviation of 13.4, and a median of 1.4.

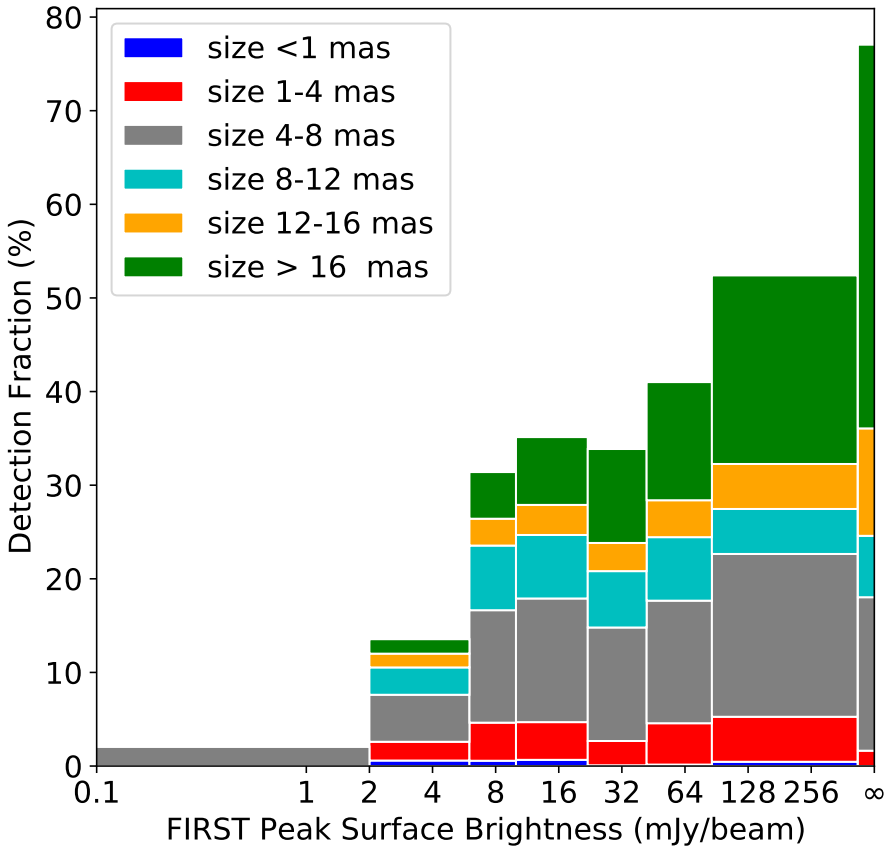
Although the mean values of the above parameters tend to be consistent for the samples of detections and non-detections on VLBI-scales, given the large scatter for each parameter, the median values point toward compact sources being more likely to be detected. Therefore, we now investigate the distribution of sources detected, as a function of their VLBI size,  $V_{\text{size}}$ . In Fig. 2.5, we show the mJIVE–20 survey detection fraction as a function of FIRST peak surface brightness, where each bin is also divided into bins of  $V_{\text{size}}$ . We consider all objects with a major axis larger than 16 mas as extended on VLBI-scales, since this is equivalent to the average beam size of the mJIVE–20 survey. It is interesting that in the low surface brightness regime ( $< 2$  mJy beam<sup>-1</sup> in FIRST), only objects with a  $V_{\text{size}}$  of  $< 8$  mas have been detected. Overall, extremely compact objects ( $V_{\text{size}} < 1$  mas) have the least contribution to the detection fraction compared to the more extended objects, implying that almost all radio sources are partially-resolved with the VLBA, that is, they have some level of VLBI structure. Another point from Fig. 2.5 is that for a surface brightness  $> 8$  mJy beam<sup>-1</sup>, the contribution of objects with  $1 < V_{\text{size}} < 16$  mas is almost constant. However, the contribution of those objects with  $V_{\text{size}} > 16$  mas increases.



**Figure 2.3** – The compactness ( $F_{\text{comp}}$ ; upper panel) and major-axis size ( $F_{\text{size}}$ ; lower panel) in the FIRST survey for the detection (purple) and non-detection (blue) samples in the mJIVE-20 survey.



**Figure 2.4**— The compactness ( $V_{\text{comp}}$ ; upper panel) and major-axis size ( $V_{\text{size}}$ ; middle panel) for objects detected in the *mJIVE*–20 survey. Also shown is the ratio of the *FIRST* and *mJIVE*–20 survey peak surface brightness ( $P_{F|V}$ ; lower panel).



**Figure 2.5** – The detection fraction for the mJIVE-20 survey, binned by peak surface brightness in FIRST. The different colours show the detection fraction in bins of  $V_{\text{size}}$  (deconvolved major axis of a single 2-dimensional elliptical Gaussian fit to the imaging data).

This is likely due to the signal-to-noise ratio of the brightest radio sources also being high, and therefore, we are more sensitive to any possible extended emission. Similarly, extended objects make up a smaller fraction of the lowest surface brightness part of the distribution because here we are mainly sensitive to the most compact emission from an object.

Overall, we find that VLBI detections are more likely from the most compact radio sources observed at arcsec-scales, and that the emission observed on VLBI-scales for around 65 to 80 per cent of the objects has a size in the range of 1 to 16 mas.

## 2.3 VLBI-detected radio source counts

In this section, we determine the source counts of radio sources detected during the wide-field mJIVE–20 survey. Our aim is to estimate the number of VLBI-detected radio sources as a function of integrated flux density, so that we can make robust estimates for the likely source counts from future surveys with current (EVN and VLBA) and next generation (SKA-VLBI and ngVLA) VLBI arrays.

We first calculate the sky-area observed during the mJIVE–20 survey, which due to the primary beam attenuation of the VLBA antennas is also a function of source surface brightness. Next we determine the completeness of the catalogue by making simulations of mock mJIVE–20 survey data and running the source detection procedure. With the sky area and an estimate of the completeness in hand, we then determine the normalized Euclidean and differential source counts for VLBI-detected radio sources. Here, we refer to the entire sample of 24 903 radio sources observed during the mJIVE–20 survey as the (FIRST) parent population, and the 4 965 radio sources detected with the VLBA as the mJIVE–20 population.

### 2.3.1 The mJIVE–20 survey sky area

Calculating the radio source density on the sky requires some knowledge of the sky area being investigated. As can be seen in Fig. 2.1, each mJIVE–20 survey observation consisted of four different pointing positions around a central calibrator source (Deller & Middelberg 2014). This resulted in the effective rms noise across the field being non-uniform, due to the primary beam attenuation of the individual antennas and the combination of data from different individual pointings and repeated observations. Also, two different pointing configurations were employed during the mJIVE–20 survey. The first used an offset from the central calibrator of  $\pm 12$  arcmin in RA and Dec, to include all parent population radio sources within 20 arcmin of each pointing centre, and the second used offsets of  $\pm 9.6$  arcmin in RA and Dec, and included all parent population radio sources within 17 arcmin of each pointing centre. A total of 306 unique calibrators were observed, with 252 observations using the set-up with  $\pm 12$  arcmin offsets, and 54 observations using the set-up with  $\pm 9.6$  arcmin offsets from the calibrator.

To measure the sky area, we use a Monte Carlo integration by stone throwing technique, which is based on the mean value theorem designed for numerical integration using random numbers. In this method, the sky area is calculated by the average number of hits in different iterations. The error in measuring the sky area with this approach goes as  $1/\sqrt{N}$ , where  $N$  is the number of thrown stones. We first define a

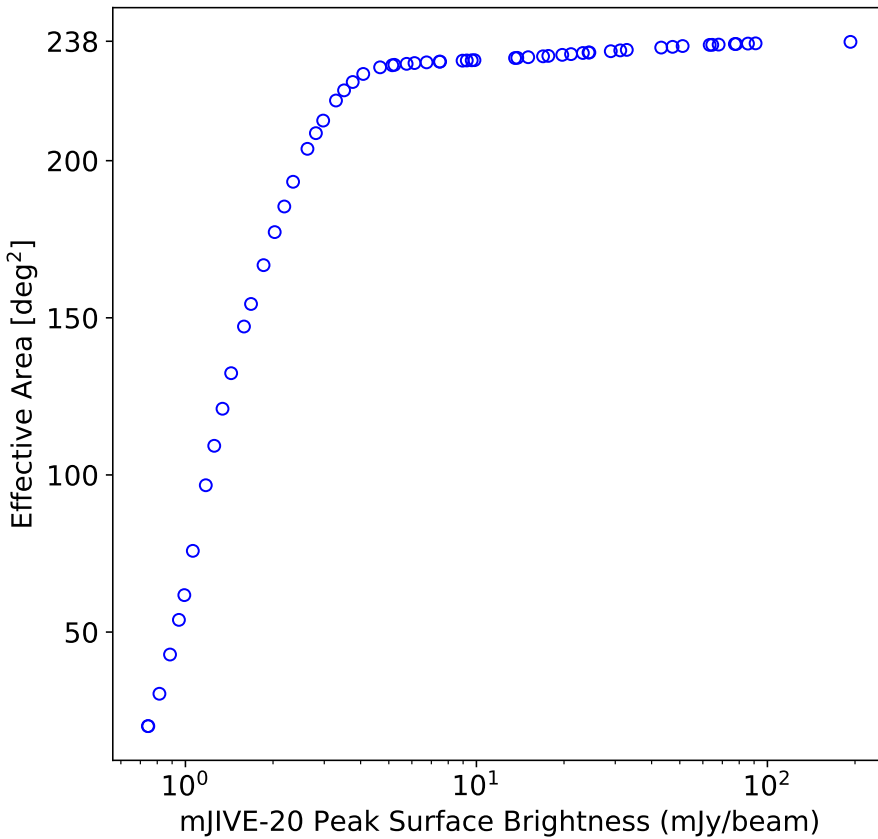
box around the intersected primary beam footprints (circles), shown for example in Fig. 2.1, and measure the area of that box. Then, we throw 1 million points into the box to estimate the size. The ratio of the number of points that are inside any of the circles to the total number of points in the box gives an estimation of the area inside the circles. Considering the area within the individual observation footprints, and the different pointing configurations, the total sky area of the mJIVE–20 survey is found to be  $237.95 \text{ deg}^2$ .

In Fig. 2.6, we show how the effective area changes as a function of the minimum peak surface brightness of an mJIVE–20 population radio source (assuming a local signal-to-noise ratio of 6.75). We see that at a surface brightness of around  $4 \text{ mJy beam}^{-1}$ , there is a knee in the effective sky area (corresponding to  $227 \text{ deg}^2$ ; 95 per cent of the total sky observed), which is roughly equivalent to 25 times the typical rms map noise of  $150 \mu\text{Jy beam}^{-1}$  at each pointing centre.

### 2.3.2 The mJIVE–20 survey completeness

In order to calculate the source counts of the mJIVE–20 population to the lowest flux densities, we also have to consider the completeness of the catalogue, that is, estimate how many sources may have been missed due to observational effects. To do this, we have made mock visibility data sets, which then go through the same imaging and source detection process as the mJIVE–20 survey data. By comparing the input and output mock catalogues, we determine the completeness as a function of flux density.

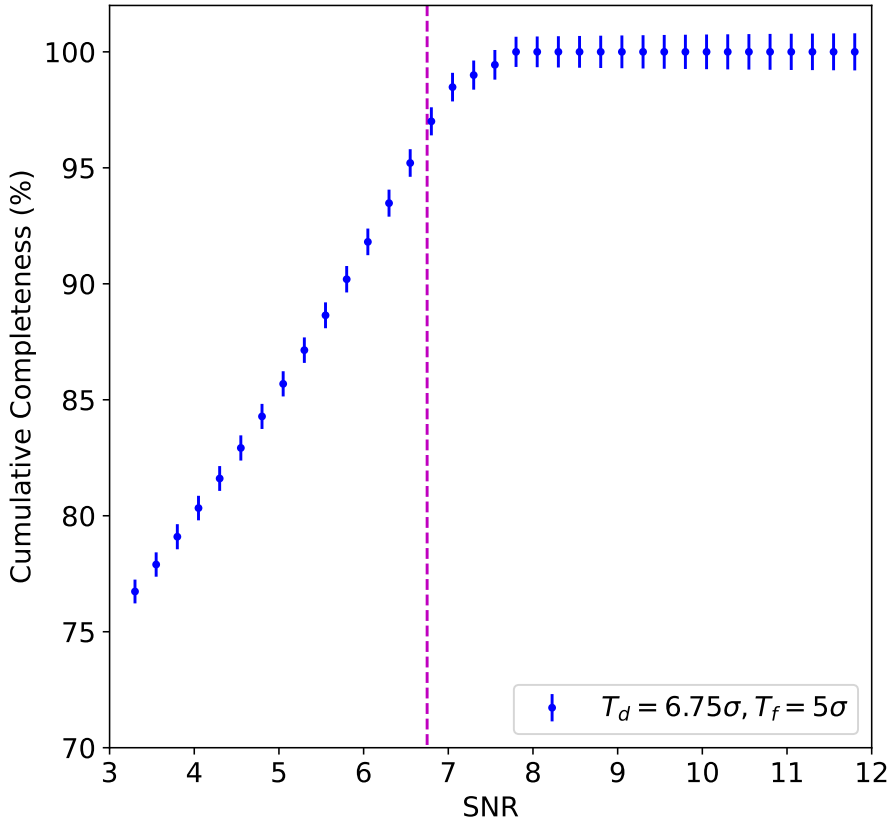
The simulated data set needed to have similar properties to the radio sources in the mJIVE–20 population catalogue. Therefore, we implemented a Monte Carlo approach to define the simulated radio source size (either delta or 2-dimensional elliptical Gaussian function), peak surface brightness, position angle, and a random  $x, y$  position with respect to the phase centre, given the actual distribution of these parameters in the mJIVE–20 catalogue (see Chapter 3). The generated sources were then converted to mock visibilities using the Common Astronomy Software Applications (CASA; McMullin et al. 2007) package. The simulation tool in CASA allows as an input an existing interferometric visibility data set that can be used to obtain the position of the antennas and other observational settings (such as frequency and time sampling, total integration time). Therefore, we used the actual mJIVE–20 survey interferometric visibility data sets when converting the simulated model radio sources into mock visibility data. The final step in the process used the `tclean` task within CASA to produce de-convolved clean images. As with the mJIVE–20 survey data, the images were centred on the position of the surface brightness peak, had a size of  $1024 \times 1024$  pixels and a pixel-scale of  $0.75 \text{ mas pixel}^{-1}$ . The dirty images



**Figure 2.6** – The effective area of the mJIVE–20 survey, as a function of minimum detectable radio source surface brightness (for a detection threshold of  $6.75\sigma$ ). The total observed area in the mJIVE–20 survey is  $237.95 \text{ deg}^2$ , but changes as a function of source brightness due to the primary beam attenuation and the combination of the visibility data from different pointings and repeated observations.

were cleaned for a maximum of 1000 iterations or until a threshold of  $0.2 \text{ mJy beam}^{-1}$  was reached, while masking was applied to the centre of the image, with a radius of 25 pixels.

To identify and measure the flux density of the sources in our mock imaging data, we have used BLOCAT (Hales et al. 2012), the same post-processing object detection package used as part of the mJIVE–20 survey. BLOCAT searches for islands of pixels that could represent a possible object considering the signal-to-noise ratio of a given pixel. This is accomplished using the surface brightness distribution of the 2-dimensional imaging data and the background rms noise as the input parameters. It requires the user to set a threshold for the minimum detection signal-to-noise ratio



**Figure 2.7** – The cumulative completeness for a set of simulated mock VLBA observations that have the same characteristics as the mJIVE–20 survey data. These have been analyzed using the BLOBCAT object detection algorithm, with the same parameters for the detection thresholds ( $T_d$  and  $T_f$ ; see text for details) used for the mJIVE–20 survey.

(known as  $T_d$ ) as well as a cutoff ratio (known as  $T_f$ ) for flooding the islands. Based on simulations carried out by Deller & Middelberg (2014),  $T_d$  and  $T_f$  were set to  $6.5\sigma$  and  $5\sigma$ , respectively, during the mJIVE–20 survey. However, only those sources with a peak surface brightness above  $6.75\sigma$  were added to the final mJIVE–20 survey catalogue, so we adopt that value for  $T_d$  in our simulations. In Fig. 2.7, we show the cumulative completeness of the mJIVE–20 survey as a function of signal-to-noise ratio from 13 500 simulated radio sources that have a signal-to-noise ratio between 3 and 15. The magenta dashed line shows the  $6.75\sigma$  detection threshold of the mJIVE–20 survey. We find from our simulations that BLOBCAT is 97 per cent complete at a threshold of  $6.75\sigma$  and reaches full completeness at a threshold of  $7.8\sigma$  for this data set.



From Fig. 2.7 we are able to calculate the completeness of the mJIVE–20 survey, given the observational set-up (signal-to-noise ratio,  $uv$ -coverage, image de-convolution) and the ability of the object detection algorithm to identify a representative sample of realistic radio sources. This allows us to define a completeness correction factor for each flux density bin, where those bins below 100 per cent completeness are divided by the completeness given in Fig. 2.7; for this, we assume that the radio sources are mostly unresolved, which is consistent with the results presented in the previous section.

### 2.3.3 A note on the resolution bias of the parent and mJIVE–20 population samples

As the mJIVE–20 survey is not a blind search on VLBI-scales, but is instead a targeted search of known FIRST radio sources, the completeness of that parent population can also affect the final source counts of VLBI-detected radio sources. In fact, it is known that below a flux density of around 2 mJy, the completeness of the FIRST survey becomes progressively worse (White et al. 1997), and the observed source counts dramatically drop-off from what is predicted from models of radio source evolution and observations of deep fields at the same frequency (e.g., Prandoni et al. 2018). This results in the source counts of FIRST being lower by a factor of between 0.9 and 0.4 towards the faint end of the flux density distribution. However, as this effect is thought to be due to faint and extended radio sources being resolved out by the 5.5 arcsec synthesised beam of the VLA (B-configuration; see White et al. 1997 for further discussion), it is almost certainly the case that such objects would also be resolved out at VLBI-scales (in the absence of significant variability). Therefore, we do not correct for the resolution bias of the FIRST survey.

However, there will be a resolution bias associated with the mJIVE–20 survey data, as the detectability and measured flux density of a given object will be dependent on the structure (compactness) and also the  $uv$ -coverage of the observations. For this reason, the source counts from deep VLBI observations have tended to use the total flux density from lower resolution imaging (e.g. the VLA; Middelberg et al. 2013; Herrera Ruiz et al. 2018) when discussing VLBI-detected radio sources with respect to the radio source population in general. Here, we do use the flux density recovered on VLBI-scales, as we are not primarily interested in testing galaxy formation models or to separate AGN and star-forming galaxies, but instead, our goal is to understand the number of radio sources that could be detected with wide-field VLBI surveys. In Section 2.4, we will discuss the effect of VLBI resolution bias further and present ways to mitigate it in the future.

### 2.3.4 Euclidean-normalized and differential source counts

We now calculate the source counts of the radio sources detected during the mJIVE–20 survey. We also compare these with the source counts obtained from studies of compact radio sources selected at higher frequencies. In Table 2.1, we present the number of sources per flux-density bin, and the correction factor that is determined by taking the effective sky area and completeness of the survey into account. Here, we also present the (corrected) Euclidean-normalized source counts for the mJIVE–20 survey.

In Fig. 2.8, we present these source counts as a function of flux density, which we see are well approximated by a power-law down to flux densities of around 2 mJy, below which there is evidence of a knee in the distribution. This downturn is likely associated with the change in the parent population of FIRST radio sources described in the previous section. Below 1 mJy, the Euclidean-normalized source counts become flatter, but are also quite noisy. At this stage, it is not clear if this is a systematic or a real effect, similar to the flattening of the radio source counts seen in lower angular resolution observations at this frequency; this is attributed to a change in the radio source population from AGN to star-forming dominated systems (Condon et al. 2012; Prandoni et al. 2018). From a power-law fit to the data presented in Fig. 2.8, we find that the Euclidean-normalized source counts can be described by,

$$n(S) = (2.4 \pm 0.1) \left( \frac{S_{1.4}}{1 \text{ mJy}} \right)^{(0.695 \pm 0.013)} \text{ Jy}^{1.5} \text{ sr}^{-1}, \quad (2.1)$$

where  $n(S)$  is the differential number of sources (Euclidean-normalized) and  $S_{1.4}$  is the flux density at 1.4 GHz.

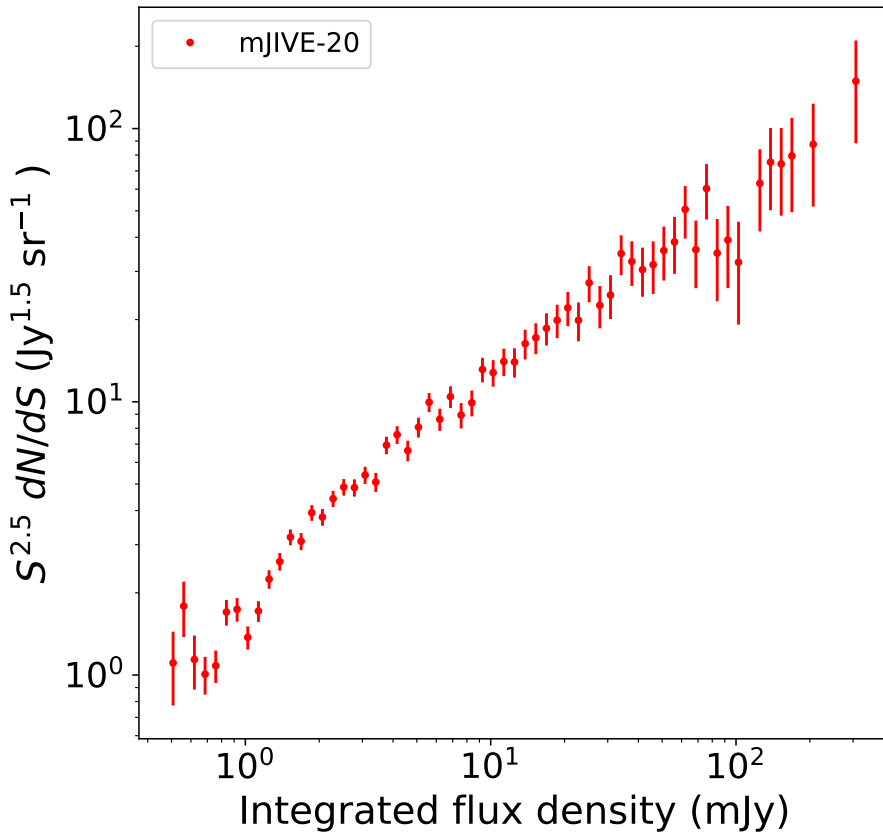
In Fig. 2.9, we present the differential source counts as a function of flux density for the FIRST parent population and the mJIVE–20 population. From a fit to the data between 1 and 100 mJy, we find that the FIRST parent population can be described by the power-law,

$$n(S) = (80.6 \pm 5.5) \left( \frac{S_{1.4}}{100 \text{ mJy}} \right)^{-(1.77 \pm 0.02)} \text{ mJy}^{-1} \text{ sr}^{-1}, \quad (2.2)$$

and that the mJIVE–20 parent population can be described by the power-law,

$$n(S) = (19.4 \pm 1.4) \left( \frac{S_{1.4}}{100 \text{ mJy}} \right)^{-(1.74 \pm 0.02)} \text{ mJy}^{-1} \text{ sr}^{-1}. \quad (2.3)$$

As in the previous case,  $n(S)$  is the differential number of sources and  $S_{1.4}$  is the



**Figure 2.8** – The Euclidean-normalized source counts for VLBI-detected radio sources from the mJIVE-20 survey.

flux density at 1.4 GHz. From these fits, and by inspecting Fig. 2.9, we see that the power-laws for the parent and mJIVE-20 populations are diverging toward lower flux densities, which would be consistent with the change in the detection fraction seen in Fig. 2.2.

Taking the result for the mJIVE-20 survey, we compare these source counts with those for a complete sample of 117 flat-spectrum radio sources selected and observed at 4.85 GHz with the VLA (McKean et al. 2007). Such sources are assumed to be compact (size < 170 mas; Myers et al. 2003), given their flat-radio spectra (due to the super-position of synchrotron self-absorption from many homogeneous emitting regions) and the higher selection frequency tending to favour core-dominated radio sources. Therefore, flat-spectrum radio sources are prime targets for VLBI observing programmes as they are expected to dominate the compact radio source population.

We find that the slope of the source counts is steeper [ $\eta = -2.06 \pm 0.01$ ; where  $n(s) = k S^\eta$ ], and the normalization is smaller ( $k = 6.91 \pm 0.42$ ) for flat-spectrum radio sources selected at 4.85 GHz (Jodrell Bank Flat-spectrum radio source sample; JBF; McKean et al. 2007), when compared to the results for the VLBI-detected radio sources at 1.4 GHz. However, by assuming a mean spectral index of  $\alpha_{1.4}^{4.85} = -0.09$  (where  $S_\nu \propto \nu^\alpha$ , and  $\nu$  is the observing frequency; McKean et al. 2007) for the JBF sample, we have calculated the expected number of compact radio sources at 1.4 GHz down to a flux density of 1 mJy. Interestingly, we find that the sky-density of flat-spectrum radio sources selected at 4.85 GHz is very similar to that of the VLBI-detected radio sources from the mJIVE-20 survey, but is a factor  $1.22 \pm 0.12$  higher. Such similar source counts suggests that both samples are drawn from a similar population, even though one is selected based on the radio spectra, and the other on compactness at VLBI-scales. This has implications when choosing an appropriate observing frequency for an all-sky VLBI survey, which we discuss in the next section.

$S_l$ (mJy)	$S_u$ (mJy)	$\bar{S}$ (mJy)	Number	Corr.	Counts (sr <sup>-1</sup> Jy <sup>1.5</sup> )
0.51	0.56	0.534	11	11.858	1.107±0.334
0.57	0.62	0.595	19	11.858	1.787±0.410
0.63	0.68	0.657	20	7.836	1.140±0.255
0.69	0.75	0.725	40	5.553	1.007±0.159
0.76	0.83	0.799	54	4.416	1.082±0.147
0.84	0.92	0.884	86	3.853	1.700±0.183
0.93	1.02	0.976	102	3.138	1.741±0.172
1.03	1.13	1.080	106	2.460	1.373±0.133
1.14	1.24	1.190	131	2.178	1.715±0.150
1.25	1.38	1.313	163	1.966	2.244±0.176
1.39	1.52	1.454	185	1.798	2.603±0.191
1.53	1.68	1.603	224	1.616	3.198±0.214
1.69	1.86	1.774	194	1.541	3.087±0.222
1.87	2.06	1.961	230	1.427	3.927±0.259
2.07	2.27	2.164	203	1.342	3.785±0.266
2.28	2.51	2.394	213	1.283	4.420±0.303
2.52	2.78	2.652	207	1.231	4.875±0.339
2.79	3.07	2.930	190	1.168	4.852±0.352
3.08	3.40	3.227	188	1.140	5.394±0.393
3.41	3.75	3.582	153	1.118	5.083±0.411
3.76	4.15	3.959	186	1.086	6.940±0.509

4.16	4.59	4.360	179	1.070	7.578±0.566
4.61	5.07	4.854	134	1.057	6.632±0.573
5.08	5.61	5.349	143	1.045	8.073±0.675
5.62	6.20	5.902	154	1.036	9.964±0.803
6.21	6.85	6.527	115	1.033	8.633±0.805
6.86	7.57	7.220	119	1.031	10.441±0.957
7.58	8.37	7.982	88	1.032	8.939±0.953
8.38	9.25	8.791	85	1.030	9.920±1.076
9.26	10.23	9.797	94	1.029	13.135±1.355
10.24	11.29	10.736	81	1.028	12.787±1.421
11.33	12.47	11.896	76	1.028	14.038±1.610
12.51	13.80	13.073	66	1.026	13.986±1.722
13.86	15.25	14.573	64	1.026	16.320±2.040
15.28	16.85	15.982	60	1.026	17.161±2.215
16.90	18.65	17.780	55	1.026	18.571±2.504
18.71	20.59	19.624	51	1.022	19.879±2.784
20.68	22.78	21.701	49	1.023	22.077±3.154
22.85	25.15	23.926	38	1.022	19.873±3.224
25.31	27.84	26.594	44	1.020	27.246±4.108
27.98	30.57	29.148	32	1.020	22.553±3.987
30.80	34.00	32.246	30	1.018	24.583±4.488
34.13	37.52	35.756	36	1.017	34.863±5.811
37.66	41.52	39.615	29	1.016	32.614±6.056
41.62	45.72	43.285	24	1.015	30.452±6.216
45.97	50.07	48.109	21	1.015	31.741±6.926
50.93	55.93	53.434	20	1.012	35.774±7.999
56.82	61.93	59.833	18	1.013	38.466±9.066
62.29	68.49	65.218	21	1.011	50.615±11.045
69.47	74.11	71.762	13	1.008	36.042±9.996
75.97	83.62	79.061	19	1.006	60.303±13.834
84.83	90.96	87.906	9	1.005	34.991±11.664
92.70	100.87	96.787	9	1.007	39.090±13.030
104.92	112.86	109.495	6	1.005	32.393±13.224
125.63	137.61	132.954	9	1.003	63.024±21.008
140.59	151.53	147.273	9	1.003	75.359±25.120
153.07	166.88	160.351	8	1.003	74.221±26.241
169.90	184.55	179.800	7	1.003	79.416±30.017
210.57	225.90	217.562	6	1.004	87.566±35.749

310.54	339.24	321.093	6	1.001	149.152±60.891
--------	--------	---------	---	-------	----------------

**Table 2.1** – The Euclidean normalized source counts for the mJIVE–20 survey. The columns from left to right are, the lower ( $S_l$ ) and upper ( $S_u$ ) bounds of a flux density bin, the mean flux density of a bin ( $\bar{S}$ ), the number of sources in each bin ( $N$ ), the correction factor (taking into account the sky area and the completeness; *Corr.*) and the resulting Euclidean normalized source counts (*Counts*). Note that 33 sources with flux densities  $> 340$  mJy are not included here.

## 2.4 Prospects for all-sky VLBI surveys

In this section, we use the differential source counts determined above to estimate the number of radio sources that could be detected from an all-sky survey at mas-scale resolution with VLBI. We focus primarily on what would be achievable with the current VLBA and EVN (using only the smaller dish radio telescopes of the array in a hypothetical survey mode), and briefly discuss the expectations from next generation instruments, like SKA-VLBI. For our calculations, we adopt a survey strategy that maximizes the number of radio sources detected on VLBI-scales.

### 2.4.1 Characteristics of the arrays

Our ability to survey the sky is dependent on the effective field-of-view of the individual dishes and their sensitivity. This requires some knowledge of the primary beam models, which we assume are consistent with a circular aperture that has a Gaussian illumination pattern, such that the primary beam diameter at the 50 per cent attenuation point is given by

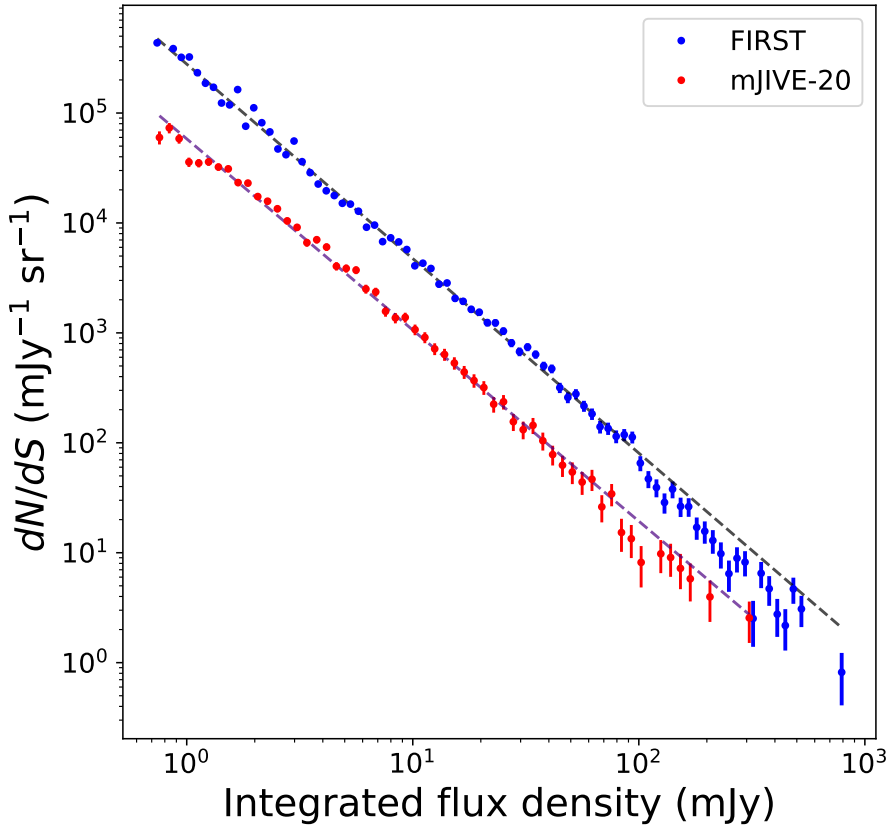
$$\theta_{\text{FWHM}} = 1.22\lambda/D, \quad (2.4)$$

where  $D$  is the physical diameter of the telescope and  $\lambda$  is the observing wavelength.

The sensitivity of an interferometric array is defined from the rms thermal noise fluctuations, given by

$$\sigma_{\text{rms}} = \frac{1}{\eta_{\text{cor}} A_{\text{eff}} \sqrt{N_p N_A (N_A - 1) \Delta\nu \Delta t_{\text{int}}}}, \quad (2.5)$$

where  $T_{\text{sys}}$  is the system temperature (assuming identical antenna systems),  $k$  is the Boltzmann constant,  $\eta_{\text{cor}}$  is the correlator efficiency (assumed to be 0.89 for 2-bit sampling),  $A_{\text{eff}}$  is the effective collecting area of the filled aperture,  $N_p$  is the number of polarisations (assumed to be 2),  $N_A$  is the number of antennas in the array,  $\Delta\nu$  is the bandwidth in frequency and  $\Delta t_{\text{int}}$  is the integration time. Typically, the antenna



**Figure 2.9** – The differential source counts for VLBI-detected radio sources using the *mJIVE-20* survey (red), with the *FIRST* parent population (blue) for comparison. The dashed lines are the best-fit power-law models to the *mJIVE-20* survey and *FIRST* data.

forward gain is defined by the System Equivalent Flux Density (SEFD), which is given by

$$\text{SEFD} = \frac{2kT_{\text{sys}}}{A_{\text{eff}}}. \quad (2.6)$$

In the cases where the antenna systems are not identical, the SEFD on a given baseline between antennas 1 and 2 is expressed as  $\sqrt{\text{SEFD}_1 \times \text{SEFD}_2}$ . Note that the VLBA and EVN are assumed to operate at central observing frequencies of 1.4 and 1.7 GHz, respectively, given the available frequency coverage of the individual receiver systems.

We see from equation 2.4 that the observable solid angle is proportional to  $\lambda^2$ , and therefore, surveys of radio sources at longer wavelengths are typically more efficient.

This is also because radio sources tend to have a higher flux density toward lower frequencies. However, we see from our comparison of the source counts for the mJIVE–20 and JBF surveys above that the expected number of detectable radio sources on VLBI-scales is likely quite similar at 1.4/1.7 and 5 GHz. Therefore, either frequency is likely viable. However, the level of radio frequency interference at 5 GHz is currently much less than at 1.4/1.7 GHz, and the usable bandwidth is currently a factor of two to four times better. This, coupled with a slightly better receiver response at 5 GHz, results in around a factor of between 1.5 and 2 improvement in overall sensitivity at 5 GHz. Unfortunately, as this comes at the expense of having to carry out a factor of around 9 to 13 more observations to cover the same area of sky, we focus on 1.4/1.7 GHz as our preferred observing frequency for a potential all-sky survey with VLBI. We note that this choice is based on maximising the number of radio sources detected on VLBI-scales, but other science goals may be better served with the improved angular resolution and frequency coverage afforded at 5 GHz.

In addition, we also consider a hypothetical VLBI array that includes 20 antennas with the same characteristics as the SKA-MID (located in the Karoo desert of South Africa). We do not discuss the locations of these antennas, but we assume that they are positioned in 20 different locations in Botswana, Ghana, Kenya, Madagascar, Mauritius, Mozambique, Namibia, South Africa and Zambia (the partner African VLBI Network countries), to provide mas-scale angular resolution, with one dish per location. Also, we assume that 512 MHz of frequency bandwidth will be usable within the SKA Band 2, operating between 0.95 and 1.76 GHz, due to radio frequency interference.

In Table 2.2, we summarise the properties of the three arrays that we consider here.

### 2.4.2 Prospects for in-beam calibration

Our ability to calibrate VLBI observations is currently limited by the density of objects that are suitable for correcting the complex antenna gains as a function of time. Also, given the large changes in phase, due to the long baselines involved with VLBI observations, calibrators are often required to be very close to the target field ( $< 2$  deg separation). There are currently 17 432 compact radio sources that have been identified as calibrators for VLBI experiments (1 in  $2.37 \text{ deg}^2$  over the whole sky; Radio Fundamental Catalog; Petrov 2021). However, astrometric accuracy scales with the calibrator distance (Pradel, Charlot & Lestrade 2006), so additional calibrators will improve the overall astrometric and imaging quality. It is this rationale that made in-beam calibration ( $< 0.4$  deg calibrator-to-target separation) a key component of the mJIVE–20 survey.



Array	$N_A$	Antennas	Diameter (m)	Freq. (GHz)	$\Delta\nu$ (MHz)	FoV (deg <sup>2</sup> )	SEFD (Jy)
VLBA	10	Sc, Hn, NI, Fd, La, Kp, Pt, Ov, Br, Mk	25	1.4	256	0.262	365
EVN/e-MERLIN	18	Mc, On, Tr, W1, Nt, Sh, Ur, Hh, Sv, Zc, Bd, Ir, JB2, Cm, Da, De, Kn, Pi	25, 32	1.7	128	0.117	485
SKA-VLBI	20		15	1.4	512	0.728	235

**Table 2.2** – The properties of the two arrays considered here for an all-sky survey with VLBI. The antenna names are given using their standard abbreviation. The portion of EVN/e-MERLIN listed here has a combination of  $10 \times 25$ - and  $8 \times 32$ -m telescopes, but for our field-of-view (FoV) calculations, we use a diameter of 32 m. The SEFD is the average for the given array. We also include a hypothetical VLBI array that includes 20 antennas with the characteristics of the SKA-MID dish design, which provides mas-scale angular resolution (excluding the SKA-MID core).

With the development of wide-bandwidth observations, radio sources at lower flux densities can now be used as calibrators, when the amplitude and phase of the visibilities are well calibrated as a function of frequency. For example, the recommended phase referencing time at 1 to 8 GHz in good weather is 300 s. For the VLBA and EVN arrays presented in Table 2.2, this is equivalent to a baseline sensitivity of  $\sigma_{\text{rms}} = 0.74$  and  $0.98 \text{ mJy beam}^{-1}$  at 1.4 and 1.7 GHz, respectively. Therefore, compact radio sources with flux-densities  $> 7.4$  and  $> 9.8 \text{ mJy}$  (for  $10\sigma_{\text{baseline}}$ ) can be regarded as good phase reference sources for the VLBA and EVN, respectively. For our hypothetical SKA-VLBI array, the baseline sensitivity is  $\sigma_{\text{rms}} = 0.39 \text{ mJy beam}^{-1}$ , which corresponds to a minimum calibrator flux density of  $> 3.9 \text{ mJy}$  (for  $10\sigma_{\text{baseline}}$ ).

Using the differential source counts of VLBI-detected radio sources given by equation 2.3, we calculate the expected sky density of radio sources needed for phase referencing from pointed observations, and determine the fraction that can be used for in-beam calibration. Note that the latter requires increasing the lower limit on the flux density by a factor of 2, due to the attenuation of the primary beam of the antennas. For the VLBA, we find that the sky density of radio sources with flux densities  $> 7.4$  (pointed) and  $> 14.8 \text{ mJy}$  (in beam) is  $5.5$  and  $3.3 \text{ deg}^{-2}$ , respectively. This corresponds to an in-beam calibrator density of  $0.87 \text{ sources beam}^{-1}$ . In the case of the EVN, we expect that the sky density of radio sources with flux densities  $> 9.8$  (pointed) and  $> 19.6 \text{ mJy}$  (in beam) is  $4.5$  and  $2.6 \text{ deg}^{-2}$ , respectively, with an in-beam density of  $0.30 \text{ sources beam}^{-1}$  (for the 32-m antennas). Finally, for our hypothetical SKA-VLBI array, the sky density of radio sources with flux densities  $> 3.9$  (pointed) and  $> 7.8 \text{ mJy}$  (in beam) is  $8.8$  and  $5.3 \text{ deg}^{-2}$ , respectively. Therefore, the number of radio sources that can be used for in-beam calibration with the SKA-VLBI array would be  $3.9 \text{ sources beam}^{-1}$ .

Overall, we find that 1 in 3 pointings with the EVN will likely have an in-beam calibrator, whereas this is likely the case for almost all pointings with the VLBA. In the case of the SKA-VLBI array, we predict that there will be multiple in-beam calibrators available, which can be important for extremely precise astrometric applications of VLBI (e.g., Dodson & Rioja 2018).

### 2.4.3 Expected number of detected sources

We now calculate the number of radio sources that can be found from an all-sky survey with VLBI. For this, we integrate equation 2.3 to an appropriate flux-density limit. We see from Fig. 2.8 that our differential source counts of VLBI-detected radio sources is robust to around  $1 \text{ mJy}$ , below which the source counts become much more noisy. In principle, we could extrapolate our calculations below  $1 \text{ mJy}$ , but given the

expected change in the radio source population at these flux densities, this would likely result in an over-estimate of the total number of radio sources that we would detect. Therefore, we use a 1.4/1.7 GHz flux-density limit of 1 mJy for our calculations.

We also assume that, given the short amount of time used per observation, any survey would have similar noise properties to the mJIVE–20 survey, and that the same detection threshold would likely be needed ( $6.75\sigma$ ); this would result in a similar completeness of 97 per cent, as we determined from Fig. 2.7, which we fold into our calculations. We also assume that the individual pointings are separated by a beam width, in a similar configuration to what was presented in Fig. 2.1. In a future work, we will simulate the effect of different pointing strategies and noise realizations. However, for our current pointing strategy, we see that the absolute value of the noise changes across the field. Therefore, to make a  $6.75\sigma$  detection at a flux density (point-source) limit of 1 mJy, means having an average thermal noise of  $150 \mu\text{Jy beam}^{-1}$ . This corresponds to a thermal noise at the pointing centre (where the primary beam response is maximum) of  $130 \mu\text{Jy beam}^{-1}$  ( $1\sigma_{\text{TMS}}$ ). From the properties of the three arrays given in Table 2.2, and using equations 2.5 and 2.6, we find that the on-source integration time needed to reach this sensitivity would be 215, 225 and 15 s for the VLBA, EVN and a hypothetical SKA-VLBI array, respectively. This highlights the improvement in sensitivity that is potentially available with SKA-VLBI, given the advances in antenna and receiver design for this next generation instrument. We note that the short integration times needed to reach the required noise levels would allow for flexible scheduling of such observations. Therefore, it would also be straightforward to define optimum scheduling blocks when the target fields are visible to all of the available telescopes. This is more challenging to achieve for long-track observations that are typically used for deep-field science cases.

We also see from equation 2.3 that the slope of the differential number counts of VLBI-detected radio sources is quite shallow ( $\eta = -1.74 \pm 0.02$ ). Therefore, maximizing the number of detected radio sources is best achieved through observing the largest possible sky area, which for our calculations we assume is  $3\pi$  sr (equivalent to about  $31\,000 \text{ deg}^2$ ). Given a flux-limited survey to 1 mJy, and a completeness of 97 per cent at this limit, such a survey could potentially detect  $(7.2 \pm 0.9) \times 10^5$  radio sources on VLBI-scales, a factor of about 30 times more than is currently known.

Finally, we estimate the total time such a survey would take with the current VLBA and EVN (using only the 32-m antennas), and a next generation instrument, like SKA-VLBI. From Table 2.2, we see the effective field-of-view of each array. From this we estimate that to survey  $3\pi$  sr would take  $1.18 \times 10^5$ ,  $2.64 \times 10^5$  and  $1.18 \times 10^4$  pointings for the VLBA, EVN and a hypothetical SKA-VLBI array, respectively.

Given the integration time to reach the required noise level, this equates to about 7000, 16500 and 180 h of on-source observing time for the VLBA, EVN and SKA-VLBI, respectively. Note that about 2 in 3 of the EVN pointings will also require additional phase referencing calibration (see above), which will add up to around another 30 per cent to the total time needed to complete such a survey with that array. For this reason, such surveys with the EVN are likely prohibitive, whereas with the VLBA and SKA-VLBI they would be very much feasible.

## 2.5 Conclusions

We have analyzed the final catalogue of the mJIVE–20 survey, which observed 24 903 radio sources in the FIRST survey using the VLBA at 1.4 GHz to an rms noise level of about  $150 \mu\text{Jy beam}^{-1}$ , detecting 4 965 radio sources on VLBI-scales. Through comparing the number of detections and non-detections in the mJIVE–20 catalogue, we found that the detection fraction is a strong function of the peak surface brightness (compactness and radio source size) of the objects in the FIRST survey; the detection fraction is over 50 per cent at a peak surface brightness at  $80 \text{ mJy beam}^{-1}$  and falls steadily to about 31 per cent at  $5 \text{ mJy beam}^{-1}$ . Below this, the detection fraction falls sharply to 8 per cent at the detection threshold of both surveys. This is likely due in part to a change in the composition of the radio source population, from AGN to star-formation dominated objects, but is also due to the VLBI observations not being sensitive enough to detect low-level compact emission toward lower flux densities. We found an overall detection fraction of  $19.9 \pm 2.9$  per cent. Given the limited *uv*-coverage of the mJIVE–20 survey observations, we found that those radio sources that are detected tend to be unresolved (median VLBI compactness  $0.9 \text{ beam}^{-1}$ ), with a median size of 7.7 mas (about half a beam-size). Finally, we found that 20 to 35 per cent of the radio sources detected were resolved, with sizes  $> 16$  mas.

From an analysis of the VLBI-detected source counts, we see hints of a similar behaviour in the Euclidean-normalized distribution that has been reported on arcsec-scales, that is, a downturn around a flux density of 2 mJy and then a potential flattening below 1 mJy. Again, this could be due to the expected change in the radio source population at these flux-densities; to confirm this will require a similar number of radio sources to be observed with VLBI to a depth that is about ten times deeper than the mJIVE–20 survey. We also determine the differential number counts for VLBI-detected radio sources, finding that the total number of objects is similar to those of flat-spectrum radio sources selected at higher frequencies, suggesting that they come from a similar population.

From our analysis of the differential source counts, we found that the sky density of suitable phase reference sources is of order 2.6 to 3.3 deg<sup>-2</sup>. This should be sufficient for in-beam phase referencing in the case of around 30 and 90 per cent of the observations carried out with the EVN and VLBA, respectively. However, we found that for a VLBI facility that includes antennas with a similar specification to the SKA-MID design, the expected sky density of phase reference sources is about 5.3 deg<sup>-2</sup>, which equates to multiple in-beam calibration sources for each observation.

Finally, we investigated the number of sources that could be found from all-sky surveys carried out with the VLBA and EVN. From this analysis, we found that a factor 30 more VLBI-detected radio sources could be identified with around 7000 h of observations with the VLBA. However, in the case of the EVN, such surveys would be rather expensive, given the smaller field of view of the antennas in that array. For a hypothetical SKA-VLBI array, such a survey would take a fraction of the time, and could be completed with just 180 h of observations. Our analysis is currently limited by our knowledge of the radio source counts below 1 mJy. More focused surveys of the radio sky with VLBI, down to a limiting sensitivity of around 100 μJy beam<sup>-1</sup> (6σ detection threshold) would be extremely informative in testing models for radio source populations and making robust predictions for the expectations with SKA-VLBI.

## Acknowledgements

This chapter is based on research developed in the DSSC Doctoral Training Programme, co-funded through a Marie Skłodowska-Curie COFUND (DSSC 754315). JPM acknowledges support from the Netherlands Organization for Scientific Research (NWO) (Project No. 629.001.023) and the Chinese Academy of Sciences (CAS) (Project No. 114A11KYSB20170054). The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.

## Data Availability

The data used in this research is publicly available in the mJIVE-20 survey and FIRST databases.

## DECORAS: detection and characterization of radio-astronomical sources using deep learning

Based on "DECORAS: detection and characterization of radio-astronomical sources using deep learning"

S. Rezaei, J. P. McKean, M. Biehl and A. Javadpour

Published in MNRAS, 2022, 510, 5891

### Abstract

We present DECORAS, a deep learning based approach to detect both point and extended sources from Very Long Baseline Interferometry (VLBI) observations. Our approach is based on an encoder-decoder neural network architecture that uses a low number of convolutional layers to provide a scalable solution for source detection. In addition, DECORAS performs source characterization in terms of the position, effective radius and peak brightness of the detected sources. We have trained and tested the network with images that are based on realistic Very Long Baseline Array (VLBA) observations at 20 cm. Also, these images have not gone through any prior de-convolution step and are directly related to the visibility data via a Fourier transform. We find that the source catalog generated by DECORAS has a better overall completeness and purity, when compared to a traditional source detection algorithm. DECORAS is complete at the  $7.5\sigma$  level, and has an almost factor of two improvement in purity at  $5.5\sigma$ . We find that DECORAS can recover the position of the detected sources to within  $0.61 \pm 0.69$  mas, and the effective radius and peak surface brightness are recovered to within 20 per cent for 98 and 94 per cent of the

sources, respectively. Overall, we find that DECORAS provides a reliable source detection and characterization solution for future wide-field VLBI surveys.

### 3.1 Introduction

Machine learning, and in particular deep learning, has been widely used for solving a number of astronomical problems (see Baron 2019 for a recent review). This is because traditional approaches, such as visual inspection or model fitting, can be less effective when characterizing datasets that are growing in both size and complexity. Several machine learning frameworks can be applied in this context, but two have gained the most attention in recent years: supervised and unsupervised learning.

In a supervised learning algorithm, the data acts as an instructor, assisting the model in discovering a relationship between a collection of features and user specified labels. Models are trained to predict the properties of unseen data (Goodfellow, Bengio & Courville 2016). Supervised learning has been used to address various problems in astronomy, for example, the classification of galaxy morphologies in imaging data (Pearson et al. 2019; Nolte et al. 2019), variable star classification using light-curve representation (Becker et al. 2020), and the classification of blazar candidates (Kovačević et al. 2020).

In the case of unsupervised learning, the algorithm simply receives data without target labels or any feedback from the environment. The goal of the analysis is to identify patterns and structures within the data that are then used to make decisions, predict future inputs, or communicate inputs efficiently to another machine (Ghahramani 2004). An example of such an algorithm is an autoencoder, which is a form of unsupervised convolutional neural network (CNN). Autoencoders have been used for a variety of astronomy applications, including real-time transient detection (Sedaghat & Mahabal 2018) and the analysis of gravitationally lensed objects (Hezaveh, Perreault Levasseur & Marshall 2017). They have also been used for outlier detection; see for example Margalef-Bentabol et al. (2020), in which a network is trained with normal data before computing an anomaly score for test query samples. As another example, Pruzhinskaya et al. (2019) presented a method for detecting rare transients or completely new flaring events of unknown physical nature. These are just a few of the many applications of supervised and unsupervised learning in astronomy. For a very brief review and further references, see for example, Biehl et al. (2018).

Here, we focus on a deep learning based approach to study astronomical images that have been made using radio interferometric techniques. In contrast to optical instruments, which capture images of the sky brightness distribution directly, radio

telescopes employ interferometry to calculate the two-dimensional discrete intensity distribution of the sky, known as visibility data. A Fourier transform of the visibility data is then performed to produce an image of the sky. The result of this process is the convolution of the true sky brightness with the point spread function (PSF) of the interferometric array, which is commonly referred to as the dirty image. Due to the incomplete sampling of the interferometric visibility data, the PSF (also referred to as the dirty beam) has strong sidelobes that affect the entire image. This can make it difficult to recover the true sky brightness distribution from interferometric data. A common solution to this problem was presented by Högbom (1974), who developed the CLEAN algorithm, which iteratively performs a deconvolution of the image by representing the underlying source brightness using simple parametric models, such as delta- or truncated Gaussian functions. The final step convolves the model of the source with the clean beam (a Gaussian function), which is then added to the Fourier transform of the residual visibilities.

To characterize the object properties from interferometric images, several commonly used object detection algorithms have been developed (note that these are applied to images that have gone through a prior deconvolution process). PYBDSF (Mohan & Rafferty 2015), BLOBCAT (Hales et al. 2012) and AEGEAN (Hancock, Trott & Hurley-Walker 2018), are all examples of Gaussian fitting source detectors. PROFOUND (Hale et al. 2019) on the other hand does not force any predefined parametric model to the detected sources, but is based on the segmentation of pixels in the neighborhood of the brightest pixel. The benefit of PROFOUND over other source detection algorithms is the more accurate flux recovery for extended sources, as the detection is not based on a specific morphology. From an analysis of simulated observations that match the instrument properties of the Very Large Array (VLA), the completeness and reliability of PYBDSF, AEGEAN and PROFOUND for compact objects detected with a signal-to-noise ratio  $> 4.3$  was found to be less than 85 per cent (Hale et al. 2019).

Machine learning, and in particular, CNNs have already been widely used in the analysis of radio interferometric data. For example, they have been used to classify radio galaxies (Bowles et al. 2021), to determine galaxy morphologies (Cheng et al. 2020a), and to select pulsar candidates (Zeng, Li & Lin 2020). More specifically, CNNs have been employed to detect astronomical sources within the CONVOUSOURCE (Lukic, de Gasperin & Brügger 2019), DEEPSOURCE (Vafaei Sadr et al. 2019) and Point Proposal Network (PPN; Tilley et al. 2020). Compared to traditional source detection algorithms, which can fail to detect sources when the signal-to-noise ratio is low, or can make false detections in regions of the images where the noise is highly correlated, the learning process in CNN-based source detectors has been shown to



generate more accurate results. Vafaei Sadr et al. (2019) and Lukic, de Gasperin & Brügger (2019) have shown that CNN based source detection algorithms are more complete down to a signal-to-noise ratio of 4 in detecting compact sources when compared to PyBDSF. However, these CNN based algorithms are all optimised for the analysis of images that have been deconvolved.

In this chapter, we investigate a deep learning based object detection algorithm that characterizes the source properties from images that have not undergone any prior deconvolution. This is partly due to the complexity of applying deconvolution methods, like CLEAN, to large datasets. Furthermore, such an algorithm could in principle also be applied directly to the visibility data via a Fourier transform, which would remove the need for any imaging step. We focus our analysis on dirty images that are produced from a sparse radio interferometric array, namely the Very Long Baseline Array (VLBA). This is because we are interested in developing a new object detection algorithm that can be applied to wide-field Very Long Baseline Interferometric (VLBI) observations with instruments like the VLBA in the future, such as those discussed in Chapter 2..

The presented approach for DEtection and Characterization Of Radio-Astronomical Sources (DECORAS) using deep learning consists of four main steps. The first step uses encoder-decoder networks to remove the noise and dirty beam from the given dirty images (the Fourier transform of the observed interferometric visibility data). The predicted model images at the output of the encoder-decoder are used in the post processing step to find the position of the source. In the third step, another encoder-decoder is used to characterize the source structure. Finally, the extracted latent variables of the trained encoder-decoder network are used to recover the source surface brightness distribution. Unlike DEEPSOURCE and PPN, which have thus far only addressed the detection of unresolved objects, DECORAS is trained on both point and extended source detection and characterization.

This chapter is arranged as follows. In Section 3.2, the training/verification data and the detailed methodology of the source detection and characterization algorithm is presented. In Section 3.3, we evaluate our results by applying the algorithm to test data and compare with the results from using a traditional source detection code (BLOBCAT). In Section 3.4, we investigate how well our algorithm can recover the source properties, such as source position, major axis and the true source surface brightness distribution. Finally, the results from this work are discussed and we present our concluding remarks on the methodology and future prospects in Section 3.5.

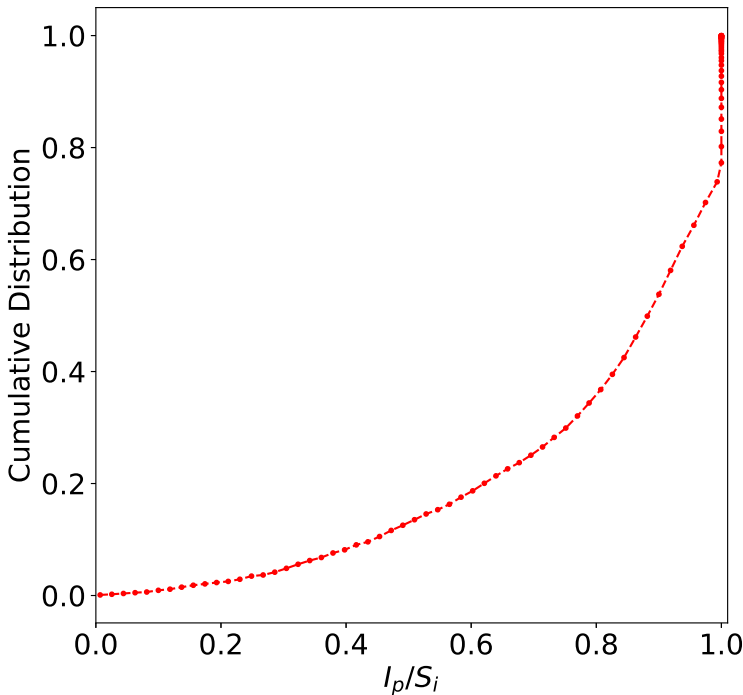
## 3.2 Method

This section presents our source detection and characterization methodology. First, the process of generating realistic simulated images is explained. This simulated dataset is used as the training and test samples for our network. Next, an overview of our approach is provided, with an explanation of the choice of loss function and specific network architecture that we have used. Then, we present our post processing object detection step, which determines the position of the source. Finally, our source characterization methodology is presented, which provides the structure and surface brightness distribution of the detected objects.

### 3.2.1 Simulating a representative training and testing dataset

Generating realistic images to train the network is one of the main steps for developing a source detection and characterization platform. This is because the network must learn the key properties and features of the data. Also, a simulated dataset can be used to test the robustness and completeness of the methodology, providing these data are unseen by the network during the training stage. Our goal is to develop a network that is applicable to data from sparse interferometric arrays, which are typical of VLBI observations. For our simulations, we have chosen to use the VLBA at an observing wavelength of 20 cm as our training and test dataset, the reasons for which we describe below. However, we see no obvious reason why our methodology cannot be used with other VLBI arrays that observe at other wavelengths, for example, the International LOFAR Telescope (ILT), the European VLBI Network (EVN), the Atacama Large Millimetre Array (ALMA), or the Square Kilometre Array (SKA-VLBI), which (will) operate from m to sub-mm wavelengths.

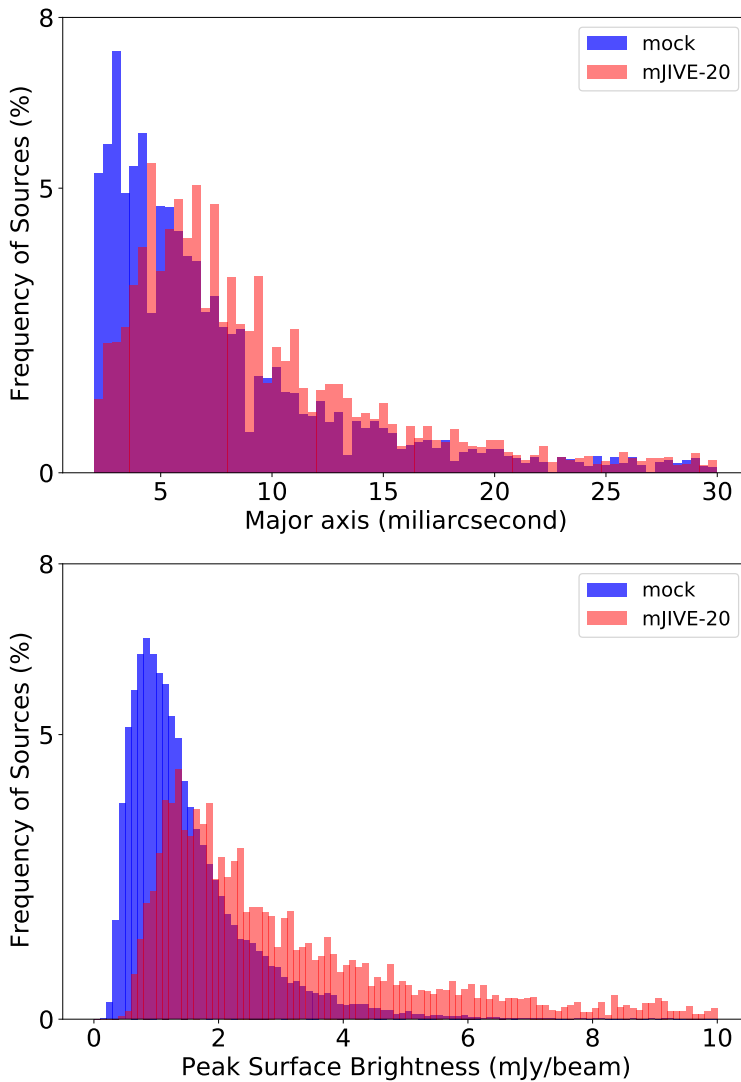
We have chosen the VLBA as our test interferometer as it is a homogeneous array of ten 25-m radio telescopes, with separations between 236 and 8611 km. This means the generated datasets will sparsely sample the visibility plane (45 measurements per time and frequency interval), will have uniform and predictable noise properties, and will be  $\sim 2$  MB in size per simulated observation. By using the VLBA at 20 cm, we can also utilize actual observations from a large statistical sample of radio sources to provide realistic observational conditions and representative source models for our simulations. For this, we have used data from the mJy Imaging VLBA Exploration (mJIVE-20) survey (Deller & Middelberg 2014), which targeted 24 903 radio sources in an area of 238 square degrees from 306 unique observations. The total observing time was 1 h for each observation, which was further divided into a set of four sub-pointings around a bright calibrator source. Within  $\sim 20$  arcmin of the central



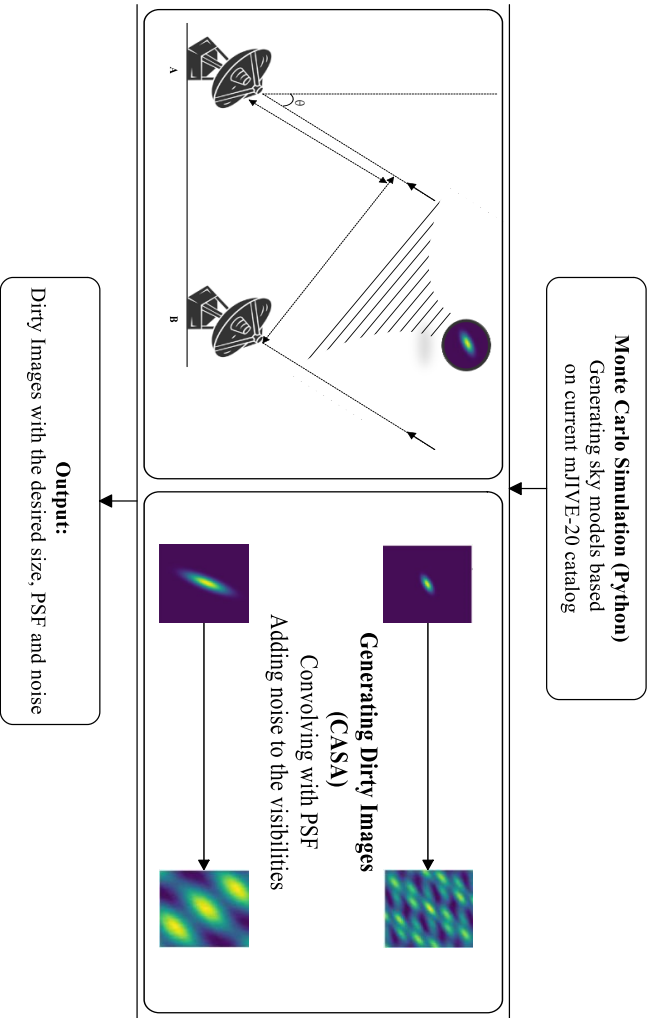
**Figure 3.1** – Cumulative distribution of peak surface brightness divided by the integrated flux density (in units of  $\text{beam}^{-1}$ ) for detected objects in the mJIVE-20 survey (Deller & Middelberg 2014). Around 20 per cent of the detected sources in the mJIVE-20 survey have an equivalent peak surface brightness and flux density (unresolved sources). In more than 65 per cent of all detected sources in the mJIVE-20 survey, the peak surface brightness is about 80 per cent of the total flux density of the source, meaning that the majority of the sources detected in the mJIVE-20 survey are compact.

calibrator, the rms noise in the deconvolved images is about  $150 \mu\text{Jy beam}^{-1}$ . The detection threshold for the mJIVE-20 survey was set to  $6.75\sigma$  (based on simulations to determine the completeness), where  $\sigma$  is the rms map noise of a given observation. This led to 4965 radio sources being detected in the deconvolved images using BLOBCAT.

We have developed a framework with three steps to create mock training and test data with similar properties to those obtained by the mJIVE-20 survey. The first step defines the images in terms of angular size and number of pixels. The size of the individual pixels is set such that the visibility data are at least Nyquist sampled in the image plane, which for VLBA observations at 20 cm with a PSF sampling of four gives a pixel size of 1.25 mas. The majority of the detected sources in the mJIVE-



**Figure 3.2** – A comparison of the measured size and peak surface brightness of the injected mock and the real sources detected as part of the mJIVE–20 survey. The upper plot compares the distribution of fitted Gaussian major axis. The lower plot shows the distribution of peak surface brightness for the two populations. The implemented Monte Carlo approach based on the mJIVE–20 survey catalog, has made it possible to create a training dataset that follows the physical characteristics of real sources.



**Figure 3.3** – Flowchart of the mock data simulator, built within CASA. First, mock sources are generated using a Monte Carlo simulation to mimic the physical characteristics of the real sources in the mJIVE-20 survey. The generated sky model images are imported to CASA, using the simulator tool to generate the corresponding uv-datasets and dirty images.

20 survey are compact (Deller & Middelberg 2014); here we define compactness ( $C$ ) as the ratio of the integrated flux density ( $S_i$ ) and peak surface brightness ( $I_p$ ), where those objects with  $C < 1.25$  beams are classified as being compact. Fig. 3.1 shows the cumulative distribution of the peak surface brightness to integrated flux density ratio of detected sources in the mJIVE–20 survey, which demonstrates that the majority of the sources are unresolved. Considering the characteristics of the detected sources in the mJIVE–20 survey, we have chosen the size of the input image to be  $256 \times 256$  pixels, which is equivalent to an angular size of  $320 \times 320$  mas on the sky. Larger images would increase the number of learning parameters in our encoder-decoder model. This means that the learning procedure would require more time and memory.

Our simulated datasets are made using the Common Astronomy Software Applications (CASA; McMullin et al. 2007) package and custom-written Python scripts. The mock sources were generated using either delta or Gaussian functions, with a defined peak surface brightness, size, ellipticity and position angle, and a random position within each image. The defined source properties were determined from a Monte Carlo simulation of all of the sources observed as part of the mJIVE–20 survey. Fig. 3.2 shows a comparison of the injected mock sources and the actual mJIVE–20 survey sources in terms of their size (major axis) and peak surface brightness. Note that the number of injected mock sources is greater than the actual number of detected sources in the mJIVE–20 survey, as we have injected mock sources to all of the 24 903 phase-centres, while the catalog only contains information for 4 965 sources. We have also included fainter sources in the generated mock data to test our detection algorithm for sources with a low signal-to-noise ratio. The output of this step is the sky model interferometric datasets that will be used to generate dirty images.

CASA stores interferometric visibility data in a format called MeasurementSet. While it is possible to create a simulated MeasurementSet from scratch, the CASA simulator tool can use an existing MeasurementSet to obtain the position of the antennas and other observational settings (frequency and time sampling). Using actual MeasurementSets is a good choice for this work as we aim to train our network with mock data that is representative of real data. In this way, the simulator samples data with the correct  $(u, v)$  coordinates, considering the current model image with the mock source in it. We have used the existing MeasurementSets of the actual mJIVE–20 survey observations to generate simulated visibility datasets. To take the thermal noise into consideration, we have added Gaussian noise to the visibility data that is representative of the noise properties of the mJIVE–20 survey. We have not included any systematic errors to the real or imaginary component of the visibilities. Finally, the dirty image is generated

as the result of taking the Fourier transform of the visibility data, and gridding using the pixel size and number of pixels described above. Fig. 3.3 presents a flowchart that summarizes how the simulated dataset of model and dirty images is generated using CASA.

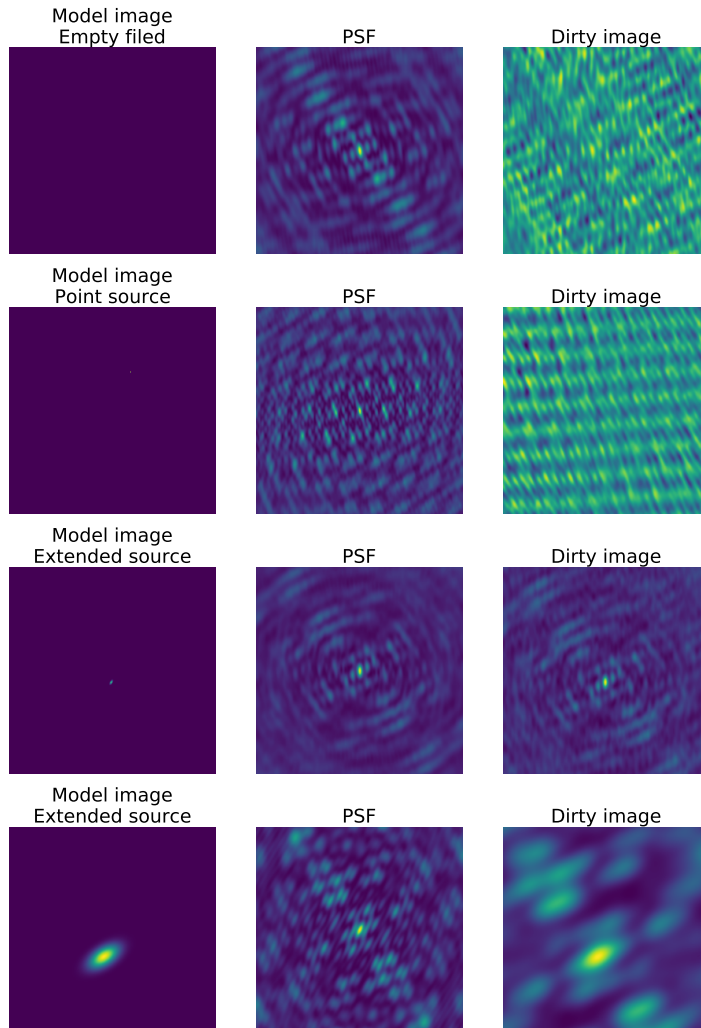
In total 50 000 simulated images, from 306 observations were made for training the network. In Fig. 3.4, we provide a few examples of the model sources, the PSF of the individual observations, and the final dirty images used to build the training and validation dataset. The dirty images are the input to the network, while the model images are the output of our encoder-decoder network.

### 3.2.2 Overview of DECORAS

Given any input dirty image, DECORAS is trained to deconvolve the PSF, remove thermal noise, locate the possible source in the image and characterize the source structure and surface brightness distribution. This process is summarised in the flowchart presented in Fig. 3.5.

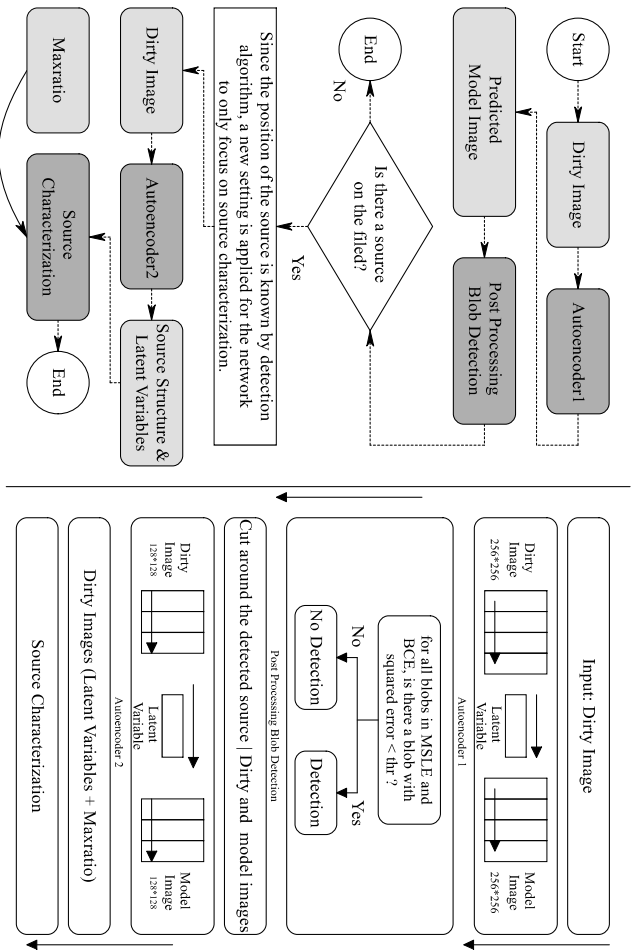
The first step of DECORAS consists of encoder and decoder parts. It is trained to recognize beam effects, correlated noise, or other sources of contamination in the dirty image, and to remove them from the predicted output. The first step of DECORAS is hereafter referred to as Autoencoder1. The output of Autoencoder1 is then passed to the Post Processing Blob Detection function to find the position of the source in the predicted model image. At the end of this step, the existence of the source and its position are known to the algorithm. To investigate the physical characteristics of the source, we crop around the region of interest in the field to form a smaller model and dirty image, with the source at the centre. The cropped images are fed to Autoencoder2 to investigate the physical characteristics of the detected source. Our experiments show that using Autoencoder2 yields a higher accuracy than with Autoencoder1 when extracting physical properties of the underlying source. Autoencoder2 has the same basic structure as Autoencoder1. However, due to the smaller image size, Autoencoder2 requires fewer convolutional layers in the encoder and decoder parts.

The right panel of Fig. 3.5 provides more information on the grey boxes shown in the left panel. For example, Autoencoder1 and Autoencoder2 are shown with four and three symbolic convolutional layers for the encoder and decoder, respectively. For the post processing step, the detection strategy is provided. More information about the Post Processing Blob Detection is presented in Section 3.2.6.



**Figure 3.4** – Each image contains  $256 \times 256$  pixels and is equivalent to a sky-area of  $320 \times 320 \text{ mas}^2$ .





**Figure 3.5** – The left panel shows the general flowchart of DECORAS. The grey rectangles in the flowchart represent the functions or the developed algorithms in which the source detection and characterization are designed. The transparent rectangles show input/output data. In the right panel, further details on each of the grey boxes in the flowchart are given. A more detailed view of Autoencoder1 is represented in Fig. 3.6, where four convolutional layers that transforms dirty images of  $256 \times 256$  pixels into the model images of the same size are shown. The Post Processing Blob Detection is responsible for locating the injected source in the predicted model image. To characterize the physical properties of the source, Autoencoder2 has been used. It has the same network structure as Autoencoder1, except the input shape is  $128 \times 128$  pixels, and therefore, the number of convolutional layers is smaller. The source characterization uses the generated latent variable and predicted model images by Autoencoder2 as well as the maxratio to estimate the source surface brightness distribution.

### 3.2.3 Preprocessing

It is important to keep the pixel values of all the images in the same range of (0, 1) in order to optimize the learning process while minimizing the achieved loss in the network. The process of normalization used here linearly transforms the pixel values on all the images to a common range of between (0, 1). We have used a MinMax normalization according to,

$$x_{\text{normalized}} = \frac{x - \min(x_d)}{\max(x_d) - \min(x_d)}, \quad (3.1)$$

where  $x$  is the value of a given pixel and  $x_d$  represents all the pixels in the image.

As we will show below, DECORAS performs very well when the pixel values in each image are normalized in this way. However, the normalized predicted model image is not useful for recovering the absolute surface brightness of the detected sources. We have addressed this problem by analysing the latent variables generated by Autoencoder2 in the training process. This method is based on the information that the network has captured through the learning by removing the thermal noise and deconvolving the dirty images to obtain the expected model images. We discuss the details of estimating the surface brightness of the detected sources in Section 3.2.7.

### 3.2.4 Network structure

Fig. 3.6 presents the architecture of the network, which consists of two main components: an encoder and decoder in which convolution, leaky ReLU (Rectified Linear Unit) activation and batch normalization are used sequentially. We use convolution with a stride of (2, 2), which down samples the input image by a factor of 2 in each axis. A fully connected neural network is placed at the final step of our encoder. It learns the weights of the neurons for producing 256 values of latent variables. The latent variables that are generated by the encoder are the only piece of information our decoder uses to determine the output model image. This means that the compression rate of the encoder that imports the dirty images and generates the latent variables is 256. This is because the network only takes the structure and position of the source into account. All of the other information about the correlated noise, PSF and beam effects are learnt to be ignored by the network. In forthcoming work we will implement networks that can also solve for the PSF, and hence determine if there are residual calibration errors in the data. The same process, using the same parameter variables, is applied to the decoder part, but in a reverse order. Instead of using a convolution, the decoder uses the transpose convolution with a stride of (2, 2) to up-sample the input image by a factor of 2. Leaky ReLU and batch normalization

are also included, as shown by the grey arrows in Fig. 3.6. After obtaining the desired image size by up-sampling, an extra layer of convolution and Sigmoid activation is used to force the final generated pixel values to be between 0 and 1.

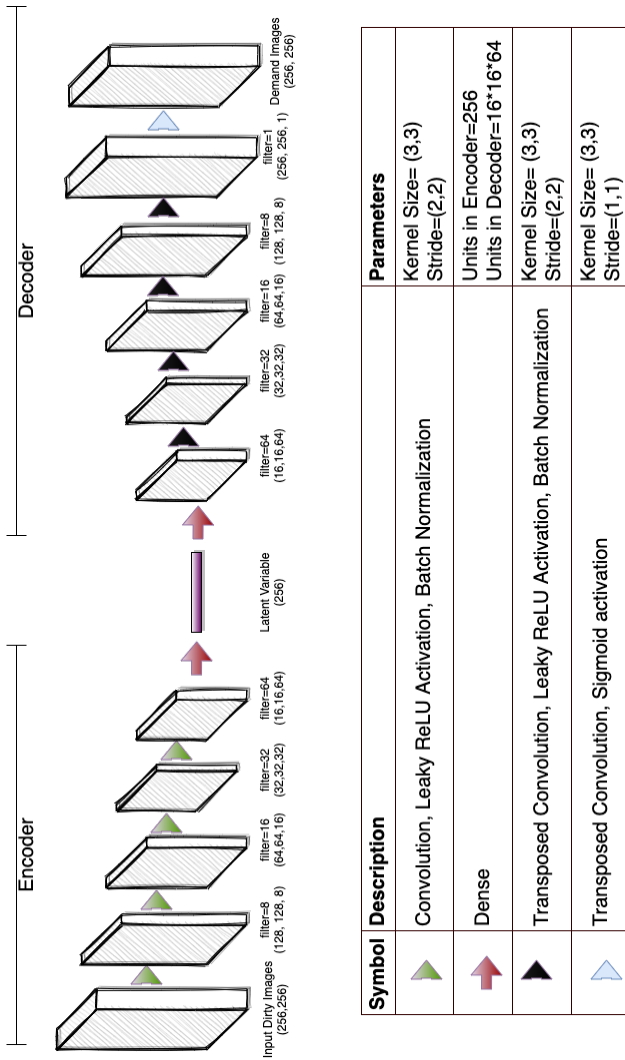
Our choice of network structure and the given input/output images can be considered as some sort of image segmentation. In such algorithms, the input image goes through different layers of convolution, and the generated latent variables are used to remap to a full output image. Instead of reconstructing the input image (the typical use of autoencoder structure), it only targets a specific segment of the image and, specifically for our case, where the source is located. Beside the location of the source, our network is sensitive to the size and structure of the source. On the other hand, as our network uses dirty images as the input and generates model images at the output, neurons on the encoder are forced to remove the effect of the PSF and correlated noise in the dirty images.

As described above, using a convolution layer with a stride of (2, 2) down samples the input image by a factor of 2. An alternative is to do so using a max pooling layer. Studies of several image-recognition benchmarks show that a convolutional layer with an increased stride can be easily substituted for a max-pooling scheme (e.g. Springenberg et al. 2014). Moreover, Ayachi et al. (2020) refers to the learnable nature of convolutional layers whereas the max pooling is a fixed function that takes the maximum value of each defined filter. The same study has also analyzed the memory efficiency of using strided convolution layers over max pooling. A batch normalization is added at the end of each convolution layer to stabilize the distribution of inputs (over a minibatch). This is achieved by keeping the mean ( $\mu_{x_k}$ ) and standard deviation ( $\sigma$ ) of the output close to 0 and 1, respectively. It also helps to decrease the importance of the weight initialization and regularizes the model (Santurkar et al. 2018). In the Keras implementation, the output of a batch normalization layer is applied to each feature  $x_k$ , such that,

$$\hat{x}_k = \gamma \frac{x_k - \mu_{x_k}}{\sigma^2(x_k) + \epsilon} + \beta, \quad (3.2)$$

where  $\epsilon$  is a constant that is being added to make sure the denominator is nonzero,  $\gamma$  (initialized as 1) and  $\beta$  (initialized as 0) are learnable parameters used for scaling and shifting purposes.

We note that based on our experiments, increasing the number of layers will increase the training time without significantly improving the TP and TN rates. Also, as part of this project, we have implemented other network structures, such as U-Nets (Ronneberger, Fischer & Brox 2015) and variational autoencoders (Kingma & Welling



**Figure 3.6** – The encoder-decoder network structure used in Autoencoder1 for source detection. There are four convolutional layers on both the encoder and decoder sections. Each convolutional layer is followed by leaky ReLU activation and batch normalization. In the encoder section, a stride of (2, 2) is defined for the convolutional layers, which are used to down sample the data. At the end of the encoder, a dense layer is used to transform the extracted set of features in the convolutional layers into the latent variables. The decoder has a similar structure, but instead of down sampling the input, it up samples the data using transposed convolutional layers. A Sigmoid activation function is placed at the end of the decoder to ensure the output pixel values are between (0, 1).

2013), to test their performance against the presented method. We found that both U-Net and variational autoencoders were less accurate compared to the presented method; the predicted model images of a trained U-Net model were not completely noise free and had a less accurate estimation of the source structure, while the latent variables for the trained variational autoencoder did not provide enough information to estimate the source surface brightness.

### 3.2.5 Loss function

In order to complete the learning process, it is necessary to define a loss function. The network uses the loss function to calculate the error between the estimated and expected model images at the end of each iteration. This error is used to update the weights and minimize the final error using optimizing strategies such as gradient decent. We have compared the performance of four different loss functions: Mean Squared Error (MSE), Mean Absolute Error (MAE), Binary Cross Entropy (BCE) and Mean Squared Logarithmic Error (MSLE).

Our tests found that the MSE and MAE loss functions are not a proper choice for our specific problem since they have difficulties in detecting the source when the demand image is a point source (characterized by a single non-zero pixel while the rest of the image is zero). In order to force the network to take into account the information provided by a single pixel, Vafaei Sadr et al. (2019) suggest to smooth the adjacent pixels around the point source to provide more non-zero pixels. On the other hand, Sedaghat & Mahabal (2018) have provided a solution that conditionally boosts the error on the non-zero pixels in the image. With this technique, the learning rate is virtually increased for only the non-zero pixels. Note that increasing the general learning rate is not a solution to the problem here as the learning procedure would generate a sub-optimal set of weights or an unstable training process. Based on the preliminary results of using the three defined loss functions with our network structure (see Fig. 3.6), and the issues that MSE and MAE have when the source structure is very small compared to the size of the image, we have decided to only consider the MSLE and BCE further.

Our encoder-decoder network is designed to solve a denoising problem. This can be interpreted as a pixel-wise classification in which each pixel (in the normalized images) is assigned a range between (0, 1). We consider a given model image  $M$  and a predicted model image  $M'$  with size of  $n \times n \times 1$ . Considering  $x_{i,j}$  as the pixel value on the  $i, j$  position of the model image and  $x'_{i,j}$  as the predicted pixel value at

the same location, the BCE and MSLE are calculated according to

$$\text{BCE}(M', M) = -\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n x_{i,j} \log x'_{i,j} + (1 - x_{i,j}) \log(1 - x'_{i,j}), \quad (3.3)$$

and

$$\text{MSLE}(M', M) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \left[ \log(x_{i,j}) - \log(x'_{i,j}) \right]^2, \quad (3.4)$$

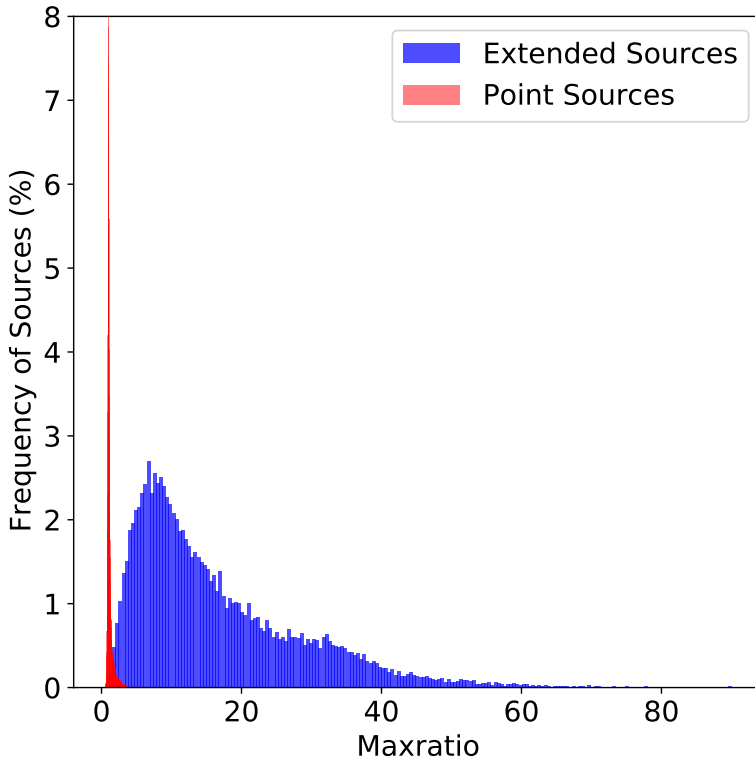
respectively. Note that the MSLE loss is similar to the MSE, but instead of calculating the loss of  $M$  and  $M'$  by computing the average, it calculates the squared error between the logarithm of the true and predicted values. Compared to MSE, MSLE penalizes underestimation more than overestimation.

### 3.2.6 Post processing blob detection

The network is designed to generate predicted model images with zero background, that is, the only non-zero pixel values in the predicted model images should be related to a detected source. In some cases, and particularly in the low signal-to-noise ratio regime, the network can get confused on where the input source is located and instead generates images with multiple components with non-zero pixel values (hereafter referred to as blobs). To identify such blobs in the images, we have used the SCIKIT Image library (van der Walt et al. 2014) in Python, which contains several blob detection solutions. We have used the `blob_dog` function to find blobs in the given predicted model images. The output of this algorithm is the  $x$  and  $y$  co-ordinates of any detected blobs, with an estimate of the uncertainty from the standard deviation of a fitted Gaussian function to the blobs. Through this process, we are able to catalog the candidate sources that have been identified by the network.

### 3.2.7 Surface brightness estimator

Losing the exact pixel values in the predicted model images due to the normalization process discussed above requires an alternative solution to estimate the absolute source surface brightness. To characterize the detected sources, DECORAS relies on two sources of information. The first is the compressed set of features that are represented by the latent variables of our encoder-decoder structure. The latent variables are the only information that the decoder uses to construct the model images with all the embedded details, like source position and the physical characteristics. Fig. 3.6 shows the position of the latent variables in our encoder-decoder structure.



**Figure 3.7** – The maxratio distribution for point and extended sources. The maxratio is defined as the ratio between the peak surface brightness in any dirty image to its corresponding model image. It is used to estimate the absolute surface brightness of the detected sources.

In general, a greater number of latent variable units will result in a more clear and sharper reconstruction of the output image.

The second source of information is based on the fact that the source peak surface brightness in the dirty image is not equal to the peak in the model image. This happens due to the process of converting the visibility data of the true sky model to the dirty images (by adding Gaussian noise and convolving with the dirty beam). Convolution of the visibility data of extended sources with the PSF increases the peak surface brightness in the predicted model image. The level of increase is dependent on the source size, PSF structure and the noise of the visibilities. We have measured this increase by calculating the ratio between the source peak surface brightness in the dirty image to the peak surface brightness of the corresponding injected source in

the model image. This parameter, which we call the maxratio, is defined as

$$\text{maxratio} = I_d/I_m, \quad (3.5)$$

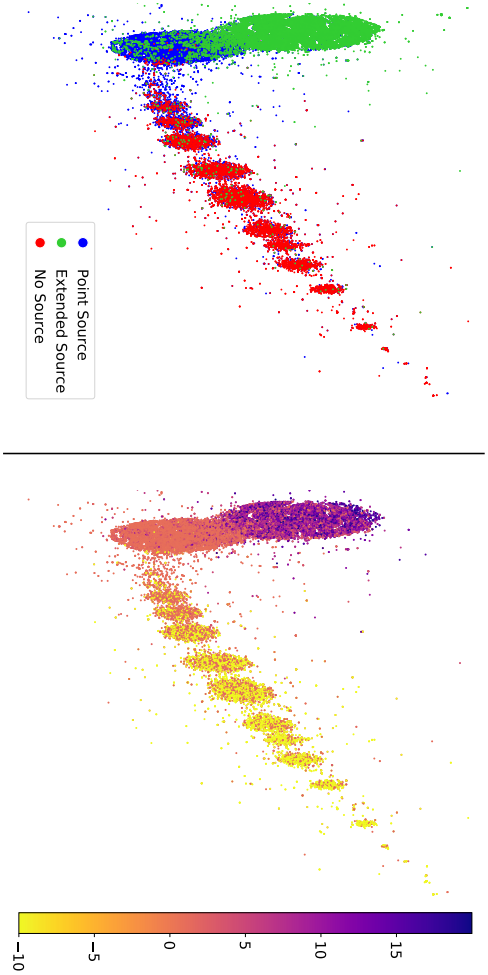
where  $I_m$  is the peak surface brightness of the source in the model image (in units of  $\text{Jy pixel}^{-1}$ ) and  $I_d$  is the peak surface brightness in the dirty image (in units of  $\text{Jy beam}^{-1}$ ). The maxratio is a measure of how much the peak surface brightness of a source has changed as it is convolved with the dirty beam to form the dirty image. Fig. 3.7 shows the distribution of maxratio for the entire simulated dataset. For point sources, the maxratio is mostly close to unity, as expected, and for extended sources the maxratio is typically larger, up to a factor of 90.

A 2D visualization of the latent variables is a powerful tool to gain insight to the structure of data. One way to visualize high-dimensional latent variables is the t-Distributed Stochastic Neighbour Embedding (t-SNE; Van der Maaten & Hinton 2008) method. Fig. 3.8 presents the t-SNE visualization of our simulated data with two different colouring schemes. TSNE is implemented using the `SCIKIT` Python package with a perplexity of 20 and the maximum number of iterations set to 400. Each data point in Fig. 3.8 represents an array of 256 values, which are the extracted latent variables of `Autoencoder1`. The left plot of Fig. 3.8 is coloured according to the classes of data: point and extended source samples, and noise realization samples with no injected source. In the right plot of Fig. 3.8, we have coloured the data points according to their corresponding maxratio. It is clear that the different source classes and their corresponding maxratios are separable using the latent variable information. Also, by determining the maxratio in this way, the absolute peak surface brightness of a given source can be recovered using equation 3.5.

Although Fig. 3.8 provides some insight to the maxratio distribution and the types of sources that are detected, there is clearly some overlap between the three classes of data. Using several regression estimator techniques, such as the  $k$ -Nearest Neighbour (KNN), XGBoost and RandomForest, we find that this overlap affects the accuracy of the source brightness estimation. Therefore, we apply a specific approach for this task; once a source is detected and located by the network, its structure is inferred more accurately when the source is positioned in the centre of a cropped image. As the structure of the source is correlated with the maxratio (see Fig. 3.7), this results in a more accurate prediction of the maxratio.

In our implementation, a square image with  $128 \times 128$  pixels is cropped around the position of the detected source in both the dirty and model images. These images are then used to train a new network (`Autoencoder2`), the latent variables of which





**Figure 3.8** – A t-SNE visualization of the 256 latent variables extracted from the encoder section of Autoencoder1. The encoder provides the compressed representation of any corresponding model image to the corresponding input dirty image. Latent variables carry the information of the source structure and accordingly the *maxratio*. The left panel is colour-coded by the class of data (point source, extended source, and noise realization). The right panel shows the same 2D representation of the latent variables coloured according to the corresponding *maxratio*. For the noise realization samples, with no injected source, the value of the *maxratio* has been assigned to  $-10$  for visualization purposes.

provide a more accurate representation of the source structure. Fig. 3.9 shows the two-dimensional t-SNE visualization of the latent variables obtained from the training data where the source is centred and the image is cropped. A comparison of Figs. 3.8 and 3.9 shows that the latent variables with the new setting yields a more distinct representation of the different classes and the maxratio.

The network architecture used for the structure estimator is similar to the encoder-decoder shown in Fig. 3.6. However, due to the smaller input image size, one convolutional layer from both the encoder and decoder sections is removed. In this reduced architecture, a layer of 64 units represents the latent variables.

### 3.3 Source detection

In this section, we present our results on training the network using the BCE and MSLE loss functions, before providing an overview of our DECORAS source detection strategy. The results from using DECORAS are compared to BLOBCAT, a traditional source detection algorithm, which was also used by the mJIVE-20 survey.

#### 3.3.1 Defining the true positive and true negative rates

In order to evaluate our results, we consider the confusion matrix presented in Table 3.1. The true positives (TP) are defined as the number of fields with an injected source that the algorithm has successfully detected, while the true negatives (TN) are the number of fields with no injected source where the algorithm correctly returns a non-detection. Conversely, the false positives (FP) are defined as the number of fields with no injected source, but the algorithm detects a source, and the false negatives (FN) correspond to the number of fields with an injected source that the algorithm fails to detect.

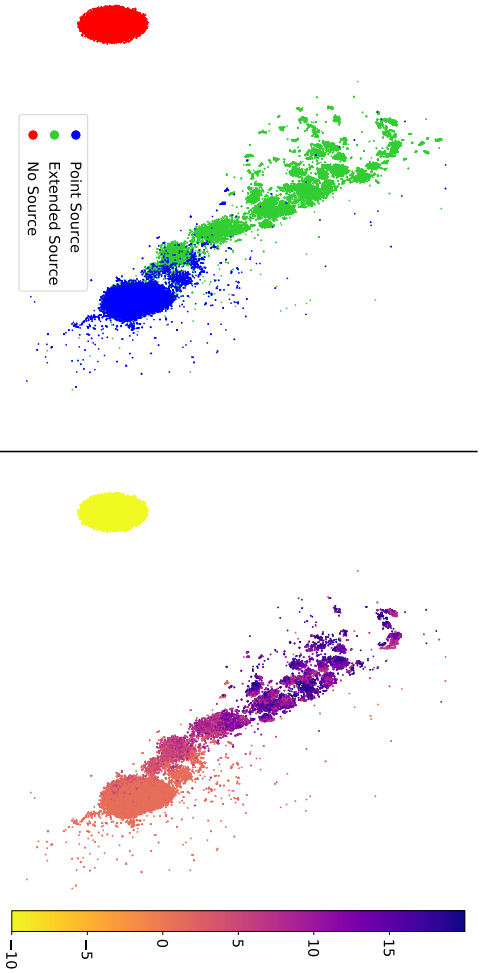
We quantify the performance of DECORAS and BLOBCAT when applied to simulated data by calculating the TP and TN rates,

$$\text{TP rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (3.6)$$

and

$$\text{TN rate} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad (3.7)$$

respectively. Given that we aim to limit the number of FP and FN events, an ideal system would achieve a TP rate (sample completeness) and TN rate (sample purity)



**Figure 3.9** – A t-SNE visualization of the latent variables extracted from *Autoencoder2*, which converts dirty images of  $128 \times 128$  pixels into model images of the same size. The position of the injected source is the centre of the image. In the left panel, the colour illustrates the class of data (point source, extended source, and noise realization). The right panel shows the same 2D visualization of the latent variables coloured according to the corresponding *maxratio*. For the noise realization samples, with no injected source, the value of the *maxratio* has been assigned to  $-10$  for visualization purposes.

		True Data	
		Source	No Source
Test Results	Detection	TP	FP
	No Detection	FN	TN
		Total	Noise Realizations

**Table 3.1** – The sample confusion matrix, showing how the TP, FP, FN and TN are defined.

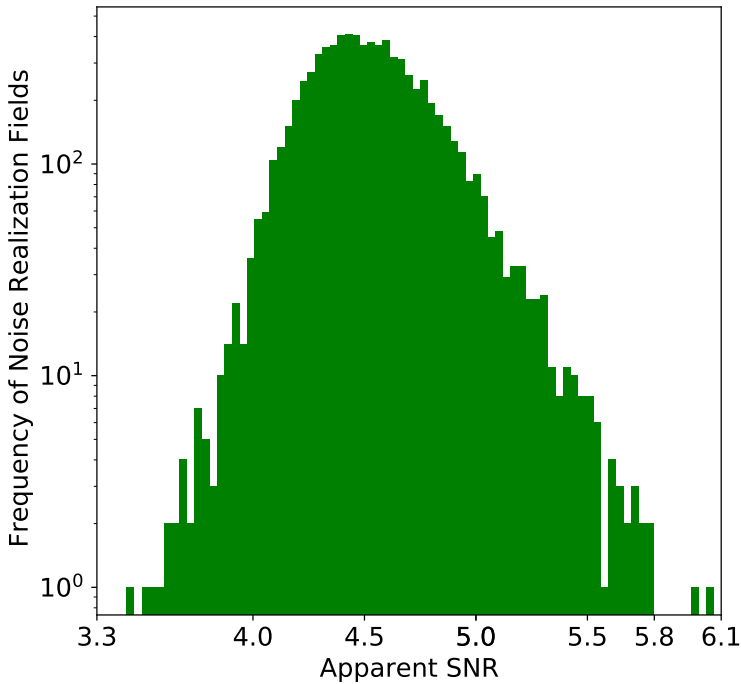
of 1.0.

The simulated test dataset consists of two components. First, we generate 8000 images in which 3000 samples correspond to point sources and 5000 samples contain (extended) elliptical Gaussian sources. To define a signal-to-noise ratio for each source, we divide the peak surface brightness of the injected source in the model image by the rms of a noise realization of the same simulated observation without any injected source. For our simulations, we use a signal-to-noise ratio of the injected sources between 1 to 16. This simulated dataset is used to determine the TP rate.

Second, we generate a dataset of 7800 noise realizations that do not contain any injected source. This dataset is used to evaluate the TN rate. To determine an apparent signal-to-noise ratio, we divide the peak surface brightness by the rms noise in each realization. The total number of noise realizations has been chosen so that we have a sufficient number of samples to test the TN rate for noise peaks at  $> 4\sigma$ . To ensure this, we decided to make images that are  $1024 \times 1024$  pixels in size, from which we cropped sub-images of  $256 \times 256$  pixels with the surface brightness peak in the centre. This results in each image having 65 653 pixels, which for a Gaussian noise distribution should, on average, have a peak surface brightness at a significance of  $4.3\sigma$ , and at least one pixel detected at the  $6\sigma$  level when all 7800 images are considered. In Fig. 3.10, we show the distribution of apparent signal-to-noise ratio of the peak surface brightness in the noise realizations. This peaks at an apparent signal-to-noise ratio of 4.4 and has at least one  $6\sigma$  noise peak. Overall, the range of signal-to-noise ratios in our noise realizations is between 3.4 to 6.1. We note that the distribution is not Gaussian, with a skew towards higher signal-to-noise ratios. This is not unexpected given that the noise is correlated in the image plane for interferometric data.

### 3.3.2 The performance of BLOBCAT

BLOBCAT is a source extraction algorithm that is designed to detect and catalog sources from pre-processed radio images (Hales et al. 2012). In order to apply BLOBCAT on



**Figure 3.10** – *The distribution of peak surface brightness-to-noise ratio for 7800 noise realizations. The distribution is consistent with the expectations, given the total number of pixels used per image, and in total for the entire test dataset.*

our test data, we first had to deconvolve the simulated visibility datasets using the TCLEAN task within CASA (down to a threshold of  $3\sigma$ ). We note that this necessary step will change the simulated images being used for our comparison (dirty versus clean images), but as the underlying input model is the same for both cases, this will still allow for a proper comparison between DECORAS and a standard source detection algorithm.

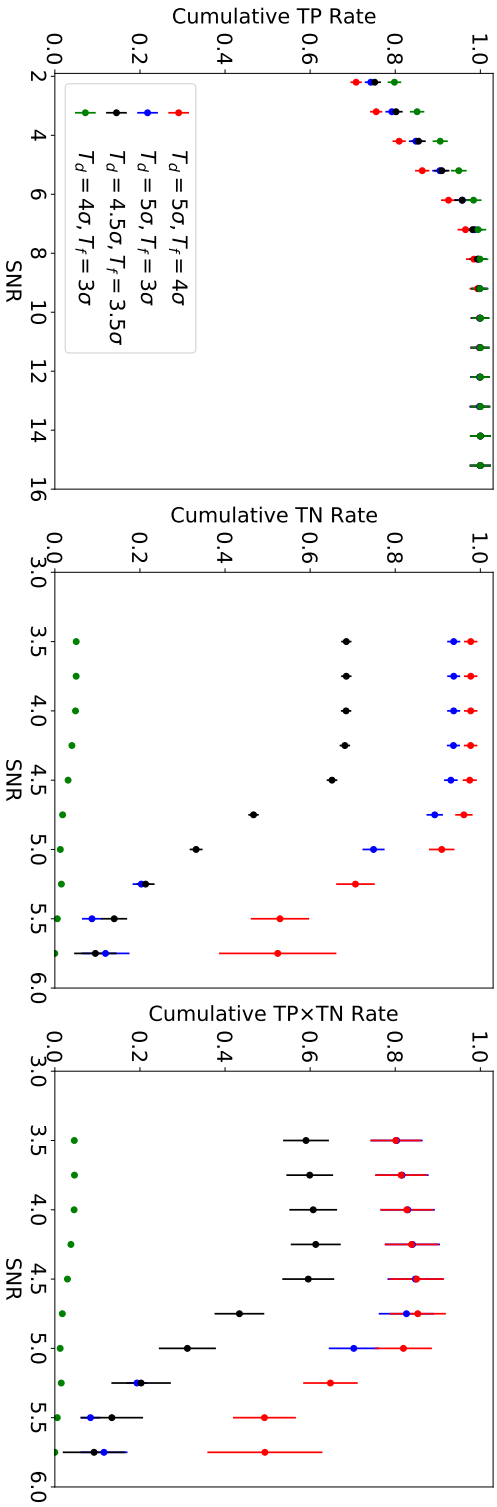
BLOBCAT works by looking for islands of pixels that might represent a source. The signal-to-noise ratio of a given pixel is the key parameter used to detect potential sources. For this, the algorithm determines the field surface brightness and the background rms noise as the input parameters. It also requires the user to set a detection signal-to-noise ratio threshold ( $T_d$ ) and a cut-off signal-to-noise ratio threshold ( $T_f$ ) for flooding the islands. The detection process starts with locating all pixels that have a higher signal-to-noise ratio than  $T_d$ . To each pixel above this limit, an island of adjacent pixels are added that are above  $T_f$ . For our simulations, we consider the pair values of  $(5\sigma, 4\sigma)$ ,  $(5\sigma, 3\sigma)$ ,  $(4.5\sigma, 3.5\sigma)$  and  $(4\sigma, 3\sigma)$  for  $(T_d, T_f)$ . Any detected source is then parameterized by fitting a 2-dimensional elliptical Gaussian

and compared with the input model.

In Fig. 3.11, we present the cumulative TP and TN rates for BLOBCAT as a function of (apparent) signal-to-noise ratio. As expected, the performance of BLOBCAT depends on the choice of  $(T_d, T_f)$ . We find that values of  $(5\sigma, 4\sigma)$  have the highest TN rate (0.98), whereas using  $(4\sigma, 3\sigma)$  returns the lowest TN rate (0.05), when the full sample of noise realizations are considered. When  $(T_d, T_f)$  are set to  $(4.5\sigma, 3.5\sigma)$ , we find that there is a transition at a signal-to-noise ratio of 4.5, due to the fraction of FP detections decreasing and the fraction of TN detections increasing, as expected. Note that the TN rate for sources with an apparent signal-to-noise ratio below  $T_d$  is high as the algorithm does not consider any blob below this significance as a potential source. We also see that the TN rate of BLOBCAT drops drastically for fields that have a higher apparent signal-to-noise ratio, due to the choice of thresholds that have been used.

We find that for a VLBI-like array, such as the VLBA, the point source catalogue has a TP rate of 0.91 at the  $4.2\sigma$ -level using a setting of  $(4\sigma, 3\sigma)$  for  $(T_d, T_f)$ . This increases to 0.95 at the  $5.2\sigma$ -level and is complete at signal-to-noise ratios  $> 6.2$  (we define the 100 per cent completeness as the signal-to-noise ratio where the first false negative is returned). However, below this, the fraction of FN detections increases, and the overall TP rate decreases. Also, as expected, the TP rate decreases faster when a combination of higher thresholds are used. For example, the  $(5\sigma, 4\sigma)$  setting for  $(T_d, T_f)$  is complete at signal-to-noise ratios  $> 8.4$ . Finally, we note that the performance of BLOBCAT is slightly better than the other source detection algorithms described above, with a completeness that is higher than 80 per cent at the  $4\sigma$  level.

As already discussed, the settings for both  $T_d$  and  $T_f$  will affect the TP and TN rates determined by BLOBCAT. For example, using pair values of  $(5\sigma, 4\sigma)$  generates a TN rate of 0.52 at the  $5.5\sigma$  level, while the  $(5\sigma, 3\sigma)$  setting has a TN rate of only 0.09 at the same apparent signal-to-noise ratio. Ultimately, the user's choice of  $T_d$  and  $T_f$  will depend on the scientific goal of the observation. For example, one might need to reach 100 per cent completeness at a specific signal-to-noise ratio, while for some other science cases a higher purity is important. In this regard, it is the combination of TP and TN rates that matters when evaluating the general robustness of a source detection algorithm. Therefore, we have calculated the combined TP  $\times$  TN rate, which is also shown in Fig. 3.11. We find that the TN rate is the dominant factor when comparing the performance of BLOBCAT with different  $(T_d, T_f)$  values. Also, from Fig. 3.11, we see that the highest catalog completeness and purity is obtained for the  $(5\sigma, 4\sigma)$  setting of  $(T_d, T_f)$ . Therefore, we will use this setting for comparing our results with DECORAS.



**Figure 3.11** – The cumulative true positive (TP; left), true negative (TN; middle) and TP  $\times$  TN (right) rates, for BLOBCAT using different values for the detection ( $T_d$ ) and flooding ( $T_f$ ) thresholds. Setting  $T_d = 4\sigma$  and  $T_f = 3\sigma$  generates the highest TP rate (high catalog completeness), but at the same time, has the lowest TN rate (low catalog purity).

		True Data	
		Source	No Source
Test Results	Detection	84.2%	81.2%
	No Detection	15.8%	18.8%
Total		8000	7800

**Table 3.2** – The confusion matrix when only using the BCE as the loss function.

		True Data	
		Source	No Source
Test Results	Detection	78.3%	73.3%
	No Detection	21.7%	26.7%
Total		8000	7800

**Table 3.3** – The confusion matrix when only using the MSLE as the loss function.

### 3.3.3 Comparing the performance of BCE and MSLE

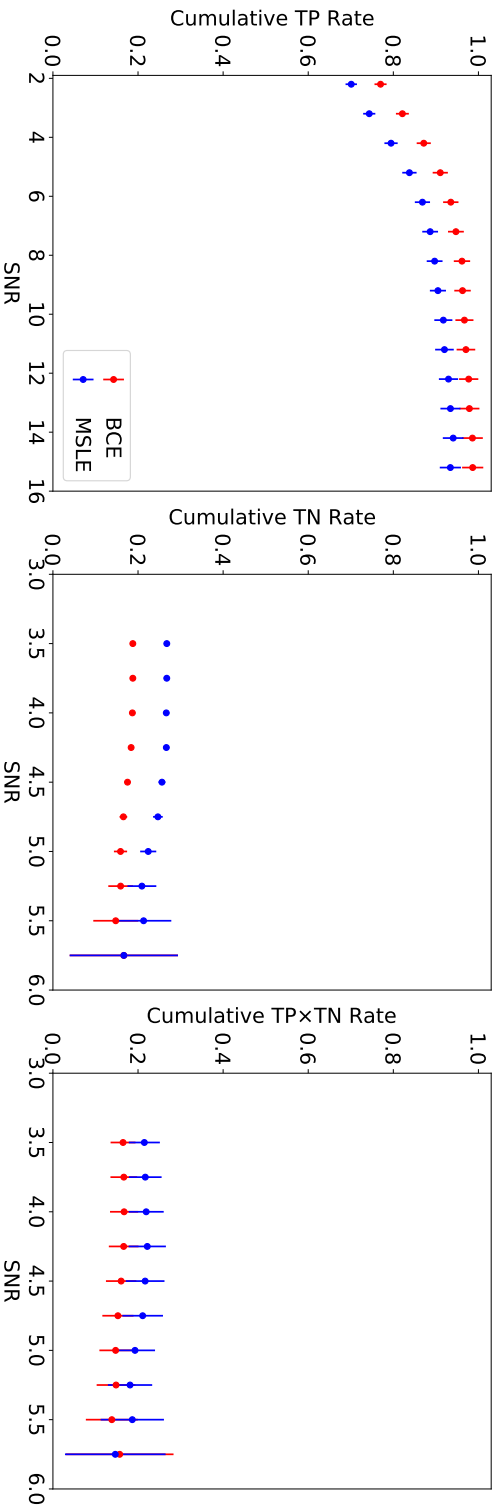
Fig. 3.12 presents the TP, TN and  $TP \times TN$  rates as a function of (apparent) signal-to-noise ratio when either the BCE or MSLE are used as the loss function in our encoder-decoder network. Each model is trained over 800 iterations with the Adam optimizer in KERAS. The learning rate is set to  $10^{-3}$  as it provides lower loss values compared to when learning rates of  $10^{-2}$  and  $10^{-4}$  are used. The initial weights were set randomly from a uniform distribution. Although a different initiation of the weights can affect the training process at early epochs, our experiments find that the network converges shortly after the early epochs. In other words, the observed performance does not vary significantly with the initialization.

For the TP rate, we find that the BCE detects more genuine sources when compared to the MSLE for all signal-to-noise ratios. We find that the MSLE has a higher TN rate when compared to the BCE, which has a higher fraction of FP detections at all signal-to-noise ratios tested here. Similar to above, the  $TP \times TN$  rate is dominated by the TN rate. The confusion matrices from applying the trained model using the BCE and MSLE loss functions to the test dataset are shown in Tables 3.2 and 3.3, respectively. Based on this comparison of the BCE and MSLE loss functions, we can conclude that neither provide a satisfactory performance individually.

### 3.3.4 DECORAS source detection strategy

Motivated by the results from using the BCE and MSLE loss functions individually, we now present a new strategy that is based on using both loss functions together. As described above, when the algorithm is not confident on where the source is located, more than one blob emerges in the predicted model image. In the previous section,





**Figure 3.12** – The cumulative true positive (TP; left), true negative (TN; middle) and  $TP \times TN$  (right) rates for the BCE (red) and MSLE (blue) loss functions when used individually. The BCE yields a higher completeness than the MSLE for all considered signal-to-noise ratios. The overall TN rates are quite low (0.19 to 0.27) when the two loss functions are used individually. Similar to the performance of BLOBCAT, the TN rate dominates the combined  $TP \times TN$  rate.

we have considered all such low-confidence samples as non-detections when using the BCE or MSLE loss functions individually. However, using the BCE and MSLE together provides the possibility of finding the correct source, even for those fields with low signal-to-noise ratio detections. We consider a blob in a low confidence image as a detected source if both trained models, using the BCE and MSLE loss functions, agree on the existence and position of the source. A distance-threshold criteria is applied to the positions of the two detected blobs from using the BCE and MSLE individually. This threshold is defined as the maximum acceptable distance between the detected positions obtained with both loss functions, where the distance  $R$  is calculated using,

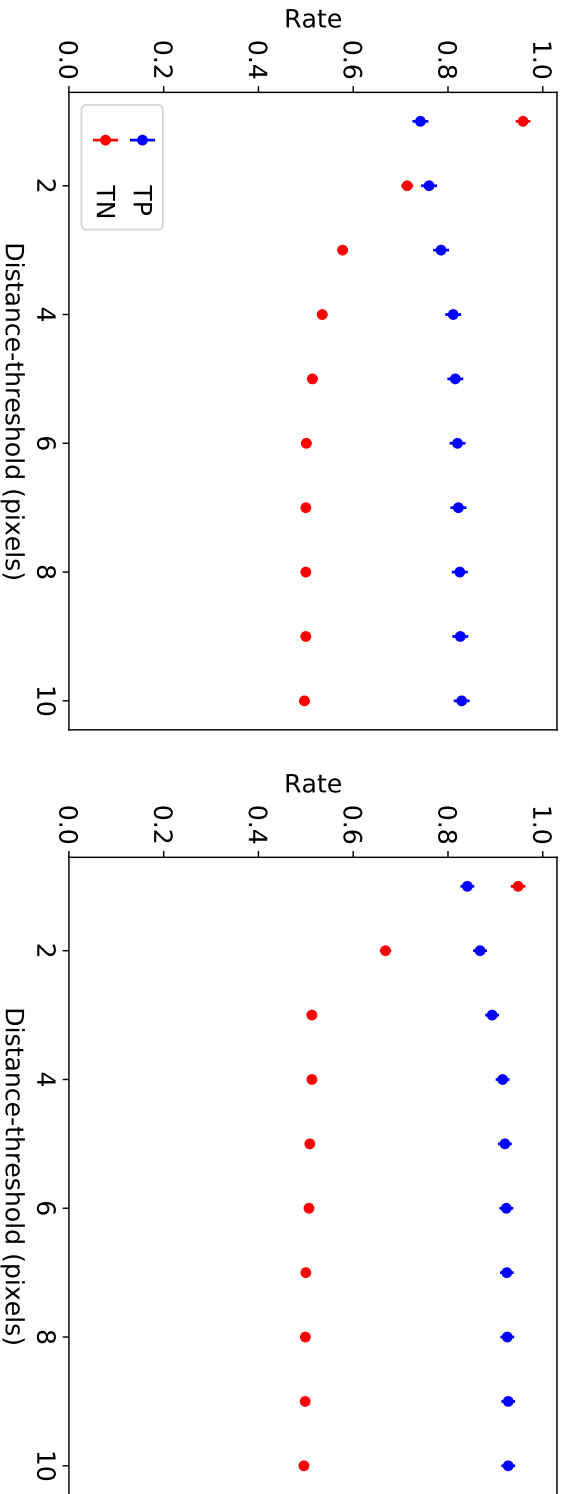
$$R = \sqrt{(x_{\text{MSLE}} - x_{\text{BCE}})^2 + (y_{\text{MSLE}} - y_{\text{BCE}})^2}. \quad (3.8)$$

Here,  $x$  and  $y$  are the co-ordinate positions obtained with the BCE and MSLE loss functions.

An alternative solution is to train the network with a combination of the BCE and MSLE loss functions, that is, BCE+MSLE. Considering the results of applying the BCE and MSLE loss functions on test data, we expect that simply adding up the loss values would not provide reliable results. As is shown in Fig. 3.12, the BCE performs better over the MSLE for the TP rates, while the MSLE provides more reliable results when it comes to the TN rates. Therefore, if we train the network with a BCE+MSLE loss function, it would not result in either complete and reliable detections. A more sophisticated approach is to define a linear combination of the BCE and MSLE loss functions (e.g.  $\alpha\text{BCE} + \beta\text{MSLE}$ ) in which  $\alpha$  and  $\beta$  are the hyper-parameters that need to be defined and optimized for given our specific problem. The simple network structure and the short training time could justify the use of two separated loss functions. Finding optimal values for  $\alpha$  and  $\beta$  can be addressed in future work.

In Fig. 3.13, we show the TP and TN rates as a function of the distance-threshold  $R$  when the BCE and MSLE loss functions are used together. For this, we have determined the rates for detections at a signal-to-noise ratio  $> 2$  and  $> 4$  separately. In both cases, we find that the TP rate is rather flat and only marginally changes as the distance-threshold is increased (up to 10 pixels, or 2.5 beam sizes). However, the TN rate changes drastically from 0.96 at a distance-threshold of 1 pixel to around 0.5 at 10 pixels, with the largest change occurring between 1 and 3 pixels.

In Fig. 3.14, we again show the TP, TN and  $\text{TP} \times \text{TN}$  rates as a function of (apparent) signal-to-noise ratio, but for the case when both loss functions are used together. We also restrict our results to distance-thresholds of 1, 2 and 3 pixels, as for larger



**Figure 3.13** – The true positive (TP) and true negative (TN) rates as a function of the accepted distance-threshold between the detected position using the BCE and MSLE for a signal-to-noise ratio of  $> 2$  (left) and  $> 4$  (right). Although the TP rate varies only slightly as a function of distance-threshold (0.74 to 0.82 and 0.84 to 0.93 between 1 and 10 pixels), the TN rate is highly dependent on the distance threshold.

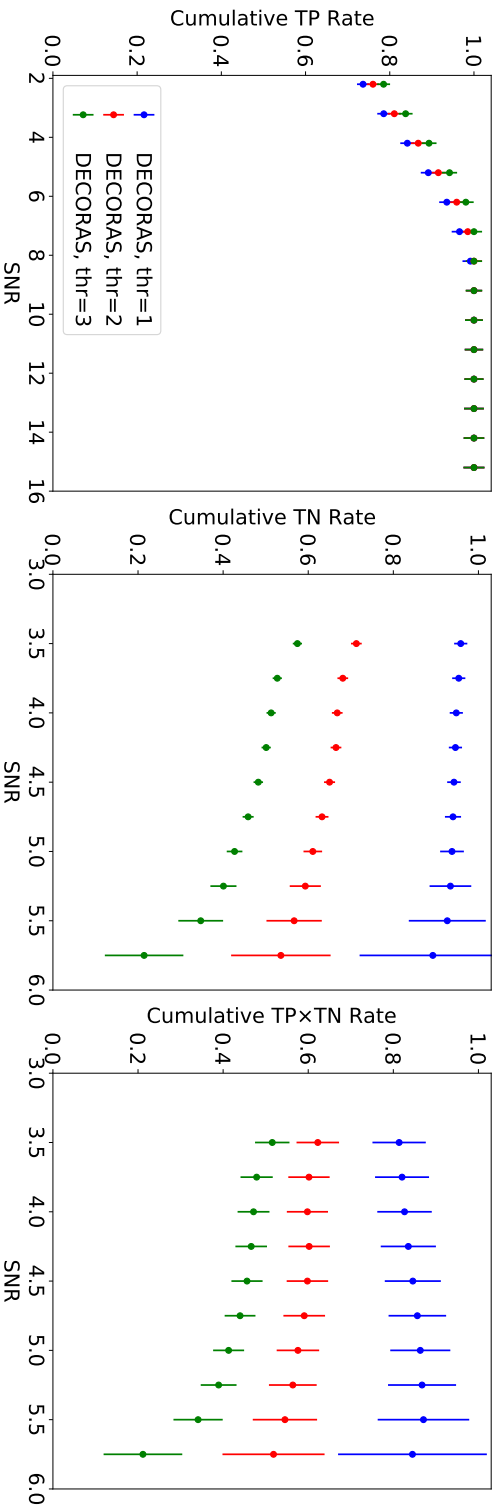
values of  $R$  the TP and TN rates are essentially constant. Here, when multiple blobs have been detected by the loss functions, that is, the low confidence cases, we use a distance-threshold to 1, 2 and 3, and for those cases with a single detected blob, we keep the distance threshold fixed to 3. From comparing with Fig. 3.12, we see that using the two loss functions together significantly improves the performance, particularly for the TN rate. We find that the TP rate has a similar behaviour for each threshold used, with slightly higher rates obtained for higher values of the distance-threshold. DECORAS is complete at signal-to-noise ratios of  $> 7.5$ ,  $> 6.9$  and  $> 6.0$  for distance thresholds of 1, 2 and 3 pixels, respectively. For the TN rates, the values are significantly higher when a distance-threshold of 1 is used, ranging from 0.97 to 0.93 between apparent signal-to-noise ratios of 3.5 and 5.5. Finally, we also see from Fig. 3.14 that the  $TP \times TN$  rates are highest when the threshold is conservatively set to 1 for the less confidence cases. Therefore, for our comparison with BLOBCAT, we will consider only the results from setting the distance-threshold to 1 pixel for those samples with multiple detected blobs, and 3 for those samples with a single blob detected by both loss functions.

### 3.3.5 Comparing DECORAS with a traditional source detection algorithm

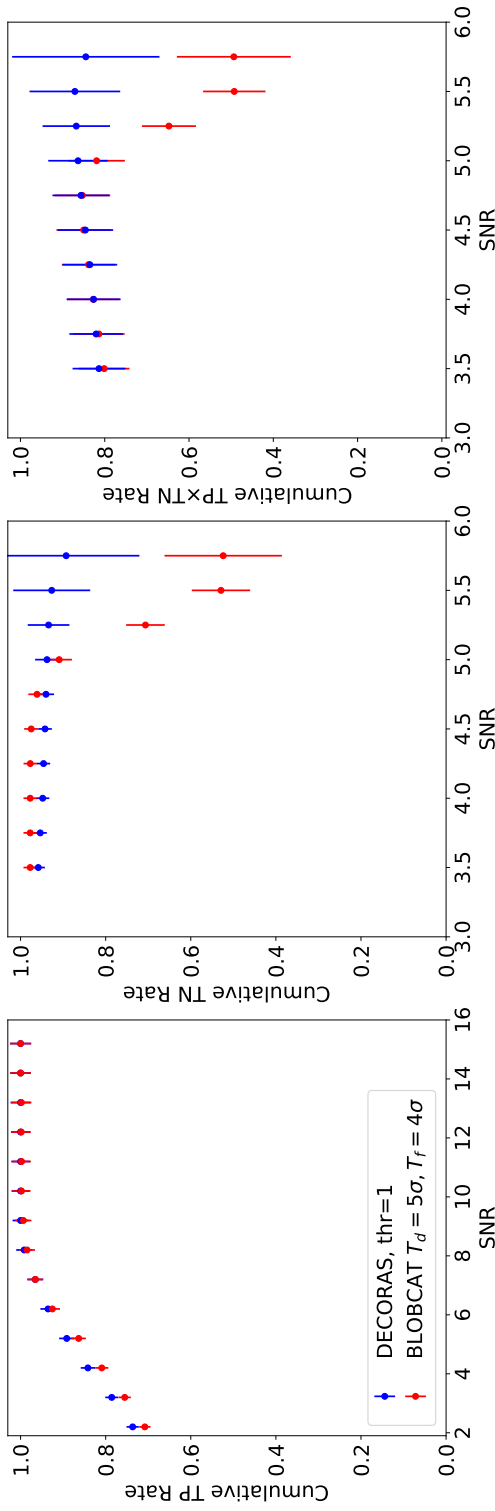
We present the comparison between DECORAS and BLOBCAT in Fig. 3.15. Overall, the TP rate (or completeness) of DECORAS is either equal to or marginally better than that of BLOBCAT at all signal-to-noise ratios. As described above, DECORAS is complete at the  $> 7.5\sigma$ -level, whereas BLOBCAT is complete at  $> 8.4\sigma$ . For the TN rates, BLOBCAT performs better at apparent signal-to-noise ratios  $< 5$ , owing to the cut-off detection threshold, but above this, DECORAS is better, with an almost constant TN rate as there are less FP detections returned at higher signal-to-noise ratio. Therefore, we conclude that both methods have a similar completeness level, but DECORAS is expected to have a better catalog purity. This is further demonstrated in the  $TP \times TN$  rate, where DECORAS out-performs BLOBCAT by an almost factor of two at signal-to-noise ratios  $> 5.5$ .

## 3.4 Source characterization

In this section, we present an analysis of how well DECORAS can recover the input source surface brightness distribution, which we parameterize as the position, size and peak surface brightness. We compare these properties with those of the input source models used to generate the test dataset discussed in Section 3.3. This is done by fitting two-dimensional elliptical Gaussian components to the predicted and input



**Figure 3.14** – The cumulative true positive (TP; left), true negative (TN; middle) and  $TP \times TN$  (right) rates, when the BCE and MSLE loss functions are used together, for distance-thresholds of 1 (blue), 2 (red) and 3 (green). The TP rates are very similar for each distance-threshold, with a value of 3 having the overall best performance. Conversely, the TN rate performance changes significantly, with a threshold of 1 having the best performance.



**Figure 3.15** – The cumulative true positive (TP; left), true negative (TN; middle) and  $TP \times TN$  rates for DECORAS (blue) and BLOBCAT (red). The TP rates are very similar for both methods, but the TN rate of DECORAS has a much better performance than BLOBCAT at  $> 5\sigma$ . This results in DECORAS having an almost factor of two better performance in combined catalog completeness and purity at signal-to-noise ratios  $> 5.5$ .

(true) model images of the sources.

### 3.4.1 Recovering the source position

Determining a reliable source position, for example, to compare the detected emission with other multi-wavelength datasets, is clearly an important aspect of any source characterization platform. In Fig. 3.16, we show the difference in the measured and expected position of the detected sources by DECORAS in both Right Ascension (RA) and Declination (Dec), such that,

$$\Delta RA = RA_{\text{DECORAS}} - RA_{\text{true}}, \quad (3.9)$$

and

$$\Delta Dec = Dec_{\text{DECORAS}} - Dec_{\text{true}}. \quad (3.10)$$

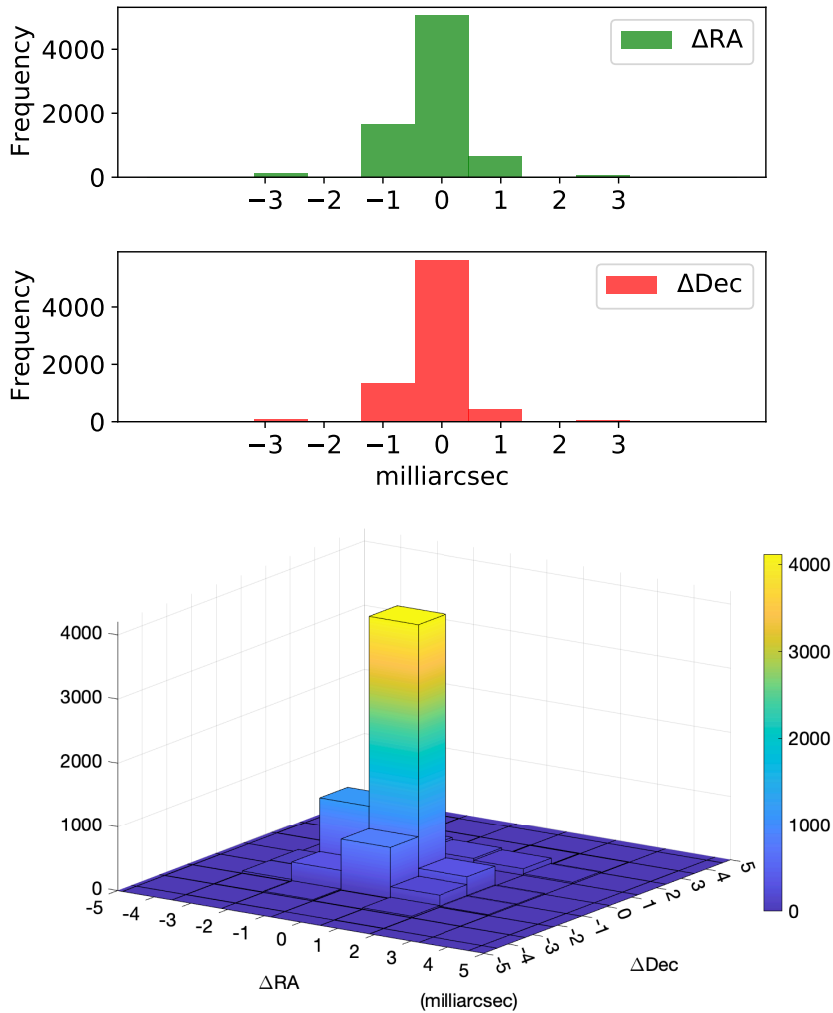
We find that the mean offset and standard deviation is  $-0.19$  and  $0.80$  mas in RA, and  $-0.16$  and  $0.68$  mas in Dec, respectively. The mean absolute offset is  $0.61$  mas, with a standard deviation of  $0.69$  mas. These results demonstrate that, the recovered source positions are consistent with the input (true) model position.

### 3.4.2 Recovering the source structure

We now analyze the performance of DECORAS with respect to recovering the true underlying structure of the source, which we parameterize as the source size. We have applied two different measures for this evaluation.

First, we calculate the MSLE between the predicted and input (true) model image. Fig. 3.17 presents the results of measuring the MSLE for the test dataset as a function of signal-to-noise ratio, for all sources detected with DECORAS. We find that the MSLE is of order  $10^{-5}$  in the majority of cases, for both point and extended sources, which translates to an acceptable source structure recovery. For example, in Fig. 3.18 we show four of the worst-case sources, where the MSLE is highest ( $> 3 \times 10^{-4}$ ). A simple visual comparison demonstrates that the model and predicted source structures are in good agreement, even for these outliers.

Second, to quantify the performance of DECORAS in recovering the source structure, we compare the major and minor axis of the 2-dimensional elliptical Gaussian fitted to the surface brightness distribution of each predicted and input (true) model image. Fig. 3.19 presents a comparison of the predicted and the ground truth major axis for each source, where we again see that there is good agreement. We have also calculated



**Figure 3.16** – Histograms of the relative offset, in Right Ascension ( $\Delta RA$ ) and Declination ( $\Delta Dec$ ), between the predicted and input (true) position of the sources detected by DECORAS. In both directions, the mean source position is consistent with being coincident with the input model position. Note that, among all the detected sources, 0.44 and 0.40 per cent of the sources have a  $\Delta RA$  and  $\Delta Dec$  that is higher than 5 milliarcsec, respectively.



the relative error between the true and predicted major axis, such that,

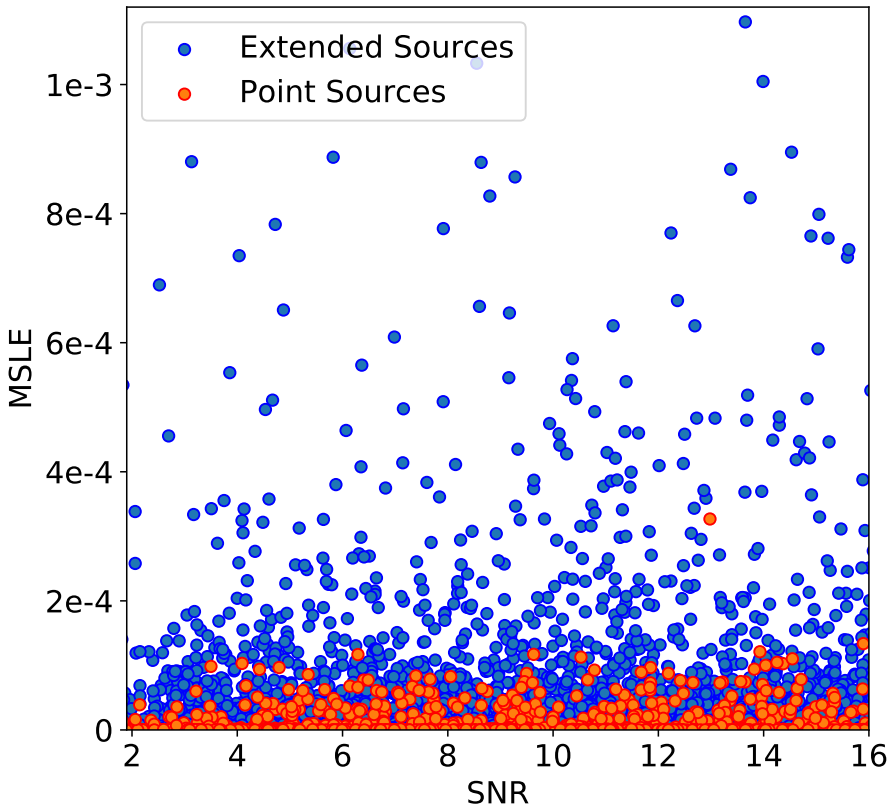
$$\text{Relative Error} = \frac{\text{True}_{\text{maj}} - \text{Predicted}_{\text{maj}}}{\text{True}_{\text{maj}}}, \quad (3.11)$$

which is also shown in Fig. 3.19. Again, we see that there is good agreement between the model and predicted source sizes, with the majority of the sources having almost no fractional difference. However, we see that there is still a scatter that extends up to 40 per cent in fractional error. We find that 88 (98) per cent of the sources have a fractional error of 10 (20) per cent in major-axis size. Only 2 per cent of the sources have fractional errors in the recovered major axis of between 20 to 40 per cent. Finally, we note that there is an excess of sources towards a positive relative error, that is, the predicted major axis is smaller than the input (true) model major axis. The reason for this is not clear, but may be related to the range of beam sizes that were used during the training.

### 3.4.3 Recovering the source peak surface brightness

Determining the absolute surface brightness is needed to measure the flux density and luminosity of the radio sources detected with DECORAS. This is important for understanding the various emission mechanisms that are at play, and for comparing the emission with other multi-wavelength datasets. Here, we use the peak surface brightness as a proxy for measuring the amplitude of the emission from the recovered sources. As discussed in Section 3.2.7, the image normalization process results in losing the absolute surface brightness information. However, this is recovered using the source brightness estimator via the maxratio (see equation 3.5) and the latent variables of Autoencoder2 (see Fig. 3.9). To do this, we must first measure the accuracy of the source brightness estimator for the test dataset. We considered the KNN, XGboost, BaggingRegressor, and the RandomForest Regressor, which all are implemented using the `scikit` package in Python (Pedregosa et al. 2012). Table 3.4 compares the performance of these different regressors in terms of the Root Mean Squared Error (RMSE) and the  $R^2$  statistic. We also give the standard deviation of the relative maxratio error distribution. We find that the XGboost has the best performance when compared to the other regressors, which we adopt for the rest of our analysis. It has been implemented using the methodology outlined by Friedman, Hastie & Tibshirani (2000); the number of gradient boosted trees is defined as 200, using a maximum depth of 7 for each of the base learners.

Fig. 3.20 compares the predicted and input (true) maxratio that is recovered for the test dataset. The result shows that using the latent variables and the maxratio can

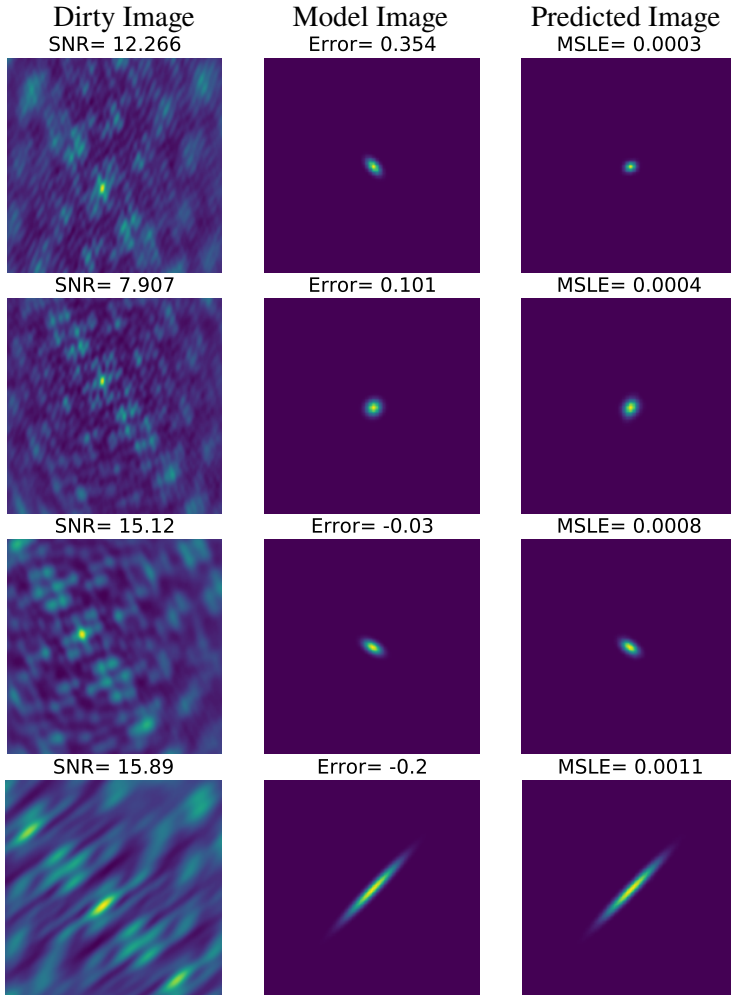


**Figure 3.17** – The MSLE between the predicted and input model images, as a function of signal-to-noise ratio for point (red) and extended (blue) sources. Overall, the MSLE values are extremely small and support our view that DECORAS reliably recovers the source structure.

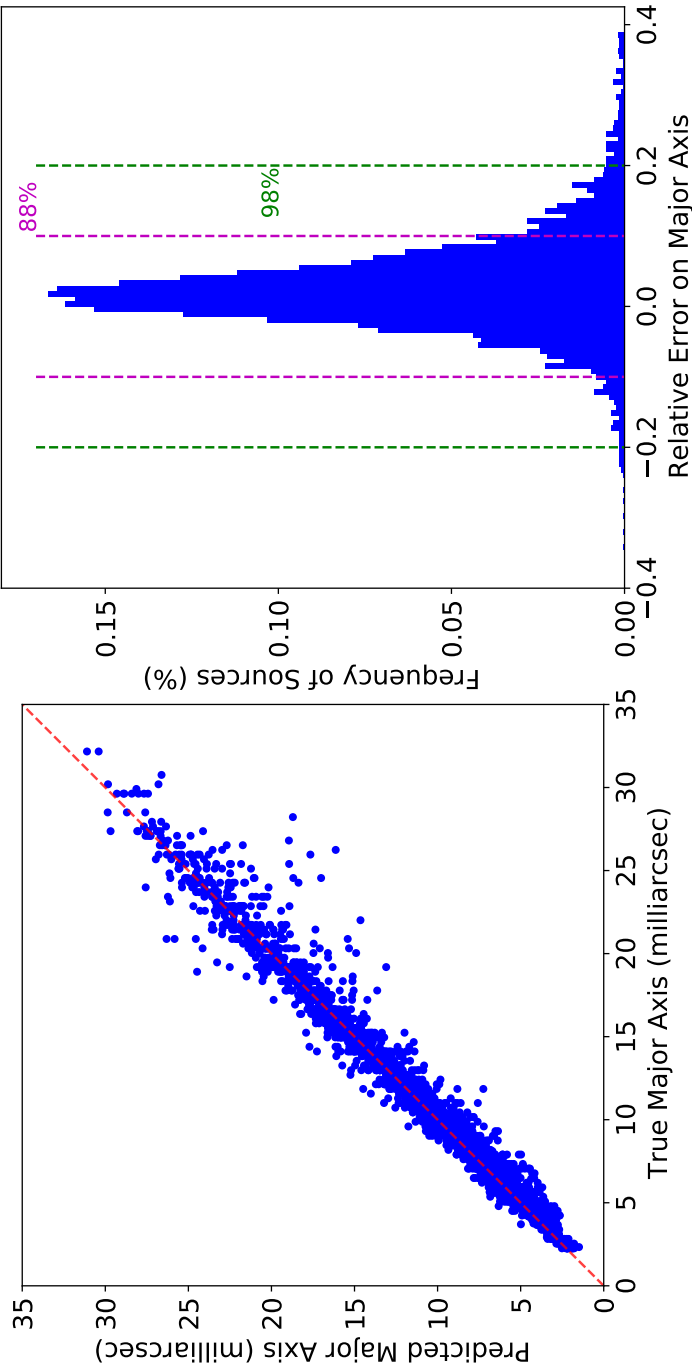
recover the model parameters. However, there is clearly a larger intrinsic scatter than in the case of the recovered major axis (see Fig. 3.19). To quantify how well the peak surface brightness is recovered, we again consider the relative error between the true and predicted parameters,

$$\text{Relative Error} = \frac{\text{True}_{\text{maxratio}} - \text{Predicted}_{\text{maxratio}}}{\text{True}_{\text{maxratio}}}, \quad (3.12)$$

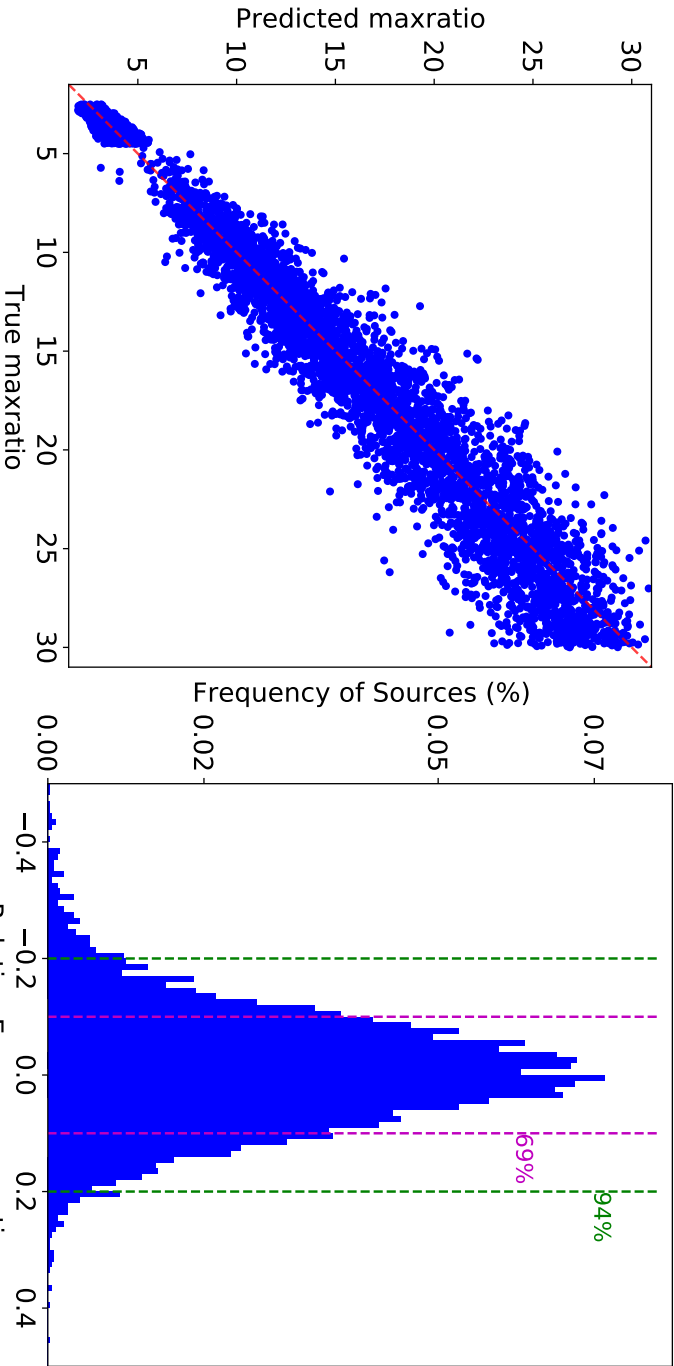
the results of which are also shown in Fig. 3.20. We find that the predicted peak brightness is almost exactly recovered in the majority of cases, but there is also a scatter that extends to a fractional error of 60 per cent, and is independent of signal-to-noise ratio. We find that 69 (94) per cent of the sources have a fractional error on their peak surface brightness of 10 (20) per cent. Given that the absolute



**Figure 3.18** – Examples of four sources with a high MSLE between the input (middle) and predicted (right) model image, with the associated dirty image (left) for reference. Such high values of the MSLE could be interpreted as a possible error in recovering the source structure. However, we see that the predicted and input images are comparable for these extreme cases. Each image contains  $256 \times 256$  pixels and is equivalent to a sky-area of  $320 \times 320 \text{ mas}^2$ .



**Figure 3.19** – In the left panel, a comparison of the predicted and true major axis of each source for the entire test dataset (blue points), is presented. The red dashed line shows when they exactly agree. In the right panel, the histogram of fractional error of the predicted major axis, with respect to the input (true) model major axis, is presented. The magenta and green dashed lines show the percentage of sources within the 10 and 20 per cent fractional error bounds.



**Figure 3.20** – In the left panel, a comparison of the predicted and true maxratio of each source for the entire test dataset (blue points), is presented. The red dashed line shows when they exactly agree. In the right panel, the fractional error of the predicted maxratio, with respect to the input (true) model maxratio, is presented. The magenta and green dashed lines show the percentage of sources within the 10 and 20 per cent fractional error bounds.

	KNN	XGBoost	Bagging	RandomForest
RMSE	2.19	2.16	2.44	2.23
$R^2$	94.3%	94.5%	92.6%	93.7%
$\sigma_{\text{Rel. Error}}$	0.15	0.13	0.16	0.15

**Table 3.4** – Evaluation of the results of applying various regressors to the source surface brightness estimator. Ideally, the RMSE and  $\sigma_{\text{Rel. Error}}$  should be as small as possible and the  $R^2$  statistic should be 100 per cent. The  $\sigma_{\text{Rel. Error}}$  is calculated by fitting a Gaussian function to the relative error distribution obtained for each method.

amplitude calibration of interferometric datasets is around 10 per cent, we conclude that our measurement errors with DECORAS will not dominate the uncertainties for the majority of the sources detected.

### 3.5 Discussion and Conclusions

Source detection and characterization will always play an important role in making new scientific discoveries, particularly in the age of large synoptic survey telescopes and interferometric arrays that will operate across all observable wavelengths. The shift to larger and more complex datasets requires robust and efficient automated approaches to be developed, many of which will employ deep learning techniques. Here, we have investigated the source detection and characterization of unresolved and extended sources in a single pipeline using machine learning techniques. Our method is designed to detect sources reliably from sparse interferometric arrays, like the VLBA, which can have highly correlated noise properties for images produced in the sky-plane domain. However, the pipeline presented here could also be used for other interferometric arrays, provided a suitable training dataset can be made. We have focused our attention on images that have not gone through a prior deconvolution process, but are instead produced from a Fourier transform of the visibility data. This was done to test how reliable such a methodology would be, as it can be extended to source detection in the visibility plane or be used to determine residual calibration errors in the visibility data (see below).

By applying our methodology to a test dataset, which is representative of observations with the VLBA at a wavelength of 20 cm, we find that the derived catalog is 100 per cent complete down to a signal-to-noise ratio of 7.5. When we used a traditional source detection algorithm, which is applied to the same dataset, but also having gone through a de-convolution process, the completeness drops from 100 per cent at a higher signal-to-noise ratio of 8.4. This improvement in detectability provided by DECORAS is equivalent to a 25 per cent decrease in the integration time needed when

compared to a traditional source detection algorithm that matches the completeness for a flux-limited sample. For example, an all-sky survey with the VLBA that reaches a similar depth to the mJIVE–20 survey would take about 7000 h to complete (see Chapter 2); applying DECORAS to such a survey could potentially save 1750 h in observing time, which is significant. Moreover, we find that DECORAS has a higher catalog purity, by almost a factor of two, when compared to a traditional source detection algorithm.

We also investigated the robustness of the source characterization using the test dataset. We found that the position of the detected sources were recovered to within 0.61 mas (0.49 pixels), with a standard deviation of 0.69 mas (0.55 pixels), from the input point source position. We also found that the peak surface brightness and size of the input sources were recovered to within 20 per cent for 94 and 98 per cent of the sources, respectively. Therefore, we conclude that the model images produced by DECORAS well represent the underlying source structure of the objects that are detected.

For recovering the source structure and source surface brightness, we had to develop a second encoder-decoder network due to the inefficiency of Autoencoder1 in recovering the source properties accurately. This is because Autoencoder1 and its latent variables were optimized to recovering the source position rather than recovering the source properties. Our experiments show that the accuracy of recovering the source surface brightness is highly dependent on the latent variables; the better the network is trained using Autoencoder2, the lower the error will be generated on the maxratio estimation. The quality of the generated latent variables also depends on the efficiency of the network structure. Looking at the maxratio distribution, we found that there was a wide range of maxratios for extended sources. However, the number of samples with a higher maxratio was significantly lower. This will need to be considered in any future learning regression model so that enough data is provided on all maxratio bins. We noted that although the source structure is correlated with the maxratio, it is not the only parameter that affects it. The other influential parameters are unknown to us at this moment, but we expect that the PSF sidelobe structure to have a lateral effect. Also, we expect that an additional framework that separates point and extended sources using the latent variables, and trains the regressors using either the point or extended source samples will help in lowering the error on maxratio estimation.

Our deep learning architecture is rather simple with only nine convolutional layers. It was designed in this way so that there was a very short training and testing time. The entire training phase of DECORAS on 50 000 samples on a GPU node takes less than 2 h (equivalent to 7 samples per second). We note that the training time does not include the time needed to produce the training dataset. However, as part

of this project, we have developed a pipeline to efficiently produce visibility datasets of realistic observations. This has formed the basis of an improved mock visibility dataset pipeline that generates training samples with more realistic source models, as opposed to the simple point and Gaussian source models tested here (de Roo et al., in prep.). Testing and refining DECORAS on this even more realistic dataset is our next goal before applying the algorithm to real observational data, for example, from the mJIVE–20 survey.

Also, through using more complicated source structures, the network can also be trained to identify the patterns associated with amplitude and phase errors within the visibility dataset. Currently, such errors would be absorbed in the derived source structure, which we plan to account for in a future implementation of DECORAS. This could be done by correcting the observed visibility's, or by simply accounting for the mis-match in the actual PSF from the expected PSF given the visibility sampling function; as our current implementation of DECORAS does not use the PSF or visibility sampling function for the analysis, testing (training) on a dataset with calibration errors should be a straightforward and potentially interesting next step.

Ultimately, we aim to expand this research by applying DECORAS to real observational data, such as from the mJIVE–20 survey. Due to the improved completeness and purity provided by DECORAS when compared to BLOBCAT, we expect to detect and better characterize more sources, and to generate a more reliable catalog for the mJIVE–20 survey. This would provide a real-world test for using machine learning techniques to detect and characterize the millions of sources to be found from the next generation wide-field surveys with SKA-VLBI.

## Acknowledgements

We thank Adam Deller for making the uvfits files for the mJIVE–20 survey available for our simulations. This chapter is based on research developed in the DSSC Doctoral Training Programme, co-funded through a Marie Skłodowska-Curie COFUND (DSSC 754315). JPM acknowledges support from the Netherlands Organization for Scientific Research (NWO) (Project No. 629.001.023) and the Chinese Academy of Sciences (CAS) (Project No. 114A11KYSB20170054). We would also like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.



## **Data Availability**

Upon reasonable request, the underlying data used for this chapter will be shared by the author.

## A deep learning based approach to gravitational lens identification with the International LOFAR Telescope

Based on "A deep learning based approach to gravitational lens identification with the International LOFAR Telescope"

S. Rezaei, J. P. McKean, M. Biehl, W. de Roo and A. Lafontaine

Submitted for publication in MNRAS

### Abstract

We present a novel machine learning based approach for detecting galaxy-scale gravitational lenses from interferometric data, specifically those taken with the International LOFAR Telescope (ILT), which is observing the northern radio sky at a frequency of 150 MHz, an angular resolution of 350 mas and a sensitivity of  $90 \mu\text{Jy beam}^{-1}$  ( $1\sigma$ ). We develop and test several Convolutional Neural Networks to determine the probability and uncertainty of a given sample being classified as a lensed or non-lensed event. By training and testing on a simulated interferometric imaging data set that includes realistic lensed and non-lensed radio sources, we find that it is possible to recover 95.3 per cent of the lensed samples (true positive rate), with a contamination of just 0.008 per cent from non-lensed samples (false positive rate). Taking the expected lensing probability into account results in a predicted sample purity for lensed events of 92.2 per cent. We find that the network structure is most robust when the maximum image separation between the lensed images is  $\geq 3$  times the synthesized beam size, and the lensed images have a total flux density that is equivalent to at least a  $20\sigma$  (point-source) detection. For the ILT, this corresponds to a lens sample with

Einstein radii  $\geq 0.5$  arcsec and a radio source population with 150 MHz flux densities  $\geq 2$  mJy. By applying these criteria and our lens detection algorithm we expect to discover the vast majority of galaxy-scale gravitational lens systems contained within the LOFAR Two Metre Sky Survey.

## 4.1 Introduction

Strong gravitational lensing occurs when the light from a distant galaxy is deflected due to the space-time curvature that is caused by another galaxy along the line of sight. As a result of this phenomenon, the foreground galaxy acts like a lens, and can produce multiple magnified images of the background galaxy (see Treu 2010 for a review).

Ever since the first gravitational lens was discovered by Walsh, Carswell, & Weymann (1979), gravitational lensing has become a powerful tool to test models for galaxy formation and cosmology. For example, over the last four decades, gravitational lensing has been used to measure the mass components of massive early-type galaxies (Treu et al. 2006; Bolton et al. 2008; Auger et al. 2009, 2010), constrain their stellar initial mass function (Spiniello et al. 2012, 2014), and determine their inner mass density profiles (Wucknitz, Biggs & Browne 2004; Koopmans et al. 2006; Spingola et al. 2018). Also, through detailed modelling of the surface brightness distribution of the lensed images, it has been possible to place constraints on the nature of dark matter (Vegetti et al. 2012, 2014; Ritondale et al. 2019; Hsueh et al. 2020; Gilman et al. 2020). When such modelling is combined with a time-delay observed between the different lensed images of a flux-variable background object, models for the expansion of the Universe and dark energy have also been tested (Suyu et al. 2010, 2013; Bonvin et al. 2017; Wong et al. 2020).

There have been several dedicated surveys to find gravitational lenses across the entire electromagnetic spectrum with both imaging and spectroscopic data (e.g. Patnaik et al. 1992; Myers et al. 2003; Bolton et al. 2006; Negrello et al. 2010; More et al. 2012; Treu et al. 2018; Spiniello et al. 2018). These surveys have discovered hundreds of examples of strong gravitational lensing, where the background galaxy is either a compact source associated with an Active Galactic Nucleus (AGN; quasar) or an extended source, primarily associated with the emission from stars in a galaxy (e.g. King et al. 1999; Browne et al. 2003; Bolton et al. 2008; Faure et al. 2008; Wardlow et al. 2013; Negrello et al. 2017; Lemon et al. 2018; Lemon, Auger & McMahon 2019; Lemon et al. 2020; Li et al. 2021).

Unfortunately, gravitational lensing by massive galaxies is quite a rare event, with

one gravitational lens found in about a thousand galaxies observed (Chae et al. 2002; Wardlow et al. 2013; Amante et al. 2020). This makes their identification from visual inspection both time consuming and prone to incompleteness (e.g. Jackson 2008; Marshall et al. 2016; More et al. 2016), as the parent population that needs inspecting tends to be of order  $10^4$  galaxies. Therefore, the vast majority of the gravitational lenses discovered thus far have been found through applying a set of selection criteria in catalogue space, based on the optical colour or radio spectral index, the total flux density and the morphology of the candidate lensed images. However, some level of visual inspection is still needed to verify potential lens candidates.

In the near future, the ever-increasing size of datasets from existing and proposed wide-field surveys necessitates sophisticated automated search techniques to identify new lens candidates. This is because with parent samples of order  $> 10^7$  galaxies, even applying various selection criteria will still result in a prohibitively large number of candidates requiring visual inspection. For example, it is expected from the large-scale imaging surveys to be carried out with the Vera C. Rubin Observatory, the *Nancy Grace Roman Space Telescope* and *Euclid* at optical/infrared wavelengths, and with the next generation Very Large Array (ngVLA) and the Square Kilometre Array (SKA) at radio wavelengths, that more than  $10^5$  gravitationally lensed galaxies will be discovered (Oguri & Marshall 2010; Collett 2015; McKean et al. 2015).

To test various identification techniques, Metcalf et al. (2019) recently carried out a lens finding challenge that focused on optical/infrared datasets. They tested a wide variety of automated techniques developed by the community, such as gravitational arc and ring finders (Cabanac et al. 2007; Sonnenfeld et al. 2018), machine learning and deep learning algorithms (Hartley et al. 2017; Petrillo et al. 2017; Schaefer et al. 2018; Lanusse et al. 2018; Avestruz et al. 2019), and also brute force visual inspection methods (Jackson 2008). They found that more than 50 per cent of lens systems could be identified, without any false positive events, using the automated approaches when certain thresholds on the lensed image brightness or size were applied. Significantly, the automated methods outperformed the brute-force visual inspection by experts in the field.

Arc and ring finder algorithms do not use any learning techniques (Cabanac et al. 2007; Sonnenfeld et al. 2018), but instead fit parametric lens models to any detected arc-like surface brightness distribution to determine the likelihood of it being lensed. The need to fit a model to each detected arc or ring can become computationally expensive, and the performance of such algorithms is limited to what we expect a lens system to look like. Moreover, the risk of a false positive detection for cases where spiral arms or tangentially elongated star forming regions appear as arc-like features

can be a problem for these types of algorithms. On the other hand, learning based techniques, in particular Convolutional Neural Networks (CNNs; Hezaveh, Perreault Levasseur & Marshall 2017; Petrillo et al. 2017, 2019; Jacobs et al. 2017, 2019a,b; Cheng et al. 2020b; Akhazhanov et al. 2021; Gentile et al. 2022) and Support Vector Machines (SVMs; Hartley et al. 2017), have been recently developed for detecting and modelling strong gravitational lens systems.

The advantage of using deep learning algorithms over traditional lens finding algorithms are multifold. They are less computationally expensive and can reduce the need for user involvement. The model-agnostic nature of learning algorithms is free to capture underlying structure that model-dependent methods may miss. However, it is still needed for users to provide the training data and label them into classes of lensed and non-lensed, leaving the machine to learn by itself the important features that describe a gravitational lens system. In this regard, deep learning is well-suited to identifying lensed features, as their surface brightness distribution is highly correlated via the lens equation. Indeed, recent applications of deep learning to lens identification in wide-field ground-based optical/infrared surveys have found hundreds to thousands of gravitational lens candidates (Li et al. 2021; Rojas et al. 2021). Although the vast majority of these candidates have still to be confirmed as genuine gravitational lenses, this highlights how powerful such techniques can be.

To date, there has been almost no research done in applying machine learning techniques for detecting gravitational lenses with radio interferometers; although, see Morningstar et al. (2018, 2019) for a discussion on using deep learning for image deconvolution and lens modelling. This is in part due to the availability of large amounts of wide-field multi-band optical imaging data from the Kilo Degree Survey (KiDS) and the Dark Energy Survey (DES), and the impending launch of *Euclid* (currently planned for early 2023). Also, as the data from interferometers is in the native visibility plane, this apparent complexity has made such studies seem additionally challenging. However, the next generation of radio interferometers will have the sensitivity and angular resolution to be excellent gravitational lens finding machines, with several unique science applications (e.g. McKean et al. 2015).

Here, we focus on developing a lens detection algorithm that can be applied to large-area surveys ( $15\,000\text{ deg}^2$ ) at high angular resolution (5 to 500 mas synthesized beam size) with next generation radio interferometers. In particular, we concentrate on lens surveys with the Low Frequency Array (LOFAR; van Haarlem et al. 2013), which mainly operates between 120 and 170 MHz, and has the sensitivity to detect the non-thermal emission from lensed radio sources (Stacey et al. 2019) and the long baselines needed to resolve their structure (Badole et al. 2022). Also, given that

the International LOFAR Telescope (ILT) has now gone through the commissioning phase (Morabito et al. 2021; Jackson et al. 2021; Bonnassieux et al. 2021; Harwood et al. 2021; Sweijen et al. 2021; Timmerman et al. 2021), the routine imaging of the survey data with the international stations of LOFAR will soon start (Sweijen et al. 2022). Therefore, developing methods for identifying lensed radio galaxies amongst the expected 15 million objects to be imaged with LOFAR is also timely.

This chapter is arranged as follows. In Section 4.2, the procedure for producing the training and verification data of simulated observations with the ILT is presented. Also, we present the detailed methodology of the lens detection algorithm. In Section 4.3, we evaluate the results of the lens detection algorithm, and the lens-parameter space that the ILT is sensitive to is determined in Section 4.4. Finally, in Section 4.5, the results from this work are discussed, and concluding remarks on the methodology and future prospects are presented.

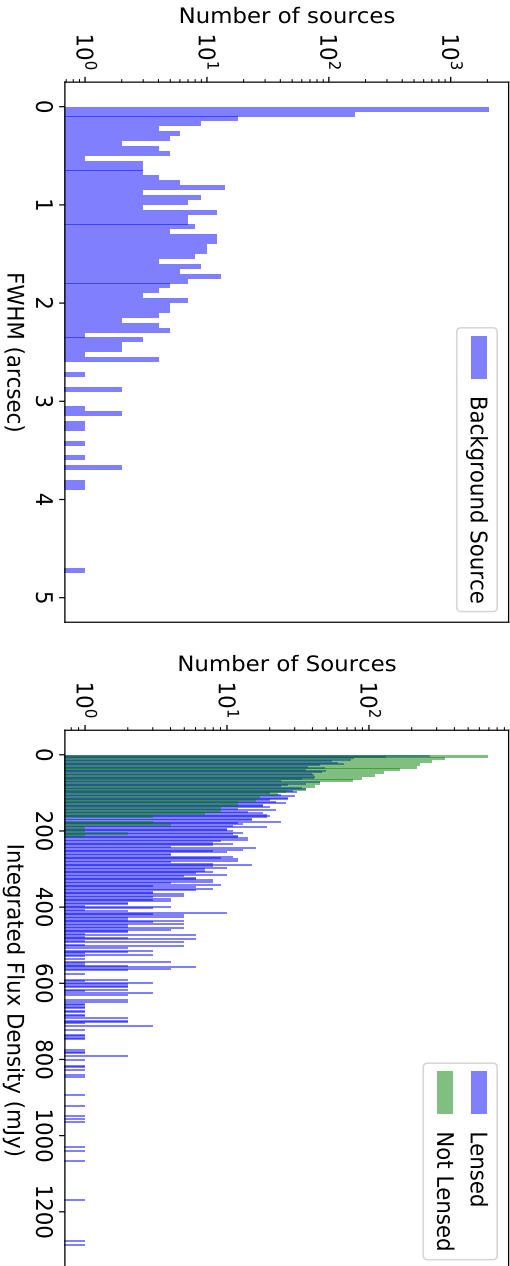
## 4.2 Method

In this section, we first summarise our pipeline for simulating realistic gravitational lensing data from the ILT (for training and testing purposes). Then, we outline the architecture of the deep learning algorithm we use to detect and rank gravitational lensing candidates.

### 4.2.1 Simulating a training dataset for the ILT

The process of simulating realistic observations of gravitational lenses with the ILT will be presented in detail by de Roo et al. (in prep.). In summary, the lensing data are created using `PYAUTOLENS` (Nightingale et al. 2021), an open source code that generates simulated images through the process of ray-tracing. Although `PYAUTOLENS` also has the functionality to produce interferometric datasets, here we have only incorporated the ray-tracing component in our pipeline to create a mock lens-plane model from a parametric source-plane model. We then make our own visibility datasets and produce deconvolved images using the Common Astronomy Software Applications (CASA; McMullin et al. 2007) package.

The lens simulation platform generates mock analytic sources and places them randomly on a grid. For simplicity, we have chosen to represent the background source surface brightness distribution with multiple elliptical Gaussian components. Fig. 4.1 shows the distribution of full width at half maxima (FWHM) for the simulated background source components in our training dataset. This distribution of source sizes is based in part on the typical sizes of radio emitting star-forming galaxies at



**Figure 4.1** – The left panel shows the distribution of full width at half maximum (FWHM) of the Gaussian background sources used for the training dataset. The right panel shows the integrated flux density of the background source population (green) in comparison to the integrated flux density of the lensed source population (blue). The effect of the magnification provided by the gravitational lensing is clearly seen in the source number counts as a function of flux density:

1.5 GHz (Muxlow et al. 2020), but as most of our sources will likely be associated with AGN activity, which can have a variety of sizes, we have also added a distribution of sources with sizes  $< 0.5$  arcsec. This upper-limit on the AGN source size was chosen to be consistent with recent ILT observations of the Lockman Hole region, where 88 per cent of the detected sources were found to be compact at 144 MHz (beam size 0.4 arcsec; rms  $25 \mu\text{Jy beam}^{-1}$ ; Sweijen et al. 2022). Note that sources much larger than 2 arcsec in size, that is, large-scale radio lobes, are rarely gravitationally lensed (Haarsma et al. 2005). Therefore, we initially chose not to densely sample source sizes larger than 2.5 arcsec (see Section 4.3.4, where we sample unlensed sources that are extended between 0.5 and 6 arcsec). As can be seen from Fig. 4.1, the majority of the simulated source components in our training dataset are  $< 0.5$  arcsec in size. The simulated background sources could have up to three Gaussian components, drawn from the distributions given in Fig. 4.1, in order to replicate a typical core and double-lobe structure, or a one-sided core-jet structure. In each case, the first component is compact and is injected at a random position from a uniform circular distribution around the lens centre; the additional components are then added co-linearly to simulate jetted radio sources. Also shown in Fig. 4.1 is the integrated flux density distributions of the background source and lensed source populations. Here, the magnification effect of gravitational lensing can be clearly seen.

The resulting mock lensed images of the simulated sources were then generated using a singular isothermal ellipsoid (SIE) mass model, with an external shear contribution. Such a model has been shown to be a good approximation for the mass distributions of massive elliptical galaxies (e.g. Koopmans et al. 2006). The parameters used to describe this mass model are the lens position (always at the centre of the grid), the axis ratio ( $b/a$ ) and position angle of the ellipsoid, the shear strength ( $\gamma_{\text{ext}}$ ) and position angle, and the Einstein radius ( $\theta_E$ ), which is used as a proxy for the lensing mass. For our initial tests (see Section 4.3), the mass models were generated using a combination of these parameters drawn from the distributions of real gravitational lens mass models (Bolton et al. 2008), except for the Einstein radius where the image-separation distribution for all known lensed quasars was used (the Einstein radius is approximately half of the maximum image separation). The resulting distributions for these lens model parameters are shown in Fig. 4.2. However, for our final tests (see Section 4.4), we use a uniform distribution for the Einstein radius and the axis-ratio to obtain an unbiased assessment of the network’s ability to identify gravitational lenses. Throughout, we only search for lens systems with Einstein radii  $\geq 0.15$  arcsec, to exclude part of the model space that the ILT is likely not sensitive to, given the resolution of the data. For simplicity, we fixed the lens redshift to be  $z_l = 0.5$  and



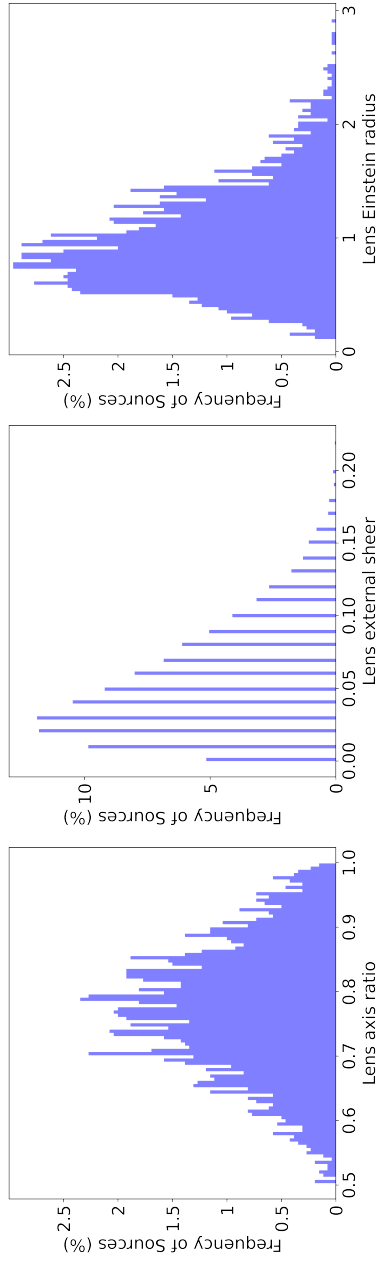
the source redshift to be  $z_s = 1.0$ . This choice will have no impact in our simulations since it is the angular-scale, as set by the varying Einstein radius that matters.

In the next step, we generated realistic interferometric visibility datasets, which was done by using *CASA*. First, we used an actual observation with the ILT of the bright radio galaxy 3C 330 (Lafontaine et al. in prep.), to provide a skeleton MeasurementSet with the correct number of stations,  $uv$ -coverage and visibility averaging time and bandwidth. This simulation consisted of the phased-core station, fourteen remote stations, and thirteen international stations of LOFAR. The total time on-source was set to 8 h, which is typical for an observation that is taken as part of the LOFAR Two Metre Sky Survey (LoTSS; Shimwell et al. 2019). Each lensed and non-lensed model surface brightness distribution was sampled in the visibility plane via a Fourier transform, and the resulting visibilities were then corrupted by adding Gaussian noise such that the final images had an rms noise of  $90 \mu\text{Jy beam}^{-1}$  (Briggs weighting; Robust = 0.5;  $uv$ -range >  $80k\lambda$ ); this is the typical rms noise for an ILT observation (Morabito et al. 2021). The resulting beam size was  $379 \times 293$  mas at a position angle of  $-16.9$  deg East of North. The simulated visibility datasets were then imaged and deconvolved using the *TCLEAN* task within *CASA*. The final images have  $64 \times 64$  pixels, where each pixel is 0.12 arcsec in size.

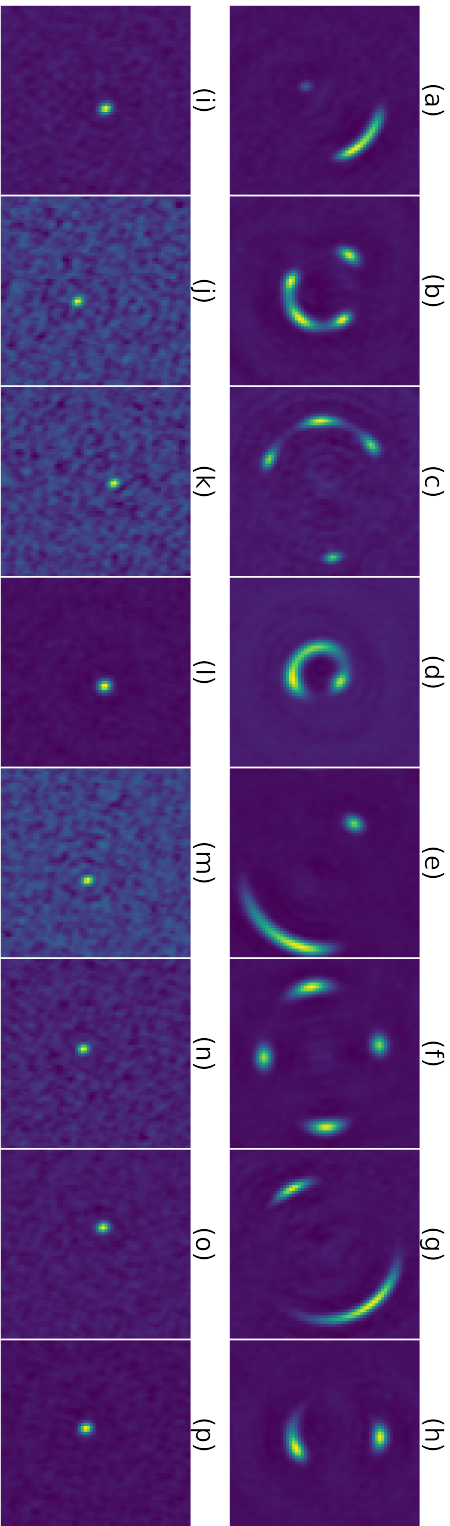
Fig. 4.3 shows a subset of the simulated gravitational lens systems and non-lensed radio sources as if they were observed with the ILT as part of LoTSS. We note that although all of the presented samples have arcs or rings, our training dataset also includes other types of gravitational lens systems, such as those producing two images with compact lensed emission. We have also included a wide range of signal-to-noise ratios to evaluate the performance of our lens detection algorithm when the source surface brightness is close to the rms noise level of the data.

### 4.2.2 Network structure

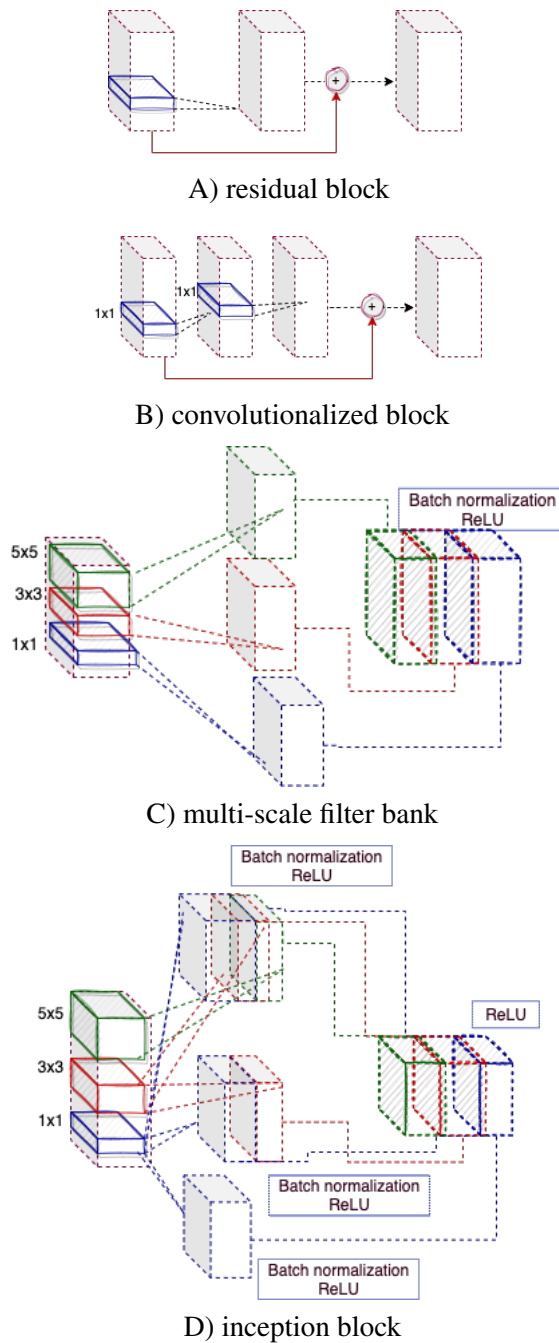
In this section, we present our CNN based approach to detect and rank strong gravitational lens systems from ILT-quality imaging data at 144 MHz. CNNs use several layers to process the input imaging data of a topological structure, and they have become the prominent approach for object detection and classification with machine learning. Each layer has a unique functionality, and different arrangements of the layers can make a variety of convolutional components. Therefore, the performance of a CNN is dependent on the implemented components, and each component may perform differently for a specific application. We have implemented four main components that are being employed in three separate network structures, which we call structure 1, 2 and 3. The three structures are then trained with the same training



**Figure 4.2** – The distribution of the lens model parameters used for the training dataset; these are (left panel) the lens axis ratio (b/a), (middle panel) the lens external shear ( $\gamma_{\text{ext}}$ ), and (right panel) the lens Einstein radius ( $\theta_E$ ). The position angles of the ellipsoidal mass distribution and the external shear were set randomly between  $\pm 90$  deg.



**Figure 4.3** – A selection of strong gravitational lens systems (upper row) and non-lensed radio sources (lower panel), as if observed with the ILT as part of LoTSS. Each image contains  $64 \times 64$  pixels and is equivalent to a sky-area of  $7.68 \times 7.68$  arcsec<sup>2</sup>. The synthesized beam size is  $379 \times 293$  mas at a position angle of  $-16.9$  deg. east of north, and the rms noise is  $90 \mu\text{Jy beam}^{-1}$  (robust weighting).



**Figure 4.4** – A block diagram showing the main components that have been used to construct the three CNNs tested here to detect and rank gravitational lens candidates in ILT imaging data.

dataset and their performance is evaluated using the same test and validation datasets. In the following, we first provide our motivation for using each of the four main components, before outlining the arrangement of these components to build the three network structures.

The training of the networks is guided by the gradient based optimization of a suitable loss function. Here, the gradient refers to the change of the loss function with respect to the weights in the network. For networks with many convolutional layers, the magnitude of the gradient in earlier layers can become small due to the multiplications imposed by the chain rule. As a result, updates are diminutive and the progress of the training can be slow. This "vanishing gradient problem" can occur faster when the number of layers in the network is increased, and it motivated the idea of using "residual blocks" to limit its effect (He et al. 2015). Part A of Fig. 4.4 shows the concept of residual blocks.

A convolutional block is presented in part B of Fig. 4.4. It has  $f$  convolutional filters, each with a filter size of  $(1 \times 1 \times d)$ , where  $d$  is the depth of the input image. As we currently use only single images produced from interferometric datasets, there is only one dimension containing information on the source surface brightness, and therefore,  $d$  is equal to 1. In the future, multiple images that take into account the source surface brightness distribution as a function of frequency will be implemented; in this case  $d > 1$ . A convolutional block with  $f$  filters is considered to be equivalent to fully connecting the input object of  $1 \times 1 \times d$  to  $f$  output nodes (Lee & Kwon 2017). Residual learning is implemented in the convolutional blocks to improve the training efficiency when extracting source components in feature space. Besides the convolutional block, we have also adapted the concept of a multi-scale convolutional filter bank (part C in Fig. 4.4) from the same study (Lee & Kwon 2017). It is used to simultaneously scan through local regions of the input image and then exploit various local spatial structures. It then concatenates the extracted feature maps to be used together. A similar concept is employed in an inception-based block, which was first introduced in GoogLeNet by Szegedy et al. (2016), and provides a deep, but also a wide structure that allows several independent paths in the model to be optimized. The block in part D of Fig. 4.4 is inspired from the inception block and provides three paths from the input. The first path starts with a filter of size  $(1 \times 1)$ , while the second path applies another convolution with a filter of size  $(3 \times 3)$ , after convolving with the filter of size  $(1 \times 1)$ . The last path contains both  $(1 \times 1)$  and  $(3 \times 3)$  sized filters that are followed by a  $(5 \times 5)$  sized filter bank. This technique has been implemented due to the varying size and structure of the gravitational lens systems that we are interested in detecting. Due to the dependence of the morphology of the lensed images to the

different input lens and source parameters, it is important to build a flexible model that can handle different configurations of the lensed images.

Considering the main components shown in Fig. 4.4 as the building blocks, we now present the three different network structures that are trained to separate and rank images of gravitational lenses from those of non-lensed samples. Fig. 4.5 shows the first network, structure 1, which contains a multi-scale filter block followed by an inception block. A dropout layer (Gal & Ghahramani 2015) followed by two fully connected layers is placed at the end. The dropout layer randomly removes units with a probability of  $p_d$  at each step during training, without updating the weights of those units. This helps to prevent overfitting by reducing the effective network complexity. The second network, structure 2, is displayed in Fig. 4.6. It contains a convolutional layer with a filter size of  $(5 \times 5)$ , followed by a multi-scale filter bank and a set of three convolutionalized blocks. At the end, there are two fully connected layers with a dropout function. Finally, in Fig. 4.7 we present the third implemented network, structure 3, which uses two sets of four convolutional layers, each with a filter size of  $(3 \times 3)$ . The two sets of convolutional layers are connected via a multi-scale filter bank. Similar to structure 2, this network contains three convolutionalized blocks, followed by two fully dense layers and a dropout layer.

### 4.2.3 Preprocessing

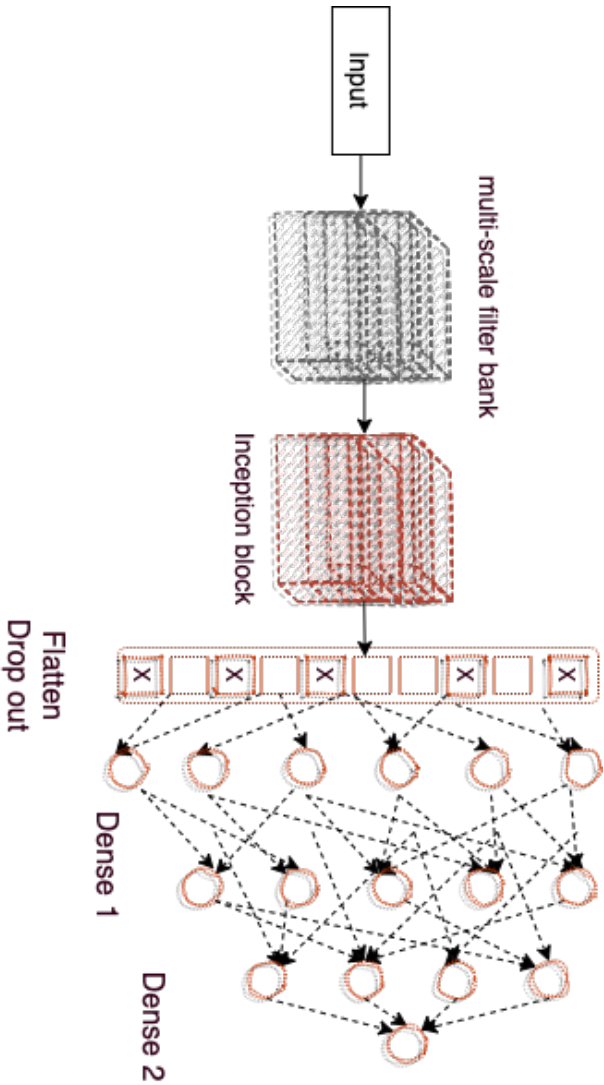
Preprocessing of the imaging data is important for optimizing the performance of the learning algorithms. Here, we keep the range of pixel values for both the lensed and non-lensed samples in the same range using a MinMax normalization, such that

$$x_{\text{normalized}} = \frac{x - \min(x_d)}{\max(x_d) - \min(x_d)}, \quad (4.1)$$

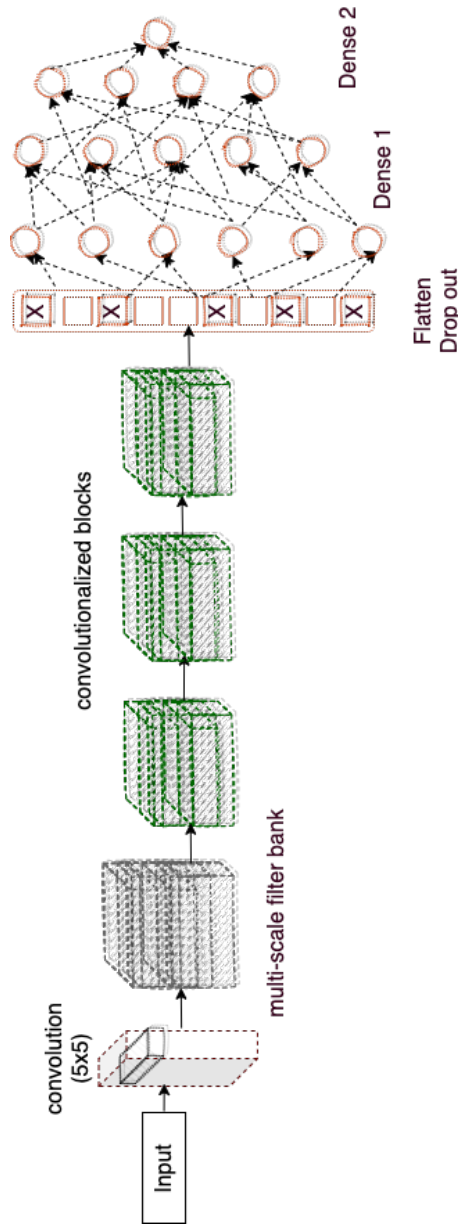
and where all pixel values in a given image are mapped to the range  $(0, 1)$ . Here,  $x$  is the value of a given pixel and  $x_d$  represents all the pixels in the image. Note that, due to surface brightness correlations introduced by the lens equation, it is the relative surface brightness of the lensed images (and non-lensed source emission) within each simulated sample that matters. Therefore, losing the absolute amplitude scaling of the data with this form of normalization will have no effect on our ability to identify lens candidates.

### 4.2.4 Loss function

Loss functions are designed to measure the performance of a network in terms of the dissimilarity between the estimated and true class labels. This difference is

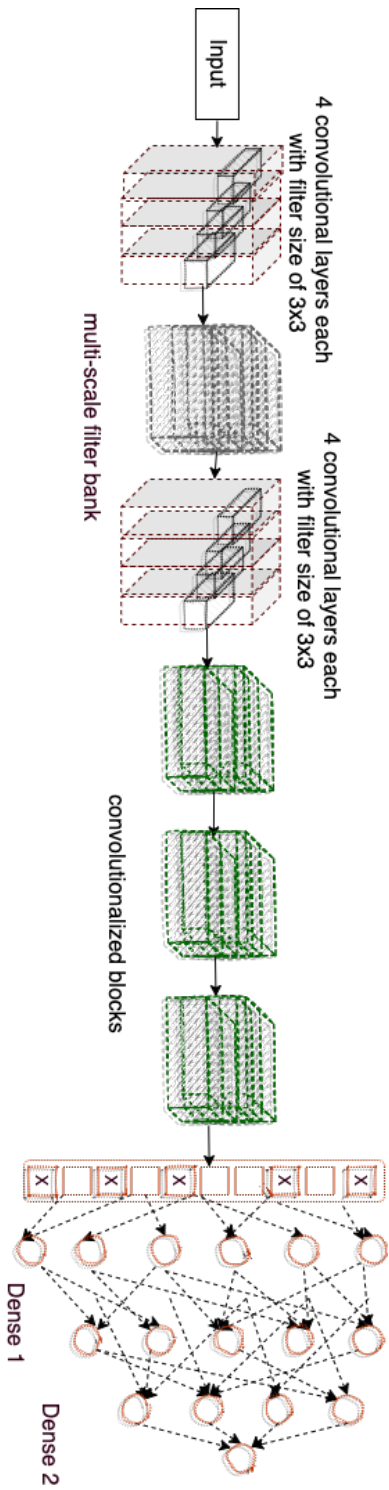


**Figure 4.5** – Structure 1: This network structure consists of a multi-scale filter bank and inception blocks that are placed next to each other. The output of the inception block is then sent to the dropout layer, which is connected to two dense layers of fully connected networks. There are in total 6 480 753 trainable parameters in this architecture.



**Figure 4.6** – Structure 2: For this structure the input is convolved with 128 filters of size  $(5 \times 5)$  before a multi-scale filter bank is applied. Next, three convolutionalized blocks are placed, where the residual blocks are used with three convolutional layers. Similar to structure 1, a dropout layer and two fully connected layers are implemented to extract the output of the network. The total number of trainable parameters in this structure is 4 538 497.





**Figure 4.7 – Structure 3:** In addition to the convolutionalized blocks and multi-scale filter bank used for structures 1 and 2, this structure also consists of two sets of 4 sequential convolutional layers, each with a filter size of  $(3 \times 3)$ . With 7 273 813 trainable parameters, this structure has the largest number of parameters.

then optimized using the gradient based Adam optimization algorithm (Kingma & Ba 2014). Since we perform a binary classification between lensed and non-lensed objects, we have chosen the Binary Cross Entropy (BCE) form of loss function for our tests. It is given as,

$$\text{BCE} = -\frac{1}{N} \sum_{i=1}^N y_i \log p_i + (1 - y_i) \log(1 - p_i). \quad (4.2)$$

For the BCE loss function, we consider a given class label  $y_i$  in a dataset of  $N$  training samples in each batch  $i$ , where  $p_i$  is the estimated probability of the input image being a lens and  $1 - p_i$  is the probability of it not being a lens.

#### 4.2.5 Detection probability uncertainty

A key goal of our lens identification algorithm is to also give some confidence on the level of detection. This is important because even though we may expect to identify many candidates with the ILT (and the ngVLA and SKA in the future), confirming the gravitational lensing hypothesis will also require additional telescope time. Therefore, it is important to have a ranked list of candidates for prioritising any follow-up observations, with some estimate of the uncertainty on the lensing probability. This is why we have also implemented a dropout component (Gal & Ghahramani 2015) in our three network structures (see Figs. 4.5 to 4.7). This method imitates Bayesian models within the deep learning algorithm without changing the network structure and the basic method of optimization.

Traditionally, dropout has been applied in deep learning algorithms to prevent overfitting and to regularize the model (e.g. Srivastava et al. 2014). During the training phase, dropout is used by randomly removing units and their corresponding connections from the network with some probability  $p_d$ . By applying dropout, a new set of neurons is eliminated at each iteration, resulting in a virtual change to the network's structure. This provides a way of approximately combining many different neural networks together with shared weights. The ratio of eliminated units at each iteration is defined by  $p_d$ . For a wide range of networks and applications,  $p_d = 0.5$  tends to generate close to optimal results (Srivastava et al. 2014).

We have also implemented a Monte Carlo dropout approach, which can be described as the approximate integration of dropout over the weights of the models. It is based on running the network for some number of iterations on a test dataset and building a distribution of the predicted probabilities. The mean and standard deviation of this distribution contain information about the most reliable prediction, as well as

the model uncertainty. Dropout has become a popular method to measure model uncertainties in gravitational lensing applications of deep learning (Hezaveh, Perreault Levasseur & Marshall 2017; Perreault Levasseur, Hezaveh & Wechsler 2017; Bom et al. 2019; Maresca, Dye & Li 2021).

#### 4.2.6 Evaluation criteria

Defining proper classification criteria is needed to interpret and compare the results of the three different network structures tested here. Since the lens detection problem can be defined as a binary classification (i.e. lensed or non-lensed), we use the reference confusion matrix presented in Table 3.1 to evaluate our results. Here, a true positive (TP) is defined as a genuine gravitational lens system that has been successfully classified as such, while a true negative (TN) is defined as when a non-lensed source has been correctly identified as not being a gravitational lens. Conversely, a false positive (FP) occurs when a non-lensed source is categorised as a gravitational lens, and a false negative (FN) corresponds to those gravitational lensing events that are missed by the algorithm, and are classified as non-lensed sources. An ideal network would detect all gravitational lens systems and reject all non-lensed samples. However, this can hardly ever be achieved in a practical real world problem due to noise in the data.

We evaluate the performance of the three network structures by using the following criteria: accuracy, precision, recall and fall out. Accuracy is the most trivial metric to determine as it is the total number of correct predictions for both the lensed and non-lensed classes divided by the size of the test dataset. Precision measures the ratio between the correctly identified gravitational lenses and all samples identified as gravitational lenses. Recall on the other hand measures the ratio of the detected gravitational lenses to the number of lensing events in the test dataset. This is also known as the completeness or TP rate. Fall-out, or the FP rate, measures the ratio of misidentified gravitational lenses to the total number of non-lensed events in the test dataset. They are each calculated using,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}, \quad (4.3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (4.4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (4.5)$$

		True Data		
		Lens	Not Lens	
Test Results	Lens	TP	FP	
	Not Lens	FN	TN	
		Total	Lens Sources	Normal Sources

**Table 4.1** – The sample confusion matrix, showing how the TP, FP, FN and TN are defined.

$$\text{Fall out} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \quad (4.6)$$

Typically, the success of a gravitational lens search algorithm is not judged purely on recall, as given the large number of gravitational lenses that are expected to be found, completeness is not important for most science goals. Instead, the quality tends to be judged on a low fall-out, as we ideally want to have a high level of genuine lens candidates in our ranked list.

Another evaluation metric is the Receiver Operating Characteristic (ROC) curve. It is used when the performance of a binary classifier can be measured as a function of some cut-off threshold. A ROC curve shows the TP rate (recall) as a function of the FP rate (fall out). In our lens detection algorithm, we have defined a threshold based on the output of the network, which is used to separate lensed and non-lensed samples. The choice of this threshold has a significant impact on all the evaluation metrics listed above, which we discuss in the next section. The ROC curve provides a qualitative measure that is independent of a specific threshold, and is therefore, also a useful metric to consider.

### 4.3 Network tests

In this section, we present the results of applying the three network structures to a set of test data, which were created in a similar way to the training dataset described in Section 4.2.1. First, we present the criteria used to determine whether a sample has been identified as a lensed or non-lensed event. Next, we investigate three scenarios that have been designed to evaluate the performance of each network structure. The first scenario uses a test dataset with a realistic distribution of lens model parameters, which allows us to test the network performance with a dataset that is similar to the training dataset. The second scenario contains test samples generated with a uniform population of lens model parameters. This allows us to test the performance of each structure to exotic lens configurations that were not necessarily well-represented in the training data. For these two tests, we trained the networks using 3000 model

lensed images, to represent the class of lensed objects, and 3000 corresponding source models to represent the class of non-lensed objects (generated as described in Section 4.2.1). In total, a maximum of 450 training epochs were used. Our third test is designed to investigate the ability of the three networks to correctly label non-lensed double-lobed radio sources. For this test, we augmented the training data set with 2000 simulated double-lobed radio sources. The test datasets for each experiment are described below.

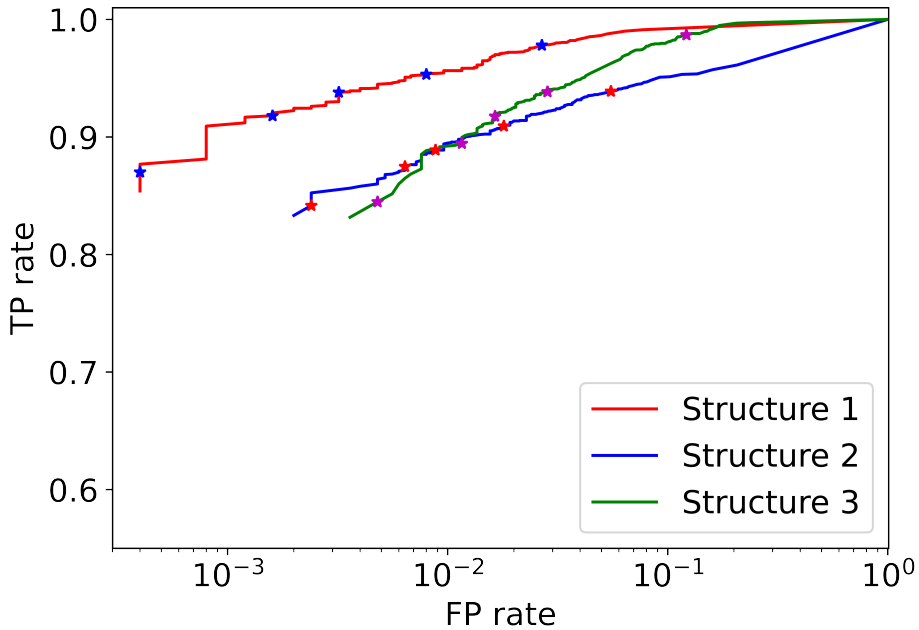
### 4.3.1 Determining the lens probability

A lensing probability in the range  $(0, 1)$  is predicted by the output of the three network structures, where the closer the output is to 1, the higher the probability of the given sample being a gravitational lens. Since we have used dropout in the training, we are able to calculate the model confidence for a specific prediction. When dropout is also active during the testing mode, the network randomly eliminates some of the connections and makes the best predictions based on the available data. Not having access to the entire trained set of weights and neurons leads to varying outputs for a given sample during testing. To calculate the most reliable output, we averaged the network prediction over 250 realizations for each single test sample to calculate the final probability. This approach also provides additional information by calculating the standard deviation of the probability.

Considering the output of the network as a continuous number between 0 and 1, we need to define a threshold according to which test results are classified into lensed and non-lensed samples. The value of the threshold affects all evaluation criteria such as completeness and number of false detections. In general, higher threshold values yield a lower detection of true lens samples, as the criteria are stricter, but this can also lower the number of false detections. We determine this threshold from inspecting the ROC plot for each network structure.

### 4.3.2 Test on a realistic lensing population

In the first testing mode, we have generated a realistic dataset with the same lens model parameters as our training data, shown in Fig. 4.2. Here, we test a dataset containing 5000 samples, where 2500 samples are non-lensed events and 2500 samples are lensed events that are created from a realistic distribution of lens models. The results for this test are shown in the ROC plots presented in Fig. 4.8. Overall, network structure 1 outperforms the two other structures in both the TP- and FP-rates at all probability thresholds. It has a TP rate of 95.3 per cent for a threshold of 0.5, while the FP rate (0.8 per cent) is significantly lower compared to structures 2 and 3. Although the TP



**Figure 4.8** – The TP rate as a function of FP rate for the three network structures, when applied to a test dataset created using a realistic distribution of lens model parameters. The stars correspond to thresholds of 0.1, 0.5, 0.75, 0.9 and 0.99, from right to left.

rates of structures 2 and 3 are high (> 90 per cent) for a threshold of 0.5, they also have a higher FP rate (1.8 and 2.8 per cent, respectively).

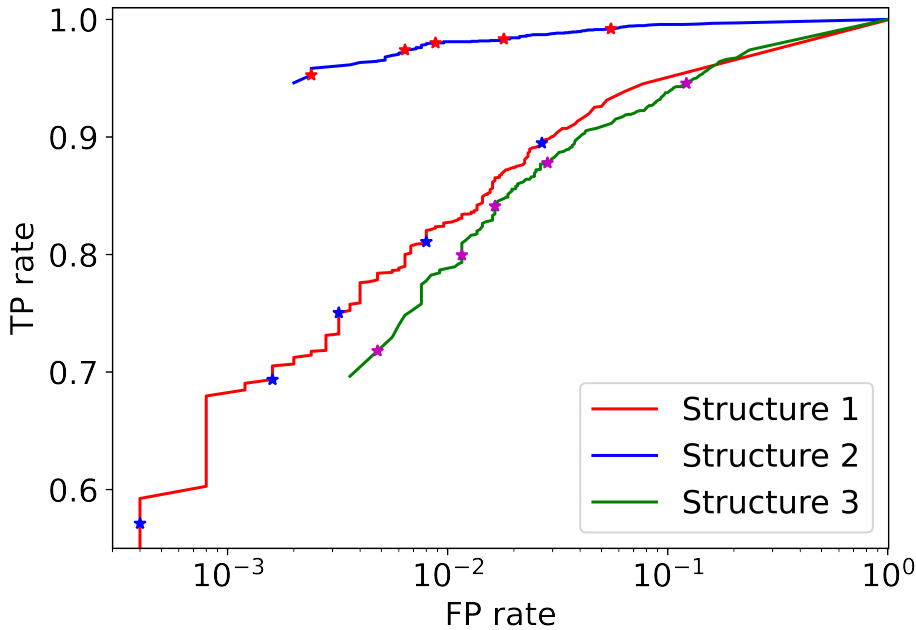
Table 4.2 contains quantitative information on the defined evaluation metrics. These metrics are calculated by considering a threshold value of 0.99, which results in lowering the TP rates by between 6.8 and 9.4 per cent, while lowering the FP rates by around an order of magnitude, when compared to the results for a threshold of 0.5. We find that network structure 1 has the best performance, with TP and FP rates of 87 and 0.04 per cent, respectively, when a realistic distribution is used for the lens model parameters and a threshold of 0.99 is applied.

### 4.3.3 Test on a lens population of uniformly selected model parameters

We now provide the test results when the trained models are applied to a test dataset using a uniform distribution of lens model parameters. Similar to the previous case, the test dataset contains 5000 samples, of which 2500 samples are lensed events and 2500 are non-lensed events. Note that the set of non-lensed radio sources is the same as used above, therefore, the FP rates are unchanged. The generated test dataset has a

	Structure 1	Structure 2	Structure 3
Accuracy	0.957	0.934	0.943
Precision	0.9995	0.9971	0.9943
Recall	0.870	0.842	0.845
Fall out	0.0004	0.0024	0.0048

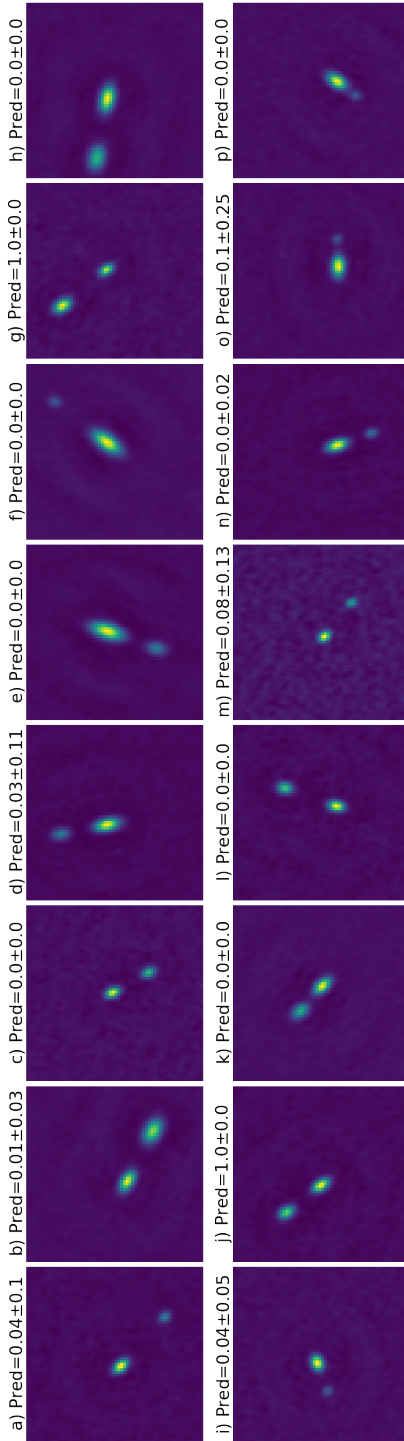
**Table 4.2** – The evaluation criteria of the three network structures when a threshold of 0.99 is used, and a test dataset created using a realistic distribution of lens model parameters is used.



**Figure 4.9** – The TP rate as a function of FP rate for the three network structures, when applied to a test dataset created using a uniform distribution of lens model parameters. The stars correspond to thresholds of 0.1, 0.5, 0.75, 0.9 and 0.99, from right to left.

	Structure 1	Structure 2	Structure 3
Accuracy	0.785	0.975	0.856
Precision	0.9990	0.9974	0.9933
Recall	0.571	0.952	0.718
Fall out	0.0004	0.0024	0.0048

**Table 4.3** – The evaluation criteria of the three network structures when a threshold of 0.99 is used, and a test dataset created using a uniform distribution of lens model parameters is used.



**Figure 4.10** – A representative sample of simulated (non-lensed) double-lobed radio sources. The second component is injected randomly within 3 arcsec distance of the first component. The size of the two components are drawn randomly from the distribution shown in Fig. 4.1. The average predicted probability and its uncertainty over 250 iterations for structure 2 is stated above each image. Each image contains  $64 \times 64$  pixels and is equivalent to a sky-area of  $7.68 \times 7.68$  arcsec<sup>2</sup>.



uniform distribution for the Einstein radius (between 0.15 and 2.8 arcsec) and the axis ratio (between 0.05 and 1). This was done to ensure that each bin in the parameter space was equally sampled, which makes the sensitivity of the three networks, as a function of the lens model used, clearer to judge. Note that we draw the shear strength from the realistic distribution, as extremely high values of the shear lead to unusual and highly exotic lensed image configurations (particularly when the axis ratio of the lens is low). As with the training dataset, we also set the position angle of the ellipsoid mass distribution and the shear to be uniformly sampled between  $\pm 90$  deg.

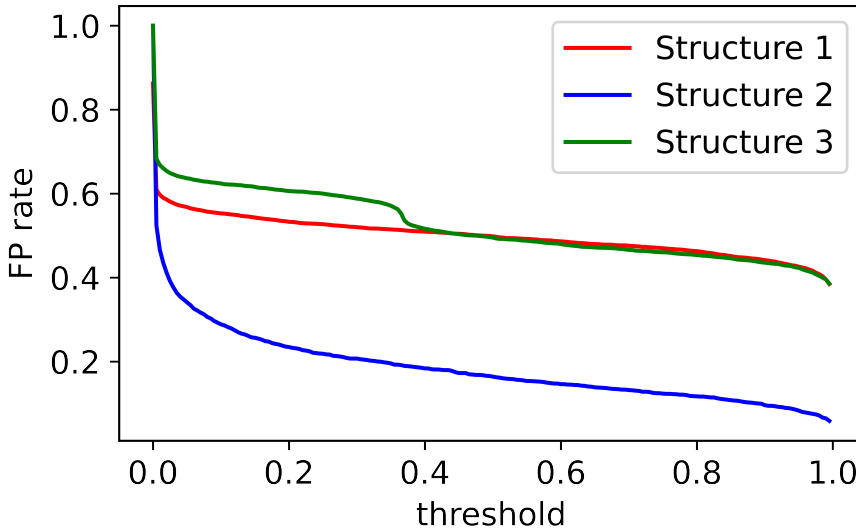
The ROC plots of the three network structures, with the probability thresholds labelled, are shown in Fig. 4.9. Comparing the results with those shown in Fig. 4.8, we find a significant drop in the TP rate at all thresholds for network structures 1 and 3, while network structure 2 performs better. The reason for this is likely due in part to each network being able to label lensed features in a different way. For example, structure 2 is better at identifying the type of lens systems generated by a uniform distribution, that is, those with smaller Einstein radii or axis ratios (see below).

In Table 4.3 we present the evaluation criteria when the trained models are tested on a dataset with a uniform distribution for the lens model parameters, and a threshold of 0.99 is applied. Compared to the previous results, we see that the TP rate of network structure 1 has decreased by 29.9 per cent, and for network structure 2 has decreased by 12.7 per cent. These results highlight that network structures 1 and 3 require the training data set to be representative of the analysis data. For network structure 2, the TP rate has increased by 12 per cent, when compared to the previous case, and now has a value of 95.2 per cent.

#### 4.3.4 Non-lensed double-lobed radio sources

As discussed above, one of the main goals of a lens detection algorithm is to limit the number of FP events. The results for when both a uniform and a realistic distribution for the lens model parameters (Einstein radius and axis ratio) is used yields a low FP rate of 0.04 per cent for network structure 1 (equivalent to 1 FP event in our test dataset of 2500 non-lensed sources). This is due to the lensed emission having a very distinctive arc-like surface brightness distribution or there being four compact lensed images detected with the expected relative positions and peak surface brightness. The lensing nature of such events tends to be rather unambiguous.

However, those cases that produce only two lensed images have less information for the networks to label them correctly as lensed events. Also, two distinct components within a few arcsec is a common morphology for double-lobed radio sources, which,



**Figure 4.11** – *The FP rate as a function of threshold for a test dataset that only includes non-lensed double-lobed radio sources.*

depending on the jet-axis with respect to the observer, can have a relative peak brightness that could mimic a lensing event. In fact, during the training of the networks, the maximum separation between the multiple components in the source-plane was just 0.2 arcsec so that the background sources always had an extent that was less than the Einstein radius of the lens (we also used the simulated radio sources for generating the non-lensed visibility datasets). Therefore, double-lobed radio sources with larger separations were not presented to the networks during training. This may result in a bias towards higher FP rates for this class of radio source when the networks are applied to real ILT imaging data.

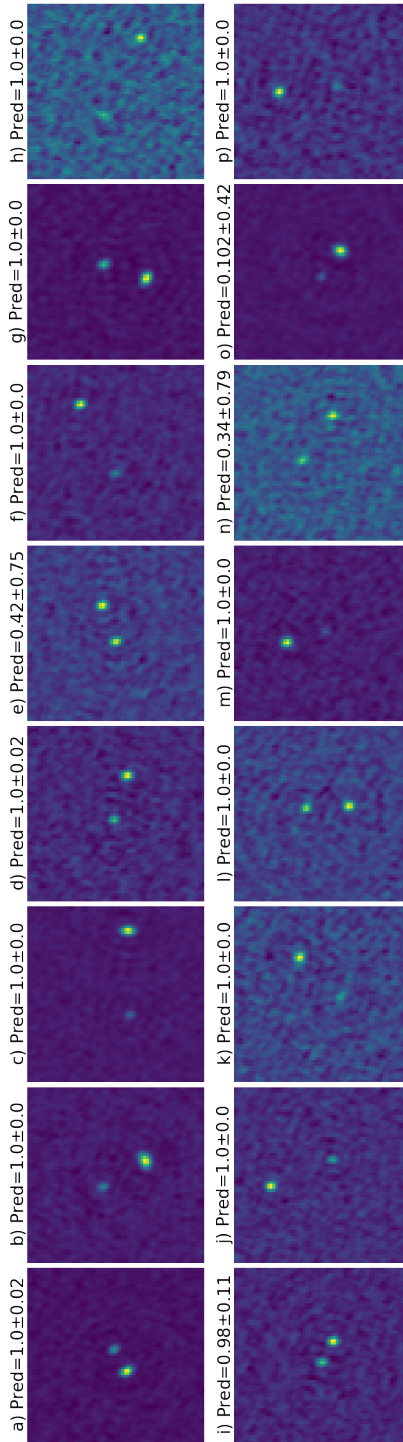
To test this, we have generated a dataset of double-lobed radio sources in which there are two co-linear components separated within a radius of 3 arcsec from the image centre. While in the original training and testing data, the first component was always compact (to represent the radio core), here, we have randomized the separation, size and ellipticity of each of the two components in this dataset. In total, we generated 2000 non-lensed events; a representative sample of these is shown in Fig. 4.10. We then apply the three network structures to determine whether these double-lobed radio sources are labelled as lensed or non-lensed events.

In Fig. 4.11, we show the FP rate of the three network structures, as a function of probability threshold. As expected, the performance of the networks to double-lobed radio sources is poor. We find that structures 1 and 3 both have a FP rate of 39.5 per cent, when the threshold is 0.99, while structure 2 has a FP rate of 6.5 per cent

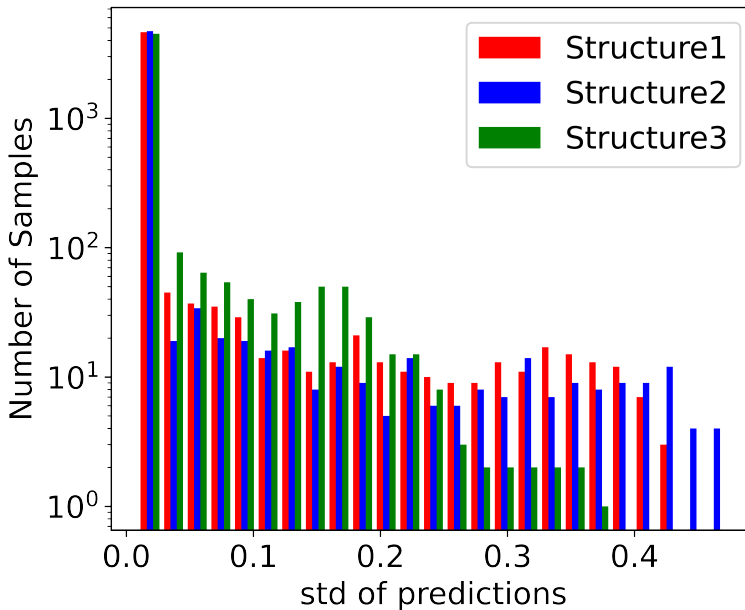
at the same threshold. This demonstrates that structure 2 identifies less double-lobed radio sources as lensing events, even when it is not specifically trained to recognise such samples. Also given in Fig. 4.10 is the predicted lensing probability and  $2\sigma$  confidence interval when structure 2 is used. We see that for this representative sample, the probability is either 0, or consistent with 0 at the  $2\sigma$ -level for 87.5 per cent of the sample (14/16 objects). For the remaining two objects (g, j), the probability is 1.0. However, for these two cases, it is clear that the surface brightness of the double-lobed radio sources is consistent with gravitational lensing; the component with the highest flux density also has the largest angular size. Therefore, it is not surprising that the network mis-labels these two objects as likely lensed events.

In order to measure the performance of the three network structures when only lensed events with two images are included in the test dataset, we have generated a sample of 2000 gravitational lens systems (with realistic lens parameters) producing two distinct components that are separated by  $> 0.3$  arcsec. We find that all three structures have a slightly poorer TP rate than before, with values of 83, 80 and 79 per cent for structures 1, 2 and 3, respectively. In Fig. 4.12, we present a representative sample of lensed events that produce two images, with the predicted lensing probability and  $2\sigma$  confidence interval, when structure 2 is used. We see that 75 per cent of the objects have a lensing probability of 1 (12/16 objects); these are all characterised with a clear surface brightness that is consistent with gravitational lensing. For 19 per cent of the objects, the probability is consistent with  $> 0.99$  at the  $2\sigma$ -level (3/16 objects), but the uncertainties are quite large, with  $\sigma$  between 0.06 and 0.40. Only one object (o) has a detection probability  $< 0.99$  at the  $2\sigma$ -level. For those cases with large uncertainties or low probabilities, the lensed images tend to have a very similar flux density or the counter-image is compact; this suggests that for these systems, the surface brightness information is insufficient to precisely label these samples as lensed events.

Our tests, using a dataset comprising double-lobed radio sources and gravitational lens systems producing two lensed images, demonstrate that structure 2 has the lowest FP rate (6.5 per cent for a threshold of 0.99) whilst having a competitive TP rate (80 per cent) to the other network structures. However, structure 2 still rates a significant number of the double-lobed radio sources as lensed events. Therefore, we re-ran our training including 2000 double-lobed radio sources, before re-testing the three networks on a dataset of 2000 previously un-seen double-lobed radio sources. We find that structures 1 and 2 returned 0 and 1 FP events (a FP rate of  $< 0.05$  and 0.05 per cent, respectively) and that structure 3 identified 2 FP events (a FP rate of 0.1 per cent). This highlights the importance of a comprehensive training sample in order to generate a reliable lens detection algorithm.



**Figure 4.12** – A representative sample of simulated gravitational lens systems with two lensed images. The average predicted probability and its uncertainty over 500 iterations for structure 2 is stated above each image. Each image contains  $64 \times 64$  pixels and is equivalent to a sky-area of  $7.68 \times 7.68 \text{ arcsec}^2$ .



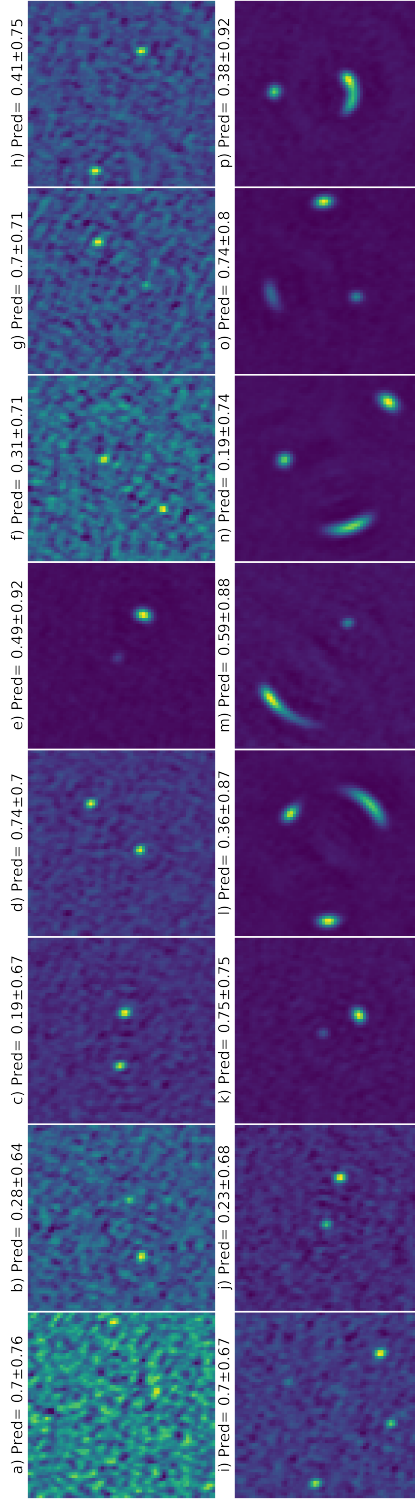
**Figure 4.13** – The standard deviation of the lensing probability for each test sample in the realistic dataset. Each sample in the test dataset has been evaluated 250 times using the dropout method.

### 4.3.5 Prediction uncertainties

As discussed in Section 4.2.5, we have used a Monte Carlo dropout technique to measure the model confidence and determine if the predicted probability of lensing is reliable or not. Human experts can then visually inspect those low reliability samples and decide if they warrant further analysis or follow-up observations. In Fig. 4.13, we show the standard deviation of the predicted probability for each sample in the test dataset made using realistic lens model parameters, as a function of network structure. Note that the test dataset for a uniform set of lens model parameters has a similar distribution, but there are more samples with a higher standard deviation due to the poorer performance of structures 1 and 3 (see Fig. 4.9).

Overall, we see that for the vast majority of cases ( $> 84$  per cent), the standard deviation in the probability is 0. This means that the same probability is returned for each iteration during the dropout process, and therefore, the network has a confident estimate of the predicted probability. We have considered a standard deviation  $> 0.3$  as non-confident cases, and have taken a closer look at those samples to understand why the network is uncertain.

Among all the lensed samples in the test data set for a realistic lens population, there



**Figure 4.14** – A representative sample of simulated gravitational lens systems with a high uncertainty in the probability of lensing. The average predicted probability and its uncertainty over 250 iterations for structure 1 is stated above each image. The upper row (lower row) is for a lens population with a set of realistic (uniform) mass model parameters. Each image contains  $64 \times 64$  pixels and is equivalent to a sky-area of  $7.68 \times 7.68 \text{ arcsec}^2$ .

are 44 samples with a high uncertainty in which 41 cases are lenses that produce two images (see upper row in Fig. 4.14 for a representative sample). The reason for these samples having such a high uncertainty is likely due to the network being confused between genuine two image systems and (non-lensed) double-lobed radio sources where there is not enough information available for the network to confidently determine the predicted probability. Therefore, excluding samples with a high uncertainty will lower the TP rate, but this will also lower the FP rate for those cases of double-lobe radio sources where the network is uncertain.

In the case of the test dataset that is drawn from a population of lenses with a uniform set of mass model parameters (see lower row in Fig. 4.14 for a representative sample), the uncertainties are at a slightly higher level. However, we also see sets of samples with large uncertainties (l, n, o) that have rather exotic lensing configurations, where an extended source is partially quadruply imaged, or the lens ellipticity is extremely high, resulting in an unusual image configuration. This is likely due to the limited size of the training dataset (see below for discussion). For all of the cases shown in Fig. 4.14, the predicted probability of lensing is less than 0.99, and so these samples would strictly not have been selected as lens candidates even though their uncertainties are high enough for the probability to be  $> 0.99$  (at the  $2\sigma$  level).

## 4.4 Lens detection with the ILT

In this section, we use the results obtained above to design and carry out a final experiment that quantifies the reliability of our network structure for finding gravitational lenses with the ILT. We then determine the parameter space that a gravitational lens survey with the ILT would be sensitive to, in terms of the depth and angular resolution of the expected imaging data.

### 4.4.1 Results for the final network test

We found from our network tests that the training and test datasets should be drawn from the same underlying lensing population, and that we must have a representative sample of non-lensed events, including both compact and double-lobed radio sources. Therefore, for our final test, we use a training dataset that has a total of 30000 samples, where 15000 are lensed events, created using a uniform distribution for the lens model parameters, and 15000 are non-lensed events, which are further divided into 11000 compact radio sources and 4000 double-lobed radio sources. This training model is then made in the same way as described in Section 4.2. Again, a maximum of 450 training iterations were used.

For testing, we have used a sample comprising 17385 lensed events that were created using a uniform distribution of lens model parameters; this was done so that we can test our trained model against a wide variety of lensing configurations, including those not necessarily seen by the network. This will likely produce more conservative, but less biased results. The sample of non-lensed events includes 15385 compact radio sources and 2000 double-lobed radio sources, which ensures that the ratio of compact-to-extended radio sources is similar to that observed with the ILT (Sweijen et al. 2022). The total number of test samples was chosen so that around 15 to 25 FP events would be returned by each of the three network structures.

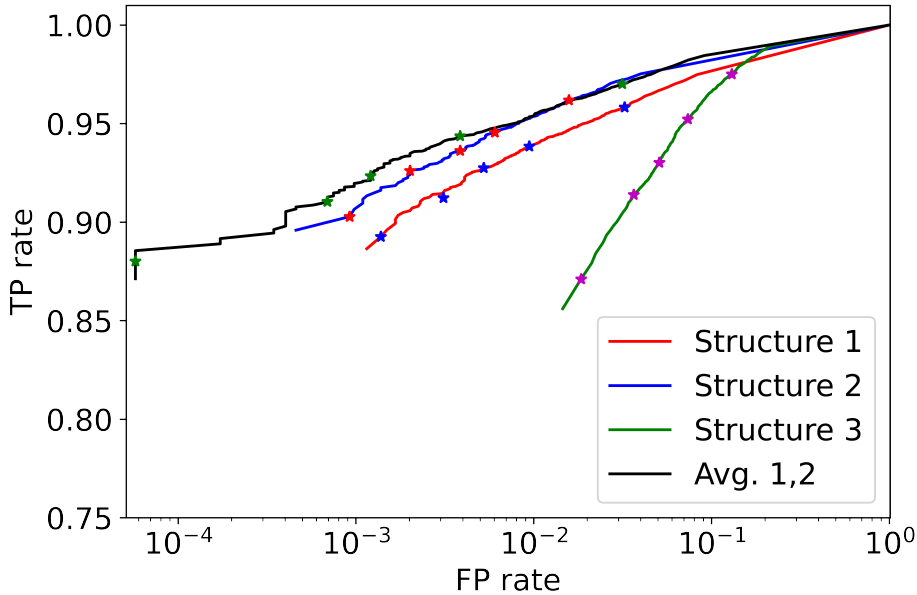
The results from this final test are shown in the ROC plots presented in Fig. 4.15, with the quantifiable information for a probability threshold of  $> 0.99$  given in Table 4.4. We see from Fig. 4.15 that the performance of the three network structures in identifying lens systems is still very good, even when the network training includes double-lobed radio sources, which can introduce an additional level of confusion between lensed and non-lensed events. We see that network structures 1, 2 and 3 have TP rates of 89.3, 90.3 and 87.1 per cent, respectively. This essentially means that one in ten lenses would be missed if the network was used for lens detection as part of an ILT survey (see below for a discussion of the completeness, given the resolution and sensitivity of the ILT with respect to the flux density and angular-separation of the lensed images). Also, we see that the FP rates are 0.14, 0.09 and 1.8 per cent for structures 1, 2 and 3, respectively, which are relatively low. However, the typical probability of galaxy-scale gravitational lensing is of order  $10^{-3}$ , that is, one lensing event in about one thousand objects observed, which is almost identical to the FP rates of structures 1 and 2. This means that there would be between a 47 and 58 per cent chance that any lens identified by the network would be a FP event when applied to real data.

To overcome this potential issue, we have combined the results of network structures 1 and 2, which is also shown in Fig. 4.15 and reported in Table 4.4. We find that this combined network structure lowers the FP rate to 0.006 per cent (note that this is equivalent to 1 FP event in our test sample), which changes the probability to 5.7 per cent that a lens identified by the network is a FP event. This improvement in the FP rate by around an order of magnitude comes at the minimal cost of lowering the TP rate by between 1.2 to 2.2 per cent.

#### 4.4.2 Parameter-space of a lens survey with the ILT

We now use the results from our final experiment to investigate the parameter-space, in terms of types of lenses and the brightness of the lensed images, that the network is sensitive to, given an input dataset from the ILT. For example, we expect the signal-





**Figure 4.15** – The TP rate as a function of FP rate for the three network structures, when applied to the final dataset created using a uniform distribution of lens model parameters. Also shown are the results when network structures 1 and 2 are combined. The stars correspond to thresholds of 0.1, 0.5, 0.75, 0.9 and 0.99, from right to left.

	Structure 1	Structure 2	Structure 3	Avg. Str. 1+2
Accuracy	0.946	0.951	0.926	0.940
Precision	0.9985	0.9990	0.9792	0.9999
Recall	0.893	0.903	0.871	0.880
Fall out	0.00138	0.00092	0.01846	0.00006

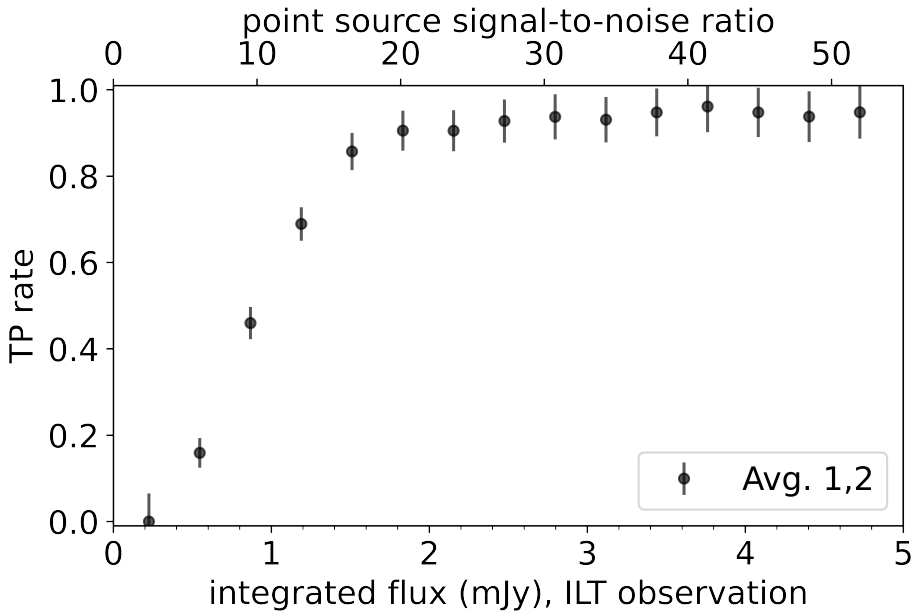
**Table 4.4** – The evaluation criteria of the three network structures when a threshold of 0.99 is used, for the final dataset created using a uniform distribution of lens model parameters. We also include the results when network structures 1 and 2 are combined.

to-noise ratio of the images to affect the reliability, as the network needs to detect multiple images to confidently predict that the object is indeed gravitationally lensed. Also, the angular resolution of the data will be an important parameter, as the lensed images need to be separable in the imaging data; not resolving all of the multiple lensed images can also lead to an inaccurate prediction of the lensing probability, particularly for exotic or high-magnification lensing events.

Understanding the detection rate as a function of signal-to-noise ratio is challenging to quantify, as the resulting lensed images from the same background source model can have a large variation in surface brightness distributions, depending on the lens model. For simplicity, we have analysed the detectability of a lensed event by considering the TP rate as a function of the total flux density in the model lensed images, which we show in Fig. 4.16 (note that to separate the effect of the angular resolution, we only include the data for those systems with an Einstein radius  $\geq 0.5$  arcsec; see below). This is a reasonable assumption, as the source counts of radio sources are always given as a function of flux density, as opposed to surface brightness.

We see that for a typical ILT observation, the TP rate is of order 90 per cent for an integrated flux density of  $\geq 2$  mJy, below which, the detectability of the lensed events steadily drops to just below 50 per cent at an integrated flux density of 1 mJy. We find that those lensed events with the lowest detectable surface brightness are usually in the form of a doubly- or a quadruply-imaged system with compact lensed images, which provide the least amount of information for the network to use. When the lensed events form Einstein rings and/or gravitational arcs, the integrated flux density needs to be higher for a detection to be made, and so, it is easier for the network to identify these cases at high integrated flux densities. Combined, these two effects result in the detectability of the lensed events being lower toward lower flux densities, but stops there being a very sharp cut-off. Also, we find that the cut-off is remarkably close to systems that are detected at the  $15$  to  $20\sigma$ -level (point-source sensitivity), which in the case of doubly-imaged systems is equivalent to a flux-ratio of between 2:1 and 3:1 for the two detected lensed images (assuming detectability of emission at the  $5\sigma$ -level), which is fairly typical for this image configuration.

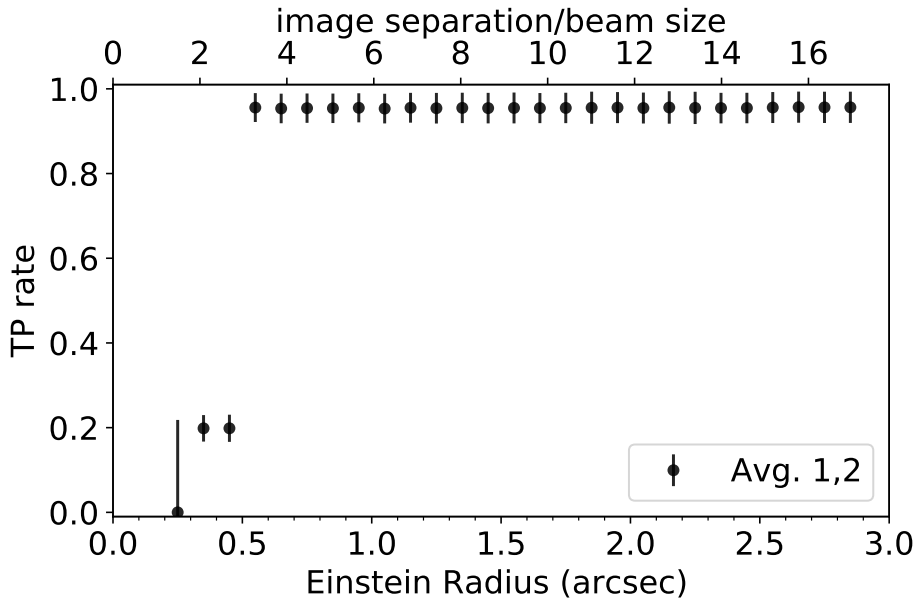
The TP rate as a function of the Einstein radius is shown in Fig. 4.17 (note that to separate the effect of the sensitivity, we only include the data for those systems with an integrated flux density  $\geq 2$  mJy; see above). We see that the overall TP rate is rather flat at around 95 per cent, down to an Einstein radius of about 0.5 arcsec, after which there is a sudden drop in the TP rate to about 20 per cent at Einstein radii between 0.3 and 0.5 arcsec. To some extent, this result makes sense, as an Einstein radius of 0.5 arcsec corresponds to an image separation of around 1 arcsec, which would be



**Figure 4.16** – The TP rate as a function of the integrated flux density of the lensed radio sources from the combined results from applying network structures 1 and 2, for an ILT-like observation (rms noise  $90 \mu\text{Jy beam}^{-1}$ ). Also shown is the point source signal-to-noise ratio for reference. Note that this is for objects with an Einstein radius  $\geq 0.5$  arcsec.

well resolved with the ILT (about 3 beam widths). Overall, this is an encouraging result, as it demonstrates that ILT-like imaging data is extremely sensitive to detecting gravitational lenses with Einstein radii  $\geq 0.5$  arcsec, which would be sufficient to detect the majority of the gravitational lenses that are currently known (see Fig. 4.2).

In summary, from our simulations we find that the ILT, when included as part of LoTSS, would be most sensitive to lensed events where the integrated flux density of the lensed images is  $\geq 2$  mJy and the Einstein radius of the lens is  $\geq 0.5$  arcsec. If we restrict our testing dataset to include only those lensed and non-lensed samples with these properties, we find that the overall TP rate is 95.3 per cent and the FP rate is 0.008 per cent, when network structures 1 and 2 are combined (the individual FP rates of structures 1 and 2 are also about an order of magnitude lower than reported in Table 4.4). Therefore, a gravitational lens survey with the ILT, which applies these criteria and uses our algorithm for lens detection, is expected to have a completeness of 95.3 per cent and a purity of 92.2 per cent. These predictions are dependent on the lensing probability of the LoTSS source population (currently assumed to be  $10^{-3}$ ), which we will discuss in a future paper.



**Figure 4.17** – The TP rate as a function of Einstein radius from the combined results from applying network structures 1 and 2, for an ILT-like observation (average beam size of 336 mas). Also shown is the ratio between the image separation and the average beam size for reference. Note that this is for objects with an integrated flux density  $\geq 2$  mJy.

## 4.5 Conclusions

In this chapter, we have presented a machine learning based approach that is designed to identify galaxy-scale gravitational lenses from imaging data obtained with the ILT, taken as part of the ongoing LoTSS survey. To do this, we first simulated realistic gravitational lensing data, based on our best understanding of the radio source population at 150 MHz and the properties of galaxy-scale gravitational lenses. With these data, we have tested multiple network structures to determine whether a single or combined network produces the best results. We find that our ability to correctly identify lensed features is highly dependent on the training dataset used, which given the large parameter-space available for both the lens and source models (and their combination), represents the most challenging aspect of our method. However, by using a combined network strategy and limiting the parameter-space of the models to include only those gravitational lenses with an Einstein radius  $\geq 0.5$  arcsec and a total flux density of  $\geq 2$  mJy for the lensed images, we find that our lens detection strategy can recover 95.3 per cent of the simulated lens systems, with a FP rate of just 0.008

per cent (equivalent to a sample purity of 92.2 per cent for a lensing optical depth of  $10^{-3}$ ).

Given the angular resolution and sensitivity of the imaging data to be taken during LoTSS, we conclude that the ILT has the requirements to be a gravitational lens finding machine, and that deep learning techniques provide an efficient route to discovering new gravitational lenses with interferometric imaging data. We find that data, with a similar resolution and noise properties to the ILT, can be used to find gravitational lenses with lensed image separations that are  $\geq 3$  times the synthesized beam width, when the lensed images are detected at the  $\geq 20\sigma$ -level (point-source sensitivity). This suggests that the SKA-MID, with baselines of up to 150 km and an angular resolution of 0.3 arcsec should also be excellent at finding new gravitational lens systems in the future. As a next step, we will carry out a similar analysis with dedicated SKA-MID gravitational lensing simulations to confirm this, and to optimise lens searches with this next generation instrument.

Our current set of simulations, although realistic, is likely the limiting factor in our analysis, and will be further improved in the future. First, we will increase the range of lens models probed to include elliptical power-law mass distributions, as opposed to focusing on only isothermal cases. We have also included a somewhat simple parameterisation of the background radio source structure, which was computationally easy to implement, but was also driven by our lack of knowledge of the low frequency radio source population. As LoTSS progresses, and the level of information about the structure of radio sources on 0.3 arcsec-scales improves, we will revisit our simulations to refine our network structure with a better model for the radio source population. This will also involve training on a larger number of lens and source model combinations to better sample the available parameter-space. Our simulations have also assumed perfectly calibrated data, and although small-scale calibration errors will exist in the real imaging data, their impact on our ability to find gravitational lenses is not clear, and will be addressed in a follow-up paper. Finally, we intend to include the frequency information (radio spectral index), given the wide bandwidth of the ILT data, to further separate lensed and non-lensed events.

Overall, our results are encouraging for finding new galaxy-scale gravitational lenses with the ILT, and in the short term we will apply our best trained model to the first tranche of data from LoTSS that includes the ILT baselines (e.g. Morabito et al. 2021; Sweijen et al. 2022) and has detected over 2500 radio sources at an angular resolution of around 350 mas. Given the expected probability for galaxy-scale gravitational lensing, there should be between 2 and 4 systems within such a dataset, which we are now focusing on finding. As more LoTSS observing fields are processed with the

ILT data included, we expect the first new discoveries of gravitational lensing with LOFAR to be made. These new systems will be used to test and further refine our network structure so that we are in the best possible position to analyse the data for the up to 15 million radio sources that will be observed during LoTSS.

## **Acknowledgements**

This chapter is based on research performed within in the DSSC Doctoral Training Programme, co-funded through a Marie Skłodowska-Curie COFUND (DSSC 754315). JPM acknowledges support from the Netherlands Organization for Scientific Research (NWO) (Project No. 629.001.023) and the Chinese Academy of Sciences (CAS) (Project No. 114A11KYSB20170054). LOFAR (van Haarlem et al. 2013) is the Low Frequency Array designed and constructed by ASTRON. It has observing, data processing, and data storage facilities in several countries, which are owned by various parties (each with their own funding sources), and that are collectively operated by the ILT foundation under a joint scientific policy. The ILT resources have benefited from the following recent major funding sources: CNRS-INSU, Observatoire de Paris and Université d'Orléans, France; BMBF, MIWF-NRW, MPG, Germany; Science Foundation Ireland (SFI), Department of Business, Enterprise and Innovation (DBEI), Ireland; NWO, The Netherlands; The Science and Technology Facilities Council, UK; Ministry of Science and Higher Education, Poland; The Istituto Nazionale di Astrofisica (INAF), Italy. The Jülich LOFAR Long Term Archive and the German LOFAR network are both coordinated and operated by the Jülich Supercomputing Centre (JSC), and computing resources on the supercomputer JUWELS at JSC were provided by the Gauss Centre for Supercomputing e.V. (grant CHTB00) through the John von Neumann Institute for Computing (NIC).

## **Data Availability**

Upon reasonable request, the underlying data used for this chapter will be shared by the author.



## Conclusions and future prospects

In this thesis, a set of deep learning solutions to address several outstanding questions in the field of high angular-resolution radio astronomy have been developed. The purpose of this chapter is to provide an overview of the main scientific conclusions, with a discussion of the future prospects and directions. In each section, a short summary of the results from each chapter is given, before the general conclusions and future work is discussed.

### 5.1 Source counts of VLBI-detected radio sources

The sensitivity of VLBI arrays are expected to improve dramatically with the construction of the Square Kilometre Array (SKA-VLBI) and the next generation Very Large Array (ngVLA). These instruments will lower the current sensitivity limits, and given their wide fields-of-view, will provide a significant increase in the numbers of radio sources detected on arcsec- to mas-scales. However, understanding their potential and planning for the opportunities that they bring requires some knowledge of the number of detectable radio sources on VLBI-scales as a function of flux density. This goal was the focus of Chapter 2, which aimed to provide an insight to the optimum strategies for all-sky VLBI surveys, based on the density of radio sources in the sky and their detectability with VLBI.

#### 5.1.1 Chapter summary

Despite the wide range of science opportunities provided by high angular-resolution imaging with VLBI, its application at cm-wavelengths is restricted to studying only



~ 25 000 radio sources with very high brightness temperatures ( $> 10^5$  K) and over a very small fraction of the observable sky. This is relatively small compared to the 5 million radio sources recently catalogued from all-sky surveys at arcsec-resolution. The small effective field-of-view of a VLBI dataset is one of the parameters that limits the detection of many radio sources per observation. Moreover, telescopes are scattered over a wide area, forming a synthesised unfilled aperture with a lack of information on various angular scales. With such configurations of antennas, only sufficiently compact and bright radio sources with measurable correlated flux on the available baselines of the array are detected.

We analysed the detection fraction and the number counts of radio sources on VLBI-scales using the publicly available catalogue from the mJIVE–20 survey. This survey was conducted with the VLBA at a frequency of 1.4 GHz, targeting 24 903 pre-identified radio sources in FIRST. A total of 4 965 radio sources were detected by the mJIVE–20 survey, resulting in an overall detection fraction of  $19.9 \pm 2.9$  per cent. Our analysis showed that the FIRST peak surface brightness is correlated with the VLBI detection fraction. Fifty per cent of the sources are detected with the VLBA, when they have a peak surface brightness of at least  $80 \text{ mJy beam}^{-1}$  in FIRST. At the detection threshold, this fraction drops to around 8 per cent. We deduced that this significant drop in the number of detected radio sources was likely due to a change in the radio source population, with an additional contribution from the limited surface brightness sensitivity of the VLBA observations. We also considered the effect of compactness at arcsec-scales on the VLBI detection fraction. Our results showed that sources with a higher compactness in FIRST were more likely to have a detection with VLBI.

Using the VLBI-detected radio sources found by the mJIVE–20 survey, we provided an estimate for the number of expected radio sources that will be detected by the next generation of wide-field VLBI surveys. Future surveys with the VLBA and the EVN in the short term ( $< 10$  yr) and with the SKA-VLBI and ngVLA in the longer term can benefit from these calculations. A measurement of the mJIVE–20 survey sky area was required to calculate the source counts. This was done by designing a Monte Carlo integration stone throwing technique, which took into account the sensitivity and pointing configuration used during the mJIVE–20 survey observations. We calculated the total sky area of the mJIVE–20 survey to be  $237.95 \text{ deg}^2$ . This calculation allowed us to measure the effective area as a function of the target minimum peak surface brightness. We also determined the completeness of the catalogue by simulating 13 500 realistic VLBI visibility data sets that were then passed through the mJIVE–20 imaging pipeline. Our analysis found that the mJIVE–20 catalogue is complete at the

97 per cent level at the detection threshold ( $6.75\sigma$ ) and reaches full completeness at a signal-to-noise ratio of 7.8.

Having determined the effective sky area and the completeness, we then calculated the normalized Euclidean and differential source counts for VLBI-detected radio sources. We found the slope of the differential number counts of VLBI-detected radio sources with flux densities  $S_{1.4\text{ GHz}} > 1\text{ mJy}$  to be  $\eta_{\text{VLBI}} = -1.74 \pm 0.02$ , which is shallower than in the case of its parent population observed by FIRST ( $\eta_{\text{FIRST}} = -1.77 \pm 0.02$ ). The calculated differential source counts slope is also shallower when compared to a population of compact radio sources selected at higher frequencies ( $\eta_{\text{JBF}} = -2.06 \pm 0.02$ ).

From our analysis, we found that all-sky ( $3\pi\text{ sr}$ ) surveys with the EVN and the VLBA have the potential to detect  $(7.2 \pm 0.9) \times 10^5$  radio sources at mas-resolution, and that the density of compact radio sources is sufficient ( $5.3\text{ deg}^{-2}$ ) for in-beam phase referencing with multiple sources (3.9 per primary beam) in the case of a hypothetical SKA-VLBI array. For the VLBA, such a survey is expected to require about 7000 h of observations to complete. However, given the sensitivity and field-of-view of a hypothetical SKA-VLBI array, the same survey would be carried out in about 180 h.

### 5.1.2 General conclusions

The main goal of this chapter was to calculate the number of detectable radio sources with VLBI, and this was achieved through determining the source counts in a reliable way. From this analysis, it is clear that, for a flux-density limit of 1 mJy, there are a large number of radio sources that could be found, given the sensitivity of the VLBA. To some extent, such a survey could also be carried out with the EVN. However, given the slightly higher observing frequency (1.7 GHz) and larger telescopes (32 m), the EVN is not as an effective instrument for an all-sky survey. We also defined a simple model for an SKA-VLBI facility that has 20 stations located in Africa. Even a simple arrangement that includes only one SKA-MID dish at each location would revolutionize surveys with VLBI, giving both the sensitivity and field-of-view needed to effectively survey the sky. The latter point is also important since the wide field-of-view also makes in-beam calibration possible, which increases the efficiency of such surveys with this design of telescope. Therefore, as a general conclusion, it can be stated that SKA-VLBI has the potential to detect around  $7.2 \times 10^5$  radio sources, and the sky density of the compact radio source population is sufficient for calibrating the system.

### 5.1.3 Future prospects

Our ability to survey the sky is dependent on the effective field-of-view of the individual telescopes and the overall sensitivity of the interferometric array. Given the sensitivity of the VLBA and the short integration times used during the mJIVE–20 survey, very few faint radio sources were detected. This made the source counts below a flux density of about 1 mJy uncertain. However, this is an important part of the parameter space, as it is where we expect a change in the radio source population from AGN to star-forming galaxies to occur. This is also expected to produce a change in the brightness of the radio emission from such sources, and so the number of detectable radio sources on VLBI-scales would also likely drop. However, we found that the normalized Euclidean source counts flattened below 1 mJy, in a similar way to those determined at a lower angular resolution. Also, for the sensitivity calculations that were made for the SKA-VLBI, it was clear that investigating the radio source population below 1 mJy should be possible. The number of radio sources that can be found determines how sensitive the SKA-VLBI should be. Therefore, it would be useful to extend the source counts derived in this thesis to at least  $100 \mu\text{Jy}$ . Unlike above, an EVN that includes the large 70 to 100-m dishes in the array will have the sensitivity to carryout such a survey over a smaller part of the sky. From this, it will be possible to characterize the compact radio source population and better inform the design and specifications of an SKA-VLBI facility.

## 5.2 Source detection and characterization

Making new scientific discoveries will always depend on the completeness and robustness of source detection and characterization algorithms. In the age of large synoptic survey telescopes and interferometric arrays, which will operate across all observable wavelengths and generate larger and more complex datasets, efficient and automated source detection algorithms need to be developed. The research presented in Chapter 3 provides a robust solution to point and extended source detection and characterization for sparse interferometric arrays, which is named DECORAS.

### 5.2.1 Chapter summary

DECORAS was designed to provide a unified pipeline for source detection and characterization for both unresolved and extended sources using machine learning techniques. The pipeline has an autoencoder structure with only nine convolutional layers. The autoencoder was trained to remove the effect of the Point Spread Function (PSF), noise structures and other sources of contamination from the given dirty

image. The focus of DECORAS is on images that are produced from a Fourier transform of the visibility data and have not gone through a prior deconvolution process (commonly known as dirty images). The purpose of this experiment was to determine the reliability of deep learning approaches on the deconvolution process.

Considering the goal of a source detection algorithm is to detect as many real sources as possible, while limiting the number of fake detections, we incorporated two autoencoder networks within DECORAS. One was trained using Binary Cross Entropy (BCE) and the other using Mean Squared Logarithmic Error (MSLE). We used the predicted output of both networks to ensure the existence of a source in a given field, and restrict the number of false detections. The generated output model images were then passed through a post processing source detection function that localized the position of the source. It also made the decision on whether there was a source in the input image. Our results showed that the position of the detected sources could be recovered to within  $0.61 \pm 0.69$  mas of the actual position, for a VLBI data set with a beam size of around 17 mas.

The pipeline characterized the detected source by passing it to another autoencoder network. This second autoencoder was designed to recover the properties of the injected source, like the effective radius and peak surface brightness. The peak surface brightness of the source was measured using the extracted latent variables from the encoder part of the autoencoder. The latent variables are the compressed set of features that the encoder extracted and passed to the decoder part of the network to generate the corresponding model image. Considering the training objective, the latent variables contained the most essential features of the given input image. DECORAS performed source characterization in terms of the position, effective radius and peak brightness of the detected sources. We found that the effective radius and peak surface brightness were recovered to within 20 per cent for 98 and 94 per cent of the sources, respectively.

In total, DECORAS was validated on 15 800 test images in which 8 000 included an injected source over a wide range of signal-to-noise ratios, and 7 800 were noise realizations. The test data were designed in a representative way to mimic VLBA observations at a wavelength of 20 cm. Moreover, the results from DECORAS were compared with a traditional source detection algorithm. However, a de-convolution process had to be applied to that test data set before the performance of the traditional source detector could be determined. The source catalogue that was generated by DECORAS was found to be fully complete at a signal-to-noise ratio of 7.5, whereas the traditional source detection algorithm reached full completeness at a higher signal-to-noise ratio of 8.4. The improvement in the detectability provided by DECORAS

was found to reduce the needed integration time by about 25 per cent, when compared to a traditional source detection algorithm. For example, an all-sky survey with the VLBA requires around 7000 h to reach a similar depth to the mJIVE–20 survey. Applying DECORAS to such a survey would save 1750 h in observing time, while detecting the same number of sources.

Also, we found that the performance of DECORAS on the noise realization samples, where no source was injected, showed significant improvement on the catalogue purity, when compared to a traditional source detector. When combining the performance of DECORAS on both catalogue completeness and purity, we found an almost factor of two improvement for sources with a signal-to-noise ratio above 5.

### 5.2.2 General conclusions

The goal of Chapter 3 was to determine whether deep learning techniques can be used to detect and characterize radio sources from sparse interferometric arrays, using the specific example of the VLBA. Overall, given the improvements in the catalogue completeness and purity obtained with DECORAS, we can conclude that deep learning techniques provide a robust alternative to traditional source detection algorithms. In particular, through being able to lower the detection threshold, in terms of more real sources being detected without the expense of a large number of fake detections, deep learning techniques can make wide-field surveys for VLBI-detected radio sources more efficient.

### 5.2.3 Future prospects

In Chapter 3, we only applied DECORAS to simulated data and only considered the ideal case with no systematic calibration errors. In the following, some potential future directions and applications are considered.

- Given the simplicity of DECORAS, it can be easily tested on representative data from other interferometric arrays, like the ILT. However, the training model for the pipeline will likely need some tuning based on the differing resolution, noise correlations and density of radio sources at the different frequencies. For example, the ILT observes a larger area of the sky and the density of faint and bright radio sources is different from that of the VLBA. These effects need to be considered when developing a smart and fast radio source extractor for the ILT. Moreover, working with the dirty images (or partially cleaned images) from the ILT could eliminate/limit the need for a deconvolution task. For example, DECORAS performs very well at the detection threshold of traditional source

detection algorithms. Therefore, it could be used to search this parameter space on partially cleaned images, where the side-lobes from the brightest radio sources have been removed. This is particularly interesting, as recovering the true sky brightness distribution from interferometric data using traditional approaches is challenging and time consuming. However, given the large fields-of-view of the ILT, the sky density of radio sources may be too high for this to work efficiently. Further tests, on complex fields with many sources must first be done to see if DECORAS can provide a novel route to image detection and characterization for the ILT.

- Using machine learning techniques, it should also be possible to provide an estimation of any interferometric calibration errors and determine their effect on the imaging quality. Also, these techniques could be used to recognize calibration errors and reconstruct the visibilities with an estimation of what the correct calibration would look like. Currently, there are a limited number of studies in the literature that examine the use of deep learning for the calibration of radio telescopes. Considering the thousands of observations, each with their own unique settings, the fine tuning of imaging pipelines could be a tedious task. Therefore, a potential use of machine learning algorithms would be to automatize the tuning of the many hyper-parameters in traditional imaging pipelines. In principle, this could be done with DECORAS if some knowledge of the PSF is given to the network, as currently, this is learned by the network to perform de-convolution. Given that calibration errors introduce an apparent change in the PSF, the network could identify these changes, and interpret them in terms of simple antenna-based amplitude and phase corrections. These can then be used to refine the calibration or provide a more robust model for the underlying source surface brightness distribution.

### **5.3 Lens detection with the ILT**

Strong gravitational lensing describes an astrophysical phenomenon in which two galaxies are along the same line-of-sight to the observer; this results in the gravitational field of the closer galaxy acting like a natural telescope by magnifying the radiation from the more distant galaxy. There are many applications of gravitational lensing, including understanding the nature of dark matter and studying the high-redshift Universe. However, there are only about 40 cases of gravitational lensing found at radio wavelengths. Typically, these have been found through visual inspection, and by applying a set of selection criteria. With an ever-increasing number of radio sources

being observed, sophisticated search techniques can potentially be used to identify many thousands of new gravitational lens candidates. Even with several selection criteria, there will still be a very large number of candidates that would need to be visually inspected. Therefore, in Chapter 4, we have developed a deep-learning approach to identify gravitationally lensed radio sources in high angular-resolution imaging data, specifically for those taken with the ILT.

### 5.3.1 Chapter summary

We incorporated four main deep learning components into three network structures to search for gravitational lenses in simulated data from the ILT. The various network components were included so that features describing lensed and non-lensed events (i.e., the surface brightness distribution) could be learned and differentiated between. All three structures were trained with the same training and validation data set for each experiment. Also, the same testing data set was used to provide a fair comparison between the performance of all three structures. The output of each network structure was defined by a lensing probability of between 0 and 1 for each test sample. Events with an output closer to 1 were classified as having a higher probability of being a genuine gravitational lens.

An ideal lens finding algorithm would identify all of the lensed events, while not selecting any of the non-lensed events. However, this rarely happens in real-world applications due to the limitations of noise and the angular resolution of the data. Given the large number of gravitational lenses that are expected to be found with the ILT, completeness will likely not be important for most science cases. Instead, it will be more important to select genuine lens candidates, by having a larger true positive (TP) rate, while returning a lower false positive (FP) rate. Our analysis demonstrated that by selecting a probability threshold of 0.99 to classify each sample into lensed and non-lensed events, we could achieve a very low number of FP detections. Although this reduced the number of TP events, our results showed that it had a greater impact on the number of FP events.

We developed different testing strategies to evaluate the performance of the three network structures. The experiments were designed to cover a wide range of scenarios that might occur during a gravitational lens search using real data from the ILT. Each experiment used a separate test dataset to validate the performance of the proposed models. However, despite using the same training and test datasets when evaluating the results, the structures were found to behave differently, meaning that they extracted a different set of features in the data. Our results showed that when training with a realistic lens population, two of the network structures failed to detect a large

portion of the lensed events that were generated by a uniform distribution for the lens parameters, which produced more exotic lens configurations that the networks had not seen. However, one of the structures was found to be less dependent on the diversity of gravitational lens systems used in the training data.

Our results for a test dataset containing non-lensed double-lobed radio sources showed a very high FP rate of around 39.5 per cent for two of the tested network structures, while one of the network structures had a FP rate of only 6.5 per cent (for a threshold of 0.99). These results were achieved when the networks were not specifically trained on non-lensed examples that included double-lobed radio sources. This motivated us to include such non-lensed radio sources in the training data. The final result, when the trained model had previously seen both non-lensed double-lobed radio sources and two-imaged gravitational lens systems, had an FP rate of  $< 0.1$  per cent for all three network structures.

The results from these experiments demonstrated the importance of a comprehensive training sample that covers a wide range of lensed and non-lensed events, which given the large variation in lens models and un-lensed source surface brightness distributions, may be a limiting factor. Therefore, we designed a final experiment, which included 30 000 training samples that were equally distributed between the lensed and non-lensed events. Among the non-lensed samples, 4 000 double-lobed radio sources were randomly injected to broaden the variety of this class. The test dataset contained 34 770 samples, in which 17 385 were lensed and 17 385 were non-lensed events. Here, 2 000 double-lobed radio sources were included in the test dataset of non-lensed samples. We found TP rates of  $> 87.1$  per cent for all three network structures. In spite of a similar performance for detecting genuine gravitational lenses, two of the structures performed significantly better at separating lensed and non-lensed events. The FP rates were 0.14 and 0.09 per cent for the more reliable structures, and 1.85 per cent for the least reliable one.

Moreover, we found that the two structures with the best performance did not agree on the specific FP events. This motivated us to consider the output of those structures simultaneously, by considering the average of the predicted lensing probability. This limited the number of FP events to only a single double-lobed radio source (1 in 17 385 samples; equivalent to a FP rate of 0.006 per cent) in the test data set. This had the minimum cost of lowering the TP rate by between 1.2 and 2.2 per cent, compared to when the two structures were considered independently.

We then analyzed what effect the sensitivity and resolution of a typical ILT observation would have on the TP rate. For this, we measured the TP rate as a function of the



integrated flux density of the lensed images, and as a function of the Einstein radius of the lensed event. We found that the TP rate reached around 90 per cent when the lensed events were restricted to those with an integrated flux density of  $\geq 2$  mJy. For lensed events with a minimum Einstein radius of 0.5 arcsec, our approach was found to have a TP rate of 95.3 per cent. Comparing these with the sensitivity ( $\sim 90 \mu\text{Jy beam}^{-1}$ ) and the angular resolution ( $\sim 330$  mas) of real imaging data taken with the ILT, we determined that this newly commissioned interferometer has the potential to be a gravitational lens finding machine.

### 5.3.2 General conclusions

The main goal of Chapter 4 was to test whether deep learning algorithms could be extended to interferometric imaging data for finding gravitational lenses. One of the main advantages of the interferometric arrays used in this thesis, namely the VLBA and the ILT, is their extremely high angular resolution. This means that they should in principle be very good for identifying gravitational lenses. However, as the noise is highly correlated in such data, it wasn't clear whether the deep learning techniques that have been successfully applied to optical data could also be used at radio wavelengths. From the simulations carried out here for the ILT, we found that it was possible to identify gravitationally lensed radio sources in such data, with a TP rate of 95.3 per cent, but more importantly, the number of FP events could be limited to just 0.006 per cent of the samples. Therefore, we conclude that deep learning algorithms do provide a novel route to lens detection with interferometric arrays. In addition, we conclude that lensed events with image separations  $\geq 3.2$  times the synthesized beam width, and with an equivalent point-source signal-to-noise ratio of  $\geq 20$ , can be identified with our lens detection algorithm. Therefore, we can state that the ILT and the SKA-MID will be extremely useful in finding gravitationally lensed radio sources, given their current specifications, when combined with deep learning techniques.

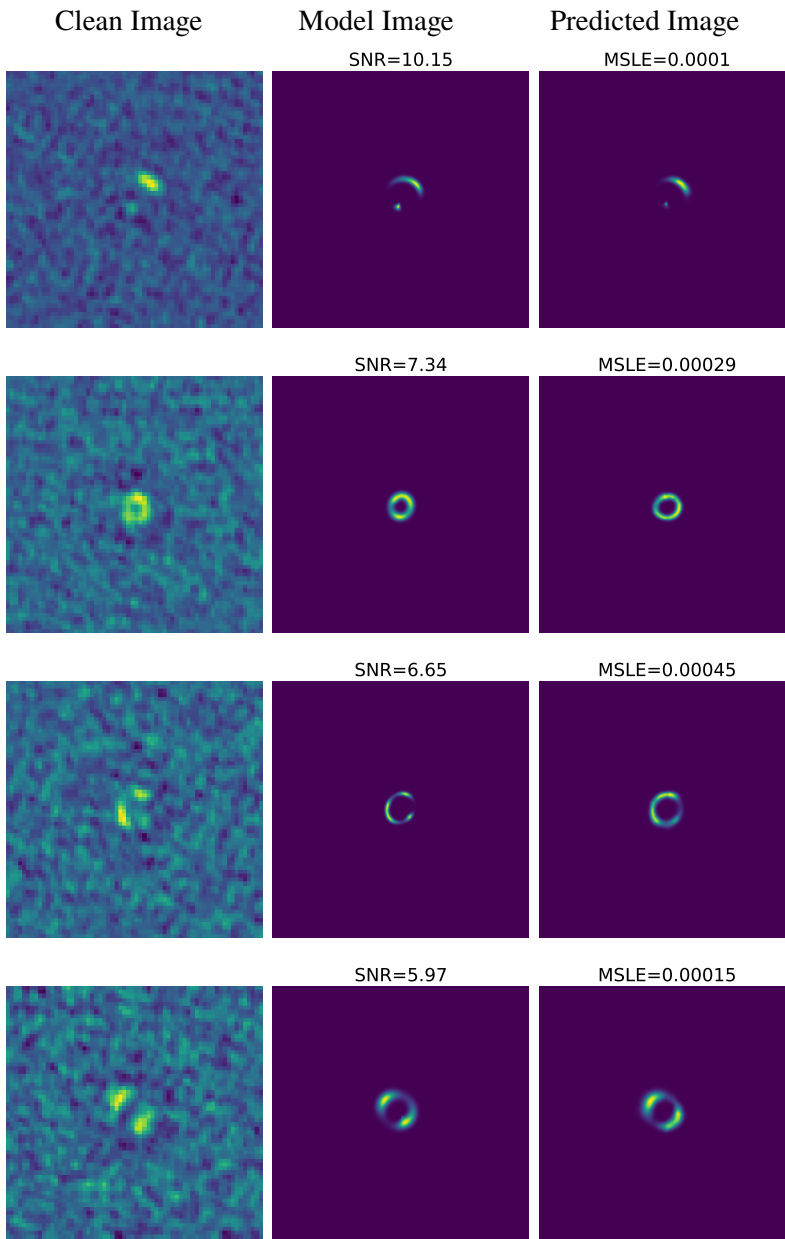
### 5.3.3 Future prospects

There are a number of steps that can be taken to improve on the lens detection algorithm presented in Chapter 4. Below is a list of ideas to be implemented in the future.

- The way that the lens detection network has been designed will allow data from multiple input images to be analyzed in the future. One way of using this feature would be to include the information from the wide bandwidth in frequency that the ILT provides (around 120 to 170 MHz). As gravitational lensing

should, in principle, not change the intrinsic spectrum of the background object, the spectral information can also be used as a constraint when discriminating between lensed and non-lensed events. For instance, a spectral index map can be extracted from an ILT image to identify if the candidate lensed images have the same spectral properties. Also, we have not considered another available channel, that is, data taken at optical/infrared wavelengths where the light from the lensing galaxy would likely be detected, but possibly not from the lensed images. This would be a further interesting test to carry out, given the similar angular-scales that are probed by the ILT and SKA-MID at radio wavelengths and, for example, *Euclid* at optical wavelengths.

- Estimating the properties of gravitational lens systems, with tens of thousands expected to be discovered with the SKA-MID in the future, will need fast and reliable solutions to be developed that automate the modelling process. Traditional approaches, such as maximum likelihood modelling, have several data preparation steps and require the user to provide additional information, such as the relative positions, fluxes and time-delays of the lensed images, and any observed properties of the lens. As of now, this effort was feasible considering the few tens of known gravitational lens samples thus far found at radio wavelengths. Therefore, including a deep learning approach within our lens detection algorithm that also models the gravitational lensing data would be extremely useful, and is an area of ongoing research that we summarize in the final section of this conclusions chapter.
- Reconstructing the un-distorted surface brightness distribution of the background radio source is another potential application of deep learning algorithms, when applied to gravitational lensing data from interferometric arrays. This could be useful for understanding the properties of a large sample of objects in the high redshift Universe that are gravitationally lensed. However, reconstructing the background source will need sophisticated techniques that combines the information from the lens model, via the lens equation, and learns the properties of the background source population. It is particularly challenging as any small error in the lens model can affect the accuracy of the background source reconstruction. The expected parameters that we would aim to recover are the intrinsic size, surface brightness and the broad band spectral energy distribution of the unlensed source.



**Figure 5.1** – Representative examples of the performance of de-convolution using DECORAS to recover the expected model image from any given cleaned image, generated from simulated ILT-like imaging data. The left panel shows the input cleaned image, the middle and right panels show the expected and the predicted model images, respectively. The signal-to-noise ratio and the MSLE are provided for each sample.

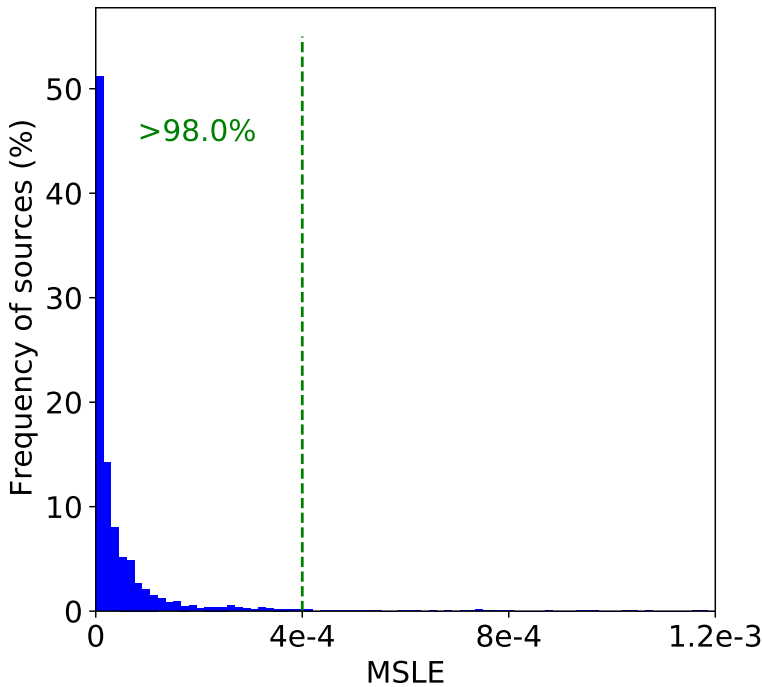
## 5.4 Preliminary results on lens modelling with deep learning

In this section, we present preliminary results from applying a deep learning technique to simulated gravitational lensing data from the ILT, with the goal of testing methods for lens modelling. This work is a synthesis of the methods obtained in Chapters 3 and 4 of this thesis, where the source characterization algorithm, DECORAS, and a lens detection algorithm were developed, respectively.

First, we have carried out a test to determine the performance of DECORAS on an ILT-like dataset. Our results are based on training the first autoencoder in DECORAS with 10 000 simulated images, which were also pre-processed (cleaned). We have used 600 training iterations. Although there are small changes to the structure of the autoencoder, it has the same core as in DECORAS. The changes are made considering the size of the input images, which instead of being  $256 \times 256$  pixels are now  $64 \times 64$  pixels, given the larger beam size and smaller field-of-view (in pixels) needed to image ILT data. This has changed the number of convolutional layers in the autoencoder to meet the specific requirements of the simulated training data set used for lens identification and modelling. The input images used to provide the following preliminary results are simulated in a similar way to the data presented in Chapter 4. However, only lensed samples are considered here, as we assume that the input data have been pre-selected as gravitational lens candidates (either using our own or another lens detection algorithm).

Fig. 5.1 shows a few representative examples of the performance of DECORAS in recovering the underlying model surface brightness distribution. Overall, this first test on cleaned images from an ILT-like data set is very encouraging, as we see a good agreement between the input and recovered model images. The apparent signal-to-noise ratio given for each sample is calculated by dividing the model peak surface brightness by the rms in the cleaned image ( $\sigma = 90 \mu\text{Jy beam}^{-1}$ ; similar to the process used in Chapter 3). We have also measured the Mean Squared Logarithmic Error (MSLE) for each of the predicted models to measure their performance (more information about the MSLE is given in Chapter 3). The distribution of the calculated MSLE for the entire test data set is shown in Fig. 5.2. It shows that the calculated MSLE for more than 98 per cent of samples in the test dataset is below  $4 \times 10^{-4}$ , which is also a promising result.

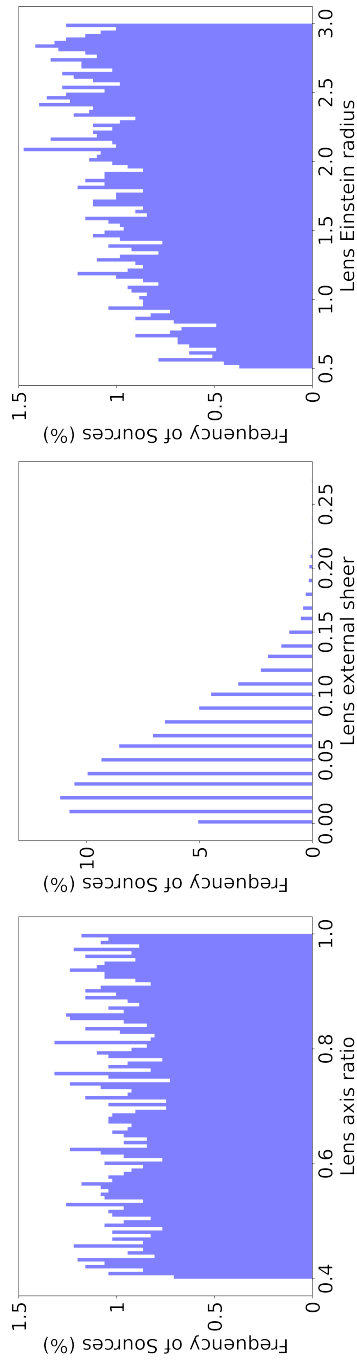
In the next step, we have used the simulated input lensed surface brightness images to predict the lens model parameters. A training data set of 7 000 lensed events has been generated using a uniform distribution of lens model parameters, which are shown in Fig. 5.3; the axis ratio and Einstein radius have (close to) uniform distributions, while



**Figure 5.2** – The distribution of the achieved MSLE by DECORAS when predicting the surface brightness distribution of the test data set. The green dashed line shows that more than 98 per cent of the test samples are recovered with a MSLE less than  $4 \times 10^{-4}$ .

the shear has a realistic distribution. Similar to the lens detection pipeline presented in Chapter 4, the input images have been passed through a pre-processing step that converts the pixel values to between (0, 1) using a MinMax normalization.

The lens modelling algorithm is shown in Fig. 5.4. It starts by employing 4 convolutional layers with filter sizes of 128, 64, 32 and 16. The output of the last convolutional layer is passed to a multi-scale filter bank (Lee & Kwon 2017). There are three parallel paths in the multi-scale filter bank, each with one single convolutional layer. The filter size in each path is  $(1 \times 1)$ ,  $(3 \times 3)$  and  $(5 \times 5)$ , respectively. The feature maps collected from each path are then concatenated and passed through an inception block, which has a similar concept and structure to the multi-scale filter bank. There are also three parallel paths in the inception block. There is a convolutional layer, with a filter of size  $(1 \times 1)$  in the first path. The second path has two convolutional layers with filter sizes of  $(1 \times 1)$  and  $(3 \times 3)$ . In the third path, a  $(5 \times 5)$  filter is applied to the output of the  $(1 \times 1)$  and  $(3 \times 3)$  convolution layers. The next component in the network structure has three convolutionalized blocks. Each convolutionalized block



**Figure 5.3** – The distribution of lens model parameters used in the training dataset.

has three convolutional filters, each with a filter size of  $(1 \times 1)$ . Residual learning (He et al. 2015) is implemented in the convolutionalized blocks to improve the training efficiency. More details are provided in Chapter 4 about the various components in this network. Before the final outputs are predicted by the network, a dropout layer and four dense layers with ReLU activation functions are used.

The network is designed in such a way that the three outputs are estimated at the end of each training iteration. The outputs correspond to the three lens model parameters that we are primarily interested in recovering at this stage, namely, the Einstein radius ( $\theta_E$ ), the axis ratio ( $b/a$ ) and the lens position angle (PA). Each output is predicted using a dense layer with a linear activation function. The performance of each predicted output is measured individually using the Mean Squared Error (MSE), which has been chosen due to the regression estimation between the true and predicted lensing parameters. It is given as

$$\text{MSE} = \sum_{i=1}^N (y_i - y'_i)^2, \quad (5.1)$$

where  $y$  is the given lens model parameter and  $y'$  is the predicted lens model parameter, in a dataset with  $N$  training samples in each batch.

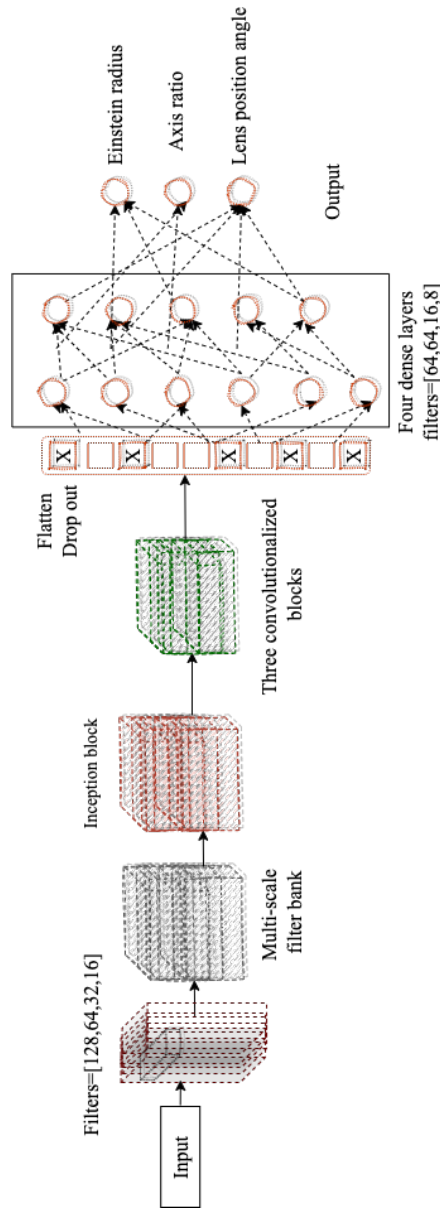
The following provides the preliminary results of applying this network structure, after training using 100 iterations, to an ILT-like test data set containing 2 000 previously unseen lensed events. The lens models for the test data are drawn from the same population that was used for the training data set, shown in Fig. 5.3. In the following, we have compared the recovered parameters to the input models, by considering the difference error or relative error, such that

$$\text{Relative Error}_{\theta_E} = \frac{\text{True}_{\theta_E} - \text{Predicted}_{\theta_E}}{\text{True}_{\theta_E}}, \quad (5.2)$$

$$\text{Relative Error}_{b/a} = \frac{\text{True}_{b/a} - \text{Predicted}_{b/a}}{\text{True}_{b/a}}, \quad (5.3)$$

$$\text{Difference Error}_{\text{PA}} = \text{True}_{\text{PA}} - \text{Predicted}_{\text{PA}}. \quad (5.4)$$

The results for the Einstein radius, axis-ratio and the position angle of the Singular Isothermal Ellipsoid (SIE) plus shear mass model considered here, are given in Figs. 5.5, 5.6 and 5.7, respectively. In each case, we show the recovered and model parameters, and the relative difference between them. We see that the Einstein radius, which is correlated with the lensed image separation, is rather well recovered when compared to the input models. We see that 95 per cent of the samples have an Einstein



**Figure 5.4** – This network structure consists of a multi-scale filter bank and inception blocks. Three convolutionalized blocks with dropout and four dense layers are added, respectively. The network generates three outputs, with a separate path used for the dense layer and the linear activation function. The performance of each output path is calculated separately.

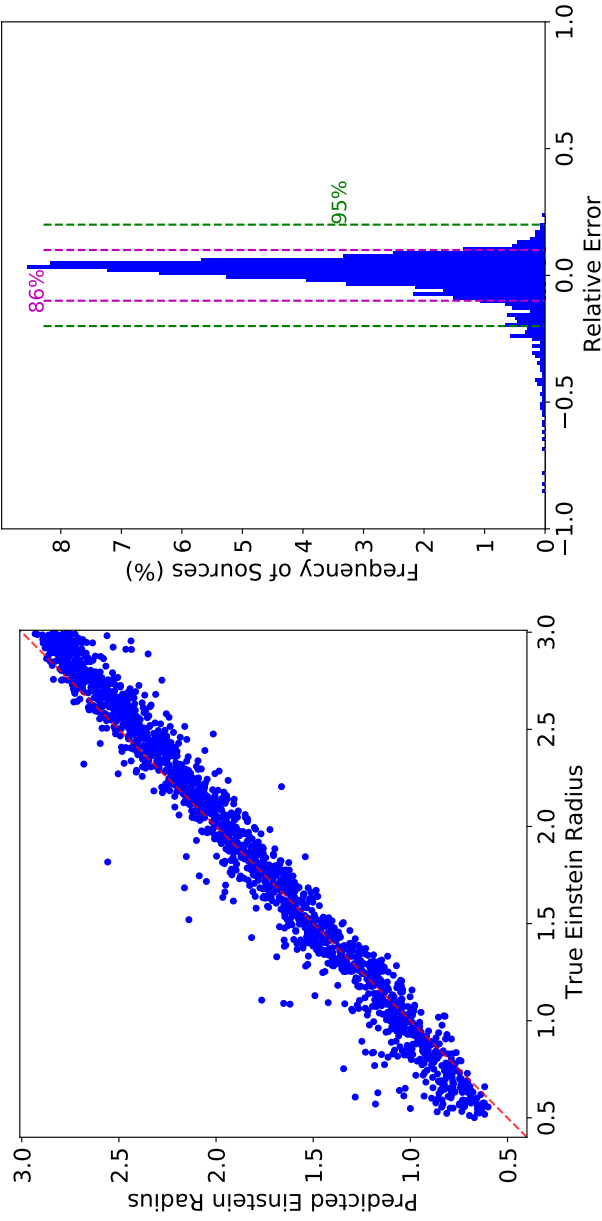


radius that can be recovered to within 20 per cent. However, the quality of the results for the predicted axis-ratio and the position angle are somewhat worse. We see from Fig. 5.6 that the axis ratio in only 78 per cent of the lensed samples are recovered to within 20 per cent of the input model. From Fig. 5.7 we see that the error on the position angle is significantly higher compared to the other parameters, given the large number of outliers. However, we find that the input and the recovered axis-ratio and position angle are highly correlated in both cases, suggesting that overall the network does well enough to give an indication of the shape and orientation of the lensing mass distribution. Although this is not sufficient for precision lens modelling, our initial results show that deep learning can be used to model lens systems from interferometric data, and be used to form sub-samples based on criteria related to the size and shape of the lens system. For example, this could lead to forming samples of near-perfect Einstein rings or highly distorted systems with complicated environments. The choice of what sub-samples to make would be up to the user, depending on their science goal.

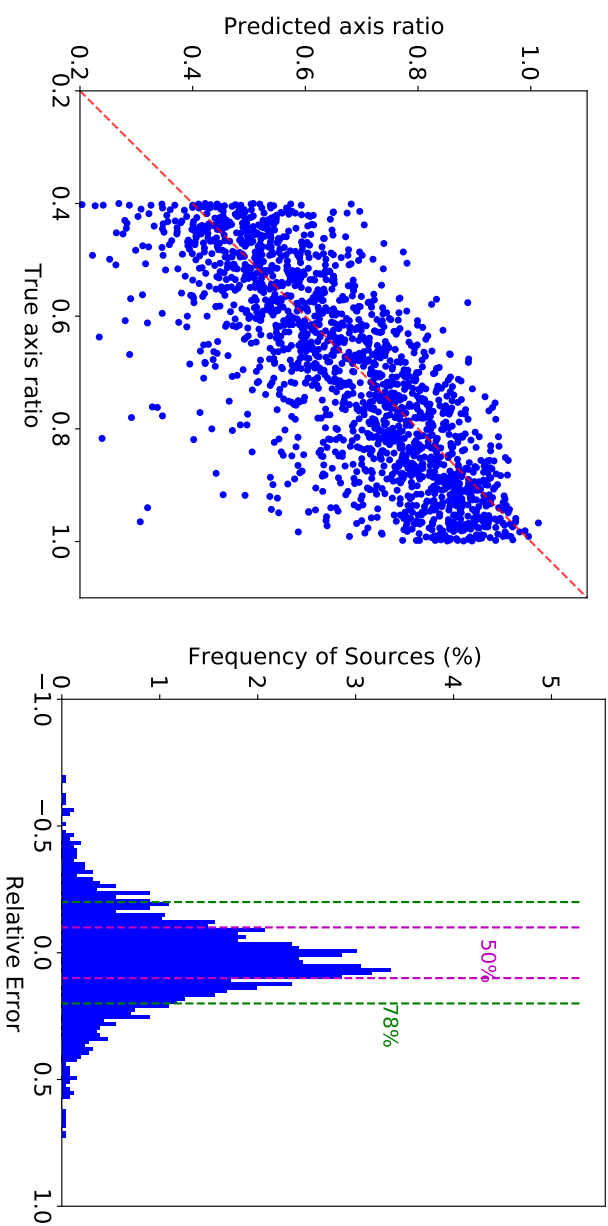
Although these results are encouraging, in the future, we plan to continue this work by using larger training and test data sets. Also, different training strategies, using state-of-the-art components in deep learning architectures will be used. We also aim to design a pipeline that can predict the other lens model parameters, such as the external shear and its position angle.

## 5.5 Final remarks

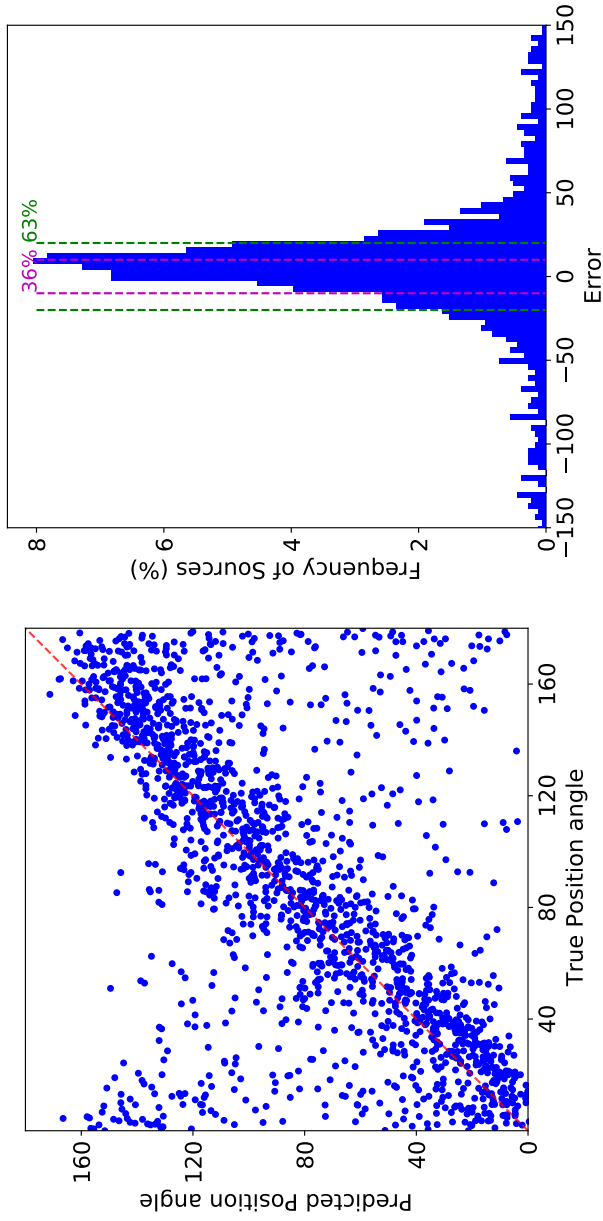
This thesis has addressed several challenges of the big data era in the field of high angular-resolution radio astronomy using machine learning algorithms. The methodologies presented in this thesis were designed with the aim of minimizing the need for human interactions, while still providing robust results. Our experience shows that beside using the right frameworks and efficient algorithms, the way that the input and output data are defined is one of the most important aspects of developing such methodologies. In other words, defining the problem in a proper way is as important as designing the various frameworks and writing the accompanying algorithms to solve a given problem. Feeding the learning-based algorithms with "good data" is part of properly defining the problem. Here, we mean the data that is representative of astronomical images that are expected to be generated through scientific experiments, and covers a wide range of possibilities that might happen in real world observations of the radio sky. We would also like to highlight the importance of collaboration in multi-disciplinary research, as writing practical algorithms and covering the sci-



**Figure 5.5** – In the left panel, a comparison of the predicted and true Einstein radius of the simulated gravitational lens systems is presented. The red dashed line shows when they exactly agree. In the right panel, a histogram of the relative error with respect to the input (true) Einstein radius is presented. The magenta and green dashed lines show the percentage of sources within the 10 and 20 per cent fractional error bounds.



**Figure 5.6** – The left panel represents a comparison of the predicted and true axis-ratio of the simulated gravitational lens systems. In the right panel, a histogram of the fractional error is presented. The magenta and green dashed lines show the percentage of sources within the 10 and 20 per cent fractional error bounds.



**Figure 5.7** – The left panel represents a comparison of the predicted and true position angle of the simulated gravitational lens systems. In the right panel, a histogram of the difference error is presented. The magenta and green dashed lines show the percentage of sources within the 10 and 20 per cent error bounds.

entific aspects in the desired application, in this case high angular-resolution radio astronomy, is only possible when groups of scientists work together, and are open to learning from each other.

Overall, we find that deep learning techniques can be very useful in the area of high angular-resolution astronomy, and we look forward to seeing the algorithms developed in this thesis used with real data from the VLBA and the ILT.

## Summary

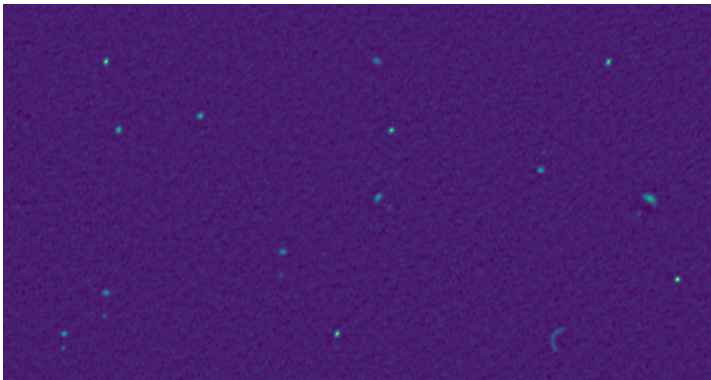
- for non experts -

### Introduction

This thesis aims to address some of the challenges facing astrophysics in the big data era. Although big data is not clearly defined, it can be considered as the amount of data that conventional methods have trouble handling. Numerous current/future ground and space telescopes make astrophysics a data-rich field that constantly presents new scientific questions with more complex data challenges. In order to make the processing of these big data possible, new methods using artificial intelligence and image processing are needed. This thesis explicitly examines the accuracy and reliability of these techniques in answering timely questions in high angular-resolution radio astronomy. It has an interdisciplinary approach and uses knowledge in computer science to advance our understanding of the radio sky. Considering Fig. I as a simulated sample of the radio sky, we can define the objectives of this thesis as the following:

- 1: to study the properties of the detected radio sources with a specific instrument;
- 2: to develop a source detection and characterization algorithm that can localize the source in any observed image from a specific radio telescope;
- 3: to design an algorithm that can find rare types of galaxies among the many observed radio emitting objects;
- 4: to characterize the properties of rare galaxies.

In the following, we first examine big data processing techniques from the perspective of computer science. In computer science, artificial intelligence can be acquired by



**Figure I** – A simulated image of the radio sky containing several radio meeting objects.

building algorithms that mimic certain functions of the human brain, such as image recognition and specification. Among the various types of artificial intelligence that can be applied to real-world problems, this thesis focuses on using data science techniques, such as machine learning (ML) and, in particular, deep learning (DL) algorithms. ML algorithms can be used for classification, regression, or similar tasks based on extracted knowledge from the data. In addition, they provide insights from the data and the problem at hand. The capability of ML for fast and reliable decision-making or forecasting is an exciting aspect of its use in many applications. DL algorithms are a subset of ML algorithms that can automatically extract properties from raw data. This quality of DL can bring us closer to working with larger and more complex datasets.

We now introduce the astrophysical challenges associated with complex astronomical observations of the radio sky. But first, we introduce radio astronomy and the definition of high angular resolution.

Visible light isn't the only form of radiation emanating from celestial bodies. There can be also radio waves. Radio receivers collect and amplify radio waves; just like ordinary telescopes that collect and magnify the image created by light waves. Examining radio waves can help us to study objects so far away that only a small amount of their energy reaches us. It also makes it possible to detect cold galaxies that mainly radiate in the radio frequency range.

The angular resolution of a telescope is defined as the smallest angle between close objects that can be seen separately. For telescopes with a lens or mirror with a diameter  $D$  observing at wavelength  $\lambda$ , the angular resolution (in radians) can be described

as approximately  $\lambda/D$ . While optical telescopes can have a high angular resolution, radio telescopes have a much lower resolution due to their longer wavelengths. In other words, at a wavelength of 1 metre (a frequency of 300 MHz), we need a single telescope with a diameter of 200 km to detect objects separated by an arcsecond. It would be costly and difficult to build such an instrument if it were not technologically impossible.

The solution to this problem is a process called interferometry, in which several radio telescopes (receivers) are connected and simultaneously observe the same astronomical object. An interferometer can achieve the resolution of  $\lambda/B$ , where  $B$  is the longest distance between the receivers. The radio waves travel a different distance to reach each one of the receivers in an interferometer. Therefore, a correction is needed for the waves to be combined in phase as constructive interference to maximize the signal, while the out-of-phase waves are subjected to destructive interference. Waves that are not entirely in or out of phase represent a pattern of moderate-intensity that may be used to detect relative phase differences.

This thesis uses the data from two interferometers with the highest angular resolution at their respective observing frequencies. One consists of 10 telescopes with a diameter of 25 metres, located in the United States and in the Virgin Islands and Hawaii. This interferometer is called the Very Long Baseline Array (VLBA) and can observe the radio sky in a frequency range of 0.3 to 83 GHz. The longest distance between two telescopes of the VLBA is 8611 km, which offers a high angular resolution. For the second and third chapters of this thesis, VLBA observations at a frequency of 1.4 GHz have been used. In the fourth chapter, we used another interferometer, which is based mainly in the Netherlands, but also has stations across Europe. The Low-Frequency Array or LOFAR, comprises many simple dipole antennas, unlike conventional diaphragm-filled telescopes. This simple design allows astronomers to observe a large sky area simultaneously. LOFAR is currently the largest radio telescope in the world and can operate at the lowest frequencies visible from Earth.

## **An overview of the chapters**

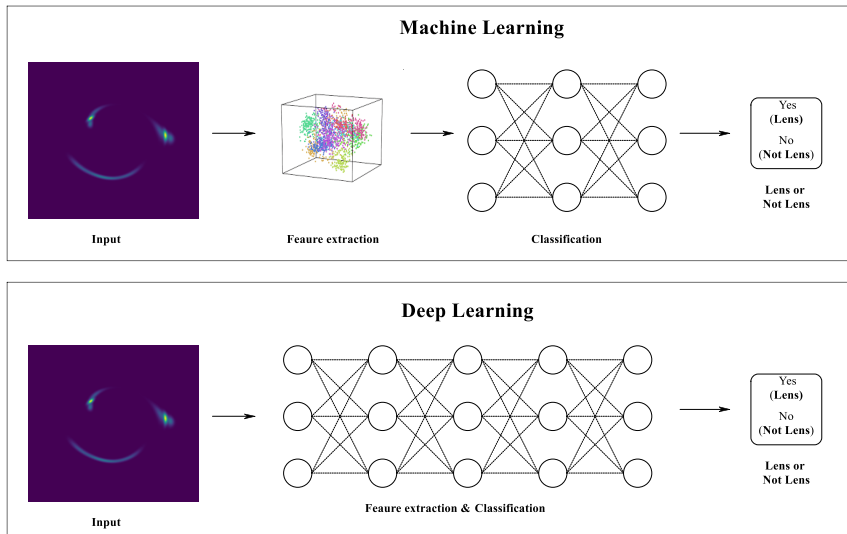
The first chapter contains the scientific introductory material to this thesis. It comprises a more detailed description of the data, the science questions, the techniques and approaches used to answer them. In the second chapter, we analyzed the radio sources observed with the VLBA. In addition, we compared the VLBA detection ratio with another interferometer that has a better sensitivity. The catalog of radio sources detected at 1.4 GHz with the VLBA is available to the public. Our experi-



ment shows that out of about 25 000 pre-identified radio sources, around 5 000 radio sources were identified with the VLBA, representing an overall detection ratio of around 20 per cent. Our analysis showed that the characteristics of the radio sources, such as their brightness and compactness, affect the chance of a detection with the VLBA. To calculate the number of existing sources in the field, we have measured the observed sky area considering the sensitivity and configuration of the observations. We also analyzed the completeness of the catalog by simulating 13 500 radio sources. The results showed that the generated catalog of the observations is complete at the signal-to-noise ratio of 7.8. These calculations were used to estimate the number of expected radio sources to be detected by the next generation of high sensitivity and high angular-resolution interferometers.

Considering the importance of identifying sources for making new scientific discoveries, we have dedicated the third chapter to a new source detection and characterization pipeline that used ML algorithms. The pipeline works with simulated images from the radio sky that contain observational noise. This chapter evaluates the use of such algorithms compared to traditional source detection techniques, in terms of the catalog completeness and the catalog purity. The catalog completeness is maximized when the detection algorithm can detect all the real sources in the test data. Catalog purity measures the number of fake sources that the algorithm mistakenly detects. Ideally, the detection algorithms should detect all of the real sources, while not detecting any of the fake sources. The source catalog made using our new pipeline was found to be fully complete at a signal-to-noise ratio of 7.5, whereas the traditional source detection algorithm reached full completeness at a higher signal-to-noise ratio of 8.4. More importantly, our catalog is significantly more reliable (almost a factor of two) in terms of not detecting fake sources for a signal-to-noise ratio above 5. In addition to source detection, the implemented pipeline presented in this chapter can remove the observational noise, restore the structure of the celestial sources and predict their properties, such as size and brightness.

In the fourth chapter, artificial intelligence techniques are used to identify a rare phenomenon in astronomy. This phenomenon, called a strong gravitational lens, occurs when two galaxies are located along an observational line of sight. The gravitational field of the nearby galaxy bends the emission from the distant galaxy at an angle that results in distorting the observed emission in the form of arcs and multiple images. While this phenomenon occurs with a probability of only 1 in a thousand, it has exciting applications for astronomers. Therefore, the discovery and study of them are of particular importance to astronomers. Fig. II shows a general overview of the lens identification pipeline developed in the fourth chapter of this



**Figure II** – An overview of machine learning and deep learning algorithms used for detecting strong gravitational lens systems. The input image is an example of a gravitational lens system that creates the illusion of a smiling face by distorting the light of a distant galaxy.

thesis. The input image of the pipeline is an example of a gravitational lens system. It shows a rare event in which the light of the distant galaxy is distorted in a way that creates an illusion of a smiling face. This figure also represents an abstract overview of the differences between machine learning and deep learning. The top panel shows a typical machine learning pipeline that consists of a separate feature extraction step. Feature extraction is the process of selecting or generating the features that are relevant to the specific goal of the pipeline. For the lens detection pipeline, as is shown for the input sample, we are looking for the existence of arcs and multiple components in the image. There are more specific requirements to classify an input sample as a strong gravitational lensing system, but these do not fit into the scope of this thesis. The observed features are used in classification algorithms to decide which class (lens or not a lens) the input sample belongs to. This paradigm has been changed using DL algorithms. As it is shown in the bottom panel, the feature extraction and classification are embedded into one step. Deep learning algorithms can learn the important features that differentiate the two classes of data.

The international stations of LOFAR can detect millions of radio sources with the

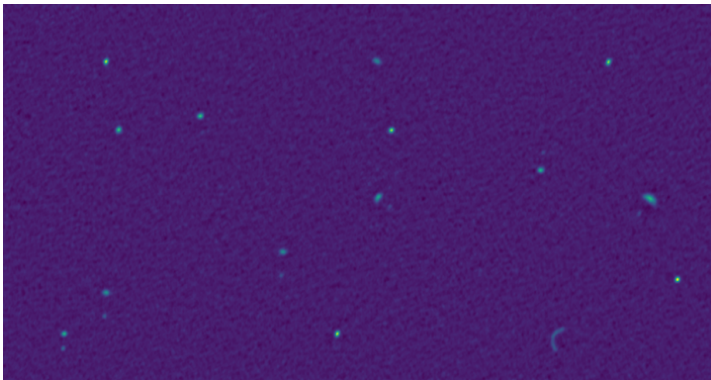
angular resolution needed to detect this phenomenon. In Chapter 4, we have tested the performance of DL algorithms on detecting strong gravitational lensing using realistic simulated data for the international LOFAR interferometer. Our results are also used to determine the sensitivity of this interferometer to gravitational lensing and determine the type of lenses that can be found according to the noise and angular resolution of the data.

## Samenvatting

### INLEIDING

Het doel van dit proefschrift is om enkele van de uitdagingen aan te pakken waarmee de astrofysica in het big data-tijdperk wordt geconfronteerd. Ondanks dat Big data geen duidelijke definitie heeft, kan het worden beschouwd als de hoeveelheid data die door conventionele methoden moeilijk verwerkt kan worden. Talloze huidige en toekomstige telescopen, op zowel de Aarde als in de ruimte, zorgen ervoor dat astrofysica een datarijk onderzoeksveld is, omdat er voortdurend nieuwe wetenschappelijke vragen gesteld worden die data van een hoge complexiteit en hoeveelheid vereisen. Conventionele dataverwerkingsmethodes kunnen deze big data niet verwerken en dus zijn nieuwe verwerkingsmethodes, die gebaseerd zijn op kunstmatige intelligentie en beeldverwerking, hard nodig. Dit proefschrift richt zich specifiek op de nauwkeurigheid en betrouwbaarheid van deze nieuwe technieken bij het beantwoorden van wetenschappelijke vragen in radioastronomie met hoge hoekresolutie. Het gaat hier om een interdisciplinaire benadering, waarbij er gebruik gemaakt wordt van kennis in de informatica om het begrip van de radiohemel te vergroten. Figuur .I beschouwend als een gesimuleerde radiohemel, definiëren wij de doelstellingen van dit proefschrift als volgt:

- 1- Om de eigenschappen van de gedetecteerde radiobronnen te bestuderen met een specifiek instrument.
- 2- Het ontwikkelen van een brondetectie- en karakteriseringsalgoritme dat het astronomische object (of de bron), waargenomen door een specifieke radiotelescoop, in



**Figuur I** – Een gesimuleerd waarneming van de radiohemel met daarin verschillende objecten.

elke omgeving kan lokaliseren.

3- Het ontwerpen van een algoritme dat in staat is om zeldzame klassen van sterrenstelsels te vinden tussen grote hoeveelheden van waargenomen bronnen.

4- De eigenschappen van de gedetecteerde zeldzame sterrenstelsels karakteriseren.

Wij bekijken eerst big data-verwerkingstechnieken vanuit het perspectief van de informatica. In de informatica spreekt men van kunstmatige intelligentie wanneer algoritmes bepaalde functies van het menselijk brein kunnen nabootsen, zoals beeldherkenning en specificatie. Van de verschillende soorten kunstmatige intelligentie die toegepast kunnen worden op problemen in de echte wereld, richt dit proefschrift zich op het gebruik van datawetenschapstechnieken, zoals machine learning (ML) en deep learning (DL) algoritmen in het bijzonder. ML-algoritmen maken gebruik van kennis onttrokken uit de onderliggende data voor classificatie, regressie of vergelijkbare taken, daarnaast geven ze inzichten vanuit de data en het probleem. ML is in staat om tot snelle en betrouwbare besluitvorming of prognoses te komen, wat maakt dat het veelbelovend is voor meerdere toepassingen. DL-algoritmen zijn een subset van ML-algoritmen die automatisch eigenschappen uit onbewerkte data kunnen halen. Dit aspect maakt dat DL veel wordt toegepast bij het verwerken van grotere en complexere datasets.

Na deze inleiding tot kunstmatige intelligentie en dataverwerkingstechnieken, introduceren wij nu de astrofysische uitdagingen die gepaard gaan met complexe astronomische waarnemingen van de radiohemel. Eerst introduceren wij echter radio-astronomie en de definitie van hoge hoekresolutie. Zichtbaar licht is niet de enige

vorm van straling die afkomstig is van hemellichamen, radiogolven zijn dit ook. Radioantennes verzamelen en versterken radiogolven; net zoals optische telescopen die optisch licht verzamelen en het waargenomen beeld vergroten. Het onderzoeken van radiogolven helpt ons om objecten te bestuderen die zo ver weg zijn dat slechts een kleine hoeveelheid van hun energie ons bereikt. Radiogolven maken het ook mogelijk om koude sterrenstelsels te detecteren, omdat deze objecten voornamelijk in het radiofrequentiebereik uitstralen.

De hoekresolutie van een telescoop is de kleinste hoek tussen twee nabij gelegen objecten die nog afzonderlijk waargenomen kunnen worden. Voor telescopen met een lens of spiegel is de hoekresolutie (in radialen) beschreven door  $\lambda/D$ , waarbij  $\lambda$  de golflengte is en  $D$  de diameter van de lens of spiegel. Radiotelescopen hebben een lagere resolutie dan optische telescopen vanwege hun langere golflengten. Met andere woorden, bij een golflengte van 1 meter (een frequentie van 300 MHz) hebben wij een telescoop nodig met een diameter van 200 km om objecten te kunnen detecteren die één boogseconde van elkaar afstaan. Het zou duur en moeilijk zijn om een dergelijk instrument te bouwen als het niet technologisch onmogelijk zou zijn.

In de radioastronomie wordt dit probleem opgelost door middel van interferometrie, waarbij meerdere aaneengesloten radiotelescopen (ontvangers) tegelijkertijd hetzelfde astronomische object waarnemen. Een interferometer kan een resolutie van  $\lambda/B$  bereiken, waarbij  $B$  de langste afstand tussen de ontvangers is. De radiogolven bewandelen voor elke ontvangers in de interferometer een ander pad. De golven worden in-fase gecombineerd om via constructieve interferentie het signaal te maximaliseren, terwijl de golven die uit-fase zijn middels destructieve interferentie verwijderd worden. Golven die niet volledig in of uit fase zijn, vertegenwoordigen een patroon van gemiddelde intensiteit, welke gebruikt kan om de relatieve faseverschillen te meten.

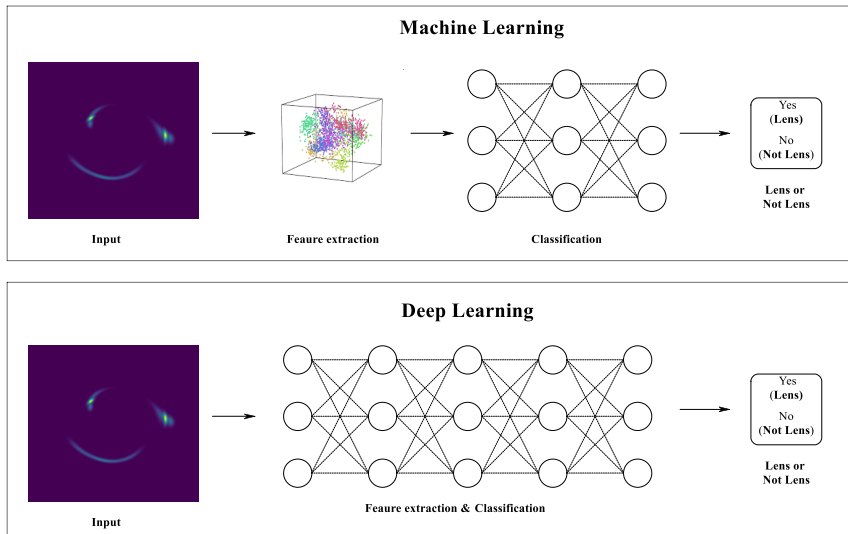
Dit proefschrift gebruikt de data van de twee interferometers met een hoge hoekresolutie. Eén bestaat uit tien telescopen elk met een diameter van 25 meter, welke zich op de Verenigde Staten, Hawaii en de Maagdeneilanden bevinden. Deze interferometer wordt Very Long Baseline Array (VLBA) genoemd en kan de radiohemel waarnemen in het bereik van 0,3 tot 83 GHz. De langste afstand tussen twee telescopen in VLBA is 8600 km, waarmee een hoge hoekresolutie bereikt kan worden (onthoud dat de hoekresolutie wordt gedefinieerd als  $\lambda/B$ ). Voor het tweede en derde hoofdstuk van dit proefschrift zijn VLBA-waarnemingen op een frequentie van 1,4 GHz gebruikt. In het vierde hoofdstuk gebruikten wij een Nederlandse interferometer. Deze interferometer, de Low-Frequency Array (LOFAR), bestaat uit veel eenvoudige (bipolaire) antennes, in tegenstelling tot conventionele met diafragma gevulde telescopen. Dit eenvoudige antenne ontwerp stelt astronomen in staat om in één keer een groot gebied

aan de hemel te observeren. LOFAR is momenteel de grootste radiotelescoop ter wereld die werkzaam is op de laagste frequenties die vanaf de aarde zichtbaar zijn.

## OVERZICHT VAN DE HOOFDSTUKKEN

Het eerste hoofdstuk betreft de wetenschappelijke inleiding van dit proefschrift en bevat een gedetailleerde beschrijving van de data, de wetenschappelijke vragen, de technieken en benaderingen die worden gebruikt de vragen te beantwoorden. In het tweede hoofdstuk analyseerden we de geobserveerde bronnen op VLBA-schaal, daarnaast hebben we de VLBA-detectieverhouding vergeleken met een andere interferometer die een hogere gevoeligheid heeft. De catalogus van de VLBA gedetecteerde bronnen op 1,4 GHz is publiekelijk beschikbaar. Onze analyses tonen aan dat, van de in totaal ongeveer 25.000 radiobronnen die geïdentificeerd werden door de andere interferometer, er ongeveer 5.000 radiobronnen werden geïdentificeerd met VLBA, wat neerkomt op een totale detectieratio van ongeveer 20 procent. Onze analyse toonde aan dat de kenmerken van de radiobron, zoals helderheid en compactheid, de mate van detectie met VLBA beïnvloeden. Om het aantal bestaande bronnen in een waarneming te berekenen, hebben we het oppervlak van het waargenomen hemelgebied bepaald, rekening houdend met de gevoeligheid en configuratie van de waarneming. Wij analyseerden ook de volledigheid van de catalogus door 13.500 radiobronnen te simuleren. De resultaten laten zien dat de gegenereerde catalogus van de waarneming compleet is bij de signaal-ruisverhouding van 7,8. Deze berekeningen kunnen worden gebruikt om een schatting te maken van het aantal bronnen die door de volgende generatie interferometers gedetecteerd zullen worden.

Voor nieuwe wetenschappelijke ontdekkingen is het identificeren van bronnen belangrijk, daarom hebben wij ons derde hoofdstuk gewijd aan een nieuwe pijplijn voor brondetectie en karakterisering met behulp van ML-algoritmen. De pijplijn werkt met representatieve gesimuleerde radio-observaties die waarnemingsruis bevatten. Dit hoofdstuk evalueert het gebruik van dergelijke algoritmen in vergelijking tot traditionele detectietechnieken op basis van volledigheid en zuiverheid van de catalogus. De volledigheid van de catalogus wordt gemaximaliseerd wanneer het detectie algoritme alle echte bronnen in de testdata (de nepdata die gebruikt wordt voor het testen van algoritmes) kan detecteren. Cataloguszuiverheid meet het aantal nepbronnen dat het algoritme detecteert. Idealiter zouden detectie-algoritmen alle echte bronnen, maar geen valse, moeten detecteren. De door onze pijplijn gegenereerde catalogus bleek compleet te zijn bij een signaal-ruisverhouding van 7,5, terwijl het traditionele brondetectiealgoritme pas volledigheid bereikte bij een hogere signaal-ruisverhouding van



**Figuur II** – Een overzicht van de machine learning en deep learning-algoritmen voor het detecteren van sterke zwaartekrachtlenssystemen. Het invoerbeeld is een voorbeeld van een zwaartekrachtlenstelsel dat de illusie wekt van een lachend gezicht, doordat het licht van achtergelegen sterrenstelsel vervormd is.

8,4. Wat nog belangrijker is, is dat onze catalogus aanzienlijk betrouwbaarder is (met bijna een factor twee) wanneer het gaat om het niet detecteren van nepbronnen voor een signaal-ruisverhouding van meer dan 5. Naast het detecteren van de radiobronnen, kan onze pijplijn waarnemingsruis verwijderen, de structuur van hemelbronnen herstellen en de eigenschappen van de objecten, zoals grootte en helderheid, voorspellen.

In het vierde hoofdstuk worden kunstmatige-intelligentietechnieken gebruikt om een zeldzaam fenomeen in de astronomie te identificeren. Dit fenomeen, de zogenaamde sterke zwaartekrachtlenst, treedt op wanneer twee sterrenstelsels zich langs één waarnemingslijn bevinden. Het zwaartekrachtsveld van het nabijgelegen sterrenstelsel buigt de emissie van de verder gelegen bron af onder een hoek, wat resulteert in vervormde emissies in de vorm van bogen en meerdere afbeeldingen. Hoewel dit fenomeen zich voordoet met een kans van slechts 1 op duizend, heeft het interessante toepassingen voor astronomen en zijn de ontdekking en studie ervan van bijzonder belang voor astronomen. Figuur.II toont een overzicht van de lensidentificatiepijplijn. Het invoerbeeld (een waarneming) voor de pijplijn is een voorbeeld van een



zwaartekrachtlenessysteem. Het toont de zeldzame gebeurtenis waarbij het licht van het verre sterrenstelsel vervormd wordt op een manier die de illusie van een lachend gezicht creëert. Dit figuur geeft ook een schematisch overzicht van de verschillen tussen machine learning en deep learning. Het bovenste paneel toont een typische machine learning-pijplijn, deze bestaat uit een afzonderlijke stap voor het afleiden van functies: het proces waarin de functies die relevant zijn voor het specifieke doel van de pijplijn worden geselecteerd of gegenereerd. Voor de lensdetectiepijplijn zoeken wij naar het bestaan van bogen en meerdere componenten in de afbeelding (zie het voorbeeld van de waarneming die gebruikt wordt als invoer voor het model). Er zijn meer specifieke eisen om een input te classificeren als een sterk lensstelsel, maar dit past niet binnen het bestek van dit proefschrift. De geselecteerde kenmerken worden vervolgens gebruikt in de classificatie-algoritmen om te beslissen tot welke klasse (lens of niet-lens) de invoer behoort. Dit paradigma is veranderd voor DL-algoritmen. Zoals weergegeven in het onderste paneel van Figuur.II, zijn de functie-extractie en de classificatie ingebed in één stap. Op deze manier kunnen DL-algoritmen leren wat de belangrijkste functies zijn die twee typen data van elkaar onderscheiden.

De internationale stations van LOFAR kunnen miljoenen radiobronnen detecteren met de hoekresolutie die nodig is om sterke zwaartekrachtlenzen te detecteren. In Hoofdstuk 4 hebben wij het gebruik van DL-algoritmen voor het detecteren van sterke zwaartekrachtlenzen getest met behulp van representatieve gesimuleerde data voor de internationale LOFAR-interferometer. Het primaire doel van dit hoofdstuk is om een nieuwe benadering te presenteren voor het vinden van zeldzame zwaartekrachtlenzen, gebruikmakend van de internationale LOFAR interferometerstations. Onze resultaten kunnen ook gebruikt worden om de gevoeligheid van deze interferometer te bepalen. Daarnaast kan op basis van de gevoeligheid en hoekresolutie van de data een uitspraak gedaan worden over het lenstype.

## Acknowledgments

This is truly an exciting moment for me as I get to wrap up one of the most amazing chapters of my life. It is almost unbelievable how my life changed since nearly four years ago. It was in 2017 that I decided to apply for PhD positions in Europe. Before that, I was a software developer in a startup company developing a Windows application from scratch. After a while as a software developer, I realized that it was not the career I wanted to pursue for the rest of my life and that it was the time to make a change. Considering my graduate studies being in computer science, I was looking for PhD positions in computer science departments. My knowledge of astronomy was limited to what I had seen in the news and in magazines. I was lucky enough to find this amazing project between the two departments of computer science and astronomy. The need for interdisciplinary projects that involve people from different sciences got me to this opportunity. For this, I am thankful not only because it has made me a better scientist with broader knowledge, but also because it has changed me as a person. I have definitely grown during these years, it feels like I have found my true self. I owe a great debt of gratitude to the people who have made this possible.

*John*, I strongly believe that the PhD experiences (also mental health of PhD students) are influenced by the personality and supervision style of the supervisor. I consider myself lucky to work with you. Your dedication, friendly supervision and continuous support have made this journey enjoyable. Thank you for teaching me radio astronomy. Your ability to explain complicated concepts in an understandable way is admirable. I look forward to more collaborations with you in the future.

*Michael*, thank you for trusting in me and hiring me to work on this interdisciplinary project. I clearly remember the day I was asked in the interview to select one of the DSSC projects. I am so glad that I chose what I chose. Thank you for your support and prompt comments on the papers. I wish you good health and hope you have a speedy recovery after the surgery.

Special thanks go out to the reading committee, I thank you for your considerable time investment. Your comments have improved the quality of my thesis.

*Reynier*, I would like to thank you for your nice presence from day one, which started with the interviews. While there were five computer scientists in the committee, you were the only astronomer there. I also appreciate your considerate and thoughtful manner over the past four years especially in the early weeks of my stay in the Netherlands.

*Pooja*, you have always been that calm, logical friend I could trust in times of need. Thank you for the lovely dinners, warm teas in the BIG mug and your comforting presence. I hope our friendship continues to grow even after we both have left Groningen. *Tirna*, you were probably the first friend in Groningen who I trusted with my life story. Thanks for being a reliable friend. Although young, you are wise and thoughtful with many interesting questions and ideas in mind. I am sure that the future holds great things for you and I cannot wait to see what they are. *Hyoyin*, you are for sure one of the most responsible people I have ever seen in my life. Thank you for being a trustworthy friend. I admire your thirst to learn new stuff, not only in science, but also in sport and music. I have enjoyed cycling with you around the city and playing games with you. The competitive spirit you brought to the games made them much more fun. *Avanti* and *Apu*, it has been nice to have you as my friends. I found your company very pleasant and hope it becomes more frequent.

*Elahe*, I remember your smile and kind personality when I first moved to the Netherlands. Although you have lived in a different city for the last four years and we didn't have much time together, I trust you like my family. *Teymoor*, I guess it took us a while until we could trust each other, but this is true for all friendships, right? Your sincerity, stubbornness, curious mind with big ideas fascinate me. I hope we get to work together sometime on a common interest. *Fateme (Fereidooni)*, I think it was the PhD day in 2018 that we first met. I have been feeling comfortable and entertained every time we hang out. Thanks for the trips and fun memories we made together. *Azadeh* and *Soheil*, you are my favorite couple. I don't know what your secret is but I have found myself being completely connected to my inner self when I am with you. I want you to know that I value your friendship greatly.

Thanks to the group members, old and new, *Cristiana, Hannah, Ruslan* and *Di*. I appreciate the conversations we had on many topics including radio astronomy. Special thanks to *Willem* for his collaboration on the lens detection project. Your lens simulation pipeline has contributed greatly in the progress of the project. To all my lovely friends and colleagues at Kapteyn whom I had meaningful and valuable conversations: *Jiwoo, Youngjoo, Nika, Katya, Anne, Kristiina, Sara, Jonas, Georg, Laurent, Pranav, Seyda, Olmo, Andrea, Enrico, Suma, Simon, Pavel, Bharat, Anastasia, Anqi, Andrés, Julia, Umit, Veronica, Niki* and many more, thank you for contributing to a lovely working environment. Kapteyn is only memorable with you being a part of it. Thanks to *Bram* for not only being a fun good friend, but also for taking care of the included dutch summary. I also would like to thank the *Intelligent Systems* and *DSSC* groups, more specifically *Stefania* for her support during last four years with administrative tasks associated with the DSSC program.

Special thanks to my old officemates. *Jorrit, Will* and *Daniel*. It was so much fun to share the office with you guys. I was bummed the day each one of you left. The way we greeted each other every morning, the lunches in the coffee corner and many fun conversations is truly memorable. I also would like to thank my new officemates *Orlin, Francisco, Bharat* and *Xin* for their nice company in the last year.

*Mina*, my best and oldest friend, our friendship is 16 years old now. We managed to keep our friendship after almost 12 years of being geographically dispersed, and for that, I am proud of ourselves. Thanks for being a shoulder to lean on, in happiness and in sadness. I hope you know how much you mean to me. Your big heart and insightful mind are only a few of my favorite qualities in you. Hope to see you soon.

*Amir*, I know you do not like to share personal feelings out loud, so I will keep it short. Thanks for taking care of me when the work load was high. Going to the gym together almost every night was truly the best stress relieving strategy that I could take when PhD life became tough. I couldn't have done it without you.

*Parmis* and *Soheila*, I cannot imagine my life without you. You make my life worthwhile. Thanks for the unlimited support and positive energy you spread for as long as I remember. Thanks to my brothers *Behzad* and *Hamid* for supporting my decisions and helping me when I needed. They are the reason I get to work on computer science, the field that I have enjoyed working on since 14 years ago. I like to thank my childhood friend *Fateme (Rezaei)* for her true friendship, which has lasted throughout the years.

*Mom* and *Dad*, I owe everything to you. Thank you for being an enormous source of energy and support for my entire life. You are to be thanked for this book, as you are

for everything else I have ever accomplished in my life. I look forward to spending more time with you in the future.

**Bibliography**

- Akhazhanov A. et al., 2021, arXiv e-prints, arXiv:2109.09781
- Amante M. H., Magaña J., Motta V., García-Aspeitia M. A., Verdugo T., 2020, MNRAS, 498, 6013
- Apicella A., Donnarumma F., Isgrò F., Prevete R., 2020, arXiv e-prints, arXiv:2005.00817
- Auger M. W., Treu T., Bolton A. S., Gavazzi R., Koopmans L. V. E., Marshall P. J., Bundy K., Moustakas L. A., 2009, ApJ, 705, 1099
- Auger M. W., Treu T., Bolton A. S., Gavazzi R., Koopmans L. V. E., Marshall P. J., Moustakas L. A., Burles S., 2010, ApJ, 724, 511
- Avestruz C., Li N., Zhu H., Lightman M., Collett T. E., Luo W., 2019, ApJ, 877, 58
- Ayachi et al., 2020, in Proceedings of the 8th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, Vol.1, Springer International Publishing, pp. 234–243
- Badole S. et al., 2022, A&A, 658, A7
- Bahé Y. M., 2021, MNRAS
- Baron D., 2019, arXiv e-prints, arXiv:1904.07248
- Becker I., Pichara K., Catelan M., Protopapas P., Aguirre C., Nikzat F., 2020, MNRAS, 493, 2981
- Becker R. H., White R. L., Helfand D. J., 1995, ApJ, 450, 559
- Biehl M., Bunte K., Longo G., Tino P., 2018, in Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2018, Verleysen M., ed., i6doc.com, pp. 307–313, available at <http://www.i6doc.com/en>
- Blandford R., Narayan R., 1992, Annual review of astronomy and astrophysics, 30, 311
- Bolton A. S., Burles S., Koopmans L. V. E., Treu T., Gavazzi R., Moustakas L. A., Wayth R., Schlegel D. J., 2008, ApJ, 682, 964
- Bolton A. S., Burles S., Koopmans L. V. E., Treu T., Moustakas L. A., 2006, ApJ, 638, 703

Bom C., Poh J., Nord B., Blanco-Valentin M., Dias L., 2019, arXiv e-prints, arXiv:1911.06341

Bonnassieux E. et al., 2021, arXiv e-prints, arXiv:2108.07294

Bonvin V. et al., 2017, MNRAS, 465, 4914

Boureau Y., Bach F., LeCun Y., Ponce J., 2010, in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2559–2566

Bowles M., Scaife A. M. M., Porter F., Tang H., Bastien D. J., 2021, MNRAS, 501, 4579

Brotten N. W. et al., 1967, Science, 156, 1592

Browne I. W. A. et al., 2003, MNRAS, 341, 13

Cabanac R. A. et al., 2007, A&A, 461, 813

Chae K. H. et al., 2002, Phys. Rev. Lett., 89, 151301

Cheng D., Zhang S., Deng Z., Zhu Y., Zong M., 2014, in International Conference on Advanced Data Mining and Applications, Springer, pp. 499–512

Cheng T.-Y. et al., 2020a, MNRAS, 493, 4209

Cheng T.-Y., Li N., Conselice C. J., Aragón-Salamanca A., Dye S., Metcalf R. B., 2020b, MNRAS, 494, 3750

Chirivì G., Yıldırım A., Suyu S. H., Halkola A., 2020, A&A, 643, A135

Cohen M. H. et al., 1977, Nature, 268, 405

Collett T. E., 2015, ApJ, 811, 20

Condon J. J. et al., 2012, ApJ, 758, 23

Condon J. J., Cotton W. D., Greisen E. W., Yin Q. F., Perley R. A., Taylor G. B., Broderick J. J., 1998, AJ, 115, 1693

Congdon A. B., Keeton C. R., 2018, Principles of Gravitational Lensing. Springer

Cover T., Hart P., 1967, IEEE Transactions on Information Theory, 13, 21

de Gasperin F. et al., 2021, Astronomy and Astrophysics, 648, A104

Deller A. T. et al., 2011, PASP, 123, 275

Deller A. T., Middelberg E., 2014, AJ, 147, 14

Dodson R., Rioja M., 2018, in 14th European VLBI Network Symposium & Users Meeting (EVN 2018), p. 86

Event Horizon Telescope Collaboration et al., 2019a, ApJ, 875, L2

—, 2019b, ApJ, 875, L1

Faure B., Bournaud F., Fensch J., Daddi E., Behrendt M., Burkert A., Richard J., 2021, MNRAS, 502, 4641

Faure C. et al., 2008, ApJS, 176, 19

Friedman J., Hastie T., Tibshirani R., 2000, The annals of statistics, 28, 337

Gal Y., Ghahramani Z., 2015, arXiv e-prints, arXiv:1506.02142

Garrett M. A. et al., 2001, A&A, 366, L5

Gentile F. et al., 2022, MNRAS, 510, 500

Ghahramani Z., 2004, Unsupervised Learning, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 72–112

Gilman D., Birrer S., Nierenberg A., Treu T., Du X., Benson A., 2020, MNRAS, 491, 6077

Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press, <http://www.deeplearningbook.org>

Haarsma D. B. et al., 2005, AJ, 130, 1977

Hale C. L., Robotham A. S. G., Davies L. J. M., Jarvis M. J., Driver S. P., Heywood I., 2019, *MNRAS*, 487, 3971

Hales C. A., Murphy T., Curran J. R., Middelberg E., Gaensler B. M., Norris R. P., 2012, *MNRAS*, 425, 979

Hancock P. J., Trott C. M., Hurley-Walker N., 2018, *PASA*, 35, e011

Hartley P., Flamary R., Jackson N., Tagore A. S., Metcalf R. B., 2017, *MNRAS*, 471, 3378

Harwood J. J. et al., 2021, arXiv e-prints, arXiv:2108.07288

Hassanat A. B., Abbadì M. A., Altarawneh G. A., Alhasanat A. A., 2014, arXiv preprint arXiv:1409.0919

Hayou S., Doucet A., Rousseau J., 2019, arXiv e-prints, arXiv:1902.06853

He K., Zhang X., Ren S., Sun J., 2015, arXiv e-prints, arXiv:1512.03385

Herrera Ruiz N. et al., 2017, *A&A*, 607, A132

—, 2018, *A&A*, 616, A128

Hezaveh Y. D., Perreault L., Marshall P. J., 2017, *Nature*, 548, 555

Högbom J. A., 1974, *A&AS*, 15, 417

Hsueh J. W., Enzi W., Vegetti S., Auger M. W., Fassnacht C. D., Despali G., Koopmans L. V. E., McKean J. P., 2020, *MNRAS*, 492, 3047

Impellizzeri C. M. V., McKean J. P., Castangia P., Roy A. L., Henkel C., Brunthaler A., Wucknitz O., 2008, *Nature*, 456, 927

Intema H. T., Jagannathan P., Mooley K. P., Frail D. A., 2017, *A&A*, 598, A78

Jackson N., 2008, *MNRAS*, 389, 1311

Jackson N. et al., 2021, arXiv e-prints, arXiv:2108.07284

Jacobs C. et al., 2019a, *ApJS*, 243, 17

—, 2019b, *MNRAS*, 484, 5330

Jacobs C., Glazebrook K., Collett T., More A., McCarthy C., 2017, *MNRAS*, 471, 167

Kim K., Lee J., Yuen R. S. H., Hannuksela O. A., Li T. G. F., 2020, Identification of lensed gravitational waves with deep learning

King L. J., Browne I. W. A., Marlow D. R., Patnaik A. R., Wilkinson P. N., 1999, *MNRAS*, 307, 225

Kingma D. P., Ba J., 2014, arXiv e-prints, arXiv:1412.6980

Kingma D. P., Welling M., 2013, arXiv e-prints, arXiv:1312.6114

Koopmans L. V. E., Treu T., Bolton A. S., Burles S., Moustakas L. A., 2006, *ApJ*, 649, 599

Kovačević M., Chiaro G., Cutini S., Tosti G., 2020, *MNRAS*, 493, 1926

Krizhevsky A., Sutskever I., Hinton G. E., 2017, *Commun. ACM*, 60, 84–90

Lacy M. et al., 2020, *PASP*, 132, 035001

Lanusse F., Ma Q., Li N., Collett T. E., Li C.-L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2018, *MNRAS*, 473, 3895

LeCun Y., Cortes C., 2010

Lee H., Kwon H., 2017, *IEEE Transactions on Image Processing*, 26, 4843

Lemon C. et al., 2020, *MNRAS*, 494, 3491

Lemon C. A., Auger M. W., McMahon R. G., 2019, *MNRAS*, 483, 4242

Lemon C. A., Auger M. W., McMahon R. G., Ostrovski F., 2018, *MNRAS*, 479, 5060

Li R. et al., 2021, *ApJ*, 923, 16



Liaw A., Wiener M., 2002, *R News*, 2, 18

Lin J. Y.-Y., Yu H., Morningstar W., Peng J., Holder G., 2020, Hunting for dark matter subhalos in strong gravitational lensing with neural networks

Lu L., 2020, *Communications in Computational Physics*, 28, 1671

Lukic V., de Gasperin F., Brüggem M., 2019, *Galaxies*, 8, 3

Marcote B. et al., 2020, *Nature*, 577, 190

Maresca J., Dye S., Li N., 2021, *MNRAS*, 503, 2229

Margalef-Bentabol B., Huertas-Company M., Charnock T., Margalef-Bentabol C., Bernardi M., Dubois Y., Storey-Fisher K., Zanis L., 2020, arXiv e-prints, arXiv:2003.08263

Marianer T., Poznanski D., Prochaska J. X., 2020, A semi-supervised machine learning search for never-seen gravitational-wave sources

Marshall P. J. et al., 2016, *MNRAS*, 455, 1171

McCulloch W. S., Pitts W., 1943, *The bulletin of mathematical biophysics*, 5, 115

McKean J. et al., 2015, in *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, p. 84

McKean J. P., Browne I. W. A., Jackson N. J., Fassnacht C. D., Helbig P., 2007, *MNRAS*, 377, 430

McMullin J. P., Waters B., Schiebel D., Young W., Golap K., 2007, in *Astronomical Society of the Pacific Conference Series, Vol. 376, Astronomical Data Analysis Software and Systems XVI*, Shaw R. A., Hill F., Bell D. J., eds., p. 127

Meneghetti M. et al., 2017, *MNRAS*, 472, 3177

Metcalfe R. B. et al., 2019, *A&A*, 625, A119

Meylan G., Jetzer P., North P., Schneider P., Kochanek C. S., Wambsganss J., 2006, *Saas-Fee Advanced Course 33: Gravitational Lensing: Strong, Weak and Micro*

Middelberg E. et al., 2013, *A&A*, 551, A97

Miyoshi M., Moran J., Herrnstein J., Greenhill L., Nakai N., Diamond P., Inoue M., 1995, *Nature*, 373, 127

Mohan N., Rafferty D., 2015, *PyBDSF: Python Blob Detection and Source Finder*

Morabito L. K. et al., 2021, arXiv e-prints, arXiv:2108.07283

More A., Cabanac R., More S., Alard C., Limousin M., Kneib J. P., Gavazzi R., Motta V., 2012, *ApJ*, 749, 38

More A. et al., 2016, *MNRAS*, 455, 1191

Morganti R., Fogasy J., Paragi Z., Oosterloo T., Orienti M., 2013, *Science*, 341, 1082

Morningstar W. R., Hezaveh Y. D., Perreault Levasseur L., Blandford R. D., Marshall P. J., Putzky P., Wechsler R. H., 2018, arXiv e-prints, arXiv:1808.00011

Morningstar W. R. et al., 2019, *ApJ*, 883, 14

Muxlow T. W. B. et al., 2020, *MNRAS*, 495, 1188

Myers S. T. et al., 2003, *MNRAS*, 341, 1

Negrello M. et al., 2017, *MNRAS*, 465, 3558

—, 2010, *Science*, 330, 800

Nieuwenhuizen T. M., Limousin M., Morandi A., 2021, *European Physical Journal Special Topics*

Nightingale J. et al., 2021, *The Journal of Open Source Software*, 6, 2825

Nolte A., Wang L., Bilicki M., Holwerda B., Biehl M., 2019, arXiv e-prints, arXiv:1903.07749

O'Brien T. J. et al., 2006, *Nature*, 442, 279  
Oguri M., Marshall P. J., 2010, *MNRAS*, 405, 2579  
Oostwal E., Straat M., Biehl M., 2019, arXiv e-prints, arXiv:1910.07476  
O'Shea K., Nash R., 2015, *An introduction to convolutional neural networks*  
Padovani P., 2017, *Nature Astronomy*, 1, 0194  
Patnaik A. R., Browne I. W. A., Wilkinson P. N., Wrobel J. M., 1992, *MNRAS*, 254, 655  
Pearson W. J., Wang L., Trayford J. W., Petrillo C. E., van der Tak F. F. S., 2019, *A&A*, 626, A49  
Pedregosa F. et al., 2012, arXiv e-prints, arXiv:1201.0490  
Perreault L., Hezaveh Y. D., Wechsler R. H., 2017, *ApJ*, 850, L7  
Petrillo C. E. et al., 2017, *MNRAS*, 472, 1129  
—, 2019, *MNRAS*, 484, 3879  
Petrov L., 2021, *AJ*, 161, 14  
Porcas R. W., Booth R. S., Browne I. W. A., Walsh D., Wilkinson P. N., 1979, *Nature*, 282, 385  
Pradel N., Charlot P., Lestrade J. F., 2006, *A&A*, 452, 1099  
Prandoni I., Guglielmino G., Morganti R., Vaccari M., Maini A., Röttgering H. J. A., Jarvis M. J., Garrett M. A., 2018, *MNRAS*, 481, 4548  
Pruzhinskaya M. V., Malanchev K. L., Kornilov M. V., Ishida E. E. O., Mondon F., Volnova A. A., Korolev V. S., 2019, *MNRAS*, 489, 3591  
Radcliffe J. F. et al., 2018, *A&A*, 619, A48  
Riechers D. A. et al., 2011, *ApJ*, 739, L32  
Ritondale E., Vegetti S., Despali G., Auger M. W., Koopmans L. V. E., McKean J. P., 2019, *MNRAS*, 485, 2179  
Rohlfs K., Wilson T. L., 2013, *Tools of radio astronomy*. Springer Science & Business Media  
Rojas K. et al., 2021, arXiv e-prints, arXiv:2109.00014  
Ronneberger O., Fischer P., Brox T., 2015, arXiv e-prints, arXiv:1505.04597  
Santurkar S., Tsipras D., Ilyas A., Madry A., 2018, arXiv e-prints, arXiv:1805.11604  
Schaefer C., Geiger M., Kuntzer T., Kneib J. P., 2018, *A&A*, 611, A2  
Scherer D., Müller A., Behnke S., 2010, in *Artificial Neural Networks – ICANN 2010*, Diamantaras K., Duch W., Iliadis L. S., eds., Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 92–101  
Schneider P., 1992, in *Gravitational Lenses*, Springer, pp. 196–208  
Sedaghat N., Mahabal A., 2018, *MNRAS*, 476, 5365  
Shimwell T. W. et al., 2019, *A&A*, 622, A1  
Sonnenfeld A. et al., 2018, *PASJ*, 70, S29  
Spingola C., McKean J. P., Auger M. W., Fassnacht C. D., Koopmans L. V. E., Lagattuta D. J., Vegetti S., 2018, *MNRAS*, 478, 4816  
Spiniello C. et al., 2018, *MNRAS*, 480, 1163  
Spiniello C., Trager S., Koopmans L. V. E., Conroy C., 2014, *MNRAS*, 438, 1483  
Spiniello C., Trager S. C., Koopmans L. V. E., Chen Y. P., 2012, *ApJ*, 753, L32  
Springenberg J. T., Dosovitskiy A., Brox T., Riedmiller M., 2014, arXiv e-prints, arXiv:1412.6806  
Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, *The*

journal of machine learning research, 15, 1929  
Stacey H. R. et al., 2019, A&A, 622, A18  
Stacey H. R. et al., 2018, Monthly Notices of the Royal Astronomical Society, 476, 5075–5114  
Suyu S. H. et al., 2013, ApJ, 766, 70  
Suyu S. H., Marshall P. J., Auger M. W., Hilbert S., Blandford R. D., Koopmans L. V. E., Fassnacht C. D., Treu T., 2010, ApJ, 711, 201  
Sweijen F. et al., 2021, arXiv e-prints, arXiv:2108.07290  
—, 2022, A&A, 658, A3  
Swinbank A. M. et al., 2009, MNRAS, 400, 1121  
Szegedy C., Ioffe S., Vanhoucke V., Alemi A., 2016, arXiv e-prints, arXiv:1602.07261  
Tilley D., Cleghorn C. W., Thorat K., Deane R., 2020, arXiv e-prints, arXiv:2008.02093  
Timmerman R. et al., 2021, arXiv e-prints, arXiv:2108.07287  
Treu T., 2010, ARA&A, 48, 87  
Treu T. et al., 2018, MNRAS, 481, 1041  
Treu T., Koopmans L. V., Bolton A. S., Burles S., Moustakas L. A., 2006, ApJ, 640, 662  
Vafaei Sadr A., Vos E. E., Bassett B. A., Hosenie Z., Oozeer N., Lochner M., 2019, MNRAS, 484, 2793  
Van der Maaten L., Hinton G., 2008, Journal of Machine Learning Research, 9  
van der Walt S. et al., 2014, arXiv e-prints, arXiv:1407.6245  
van Haarlem M. P. et al., 2013, A&A, 556, A2  
Vegetti S., Koopmans L. V. E., Auger M. W., Treu T., Bolton A. S., 2014, MNRAS, 442, 2017  
Vegetti S., Koopmans L. V. E., Bolton A., Treu T., Gavazzi R., 2010, MNRAS, 408, 1969  
Vegetti S., Lagattuta D. J., McKean J. P., Auger M. W., Fassnacht C. D., Koopmans L. V. E., 2012, Nature, 481, 341  
Walsh D., Carswell R. F., Weymann R. J., 1979, Nature, 279, 381  
Wardlow J. L. et al., 2013, ApJ, 762, 59  
White R. L., Becker R. H., Helfand D. J., Gregg M. D., 1997, ApJ, 475, 479  
Wong K. C. et al., 2020, MNRAS, 498, 1420  
Wucknitz O., Biggs A. D., Browne I. W. A., 2004, MNRAS, 349, 14  
Zeng Q., Li X., Lin H., 2020, MNRAS, 494, 3110