

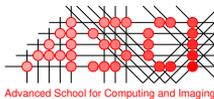
# Machine learning-based *anomaly* detection for radio telescopes



Michael Mesarcik

MACHINE  
LEARNING-BASED  
ANOMALY DETECTION FOR  
RADIO TELESCOPES

MICHAEL MESARCIK



**ASTRON** netherlands  
**eScience center**

This work was carried out in the ASCI graduate school.  
ASCI dissertation series number: 450

This work is part of the “Perspectief” research programme “Efficient Deep Learning” (EDL, <https://efficientdeeplearning.nl>), which is financed by the Dutch Research Council (NWO) domain Applied and Engineering Sciences (TTW). The research makes use of radio astronomy data from the LOFAR telescope, which is operated by ASTRON (Netherlands Institute for Radio Astronomy), an institute belonging to the Netherlands Foundation for Scientific Research (NWO-I)  
Copyright © 2024 Michael Mesarcik.

Cover Design by: Tammy Joubert  
Thesis template: classicthesis by André Miede and Ivo Pletikosić.  
Printed and bound by Ipskamp printing

ISBN: 978-94-6473-446-1

Machine learning-based anomaly detection for radio telescopes

## ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. P.P.C.C. Verbeek

ten overstaan van een door het College voor Promoties ingestelde commissie,

in het openbaar te verdedigen in de Agnietenkapel

op woensdag 24 april 2024, te 16.00 uur

door Michael Benno Mesarcik

geboren te ZAF

***Promotiecommissie***

<i>Promotores:</i>	prof. dr. R.V. van Nieuwpoort prof. dr. ir. C.T.A.M. de Laat	Universiteit van Amsterdam Universiteit van Amsterdam
<i>Copromotores:</i>	dr. ir. A.J. Boonstra dr. E.B. Rangelova	ASTRON Netherlands eScience Center
<i>Overige leden:</i>	dr. P. Grosso prof. dr. P.T. Groth prof. dr. ir. H. Corporaal prof. dr. A.M.M. Scaife dr. B.A. Rowlinson prof. dr. ir. S.J. Wijnholds	Universiteit van Amsterdam Universiteit van Amsterdam Eindhoven University of Technology University of Manchester Universiteit van Amsterdam ASTRON

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

To vertebrae C4, C5 and C7, your support is appreciated.



# CONTENTS

---

1	INTRODUCTION	1
1.1	System health management in radio telescopes	3
1.2	Machine learning-based anomaly detection	5
1.3	Real time scientific data processing	7
1.4	Thesis structure	7
1.5	Author publications related to this thesis	8
1.5.1	Peer reviewed journal publications	8
1.5.2	Peer reviewed conferences and workshops	9
1.5.3	Code	9
1.5.4	Datasets	10
2	BACKGROUND	11
2.1	The Low Frequency Array	11
2.1.1	The LOFAR central data processor	13
2.1.2	Online scientific data processing	15
2.1.3	System monitoring and control	16
2.2	Machine learning in astronomy	16
2.2.1	Representation learning in astronomy	17
2.2.2	Machine learning-based anomaly detection	17
2.2.3	Radio frequency interference detection in the deep learning era	19
3	LEARNING REPRESENTATIONS OF RADIO ASTRONOMY SPECTROGRAMS	21
3.1	Introduction	21
3.2	Data preparation	23
3.2.1	Data-set preprocessing	24
3.3	Variational autoencoder architecture	25
3.3.1	Architectural motivation and constraints	28
3.4	Model evaluation	29
3.4.1	Quantitative evaluation using simulated data	29
3.4.2	Qualitative evaluation using LOFAR data	30
3.5	Results	32
3.5.1	Simulated data	32
3.5.2	LOFAR data	36

3.5.3	Generative abilities of the LOFAR-based VAE	39
3.6	Conclusions and discussion	43
4	NEAREST NEIGHBOUR-BASED ANOMALY DETECTION	45
4.1	Introduction	45
4.2	NLN: Nearest Latent Neighbours	47
4.2.1	Motivation	47
4.2.2	Problem formulation and approach	49
4.2.3	The NLN algorithm	52
4.3	Experiments	54
4.3.1	Evaluation methodology	54
4.3.2	Datasets	55
4.3.3	Model and parameter selection	56
4.3.4	Results	57
4.3.5	Time and memory efficiency	61
4.4	Ablation study	62
4.5	Discussion and conclusions	63
5	RADIO FREQUENCY INTERFERENCE DETECTION	65
5.1	Introduction	65
5.2	Method	67
5.2.1	Model definition and training	68
5.2.2	Nearest-Latent-Neighbours for RFI detection	70
5.2.3	Architectural considerations	71
5.3	Data selection and preprocessing	73
5.3.1	Simulated HERA dataset	73
5.3.2	LOFAR	74
5.4	Results	77
5.4.1	Evaluation Methodology	77
5.4.2	HERA Results	79
5.4.3	LOFAR results	81
5.5	Conclusions	86
6	THE RADIO OBSERVATORY ANOMALY DETECTOR	87
6.1	Introduction	87
6.2	Dataset	89
6.2.1	Observation selection and preprocessing	90
6.2.2	Labelling methodology	90
6.2.3	Class imbalance	94
6.2.4	Radio frequency interference considerations	94
6.3	The radio observatory anomaly detector	95
6.3.1	Problem formulation	95

6.3.2	Self supervised representation learning	96
6.3.3	Distinguishing normal from anomalous samples	97
6.3.4	Combining classification with anomaly detection	98
6.4	Experiments	99
6.4.1	Model parameters and training	100
6.4.2	Anomaly detection and classification	100
6.4.3	Model ablations	106
6.4.4	Computational performance analysis	108
6.5	Conclusions and future work	109
7	CONCLUSIONS	111
7.1	Contributions	111
7.1.1	Representation learning for radio astronomy spectrograms	111
7.1.2	Nearest neighbour-based anomaly detection	112
7.1.3	Expert labelled datasets for anomaly detection in radio astronomy	112
7.1.4	Data inspection framework for radio observatories	112
7.1.5	Unsupervised RFI detection algorithm	112
7.1.6	Real time anomaly detector for LOFAR	113
7.2	Answers to research questions	113
7.3	Future work	115
7.4	Vision and outlook	117
	BIBLIOGRAPHY	119
	ACKNOWLEDGEMENTS	135
	SUMMARY	137
	SAMENVATTING	139

# ACRONYMS

---

AE	Autoencoder
AAE	Adversarial Autoencoder
AUPRC	Area Under Precision Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
CNN	Convolutional Neural Network
CPU	Central Processing Unit
EHT	Event Horizon Telescope
EoR	Epoch of Reionisation
FRB	Fast Radio Burst
GAN	Generative Adversarial Network
GUI	Graphical User Interface
GPU	Graphics Processing Unit
HBA	High Band Antenna
HERA	Hydrogen Epoch of Reionization Array
IF	Isolation Forrest
KNN	K-Nearest Neighbours
LBA	Low Band Antenna
LOFAR	Low Frequency Array
LTA	Low Term Archive
MISO	Multiple Inlier Single Outlier

MLP Multilayer Perceptron

NLN Nearest Latent Neighbours

OOD Out of Distribution

RFI Radio Frequency Interference

ROAD Radio Observatory Anomaly Detector

SCADA Supervisory Control and Data Acquisition

SDSS Sloan Digital Sky Survey

SETI Search for Extraterrestrial Intelligence

SIMO Single Inlier Multiple Outlier

SHM System Health Management

SKA Square Kilometre Array

SNR Signal to Noise Ratio

SSL Self Supervised Learning

SVM Support Vector Machine

TCN Temporal Convolutional Network

VAE Variational Autoencoder

VIT Vision Transformer

VLA Very Large Array



# INTRODUCTION

---

Radio astronomy is the study of celestial objects' radio frequency emissions. The analysis of the radio spectrum has allowed astronomers to peer further back in the universe than ever before and study previously unseen phenomena like hydrogen line emissions [1], the Cosmic Microwave Background [2], pulsars [3] and many more. As astronomers' goals become more ambitious so do their requirements for sensitivity and resolution of their instruments. Research indicates that the sensitivity of radio telescopes has been increasing at a rate of tenfold per decade [4], this increase in sensitivity and resolution directly affects instrument size and complexity and this is only expected to continue growing.

Telescope resolution is directly proportional to aperture size and wavelength [5]. To reach the angular resolution required for modern science use-cases, such as the recently publicised black hole image [6], single "dish" telescope would need a diameter of approximately 13000 km (i.e. greater than the diameter of earth) [7]. For reasons of practicality and cost, *arrays* of smaller telescopes are used as cheaper and more practical alternatives. Through aperture synthesis and beamforming it has been shown that it is possible and highly effective to use large arrays of smaller antennas to resolve celestial objects with extremely small angular resolutions. Figure 1.1 shows an illustration of the LOw Frequency ARray (LOFAR) [8] and the Square Kilometre Array (SKA) [9], where LOFAR has an 0.21 arcsecond resolution at its highest observable frequency [8]. Whereas the SKA, once completed, will offer a 0.002 arcsecond angular resolution [10]. Where an arcsecond angular resolution is equivalent to resolving an 18mm coin at a distance of 4 km away.

Modern radio telescopes such as LOFAR and the SKA combine thousands of receivers across thousands of kilometres to offer the required sensitivity, resolution and the myriad of receiving modes for many different science cases. These telescope systems are composed of a multitude of interconnected components, including antennas, digital signal processors, corre-



(a) A part of the LOw Frequency ARray (b) Artist's impression of the Square Kilometre Array (SKA) [9]

Figure 1.1: Illustration of two different radio telescopes

lators, data storage systems, and an extensive network infrastructure [11]. These elements work in synergy to capture, process, and analyse radio signals from celestial sources. The large-scale and distributed nature of these telescopes necessitates meticulous management and monitoring to ensure seamless operation and maximise scientific output.

Monitoring of radio telescopes is not only necessary to ensure nominal system operations but also to manage external events. There has been significant progress in the field of Radio Frequency Interference (RFI) mitigation in the context of radio astronomy. This is partially a consequence of radio telescopes' extreme sensitivity (several orders of magnitude higher than conventional communications systems [4]), but also due to the unpredictability of both the terrestrial and astronomical radio frequency spectrum [12]. In effect terrestrial RFI can in some cases saturate the receiver front-end electronics, cause unstable amplification modes or, more commonly, leave strong high power artefacts in observing bands. While some system-specific monitoring in scientific computing pipelines has been integrated as indicated by Broekema et al. [13], the current monitoring systems do not suffice, and do not detect all artefacts. These corrupt the data, which are then less useful to astronomers. The corruption may also be detected much later, making computing pipelines inefficient and expensive. This study aims to rectify this problem, by addressing interruptions to the science data processing at a system-wide level, treating all potential interruptions equally.

Furthermore, it is expected that the sensitivity, resolution and by extension data-rates of radio telescopes will grow dramatically, we therefore anticipate the demand of real-time system-wide anomaly detection and monitoring to increase. This demand stems from the impossibility of storing the vast volumes of data generated by new scientific instruments. As such,

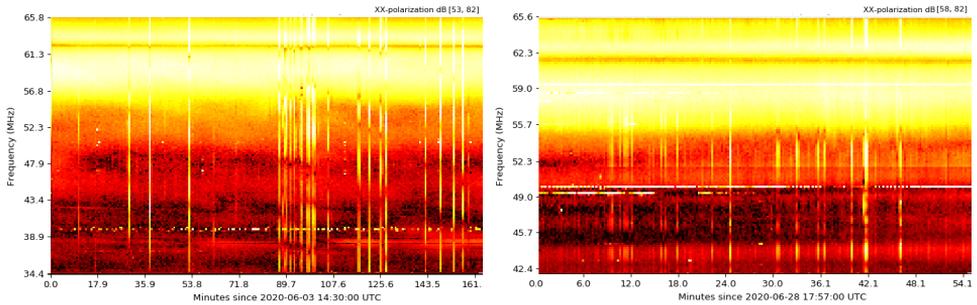
system health reporting and error localisation needs to be completed in lockstep with the other processing systems to limit buffering and potential data loss. Moreover, owing to the scale of modern radio telescopes, which consume substantial computational, networking, and signal processing resources, even small improvements in overall reliability can yield substantial gains in system efficiency. For this reason, this thesis concerns the study of System Health Management (SHM) in radio telescopes to maintain nominal telescope operations.

## 1.1 SYSTEM HEALTH MANAGEMENT IN RADIO TELESCOPES

System health management is a strategy for diagnosing and prognosing the cause and effect of system-based phenomena to maintain system uptime and ensure correct functioning. Radio telescopes offer a uniquely challenging environment for monitoring system behaviour due to their distributed scale and the variety of data processing systems. Furthermore, different science-cases require different modes of operation, such that the resolution and appearance of similar system critical-events can be vastly different across observation modes. Additionally, antennas in distant locations can experience problems simultaneously, thus resulting in potentially combinatorial effects in the data products.

At LOFAR, system health is managed by operators at the observatory who manually inspect intermediate data products and event reporting streams from each observation to detect and diagnose any potential problems. In this case, lower resolution plots are produced for each LOFAR telescope station for an observation to determine whether any potentially problematic events have occurred. Not only is expert-domain-specific knowledge necessary, but also understanding of the particularities of certain system health-threatening events. For example, events such as lightning storms and high-noise impulses can look fairly similar, especially in the low-resolution context. Lightning storms are geographically localised and impact all stations within a specific region. Moreover, lightning exhibits a high degree of temporal correlation across stations, with minimal delay between recorded events at each station. In contrast, high-noise events typically affect only one antenna at a time, with no temporal interdependence between antennas and stations

By this logic, if multiple stations in the same geographic location experience broadband high-power events across all polarisations that are temporally correlated, they are categorised as being impacted by lightning storms.



(a) Spectrogram with a lightning storm      (b) Spectrogram with a high noise event

Figure 1.2: Illustration of two different system wide anomalies from the LOFAR low band stations that appear visually similar without additional information.

Conversely, stations affected individually are typically experiencing high-noise events. The differences between the two events can be seen in Figure 1.2, for information on how these plots are generated see Chapter 2.1.

Not only does this process require significant amounts of time from domain experts, but it is fundamentally offline, with no way of ensuring the nominal system health *throughout* an observation. This is a key distinction; if failures can be detected in real-time, actions may be taken to correct, or at least mark the data as invalid. This way, the observation is not contaminated, and can still be used. Without real-time anomaly detection, (a part of) the dataset can be lost, and may need to be re-observed.

Additionally, the increasing data-rates results in operator based data inspection becoming infeasible for next generation telescopes such as the SKA. Therefore, this work develops both tools for analysis of system health data, as well as a tool that can potentially automate the whole process. This is done by creating models of normal telescope behaviour and trying to identify outliers using anomaly detection techniques. As such the first research question of this thesis is given by:

**Research Question 1:** *How can machine learning-based anomaly detection techniques be harnessed to improve system health management in radio telescopes?*

To validate our models for SHM we produced a dataset containing ten different classes comprising of system-based anomalies in radio telescopes. Here, we categorised the events into four different groups, these being astronomical anomalies, environmental effects, issues in the electronics and data processing system-based anomalies. This data was collected using observations from a variety of science cases to capture different feature morphologies and noise.

## 1.2 MACHINE LEARNING-BASED ANOMALY DETECTION

Anomaly detection refers to the process of identifying patterns or instances that deviate significantly from the norm or expected behaviour within a given dataset or system. Anomaly detection, novelty detection and out of distribution detection are closely related topics, with many overlaps in the literature. For simplicity sake, we refer to this category of methods as anomaly detection. Anomaly detection is typically a two step process. First a model is created of normal system behaviour and then some measure of deviation from the norm is defined [14]. Approaches for both modelling normality and discriminating between normal and anomalous data points have become increasingly dominated by machine learning based approaches [15].

Machine learning anomaly detection has emerged as the preferred approach over traditional methods due to its ability to generalise to a variety of complex datasets and problem types without significant modification [14]. Typical machine learning-based approaches for learning the inlying distribution are autoencoders [16–24], flow-based models [25], diffusion processes [26], generative adversarial networks [27–29] and many more. Furthermore, machine learning based discriminative approaches for separating normal from anomalous samples have been based on density [30], reconstruction [27] and statistical [31] approaches.

In this thesis, we limit our focus to only machine learning-based methods for anomaly detection. There has been consequential research in signal processing approaches for radio astronomy feature classification in sub-domains such as RFI detection, spectral line classification, transient characterisation and many more [4, 32–36]. We would expect that with sufficient effort, alternative signal-processing based approaches could be utilised to achieve similar goals of this thesis, however due to time limitations we focus exclusively on machine learning approaches for anomaly detection in radio telescopes.

Radio telescope system health management is a challenging environment for existing anomaly detection techniques for several reasons. Firstly, the variability of operating conditions of radio telescopes results in some anomalies occurring significantly more frequently than others. This means that anomaly detection models both need to be capable of detecting known, commonly occurring anomalies as well as rare, potentially unseen events. This is in contrast with typical anomaly detection which is normally posed as one-class-classification problem.

This thesis focuses on how machine learning based anomaly detection can be leveraged for detecting system health threatening events in radio observatories. As such our second research question is given by:

**Research Question 2:** *What is the most effective to way learn robust representations of normal data from radio telescopes?*

To answer the second research question we look at several different latent variable models as well as self-supervised learning approaches to find meaningful representations of the visibility spectrograms. This is challenging as the spectrograms from radio telescopes are noisy, have extremely a high-dynamic range, they combine thousands of stations for a single observation with complex-valued data that have multiple polarisations. We evaluate the learnt representations in two different scenarios. The first is the general SHM problem and the second considers RFI mitigation and how representation learning can yield improved performance over the state-of-the-art. In doing so we then evaluate the third research question:

**Research Question 3:** *Which are the most appropriate ways to discriminate between normal and anomalous samples using the learnt representations of normal data?*

We find that existing anomaly detection problem definitions are not well suited to the radio observatory use-case. Typical methods rely on the assumption that there is a single outlying class, whereas we want to classify both commonly occurring anomalies as well as detect those that are rare. Using this modified problem definition, we will answer the second

question by considering density, reconstruction and supervised methods for anomaly detection. We also investigate combining supervised and unsupervised approaches to detect both known and unknown events.

### 1.3 REAL TIME SCIENTIFIC DATA PROCESSING

In the context of low-frequency radio astronomy, scientific data processing has been successfully automated by running complex workflows that perform signal processing, calibration and imaging of interferometric data [37–40], RFI mitigation [41] and de-dispersion [42, 43] of time-domain data among many more. Additionally, continuous effort is being made to create high-performance real-time algorithms, to improve the quality and reliability of the scientific data [13, 44–47]. However, as of yet, there have been no attempts to fully automate the System Health Management (SHM) pipeline, and by virtue of the lack of work on this topic, no real-time implementations exist. This is in part due to the complexity of the challenge as well as the unavailability of SHM-specific datasets.

To demonstrate the practicality of our approach within the context of radio observatories, we assess the computational performance and throughput of the proposed system, additional details can be found in Chapter 2.1. As such we formulate the final research question:

**Research Question 4:** *To what extent is it possible to make system health monitoring techniques efficient enough to be real-time in the context of LOFAR?*

We demonstrate that our system is real-time in the context of the LOFAR data processing pipeline, requiring less than 1ms to process a single spectrogram. This is demonstrated using our open-source hand-labelled datasets that have been published in conjunction with these results.

### 1.4 THESIS STRUCTURE

The thesis is organised into several chapters, each addressing a specific aspect of system health management for radio telescopes. Chapter 2 provides an in-depth overview of the LOFAR telescope, its architecture, and the literature relating to machine learning based anomaly detection. Chapter 3

explores how representation learning methods can be used to create a diagnostic framework of system health management in radio telescopes. In Chapter 4, we explore the application of nearest neighbour reconstructions within representation learning methods for anomaly detection. This is done using both general machine learning datasets as well as simulated radio telescope data. In Chapter 5 we investigate how the previously introduced anomaly detection methods can be incorporated into RFI detection schemes. In the penultimate chapter, we propose the radio observatory anomaly detector, a supervised and self-supervised hybrid model that is capable of detecting a variety of different anomalies in the LOFAR telescope. In Chapter 7, we provide the thesis conclusion, addressing each research question and suggesting directions for future research. Through this comprehensive examination of system health management for the LOFAR telescope, this thesis aims to contribute to the advancement of radio astronomy and facilitate the continued exploration of the universe using low-frequency observations. Although we use LOFAR as the instrument to validate the approach, the methodology we propose in this thesis is generic, and can be applied to other telescopes as well.

## 1.5 AUTHOR PUBLICATIONS RELATED TO THIS THESIS

### 1.5.1 Peer reviewed journal publications

- [J1] Michael Mesarcik, Albert-Jan Boonstra, Christiaan Meijer, Walter Jansen, Elena Ranguelova, and Rob V. van Nieuwpoort. “Deep learning assisted data inspection for radio astronomy.” In: *Monthly Notices of the Royal Astronomical Society* 496.2 (May 2020), pp. 1517–1529. ISSN: 0035-8711. DOI: [10.1093/mnras/staa1412](https://doi.org/10.1093/mnras/staa1412). eprint: <https://academic.oup.com/mnras/article-pdf/496/2/1517/33483423/staa1412.pdf>. URL: <https://doi.org/10.1093/mnras/staa1412>.
- [J2] Michael Mesarcik, Elena Ranguelova, Albert-Jan Boonstra, and Rob V. van Nieuwpoort. “Improving novelty detection using the reconstructions of nearest neighbours.” In: *Array* 14 (2022), p. 100182. ISSN: 2590-0056. DOI: <https://doi.org/10.1016/j.array.2022.100182>. URL: <https://www.sciencedirect.com/science/article/pii/S2590005622000388>.
- [J3] Michael Mesarcik, Albert-Jan Boonstra, Elena Ranguelova, and Rob V. van Nieuwpoort. “Learning to detect radio frequency interference in radio astronomy without seeing it.” In: *Monthly Notices of the Royal*

*Astronomical Society* 516.4 (Sept. 2022), pp. 5367–5378. ISSN: 0035-8711. DOI: [10.1093/mnras/stac2503](https://doi.org/10.1093/mnras/stac2503). eprint: <https://academic.oup.com/mnras/article-pdf/516/4/5367/48490907/stac2503.pdf>. URL: <https://doi.org/10.1093/mnras/stac2503>.

- [J4] Michael Mesarcik, Albert-Jan Boonstra, Marco Iacobelli, Cees de Laat, Elena Rangelova, and Rob V. van Nieuwpoort. “The ROAD to discovery: machine learning-driven anomaly detection in radio astronomy spectrograms.” In: *Astronomy & Astrophysics* 680 (Dec. 2023), A74. ISSN: 0004-6361. DOI: [10.1051/0004-6361/202347182](https://doi.org/10.1051/0004-6361/202347182). eprint: <https://www.aanda.org/articles/aa/pdf/2023/12/aa47182-23.pdf>.

#### 1.5.2 Peer reviewed conferences and workshops

- [C1] Michael Mesarcik, Boonstra, Elena Rangelova, and Rob V. van Nieuwpoort. “RFI Detection using Novelty in Autoencoders.” In: ECMWF-URSI. 2022.
- [C2] Michael Mesarcik, Albert-Jan Boonstra, Marco Iacobelli, Elena Rangelova, and Rob V. van Nieuwpoort. “The Radio Observatory Anomaly Detector.” In: URSI-GASS. 2023.

#### 1.5.3 Code

- [C1] Michael Mesarcik. *mesarcik/NLN: 0.0.1*. Version zenodo. July 2023. DOI: [10.5281/zenodo.8108231](https://doi.org/10.5281/zenodo.8108231). URL: <https://doi.org/10.5281/zenodo.8108231>.
- [C2] Michael Mesarcik. *mesarcik/ROAD: 0.0.1*. Version zenodo. July 2023. DOI: [10.5281/zenodo.8108241](https://doi.org/10.5281/zenodo.8108241). URL: <https://doi.org/10.5281/zenodo.8108241>.
- [C3] Michael Mesarcik. *mesarcik/RFI-NLN: 0.0.1*. Version zenodo. July 2023. DOI: [10.5281/zenodo.8108237](https://doi.org/10.5281/zenodo.8108237). URL: <https://doi.org/10.5281/zenodo.8108237>.
- [C4] Michael Mesarcik. *mesarcik/DL4DI: 0.0.1*. Version V1.0.0. Mar. 2020. DOI: [10.5281/zenodo.3697078](https://doi.org/10.5281/zenodo.3697078). URL: <https://doi.org/10.5281/zenodo.3697078>.

1.5.4 *Datasets*

- [D1] Michael Mesarcik, Albert-Jan Boonstra, Christiaan Meijer, Walter Jansen, Elena Rangelova, and Rob V. van Nieuwpoort. *LOFAR dataset for deep learning assisted data Inspection for radio astronomy*. Version V1.0.0. Mar. 2020. DOI: [10.5281/zenodo.3702430](https://doi.org/10.5281/zenodo.3702430). URL: <https://doi.org/10.5281/zenodo.3702430>.
- [D2] Michael Mesarcik, Albert-Jan Boonstra, Elena Rangelova, and Rob V. van Nieuwpoort. *Learning to detect radio frequency interference in radio astronomy without seeing it*. June 2022. DOI: [10.5281/zenodo.6724065](https://doi.org/10.5281/zenodo.6724065). URL: <https://doi.org/10.5281/zenodo.6724065>.
- [D3] Michael Mesarcik, Albert-Jan Boonstra, Marco Iacobelli, Cees de Laat, Elena Rangelova, and Rob V. van Nieuwpoort. *The radio observatory anomaly detector dataset*. June 2023. DOI: [10.5281/zenodo.8028045](https://doi.org/10.5281/zenodo.8028045). URL: <https://doi.org/10.5281/zenodo.8028045>.

# 2

## BACKGROUND

---

Radio telescopes are complex instruments. They comprise of a variety of hardware and software systems which require different processing pipelines to run an assortment of signal- and data-processing algorithms so to produce data for astronomers. The hardware systems are made up of receiver front ends, electronics, station processing units and the central processor, whereas the software systems run a multitude of different workflows to effectively process the digitised signals. The data is transported across the telescope subsystems by dedicated high-speed links over several thousands of kilometres. The distributed system aggregates the data in a central processing cluster, where many online and offline processing algorithms are run. As a result of the distributed nature of radio telescopes and the amount of data- and signal-processing algorithms required to produce a single radio image, system complexity, data rates and system uptime are a huge concern. Therefore, this thesis focuses on machine learning driven data analysis and anomaly detection algorithms for the purpose of maintaining system uptime in radio telescopes.

Recent works that apply machine learning-based anomaly detection and diagnosis to radio astronomy have so far focused on only scientific discovery, using galaxy images, transient signals or light curves. In this work we apply machine learning methods to visibilities obtained from radio telescopes.

This chapter elaborates further on the LOFAR telescope architecture and unpacks the current landscape of machine learning-based anomaly detection and the recent developments in applying it to astronomy-related fields.

### 2.1 THE LOW FREQUENCY ARRAY

The Low Frequency Array (LOFAR) [8] is a telescope comprising of 52 stations across Europe. Each LOFAR station is an array of 96 dual polarisation

low-band antennas (LBA) in the range 10–90 MHz and 48 or 96 dual polarisation high band antenna antennas (HBA) in the range 110–250 MHz. The signals received by each antenna are coherently added in the station-level processor, resulting in each sub-band (subdivision of frequency bands) being approximately 200 kHz wide. These signals are then transported to the central processor for further processing. An illustration of the system topology can be seen in Figure 2.1.

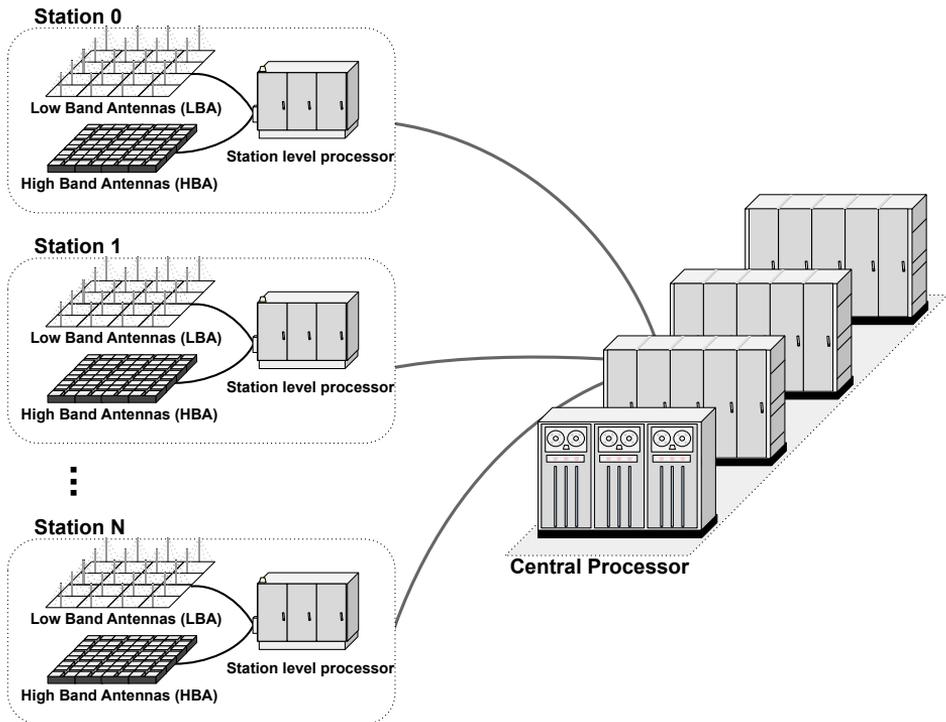


Figure 2.1: The LOFAR telescope topology and data processing chain

To obtain the radio-spectrum images and various other data formats used by astronomers, several signal- and data-processing algorithms need to be applied. The different data processing stages are illustrated in Figure 2.2. Analogue to digital conversion and filtering are first performed at the station-level processor. Then the digitised signal is *beamformed* to produce multiple independent observations of smaller regions of the sky. This signal is then transported over a dedicated network to the central processor, where several real-time signal and data processing algorithms are applied.

### 2.1.1 The LOFAR central data processor

Upon entering the central processor, delay compensation is applied to the signals. As radio waves travel at a finite speed, antennas located in different geographical locations do not receive the same signal coherently. Therefore, to correlate the same components of the received signals (which happens further down the processing chain), the signals from each station need to be compensated by the delay caused by difference in arrival time. This delay is related to the distance between the stations, as well as the direction in which they observe (and is complicated by the rotation of the earth) [48].

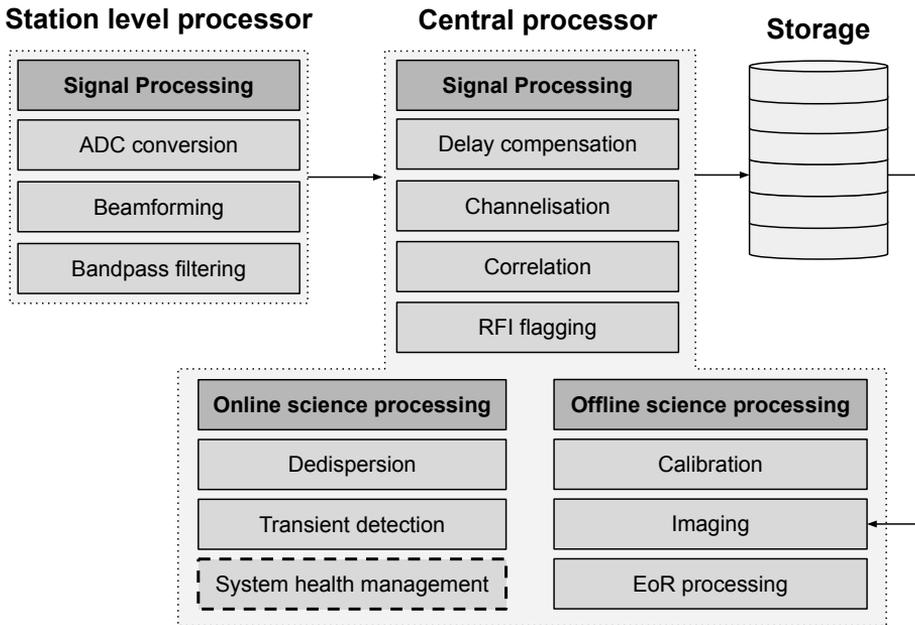


Figure 2.2: LOFAR data processing and signal processing chain

For purposes of downstream algorithms, the incoming data packets are grouped together as a set of sub-bands per station, and a poly-phase filter is applied to re-sample the data. The filter bank splits each subband into 256 frequency channels,  $N_f$ . At this point the correlator computes the cross- and auto-correlations between all pairs of stations for each channel and polarisation  $N_{\text{pol}}$ . With  $N_{\text{st}}$  telescope stations, there exist  $N_{\text{st}}$  autocorrelations and  $0.5N_{\text{st}}(N_{\text{st}} - 1)$  cross-correlations for each polarisation. The polarisations here refer to the Stokes parameters of the data given by  $(XX, XY, YX, YY)$ . This correlated data product is then integrated in time ( $N_t$ ) with the inte-

gration interval dependant on the science-case. This time-frequency data product is used for the remainder of the work and is referred to as a spectrogram. An illustration of a spectrogram from LOFAR can be seen in Figure 2.3.

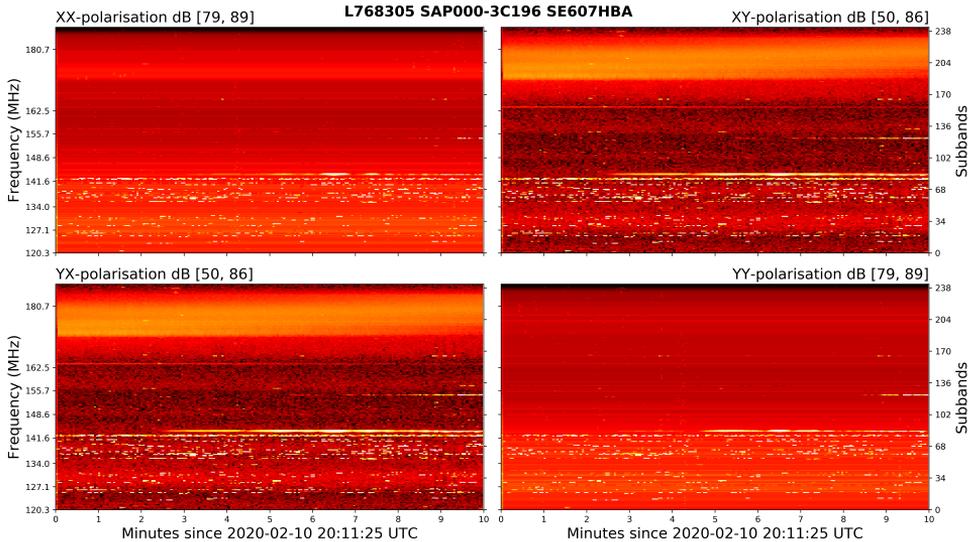


Figure 2.3: High-band autocorrelation spectrograms with four Stokes polarisations from a Swedish LOFAR station.

To obtain the data formats required by astronomers, several other processing steps need to be applied. These entirely depend on the application and dictate whether these processes need to be executed in real time or not. For example for imaging-based applications [49], multiple iterations of calibration and 2D Fourier transforms need to be applied to determine the best parameters for imaging of the celestial object [50]. This process is inherently offline and can take hours if not days to complete and validate. For other extreme use-cases, such as the Event Horizon Telescope (EHT) this imaging process took weeks complete [6]. Similarly, Epoch of Reionisation (EoR) studies integrate 100's of petabytes of data to increase the Signal-to-Noise-Ratio (SNR) thereby enforcing its offline nature [47].

The different science cases result in different observing setups, which dictate the array configuration (i.e., the number of stations used), the number of frequency channels, the time sampling as well as the overall integration time of the observing session. As already mentioned, the required resolution of modern instruments cause the data products to be relatively large. The data size of an observation is given by

$$N_{\text{obs}} = N_t N_f (N_{\text{st}} + \frac{1}{2} N_{\text{st}} (N_{\text{st}} - 1)) N_{\text{pol}} N_{\text{bitres}} \quad (1)$$

where  $N_{\text{bitres}}$  is the resolution of a sample in bits. This means that a 10 hour observation with a 1 second integration time, a 1KHz channel resolution with a 50MHz bandwidth and a 32-bit resolution of complex data can result in observations sizes in the order of 100s of terabytes.

### 2.1.2 *Online scientific data processing*

To cope with the increasing data-rates from modern scientific instruments [8, 47, 51, 52] real-time algorithms have been developed for scientific data pipelines. Real-time methods for RFI detection [44, 45, 53], calibration [54], Fast Radio Burst (FRB) detection [55] and correlation [46, 48] have been essential to modern radio telescope operations. This thesis aims offer an automated system health management workflow for the online science data processing system as shown by the dashed block in Figure 2.2.

LOFAR's correlator is IO-bound, as documented in Broekema et al. [13] and Romein et al. [48]. It handles an input bandwidth of 192Gbp/s (when considering all 64 antenna fields in parallel over 3Gbp/s links) while producing data at a bandwidth of 80Gbp/s. Where, antenna fields refer to core stations that are split, enabling two HBA groupings to be treated as smaller *stations*. Furthermore, the correlator outputs complex visibilities ranging from 32 to 4 bits in resolution, with up to 488 subbands approximately with a 1 s time resolution, encompassing both cross and autocorrelations. To ensure our system health management methodology is real time, it needs to guarantee an output bandwidth greater or equal to 80Gbp/s without significant buffering.

Very few machine learning techniques have been shown to be effective in real-time for astronomy. In seminal work by George and Huerta [56] machine learning gravitational wave detection algorithms were implemented in real time. Furthermore Muthukrishna et al. [57] shows that Temporal Convolutional Networks (TCNs) can be implemented to detect transient anomalies in real time. To demonstrate the effectiveness of our work in the context of radio observatories we investigate the computation performance and throughput of the proposed system.

### 2.1.3 *System monitoring and control*

The Monitoring and Control (MAC) system controls all operations and reporting of the LOFAR telescope. It consists of the WinCC-OA [58], which is a Supervisory Control and Data Acquisition (SCADA) system that distributes and designates the appropriate signals across the LOFAR stations and central processor to control the operations during an observation. Using the WinCC-OA system collection and monitoring of system states and health is performed. This is done in a hierarchical manner such that multiple resolutions and categorisations of operations can be reported to the telescope operators [59].

As different observations require different hardware and software workflow setups, the MAC system orchestrates the telescope configuration. This is essential as observation setups can be quite dynamic. Further, the MAC system offers a Graphical User Interface (GUI) to telescope operators that enables hierarchical inspection of both the data produced during an observation as well as diagnostic tools. For example, an operator can select specific stations and analyse the details from the station processor temperature and set particular alert levels during an observation [59].

In Chapter 3 we design and implement a machine learning based data inspection tool. It enables analysis of observation parameters in a single low dimensional prescriptive space.

## 2.2 MACHINE LEARNING IN ASTRONOMY

Machine learning has successfully been applied to astronomy in several areas ranging from galaxy morphology classification [60–63], exoplanet discovery [64, 65], detection of new and unusual celestial objects [66, 67], automatic telescope calibration [68], Radio Frequency Interference (RFI) mitigation [69–71] among many more. Machine learning has found particular successes in astronomy as the amount of data generated by modern instruments is large and seemingly ever increasing.

However, one application domain in astronomy where machine learning has been limited is in automated system health management for radio telescopes. For this reason, this thesis focuses on applying weakly-supervised methods to SHM in radio telescopes. We focus on three sub-fields; (i) representation learning for radio astronomy spectrograms, (ii) machine learning-based anomaly detection and (iii) machine learning data inspection for telescope operators.

### 2.2.1 *Representation learning in astronomy*

Representation learning refers to the process of modelling features of data in a meaningful and compact way. It involves extracting high-level structures that capture the underlying patterns and relationships within the data, enabling downstream tasks such as classification, clustering and anomaly detection. Learning informative representations helps to reduce dimensionality, enhance data understanding, improve generalisation, and facilitates more effective and efficient data analysis and decision-making processes in various domains, including computer vision, natural language processing, and signal processing [72].

Learning representations of high-dimensional data is essential to anomaly detection. For this reason, among many others, effort has been made to find methods that learn robust projections of high-dimensional data [73–76]. These successes have materialised in the astronomical community with results mostly in the galaxy classification domain. Walmsley et al. [77] show that by pretraining on the Galaxy Zoo DECaLS [78] dramatically improves model performance for several downstream tasks. Furthermore, Hayat et al. [79] show how contrastive learning can be applied to galaxy photometry from the Sloan Digital Sky Survey (SDSS) [80]. The authors show that with novel data augmentations, they can achieve state of the art results on several downstream tasks. Furthermore, several additions and modifications have been made to the reconstruction error-based loss functions of autoencoders. In work by Mesarcik et al. [81] it is shown that using both magnitude and phase information in VAEs improves performance of finding representations of astronomical data, whereas Villar et al. [82] use a recurrent adaption of a VAE to make training more suitable to light-curve data. Similarly, Melchior et al. [83] shows how the inclusion of self-attention mechanisms and redshift-priors into the latent projection of autoencoders, can improve the learnt representations of galaxy spectra.

### 2.2.2 *Machine learning-based anomaly detection*

Machine learning-based anomaly detection relies on modelling normal data and then classifying abnormality by using a discriminative distance measure between the normal training data and anomalous samples [14]. Autoencoding models are a popular approach for learning latent distributions of normal data [18, 20, 23, 84]. Anomaly detection using autoencoders can be performed either in the latent space using techniques such as

One-Class Support Vector Machines (OC-SVM) [85], K-Nearest-Neighbours (KNN) [30], Isolation Forest (IF) [31] or using the reconstruction error [86]. The use of pretrained networks to obtain latent representations of normal data have also been successful in anomaly detection [30, 87, 88]. By first training these models on an objective such as ImageNet classification [89], they are able to generalise to other tasks such as anomaly detection. Additionally, Self-Supervised Learning (SSL) has been shown to be invaluable for finding meaning representations of normal data [90–92]. Here pretext tasks allow the model to learn useful representations of training data in a self-supervised-manner that can then be used for other downstream tasks. In both the SSL and pretrained cases, KNN-based measures can be used to distinguish anomalous samples from the normal training data [30, 90].

In most machine learning-based anomaly detection performance is evaluated according to the Single-Inlier-Multiple-Outlier (SIMO) or Multiple-Inlier-Single-Outlier (MISO) [93] settings on natural image datasets such as MVTecAD-[23]. With this paradigm in mind, we find that anomaly detection in the system health management in radio telescope context is a Multiple-Inlier-Multiple-Outlier (MIMO) problem. In effect, anomaly detection formulations that make a strong assumption about the number of inliers or outliers are not directly applicable to the radio observatory setting due to the increased problem complexity. Furthermore, we find methods that rely on pretraining with natural images to be ill-suited to the spectrograms used in this work, due to differences in dynamic range and SNR as shown by [94].

Efforts have been made for detecting anomalies in light-curves and spectra in works such as Astronomy [95] and transients in Malanchev et al. [96]. Astronomy is an active learning framework developed for the classification of unusual events in imaged data or light curves at observatories to aid with scientific discovery. Nevertheless, it follows closely with generic anomaly detection methods, where normal data is first projected to a latent representation and metrics such as IF are used to distinguish normal training samples from anomalous testing samples at inference time. Although Astronomy assumes a MIMO context, it is still only able to detect unknown anomalies (or at least says all anomalies belong the same class). This is in contrast with our work described in Chapter 6, where we can both distinguish between all known anomaly classes with a high precision and detect unknown or rare anomalies.

Deep generative neural networks are also used for anomaly detection. Works by Mesarcik et al. [81], Villar et al. [82], and Ma et al. [97] have

shown that the Variational Autoencoders (VAEs) can be used for anomaly detection with astronomical data. Whereas Margalef-Bentabol et al. [98] and Storey-Fisher et al. [99] show that Generative Adversarial Networks (GANs) are effective in learning representations of normal images of galaxies thereby enabling reconstruction-error based anomaly detection. In work by Zhang et al. [100], GANs have also been shown to be effective in the Search for Extraterrestrial Intelligence (SETI) anomaly detection context. However due to problems in training stability we do not make use of such architectures.

### 2.2.3 *Radio frequency interference detection in the deep learning era*

Machine learning for RFI detection is an actively researched field. Numerous works offer a variety of radio astronomy-specific modifications to improve accuracy of detection. Semantic segmentation is at the heart of the deep learning-based RFI detection, with U-Net [101] and derivatives acting as the architectural backbone of recent research. The purpose of semantic segmentation is to determine the pixel-precise regions where a specific class exists – in this case RFI. Architecturally, U-Net is a Convolutional Neural Network (CNN), with an encoder-decoder pair that share activations between the two stages. It is trained in a supervised manner, requiring pixel-level Boolean masks per spectrogram.

The first application of U-Net to radio astronomy-based RFI detection is reported in seminal work by Akeret et al. [69]. The network is trained and evaluated on the magnitude of spectrograms obtained from both simulated data and real data from a signal antenna from the Bleien Observatory [102]. Interestingly, the models are trained using masks obtained from a classical flagging approach. We show later in Chapter 5 that this is not suitable, as supervised methods tend to over-fit to the weak-label based ground-truth. Additionally, work by Akeret et al. [69] makes use of the HIDE & SEEK radio astronomical data simulator [103]. We find the use of this simulator problematic because the ground truth needs to be determined by user-defined thresholds of the residual RFI maps as described by Sadr et al. [70]. Due to this, other works such as [71] do not describe the threshold used for evaluation making comparison extremely difficult.

To counter-act the issue of over-fitting to the potentially incorrect labels we focus on the inverse problem. We train a model to represent all non-RFI signals, such that any deviation from the learnt representations is flagged as RFI. It must be noted that this approach depends on the assumption that

the number of false positives of classical RFI detectors is higher than the number of false negatives. In other words, most RFI is flagged as RFI, but other features may be incorrectly flagged.

Generative models have also been used for RFI detection in the context of radio astronomy. Work by Vinsen, Foster, and Dodson [104] shows that Generative Adversarial Networks (GANs) can be used for RFI detection. However this research is limited in its evaluation and does not offer a practical way to obtain pixel-precise predictions of RFI. Finally, Vos et al. [105] offer a significantly different paradigm for RFI detection using GANs. Here, the authors propose a source-separation approach that uses 2 separate generators to distinguish astronomical signals from RFI. However, this method requires significant supervision, as the model needs access to the mixture as well as the separated RFI and astronomical sources during training. We find this requirement prohibitive, as to obtain these source separations for real data is extremely costly. In Chapter 5, we show that generative models can be used without the cost of supervision, by treating novelty detection as a downstream task.

# 3

## LEARNING REPRESENTATIONS OF RADIO ASTRONOMY SPECTROGRAMS

---

*This chapter focuses on how representation learning methods can be used to create low dimensional projections of radio astronomy spectrograms. In doing so, we provide an operator-in the loop tool for data inspection, for managing data quality in radio telescopes. We propose an autoencoding architecture that separately processes both the magnitude and phase components of the radio astronomy spectrograms. In this chapter we also introduce the feature compounding problem in radio astronomy based spectrograms as well as a simple method to evaluate performance using synthesised data. Overall this chapter aims to answer research questions 1 and 2.*

This Chapter is based on:

- **Michael Mesarcik**, Albert-Jan Boonstra, Christiaan Meijer, Walter Jansen, Elena Ranguelova, and Rob V. van Nieuwpoort, "Deep learning assisted data inspection for radio astronomy" [106], in *In: Monthly Notices of the Royal Astronomical Society* 496.2 (May 2020)
- **Michael Mesarcik**, Albert-Jan Boonstra, Christiaan Meijer, Walter Jansen, Elena Ranguelova, and Rob V. van Nieuwpoort, "LOFAR dataset for deep learning assisted data Inspection for radio astronomy" [107], in *Dataset on Zenodo*

### 3.1 INTRODUCTION

Modern radio telescopes generate an ever growing amount of data. To improve spatial resolution, sensitivity, and field of view, larger telescope arrays are being constructed [8, 51, 108]. The increased size and capabilities

of the instruments lead to more data, a higher system complexity, and ever-growing error rates due to malfunction.

System Health Management (SHM) is the process of detecting, diagnosing and remedying system failures to maximise system uptime. SHM and pinpointing error sources in radio telescopes today still rely on manual inspection and human interpretation of the data. This is error-prone, and large-scale spatially distributed radio telescopes such as LOFAR [8] face reliability and up-time issues. These issues stem from the scale and complexity of the systems and processing chains involved [13]. In LOFAR for example, we expect that, due its scale and its somewhat harsh operating conditions, at any given time several components in the systems will not operate correctly. The LOFAR stations in the Netherlands are exposed to high moisture levels which may result in failure of components such as antennas and amplifiers [8]. Other reasons for failures include normal wear of components, network packets being dropped, and hardware and software errors among many others.

For the next generation exascale SKA radio telescope [9] this will be an even bigger issue. The complexity of the instruments, the myriad of observational modes, and the scale of the data transport and compute platform [11] make accurate error detection [109] and complete fault localisation very difficult. Therefore, intelligent automated SHM approaches would significantly improve the quality and availability of the observational systems. This is not only beneficial for (predictive) maintenance, operations, and cost, but it is also crucial for the science results, as accurate knowledge of the state of the telescope is essential for calibrating the system [40].

The variability and amount of features found in data obtained from radio telescopes makes applying classic signal processing techniques for SHM difficult, as these techniques depend on specific feature morphologies [110, 111]. Therefore, the scale, complexity and variability of features have made machine learning approaches candidate solutions to this problem.

This chapter discusses the data formats, preprocessing as well as system design required to learn representation of radio astronomical data. In Section 3.2 we discuss and analyse the preprocessing techniques that we apply to the spectrograms obtained from LOFAR and the data obtained from the HERA simulator. Section 3.3 documents the deep learning architecture used, while Section 3.4 establishes performance evaluation metrics. Finally, in Section 3.5 we present results and in Section 3.6 the conclusions are given.

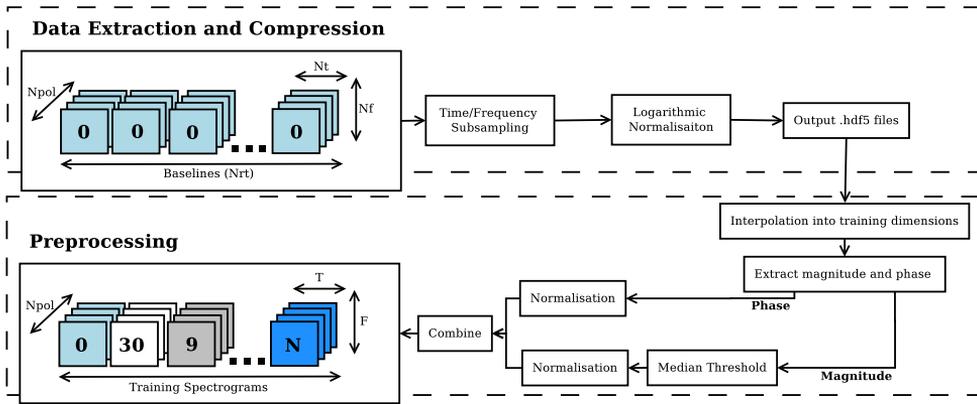


Figure 3.1: Block diagram showing the data extraction and preprocessing pipelines applied to the observations before training.

### 3.2 DATA PREPARATION

In this section, we discuss and analyse the preprocessing techniques that we apply to the spectrograms obtained from LOFAR. This step is often under-emphasised or even omitted in papers describing the application of machine learning to radio astronomy based applications. This is unfortunate, since we found this is far from trivial, as it has significant impact on the quality of the end results and is important for the reproducibility of the research.

Preprocessing is particularly challenging for radio astronomy compared to other fields. Examples of some of the preprocessing challenges faced in the field are the high dimensional data cubes; the complications of having amplitude and phase information, which must be treated separately; and the extremely noisy data with a very high dynamic range due to interference.

As described in Chapter 2.1 the LOFAR telescope produces large amounts of data that can be processed in a variety of different ways. As the amount of data produced from a LOFAR observation is rather high, data inspection is usually performed using a lower spectral resolution, and/or time averaging, and/or sub-sampling. The so-called *inspection plots* are based on either low spectral resolution auto-correlation spectra created at the stations, or on compressed spectrograms. For each LOFAR aperture synthesis observation, an additional compressed .hdf5 dataset is produced in parallel to the full measurement set. The compression parameters are tunable, but a 100 TB observation is typically reduced to the order of gigabytes.

The data compression stages can be seen in the top half of Figure 3.1. First, the the dataset is sub-sampled in time, using a regular grid. The next step is aimed at reducing the dynamic range of the data so that it fits in fewer bits. This includes taking the logarithm of the absolute value of visibilities, and normalisation by a scaling factor computed as the maximum value for each baseline, sub-band, and polarisation. The normalisation factor is stored in the .hdf5 file so that the original data can be reconstructed albeit with reduced resolution. The compressed data are stored as 8 bit integers for the real and imaginary parts of the complex visibilities.

### 3.2.1 Data-set preprocessing

The main considerations for the preprocessing of the time-frequency spectrograms are the data type, the normalisation method, the correct scaling of features and size of each spectrogram. The preprocessing scheme used in this work can be seen in the bottom section of Figure 3.1.

As shown in work by Kerrigan et al. [112], the use of both amplitude and phase components of the complex visibilities yield an increase in performance in the classification of RFI. For this reason, we deem it critical to make use of both components of the complex visibilities. The model evaluation shown in Section 3.4 demonstrates the differences in performance when changing the domain of the training data.

Due to the GPU memory constraints for training and inference, all complex spectrograms need to be constrained in size. As the extracted .hdf5 could consist of different observation durations each with a different number of sub-bands, it was necessary to resize the visibility matrices. The resizing was performed by down-sampling all dynamic spectra greater than  $128 \times 32$  and interpolating all spectrograms smaller than  $128 \times 32$  in frequency and time.

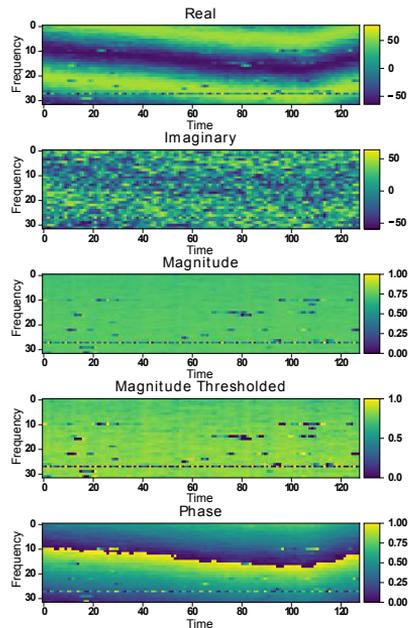


Figure 3.2: The processing stages applied to a randomly selected baseline from the LOFAR training set.

To increase the dynamic range of the astronomical features present in the magnitude component of the complex visibilities, we perform naive radio frequency interference suppression. We achieve this through the use of a median based thresholding mechanism which we only apply to the magnitude component of the spectrograms. We do not apply the scheme to the phase components of the visibilities, as they are bounded between  $-\pi \leq \phi \leq \pi$ .

The median-threshold was calculated on a per-spectrogram basis. In effect every different baseline had a unique threshold defined by its own power spectrum such that

$$|V'_{i,j}(\tau, \nu)| = \begin{cases} \sigma, & \text{if } |V_{i,j}(\tau, \nu)| \geq \sigma \\ |V_{i,j}(\tau, \nu)|, & \text{otherwise} \end{cases} \quad (2)$$

where  $\sigma$  is the first standard deviation of the visibility of antennas  $i$  and  $j$  given by  $V_{i,j}(\tau, \nu)$  where  $\tau$  and  $\nu$  correspond to the time and frequency dimensions respectively.

As a consequence of the magnitude being unbounded, we normalised the magnitude between 0 and 1 on a per baseline basis after the naive RFI suppression. Additionally, to ensure that the phase component was weighted equally in training, we normalise the phase component to the same scale as the magnitude. An illustration of the intermediate outputs of the preprocessing pipeline can be seen in Figure 3.2.

### 3.3 VARIATIONAL AUTOENCODER ARCHITECTURE

In this section we describe the design and constraints of the proposed Variational Autoencoder (VAE) based architecture, as shown by Figure 3.3.

An autoencoder is a machine learning model used to generate compressed representations of data in an unsupervised manner. The model can be considered in two parts, the encoder network,  $h = f(X)$ , that generates an intermediate representation of the input data,  $z$ , and a decoder network,  $r = g(z)$ , that regenerates the input data from the encoded representation.

VAEs are a generative branch of autoencoding models. Unlike traditional autoencoders that use purely reconstruction-based loss functions, variational autoencoders are probabilistic models. This means that they consider the input, output and latent representations as distributions rather than discrete values. This is achieved by minimising the KL-divergence between the input distribution,  $q(X)$ , and the output distribution,  $p(X)$ , given an intermediate representation  $z$ , as given by

$$\mathcal{D}_{\text{KL}}(q(z|X)||p(z|X)) = \sum_z q(z|x) \log \left( \frac{q(z|X)}{p(z|X)} \right). \quad (3)$$

As the posterior distribution is not tractable, the Evidence Lower Bound (ELBO) [113] of  $\log(P(X))$  is maximised as an approximation to the distribution, given by

$$\mathbf{max} \log(P(X)) \geq \mathbb{E}_{q(z|X)} [\log(p(X|z))] - \mathcal{D}_{\text{KL}}(q(z|X)||p(z)). \quad (4)$$

When we parameterise  $p(z)$  by  $\theta$  and  $q(z|X)$  by  $\phi$  the loss function of the VAE can be stated as

$$\begin{aligned} \mathcal{L}_{\text{ELBO}}(\theta, \phi, X) = & -\mathcal{D}_{\text{KL}}(q_{\phi}(z|X)||p_{\theta}(z)) \\ & + \mathbb{E}_{q_{\phi}(z|X)} [\log(p_{\theta}(X|z))]. \end{aligned} \quad (5)$$

In the case of the autoencoder-based structure, the  $q_{\phi}(z|X)$  term may be considered as the encoder, mapping the input distribution  $q_{\phi}(X)$  to the latent projection given  $p_{\theta}(z)$ , whereas the  $p_{\theta}(X|z)$  term may be considered the decoder, mapping the output distribution  $p_{\theta}(X)$  given  $z$  [114].

Typically, the prior distribution,  $p_{\theta}(z)$ , is obtained by sampling it from a Gaussian. However, as the sampling operation is not differentiable, a *reparameterisation trick* is used. The reparameterisation removes the non-differentiable sampling operations from the network and replaces them with differentiable operations.

The encoder and decoders are convolutional neural networks (CNNs). Structurally each convolutional layer in the encoder is followed by a max-pooling layer and batch normalisation, this is done to minimise the receptive field of the each subsequent convolution. The decoder uses an upsampling layer after each convolutional layer to restore the correct dimensionality of the output. In the case of the architecture used, the number of filters was decreased by a factor of 2 for each sequential convolutional layer in the encoder, and is up-sampled by a factor of 2 in each consecutive convolutional layer in the decoder.

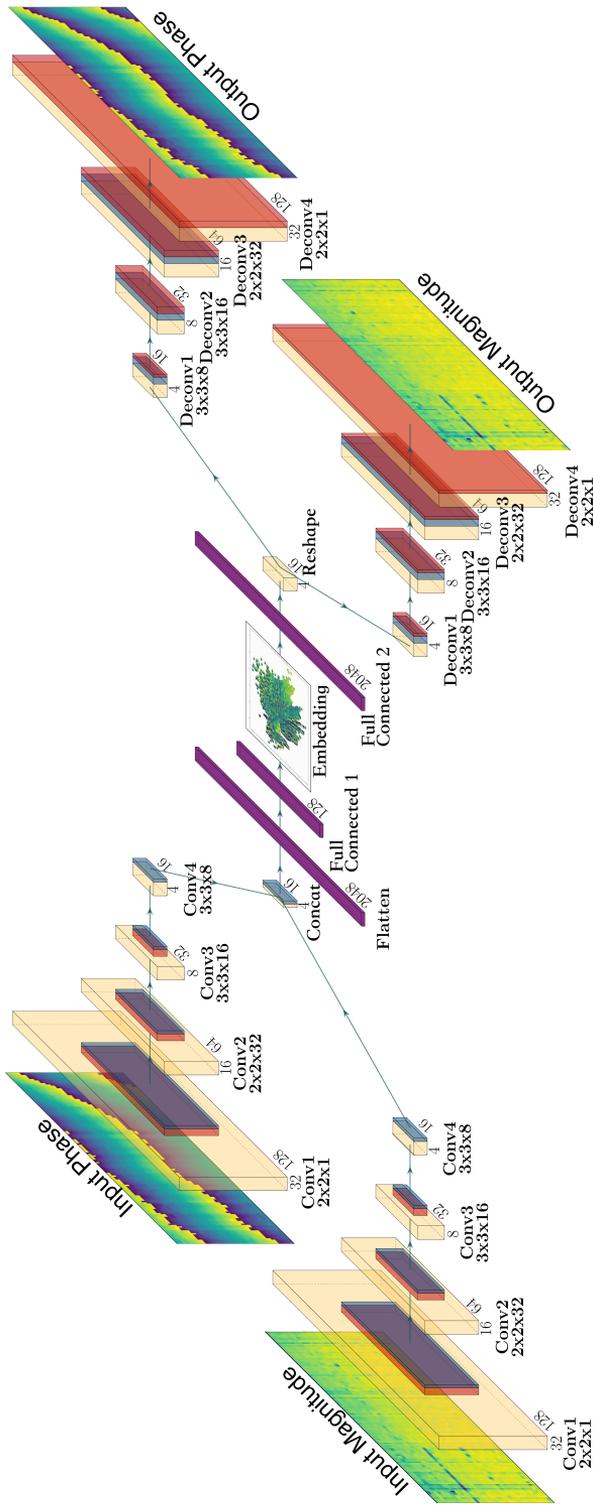


Figure 3-3: The architecture of the magnitude and phase-based VAE used for the projection of the astronomical spectrograms. The yellow layers correspond to convolutional layers, the green and blue layers correspond to max pooling and upsampling layers respectively, whereas the mauve layers correspond to batch-normalisation layers. For the intermediate representation, purple is used to designate the flattening operations, fully-connected layers as well as the latent representation.

### 3.3.1 *Architectural motivation and constraints*

The major consideration that was made regarding the VAE architecture was the structural separation between the magnitude and phase components of the complex visibilities. We deem it critical to make use of both components as they contribute differing representations of particular astronomical and terrestrial phenomena. In section 3.5.1, we discuss a performance comparison of network architectures that use different combinations of real, imaginary, magnitude and phase components.

A result of using the magnitude and phase components for training the VAE-based model is that the learnt representations from both the phase and magnitude encoders need to be joined together. As features appear differently in each domain, we determined that normalisation after each convolutional layer was necessary. The normalisation ensured that the independent magnitude and phase activations are maintained between 0 and 1, so that no higher magnitude activations, such as RFI, may take precedence over the learnt representations. They can be seen in mauve in Figure 3.3. Similarly to Kerrigan et al. [112], we found that the concatenation of the learnt normalised magnitude and phase activations yields increased performance. In the case of the VAE architecture, we concatenate the magnitude and phase components after the last convolutional layer of the encoder such that their joint-embedding could be determined.

As the objective of this work is a data inspection and visualisation tool, it was necessary to reduce the dimensions of the input spectrograms to 2 dimensions. This is critical, as visualisation and navigation of a two dimensional space is intuitive for the end users of this system at observatories. The impact of this is that the 2-dimensional embedding of the 128x32 sized spectrograms results in a 97% reduction in dimensionality which limits the VAE's reconstruction abilities. In other literature that use autoencoders for dimensionality reduction [115, 116] and representation learning [62], higher dimensional latent spaces are often used to encode high-order features such as affine transformations. However, in our work, the latent representation is limited to 2-dimensions. The effect of this is that some higher-order features are lost and cannot be regenerated by the decoder network.

A benefit of the architecture we use is the ability of the VAE to generate new samples from the learnt distribution of the complex radio astronomical spectrograms. The ability of VAEs to generate new labelled data is described in more detail in Pu et al. [117]. Although the generative aspect of these networks are not used in this work, it is an interesting considera-

tion as this model can be used to generate new labelled training data given that the latent projection can be labelled. We intend to investigate if we can exploit this feature to automatically classify anomalies in future work as described in Akcay, Atapour-Abarghouei, and Breckon [16].

For training we used the Adam optimiser with a learning rate of  $1e-4$  when used cross entropy for reconstruction term of the VAE loss. Furthermore we experimentally selected a batch size of 256 and 200 epochs for training.

### 3.4 MODEL EVALUATION

In representation learning for radio astronomy, challenges arise from the combinatorial effects from multiple features possibly being present in a spectrogram. For example, a given spectrogram from LOFAR may contain several features such as narrow band RFI, point sources, solar storms, Casiopeia in the antenna side-lobes and many more in a particular spectrogram. This means that for a single observation the total number of combinations of features is given by

$$n_c = \sum_{k=1}^n \frac{n!}{(n-k)!k!} + 1 \quad (6)$$

where  $n$  is the total number of possible features, and  $n_c$  designates the number of different clusters, each with a particular combination of astronomical features. In the case of 10 features, the number of combinations of features will yield 1024 different classes, given that the  $\emptyset$ -set is included.

To create a controlled environment to evaluate the projection, we use a radio astronomy simulator so that the number of features present in a particular spectrogram could be designed in a predictable manner. Additionally, the simulator allows feature labels to be generated with each spectrogram, such that the accuracy of the separation of the features can be measured, allowing a quantitative analysis. We later perform a manual qualitative analysis with real LOFAR data.

#### 3.4.1 Quantitative evaluation using simulated data

We use the HERA [118] radio astronomical data simulator for quantitatively evaluating the magnitude and phase-based VAE, as it is capable of generating various astronomical and terrestrial events. Features such as point sources, pseudo-sky models and various forms of RFI can be generated.

Through the use of the simulator, we generated 20,000 training spectrograms with multiple baselines of a *pseudo-sky model*. We added random combinations of different features to this. These features were: point sources, narrow and broadband RFI, gain fluctuations and antenna cross talk. The features and their parameters that we used can be seen in Table 3.1 and the exact model specifications of these features are described in work by Kerri-gan et al. [112].

The data obtained from the HERA simulator has far fewer features and significantly smaller variability of parameters of those features compared to the real-world LOFAR data. Therefore, the results from the simulated data do not guarantee performance in geometric-separability when applying the same model to the LOFAR data. However, by evaluating the model performance both prescriptively on the HERA data and descriptively on the LOFAR data, it is possible to ensure a degree of confidence regarding the obtained results. In addition to this, the use of both synthesised HERA data and real-world LOFAR data prove the generalisability of the model for different astronomical instruments and data sets.

We use a linear Support Vector Machine (SVM) to measure the separability of features in the latent space of the model. Linear SVMs are a supervised machine learning technique that classifies labelled data by segmenting the N-dimensional mappings using hyper-planes. The classification output generated from the SVM is used as an evaluation metric of the VAE's representation learning performance, because the classification accuracy increases proportionally to the euclidean separation of spectrograms with the same multi-feature labels in two dimensional space [119].

Since the focus of this work is on visualisation of high dimensional spectrograms, and not on classification, we chose a naïve classifier. Naturally, the choice of a more sophisticated classification technique would yield improved performance in many cases. However, by illustrating the effectiveness of the naïve classifier, we prove that the VAE is clearly capable of separation of features in the low dimensional embedding.

#### 3.4.2 *Qualitative evaluation using LOFAR data*

To demonstrate the real-world efficacy of the model, we evaluate the learnt representations descriptively using data obtained from the LOFAR telescope. A prescriptive evaluation was not possible on the LOFAR data projection as there exist a multitude of features that can be simultaneously present in a given spectrogram. Moreover, these features can vary signif-

Table 3.1: Parameters of the HERA simulator used to synthesise training data.

Feature Type	Value
Baseline geometric delays	5 ns, 10 ns, 20 ns, 100 ns, 500 ns
Number of frequency channels	64
Number of time samples	128
Number of uncorrelated sources	200
Synthesised noise type	Gaussian

icantly across different baselines. The variability of features is a result of the geometric localisation of the terrestrial or astrophysical phenomena to a particular station. The effect of this is that localised sources will appear significantly different in geographically distant stations. Consequentially, different labels would have to be given to the same feature for different baselines which presents a problem when trying to evaluate separability of features in a prescriptive evaluation of results. Additionally, the effort of manually labelling enough data with many combinations of features is prohibitive. These limitations are addressed in later chapters of this thesis, where expertly labelled data sets have been produced for both anomaly detection and radio frequency interference mitigation.

For this reason and the compounding of features shown in Equation 6, traditional clustering evaluation methods such as accuracy and normalised mutual information (NMI) [114] are not suitable metrics in this context. We aim to perform further analysis regarding anomaly and outlier detection in future work.

We evaluate the LOFAR results through visual inspection of the embedding plots, scatter plots with the magnitude spectrograms superimposed on the points, as well as considering the difference in the generative abilities of the input and output spectrograms.

To evaluate the LOFAR-based results, we trained the VAE-based model using 327 unique observations, each consisting of between 600 and 3000 baselines from both the LBA and HBA stations. From these observations, 256 baselines were randomly sampled from each file. We preprocessed the data using the preprocessing pipeline described in Section 3.2. The result of this was that we trained the model using 60672 32x128-sized complex spectrograms.

Finally, we integrated the embedding plots into a data inspection environment used by the data commissioning astronomers at the ASTRON ob-

servatory <sup>1</sup>. With the use of experts in the loop, the system was iteratively evaluated whilst in the development phases.

### 3.5 RESULTS

In this section, the effectiveness of the designed variational autoencoder is shown. This is done by evaluating the model's ability to give meaningful low dimensional projections for telescope operators diagnosing system health. In all cases the test-train split ratio of the training data was 80:20. For the quantitative evaluation of the HERA data, the model's classification accuracy is used as the primary performance metric unless otherwise specified. The LOFAR-based training data can be found online <sup>2</sup>.

#### 3.5.1 *Simulated data*

To evaluate how best to represent the input data to the VAE-based model we consider six different formats. Here we train a VAE on the real component only, the imaginary component only, both the real and imaginary components, on the magnitude only, the phase only and on both the magnitude and phase components of the complex data. From this, we evaluated the classification performance for each of the different architectures.

It is clear from the results shown in Figure 3.4 that using both the magnitude and phase components of the complex visibilities yields the best classification accuracy. Albeit a small performance increase, in comparison to using the magnitude component only, it is comparable to those results reported by [112]. Furthermore, even though the real and imaginary representation contains the same information as the magnitude and phase representation, using the latter shows a significant performance gain. This can be explained by the complex time-frequency data being represented in a more interpretable manner for the convolutional layers of the VAE. For this reason, the magnitude and phase-based network architecture is used from this point onward in the experiments of this chapter.

To draw performance comparisons between the models trained on HERA and LOFAR data, we ran experiments to determine the HERA-trained model's classification accuracy as the number of features increases. We did this by varying the number of features,  $n$ , in a given HERA data training set from 3 to 7 and calculating the classification accuracy of the model. In Figure 3.5

---

<sup>1</sup> <https://www.astron.nl/>

<sup>2</sup> <https://zenodo.org/record/3702430>

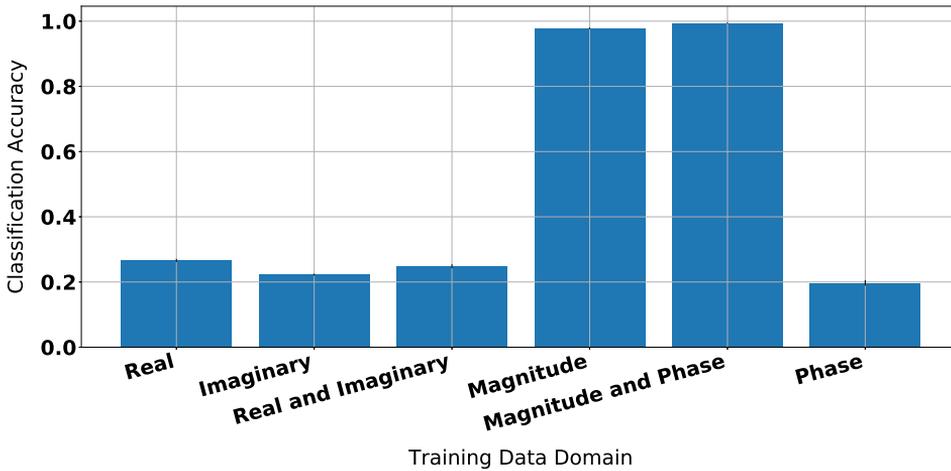


Figure 3.4: Classification accuracy of the SVM classifier applied to the learnt representations when the VAE was trained on data with 5 different features presented in different domains.

we can see that as the number features increases, the classification accuracy of the network decreases significantly.

There is a significant classification performance loss from 5 features onward, however the use of only classification accuracy as an evaluation method does not perfectly capture the performance of the model. The linear SVM-based classifier's performance requires the embedded spectrograms to be linearly separable in their latent projection, however it shows that as the number of features increases the linear separability decreases. This is due to the fact that pronounced features such as persistent narrow band RFI tend to take precedence over lower power features such as astronomical sources such that when compounded, features consisting of a weak and strong feature are localised to the same embedding location.

This being said, when visually inspecting the embeddings obtained from models trained on data with a high number of features it is clear that certain features are well separated. Figure 3.6 (a) shows the 2 dimensional embedding of the input spectrograms with six features, namely narrow-band radio station-based and digital television broadcasting signals (RFI), broadband impulsive RFI, gain fluctuations, source structure and Gaussian noise. In this experiment the labels associated with each compounded feature are coloured differently. In Figure 3.6 (b) the same embedding is shown, except rather than colouring each point by its respective label, the magnitude component of the input spectrogram is overlaid onto each of the embed-

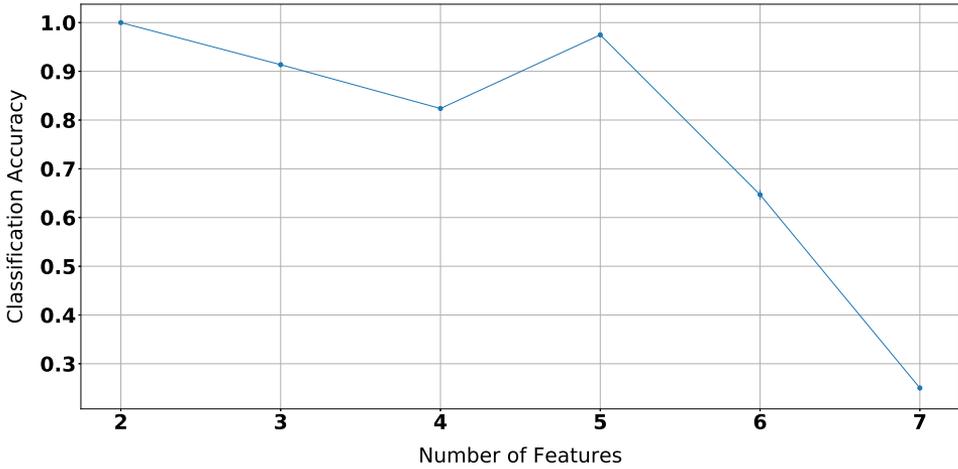


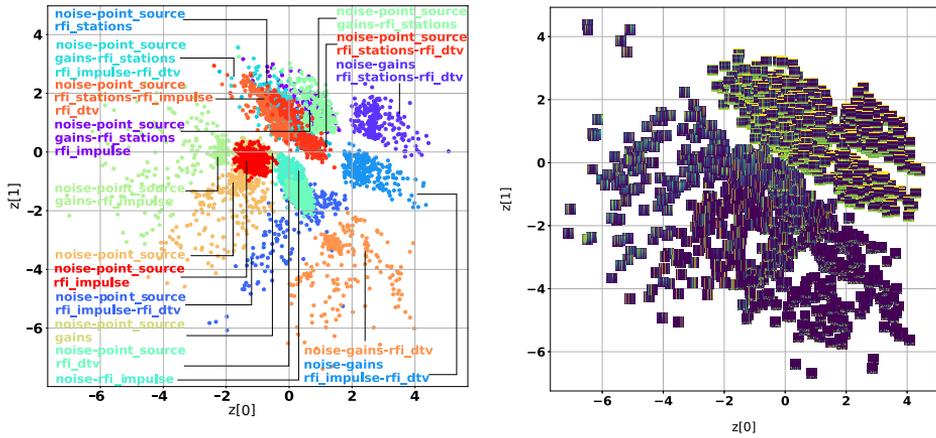
Figure 3.5: Classification accuracy of the SVM classifier applied to the learnt representations while varying the number of features in the HERA simulator.

ding points associated with that input. It can be seen that the features of associated with each input are clearly visually separated, additionally the segmentation of the label-coloured inputs shown in Figure 3.6 (a) is clearly reflected in this plot.

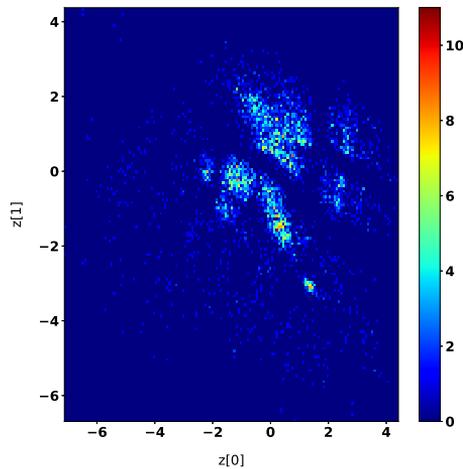
Notably, there is a clear separation in the features that contain narrow-band radio station-based RFI, shown in the top-right half, and those that do not, shown in the bottom half. Furthermore, within the station based-RFI contaminated region, there is a clear separation between the spectrograms that contain source structure, in the top left-most region, and those that do not, shown in the top right-most region. Similar segmentations can be seen in the bottom-most region that do not contain radio station-based RFI, in that there is a clear separation between those spectrogram that contain sources and those that do not.

Furthermore, the 2D histogram shown in Figure 3.6 (c) reflects how the input spectrograms are evenly distributed within the two dimensional embedding. It can be seen that there is a uniform distribution of data within the 2 dimensional space, with no feature occupying a single point on the grid.

Although the designed visualisation system is constrained to two dimensions, it is useful to consider the model's classification performance (or higher dimensional geometric separability) for sake of explainability and generalisability of the system. For this reason experiments were run to measure the SVM classification accuracy when the dimensionality of latent pro-



- (a) The embedding of the HERA-trained VAE with the colours corresponding to the compounded labels for each feature.
- (b) The embedding of the HERA-trained VAE with the magnitude of the input spectrograms superimposed onto each of the corresponding points of the embedding.



- (c) 2D histogram showing the distribution of data within in 2D latent embedding.

Figure 3.6: The hierarchical feature separation for the magnitude and phase-based VAE trained on simulated HERA data with 6 features.

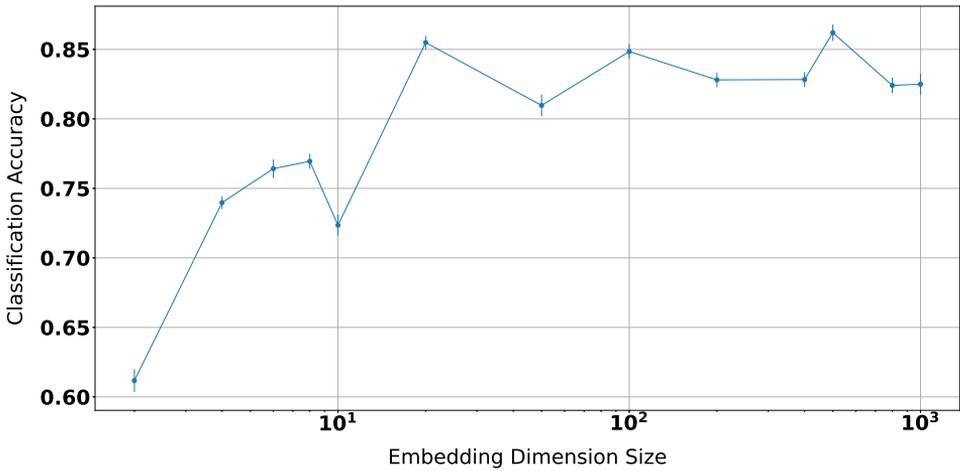


Figure 3.7: Classification accuracy of the SVM classifier when increasing the dimensionality of the latent projection and the number of features is fixed to 6.

jection is varied between 2 and 1000 while the number of features is fixed to 6. Six features were chosen as it reflects the point where the classification accuracy when using a two-dimensional embedding deteriorates. For this reason it is of interest to see the behaviour of classifier as the dimensionality of the latent projection increases. The result of this may be seen in Figure 3.7.

It can be seen that as the dimensionality of the embedding vector increases so does the model’s ability to classify compounded features, however the high dimensional latent space does not enable easy visualisation of data. This being said, using higher dimensional embeddings may enable anomaly detection and aid the generative abilities of the VAE in future work. Additionally, this result further confirms the limitation of the model ability to learn more complex compounded features.

### 3.5.2 LOFAR data

Here we describe the qualitative results of the model trained and evaluated on LOFAR data. This subsection uses descriptive methods to show the generalisability of the model to LOFAR data where the number of features and their compounding become more apparent.

As shown in Section 3.5.1 the inclusion of both magnitude and phase information yields improved performance. This being said, if we were to

train the LOFAR-data based model on magnitude components only, then it would be expected that the embedding would appear significantly different. It is expected that the geometric separability of features such as auto-correlation will be worse, as auto-correlations in phase appear as zeros in their corresponding spectrograms, which is an easily represented feature in the phase based embedding.

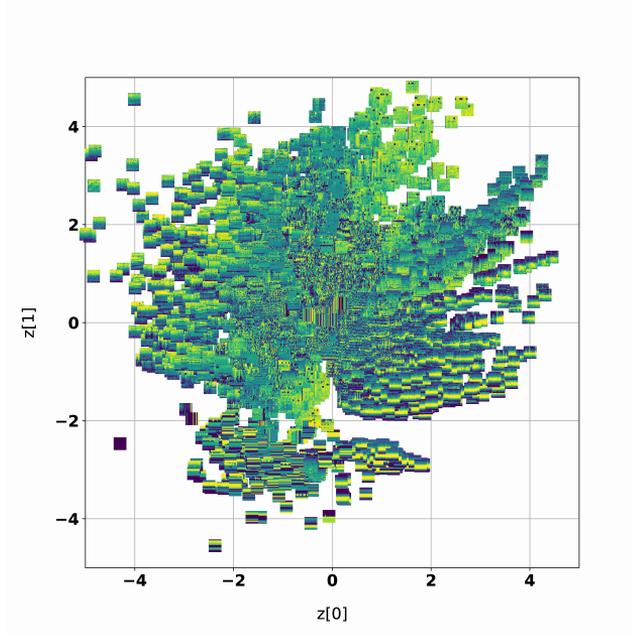
The evaluation of the LOFAR-trained VAE was performed using 100 randomly selected .hdf5 files, that were not in the training set, with 256 baselines sampled from each file. Each spectrogram used for testing contained a number of features compounded in different ways. The projection of the test data is shown in Figure 3.8, where the magnitude and phase components of the input spectrograms are superimposed on each of the points that they are projected to in the 2D latent space.

It can be seen that the generated embedding is clearly capable of grouping together visually similar features as well as geometrically separating distant features. Admittedly, the interplay between the magnitude and phase components do make interpretability of results challenging, however it is clear that the spectrograms from the test data lie on a low dimensional projection, or a manifold, with multiple vertices.

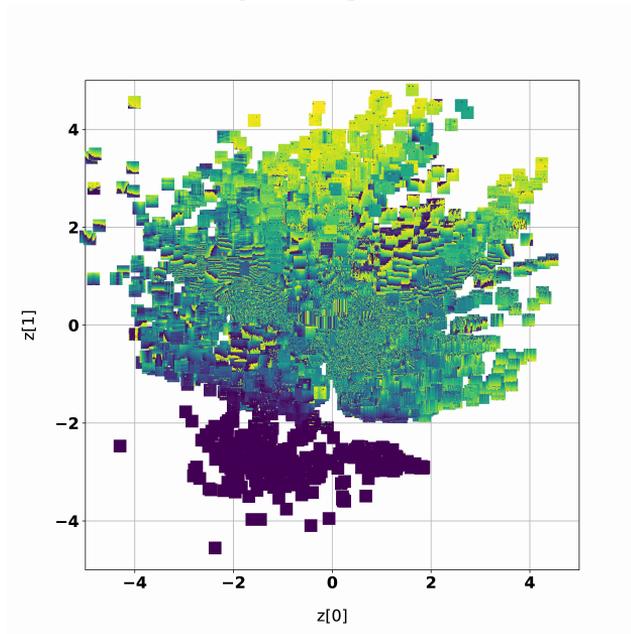
In Figure 3.8, it can be seen that spectrograms of the auto-correlations are embedded to the bottom half of the latent space. This is shown by the normalised phase component of 0 in Figure 3.8 (b), whereas spectrograms with high normalised phase components and *block-like* RFI are placed toward the top of the embedding. Similarly, when considering the bottom-right region of Figure 3.8 (a) it can be seen that particular source structures with a high magnitude component in high frequency bands are grouped together. It can be observed that spectrograms that contain zero magnitude and phase information are geometrically separated from the manifold. This is shown in both Figures 3.8 (b) and 3.8 (a) approximately at coordinates  $(-4.2, -2.5)$ .

More generally, Figure 3.9 shows the decoded output of each point in the latent space superimposed onto its corresponding 2D coordinate when each spectrogram is separated into magnitude and phase components. This plot shows the geometric separability of each feature in its latent projection, which clearly reflects the results shown in Figure 3.8.

This being said, the plots in Figure 3.9 reflect the learnt representations of the input spectrograms rather than *ground-truth* of these inputs. This aspect of the generative abilities of the model is discussed in Section 3.5.3. Figure 3.9 (a) shows clear geometric separation between various features in the latent space. It can be seen that sparse RFI-based features are grouped



(a) Magnitude representation



(b) Phase representation

Figure 3.8: The embeddings of the VAE-based model for preprocessed testing data obtained from LOFAR sampled from 100 random observations. Where the magnitude and phase components are overlaid onto each point in the 2D latent space. It is shown that the model can effectively separate various source structures with an even distribution of data.

toward the origin of the embedding, whereas toward the top-region of the plot higher magnitude representations are learnt.

The corresponding learnt phase-representations of the complex spectrograms are reflected in Figure 3.9 (b). It can be observed that the bottom region of the embedding corresponds to auto-correlations which have a constant 0-phase. Furthermore, it is shown that the VAE learns that spectrograms projected around the point  $(-1.7, -0.4)$  contain particular low-frequency source structure.

This result gives insight into the magnitude and phase based performance gains that were previously discussed. In this case, the lack of detail in the learnt representations are attributable to the VAE under-performing in its ability to learn phase representations of the data. However the lack of detail in the learnt phase representations is to be investigated in future work.

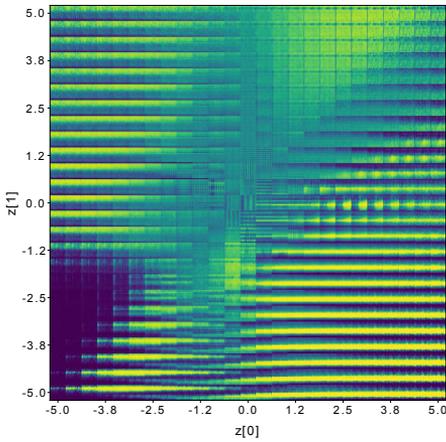
### 3.5.3 Generative abilities of the LOFAR-based VAE

Figures 3.9 (c) and 3.9 (d) show the differences between input spectrograms separated into their respective magnitude and phase components and the learnt magnitude and phase representations of the VAE. These figures clearly show the limitations of the network to accurately learn the correct representations of the LOFAR data. It can be observed that the network has learned more refined representations of the phase component of the spectrograms as seen in Figure 3.9 (d).

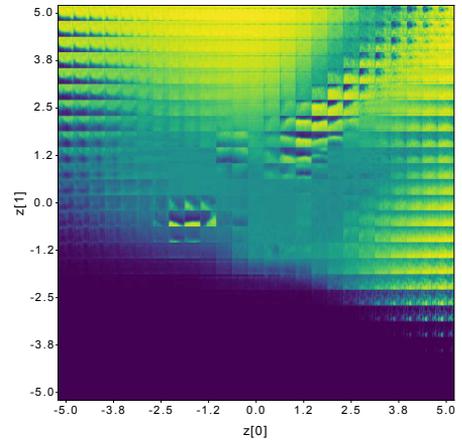
The top-most plot of Figure 3.9 (d) illustrates that the model learns a *general* structure of the input. This is reflected in the correct dynamic range scaling, however finer detail features in the input spectrogram located at time-samples 60 and 80 are not regenerated. Furthermore, it can be seen in the bottom-most plot that the network seems to learn a *blurred* sinusoidal representation of the input.

An easy way to improve the performance of this model would be to increase the number of dimensions in the latent vector, however this would defeat the purpose of this work.

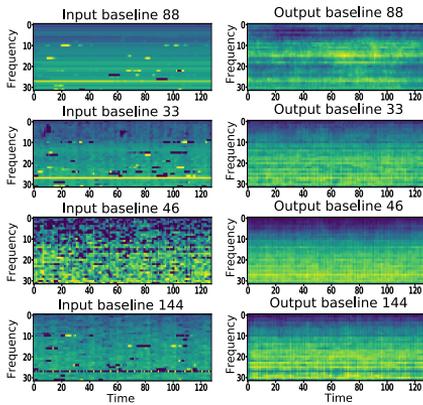
Interestingly, Figure 3.9 (c) shows how the VAE learns the intensity of the magnitude component of the spectrograms, however fails to learn more chaotic features such as spurious RFI. In contrast, the phase representation, shown in Figure 3.9 (d), the model seems to learn the narrow band RFI in the input spectrogram from the 33rd baseline.



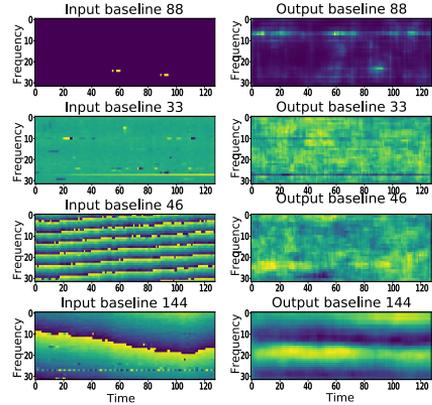
(a) Magnitude embedding plot, where each point in the embedding corresponds to the learnt magnitude representations of the VAE.



(b) Phase embedding plot, where each point in the embedding corresponds to the learnt phase representations of the VAE.



(c) The learnt magnitude representations, where the left column corresponds to the input spectrogram and the right column corresponds to the representation learnt by the VAE.



(d) The learnt phase representations, where the left column corresponds to the input spectrogram and the right column corresponds to the representation learnt by the VAE.

Figure 3.9: The learnt magnitude and phase representations of the VAE that are superimposed onto each point of the latent space and the difference in magnitude and phase components of the input spectrograms relative to the representations that the VAE has learnt.

As already mentioned, the learnt source representations are clearly incorrect, yet they show convincing uses for the generative abilities of the VAE to correctly embed the spectrograms obtained from LOFAR.

### 3.5.3.1 *Prototype data inspection environment*

We integrated the VAE into a flask-based <sup>3</sup> web-application to make the model accessible to astronomers. A screen-shot of the interface is shown in Figure 3.10. The web-interface displays the embedding of a selected observation. Moreover, it enables the filtering of displayed results based on various criteria such as correlation-type, station location and others.

The interface enables operators to use the embedding as a diagnostic tool. They can detect outliers in the embedding and trace the errors back to a particular station. The system is currently being evaluated at the ASTRON observatory.

---

<sup>3</sup> <https://flask.palletsprojects.com/en/1.1.x/>

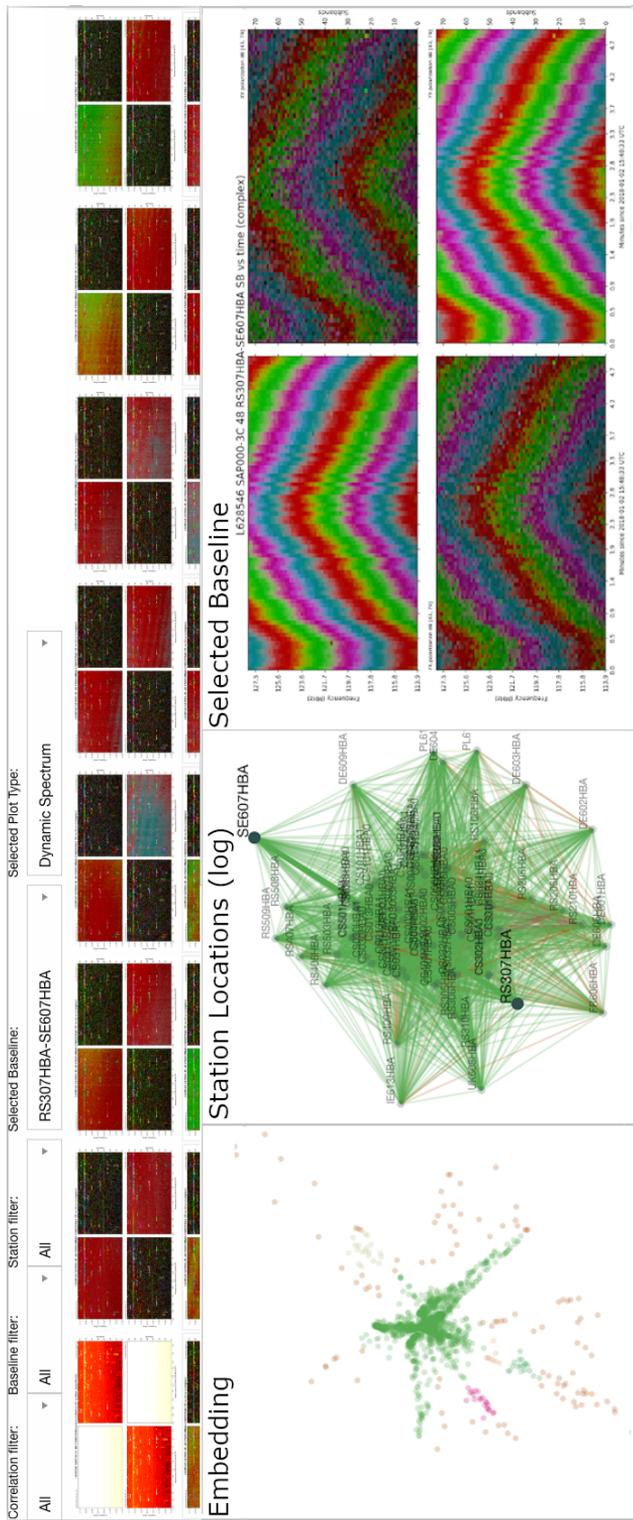


Figure 3.10: Screenshot of the prototype-web interface being used at the ASTRON observatory. The top-most selection boxes correspond to the filters on the baselines and the embedding that will be displayed. The row of images below the filters correspond to all spectrograms within a particular filter. The left-most image in the bottom row corresponds to the embedding of the particular observation with the filters applied, whereas the centre plot on the bottom row is the log-scale station location plot. Finally the right-most plot on the bottom row is the spectrogram corresponding to the baseline selected in the embedding.

### 3.6 CONCLUSIONS AND DISCUSSION

In this chapter we have shown how VAEs are effective in learning representations of radio astronomy spectrograms. We proved that, through the use of a simple preprocessing pipeline and a relatively small amount of data (compared to the amounts being generated on a daily basis by LOFAR), our VAE is capable of learning low dimensional embeddings of the high dimensional features from both simulated HERA data as well as real LOFAR data. We showed that a convolutional VAE that uses both magnitude and phase components of the complex visibilities is capable of producing low dimensional embeddings to aid astronomers with data quality analysis and system health management.

We quantitatively proved how the various parameters of these models effect the performance in geometric separability of data obtained from a simulation from of the HERA telescope. We showed the limits of the model by considering the dimensionality of the projections and the accuracy of a SVM classification algorithm. Through this experimentation, we showed that this model is capable of generalising to the unlabelled real data from LOFAR. The VAE and SVM combination scores between 65% and 90% accuracy, and the geometric separability of features in two dimensions follows a hierarchical compounding scheme. This combination is proven a useful method for assisting human operators to diagnose failure. We showed the integration of the model into a diagnostic web-framework enables telescope operators to pinpoint system failures.

We do recognise that the model is limited by its generative abilities to learn the complex compounded features obtained from the LOFAR telescope. We show that there is a trade-off when trying to project the data to a two dimensional latent space and maintain the high fidelity generative abilities of the VAE. Additionally, a more sophisticated generative model could be used to overcome the short-comings of the VAE as they have been shown to be more effective in their abilities to generate new samples from a learnt distribution of data [16]. In the upcoming chapters we illustrate the anomaly detection abilities of the VAE-based method. We plan to integrate this visualisation tool into the LOFAR observatory pipeline, which may grant the opportunity to investigate active learning in the context of LOFAR.



# 4

## NEAREST NEIGHBOUR-BASED ANOMALY DETECTION

---

*This chapter focuses on how the representation learning methods in the previous chapter can be harnessed for anomaly detection. It does this by considering the anomaly detection general problem proposed by the machine learning community and evaluates performance using both natural images and simulated astronomical spectrograms. In this manner, all complexity and unknowns associated with collection and preprocessing of astronomical data can be avoided when selecting the appropriate anomaly detection method to be used in the system health management context. In this chapter we illustrate a simple, yet effective method, of improving anomaly detection performance of existing autoencoders and generative adversarial networks. This chapter tackles research question 3.*

This Chapter is based on:

- **Michael Mesarcik**, Albert-Jan Boonstra, Elena Ranguelova, and Rob V. van Nieuwpoort, "Improving anomaly detection using the reconstructions of nearest neighbours" [120], in *In: Array 14* (2022)

### 4.1 INTRODUCTION

Anomaly detection is an important field of research as identifying previously unknown behaviours in systems is critical for their maintenance and smooth operation. It is the procedure in which a model is able to identify new classes of data that it has not been exposed to before. Anomaly detection is a far-reaching topic having been applied extensively in fields such as manufacturing [121], cyber-security [122], biomedical analysis [27, 123], astronomy [98] and many more [124].

Novelty, anomaly, outlier, abnormality and out-of-distribution (OOD) detection are closely related topics [125]. The distinction between them is vague across variety of literature studies [16, 19, 20, 27, 126]. For clarity purposes, we consider anomaly detection to be the overarching paradigm, and this terminology will be used throughout the rest of this work.

Approaches for anomaly detection can be divided into a number of categories [14, 124, 127, 128]. In this work we exclusively focus on autoencoder-based anomaly detection. As it offers a data agnostic method that does not rely on significant data augmentation [91], finding negative samples [129, 130] or pretraining on large labelled datasets [30] such as ImageNet [89].

Autoencoders (AEs) are widely used as anomaly detectors [16–24]. The underlying mechanism that governs the AE’s anomaly detection abilities is that they are first trained on data without abnormal, anomalous or outlying samples. Then, during inference, the AE is exposed to anomalous samples which result in higher errors thus enabling anomaly detection. Methods such as mean-square-error (MSE) [21], residual error [27], structural-similarity (SSIM) [84] or feature consistency [131] are used to calculate the pixel-wise difference.

A common problem with using autoencoding methods for anomaly detection is that AEs can generalise to unseen classes thereby performing poorly as anomaly detectors [132]. In Perera, Nallapati, and Xiang [19], this issue is addressed by placing a classifier in the training path of a multi-discriminator based autoencoder, which results in a fairly complicated and costly training procedure. Alternatively, we propose the Nearest-Latent-Neighbours (NLN) algorithm which uses the reconstructions of the nearest-neighbours in the latent space of autoencoders in-order to combat the aforementioned generalisation problem.

Unlike existing nearest neighbours methods [133], our NLN algorithm uses both the *reconstruction error* between a given sample and it’s neighbours in the latent space as well as the average latent-distance to its neighbours. Figure 4.1 illustrates how a vanilla autoencoder generalises to reconstruct unseen samples whereas the reconstructions of an input’s nearest-latent-neighbours more closely resemble the non-anomalous training set thereby offering improved anomaly detection.

We evaluate the proposed method using the anomaly detection framework described in Burlina, Joshi, and Wang [93] and prove it’s effectiveness in a two-stage testing strategy. Firstly by comparing different architecture’s performance with and without the use of our NLN algorithm. Secondly we compare our best performing model with the current state-of-the-art

AEs. We show that NLN is competitive with the state-of-the-art methods across a number of datasets. Furthermore, we consider the radio astronomy anomaly detection use-case through experimentation on the simulated HERA dataset described in the previous chapter.

In summary, the main contributions of this chapter are: (1) a novel nearest-neighbour based algorithm that harnesses the reconstruction error of a given sample's nearest-latent-neighbours and their latent-neighbour distances. (2) The formulation of the NLN algorithm applied to a variety of autoencoding architectures using several different error calculation methods. (3) Improved performance to the state-of-the-art autoencoders using NLN, a fairly simple, cheap and intuitive method, across a number of standard datasets.

## 4.2 NLN: NEAREST LATENT NEIGHBOURS

Here we present our anomaly detection framework for autoencoders. We show that using a simple addition to existing autoencoding architectures we can significantly increase their anomaly detection performance.

### 4.2.1 *Motivation*

In Perera, Nallapati, and Xiang [19] and Hong and Choe [132] the generalisation problem of autoencoders when used for one-class anomaly detection is described. They show that when an AE is trained on the relatively complex 8-class from the MNIST dataset [134], the AE is able to implicitly learn the representations of digit classes such as the 1, 3, 6 and 7. In effect, reconstruction-based anomaly detectors are prone to misidentify these implicitly learnt classes.

In order to solve this problem, Perera, Nallapati, and Xiang [19] propose placing a classifier in the training path of a multi-discriminator-based AE to decrease the training signal for the reconstructions of implicitly learnt anomalous classes. Conversely, we show that if we consider both the distance to, and the reconstruction of, a given sample's nearest latent neighbours we can effectively mitigate this issue, as demonstrated in Figure 4.1.

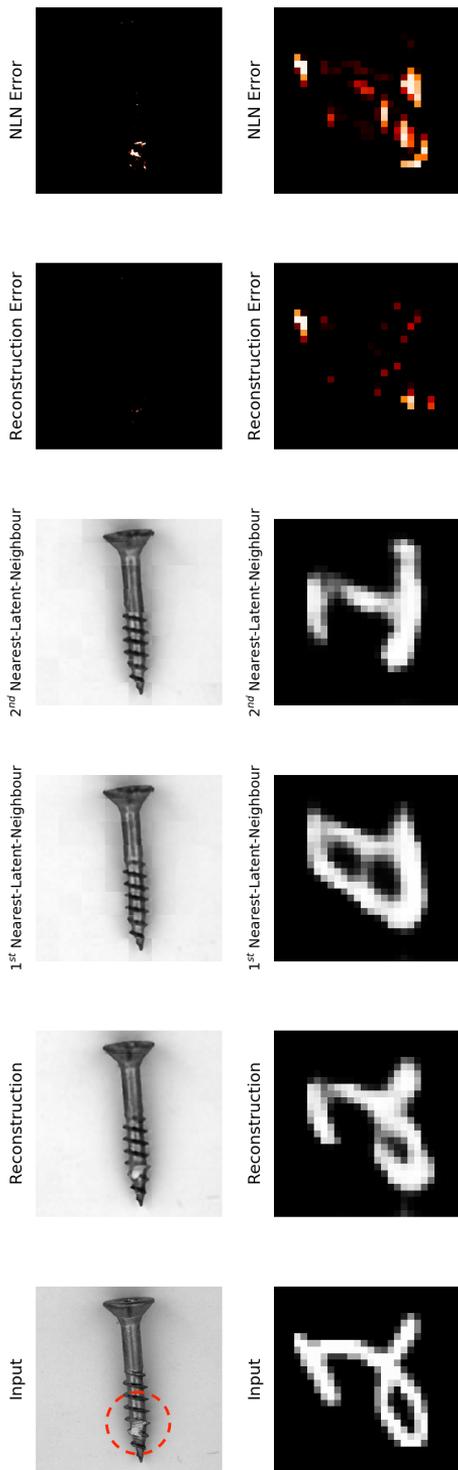


Figure 4.1: Comparison between the MSE and the Nearest-Latent-Neighbours (NLN) based error using a vanilla AE. The top row shows an AE trained on the non-anomalous *screw* images of the MVTec-AD dataset, with the anomaly in the input circled in red, and the bottom row illustrates an AE trained on all MNIST digits except for "2". The first column is the input to the AE, the second is the AE's output and the following two columns show the reconstructions of the input's NLNs. The final two columns show the difference between the MSE and the NLN-based error, where in this work maximising the error on anomalous classes effectively performs anomaly detection. It is clear that the AE learns to reconstruct unseen classes whereas the reconstructions of the NLNs do not.

In addition to the improved performance over Perera, Nallapati, and Xiang [19] and Hong and Choe [132] shown in Table 4.4 we find that the AE-backbones of our NLN algorithm have significantly better training stability and are less prone to mode-collapse [135]. Furthermore, we motivate our focus on AEs for anomaly detection as they are applicable to a variety of datasets without significant augmentation [91, 130], do not need pretraining on large labelled datasets [30, 136] and require far fewer network parameters [137]. Additionally, their structure provides segmentation maps for free without the need of many small patches [90] that result in a significantly more expensive KNN search or additional networks for segmentation [91]. For example, Yi and Yoon [90] perform 2 KNN searches on patches sizes of 32 and 64, whereas we only need a single lookup for patches of size 128.

#### 4.2.2 Problem formulation and approach

Considering an autoencoding model with encoder,  $f_{\theta_f}$ , and decoder,  $g_{\theta_g}$ , then

$$f_{\theta_f}(\mathbf{x}) = \mathbf{z}, \quad f_{\theta_f} : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^l \quad (7)$$

where  $\mathbf{x}$  is the input,  $\mathbf{z}$  is the input's latent representation and  $\theta_f$  are the parameters of the encoder. Additionally,  $\mathbb{R}^{n \times m}$  is the  $n \times m$ -dimensional image-space and  $\mathbb{R}^l$  is the  $l$ -dimensional latent-space. Now consider the decoder with an input,  $\mathbf{z}$ , and a reconstructed output  $\hat{\mathbf{x}}$  such that

$$g_{\theta_g}(\mathbf{z}) = \hat{\mathbf{x}}, \quad g_{\theta_g} : \mathbb{R}^l \rightarrow \mathbb{R}^{n \times m} \quad (8)$$

where  $\theta_g$  are the decoder's parameters, such that the decoder maps from the  $l$ -dimensional latent space to the  $n \times m$ -dimensional image space. The encoder and decoder pair is trained in an end-to-end manner using a loss function such as Mean-Square-Error (MSE) or Binary-Cross-Entropy (BCE). Once trained, the AE's novelty score ( $\eta$ ) is computed for the  $i^{\text{th}}$  sample using

$$\eta = \frac{1}{NM} \sum_{n=1}^N \sum_{m=1}^M (\mathbf{x}_i[n, m] - \hat{\mathbf{x}}_i[n, m])^2 \quad (9)$$

where  $n$  and  $m$  are the pixel-indexes for an image of size  $N \times M$ . This score is typically thresholded in order to determine whether a sample is anomalous and the threshold is calculated using AUROC-based methods that are explained in more detail in Section 4.3.

In order to motivate our use of nearest-neighbours to solve the generalisation problem of AEs, we assume that the high-dimensional training data is concentrated on a low-dimensional data manifold in  $\mathcal{R}^1$  that we attempt to learn using an autoencoder [113]. The learnt manifold is illustrated in Figure 4.2. Here we demonstrate that closely-connected regions on the learnt manifold contain points similar to non-anomalous inputs and dissimilar to those which are anomalous. We exploit this fact to improve the anomaly score robustness by including the nearest-latent neighbours into the reconstruction error.

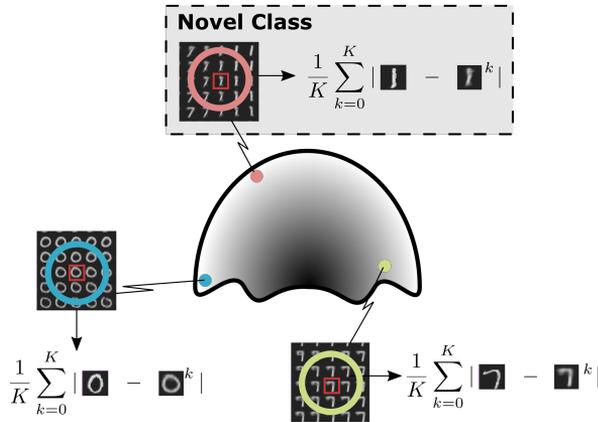


Figure 4.2: Illustration of the learnt MNIST data manifold trained without the class of  $1$ 's (the class of  $1$ 's are anomalous). The closely-connected regions of the anomalous class of  $1$ 's contain dissimilar digits resembling  $7$  and  $9$  whereas the non-anomalous classes consisting of  $0$  and  $7$  do not.

This is done by including the neighbours of the  $i^{\text{th}}$  test sample in the latent space  $\mathbb{R}^1$  in the calculation of the novelty score ( $\eta_{\text{nlm}}$ ). Such that

$$\begin{aligned} \eta_{\text{nlm}} = & \frac{\alpha}{KNM} \sum_{k=0}^K \sum_{n=0}^N \sum_{m=0}^M |x_i[n, m] - g_{\theta_g}(\mathbf{z}_i^k)[n, m]| \\ & + \frac{1 - \alpha}{K} \sum_{k=0}^K |\mathbf{z}_i - \mathbf{z}_i^k| \end{aligned} \quad (10)$$

where  $k$  is the neighbour index such that  $\mathbf{z}_i^k$  is  $\mathbf{z}_i$ 's nearest neighbours in the latent space.  $K$  is the maximum number of latent neighbours and  $\alpha$  is the hyper-parameter ( $\in [0, 1]$ ) used to tune the contribution of latent-space and image-space based distances respectively.

It must be noted that Equation 10 shows the critical difference between [30, 133] and our work. We propose using the reconstruction error in the image

space,  $\mathbb{R}^{n \times m}$ , whereas earlier work only use the difference of extracted feature vectors in  $\mathbb{R}^l$ . We find that there is additional information that can be leveraged for anomaly detection in the image space of autoencoding methods, this is shown experimentally by the results in Tables 4.3 and 4.4. Furthermore, for purposes of anomaly segmentation as in done in the MVTEC-AD dataset [30] the latent space error cannot give pixel-level segmentation maps whereas the NLN-algorithm can.

#### 4.2.2.1 Discriminative Considerations

Discriminative autoencoding models use discriminators in the training of autoencoders. This is done to either improve the *realism* of the AE's outputs or to regularise the latent space to a prior distribution. In this work we focus on the former case. Given a discriminator  $d_{\theta_d}$ , trained on inputs  $\mathbf{x}$  and  $\hat{\mathbf{x}} = g_{\theta_g}(\mathbf{z})$  then

$$d_{\theta_d}(\mathbf{x}) : \mathbb{R}^d \rightarrow [0, 1]. \quad (11)$$

Where the discriminator on  $\mathbf{x}$  maps between the image space and a value on the interval between 0 and 1. It returns 0 or 1 based on whether the sample  $\mathbf{x}$  is taken from the training set or if it is generated by the decoder,  $g_{\theta_g}$ . The discriminator's training objective is stated as [138]

$$\mathcal{L}_{\text{disc}} = \mathbb{E}[\log(d_{\theta_d}(\mathbf{x}))] + \mathbb{E}[\log(1 - d_{\theta_d}(\hat{\mathbf{x}}))] \quad (12)$$

In addition to improving the regularisation, discriminators can also be used for anomaly detection. Novelty is calculated through the difference between the representations of a sample  $\mathbf{x}_i$ , and its respective decoded output  $\hat{\mathbf{x}}_i$ , from an intermediate layer,  $q$ , of  $d_x$ . This is also referred to as the residual error [27] and we include the nearest-latent-neighbours by

$$\begin{aligned} \eta_{\text{res}} = & \frac{\alpha}{HK} \sum^H \sum^K |q_{\theta_d}(\mathbf{x}_i)[h] - q_{\theta_d}(g_{\theta_g}(\mathbf{z}_i^k))[h]| \\ & + \frac{1 - \alpha}{K} \sum^K |\mathbf{z}_i - \mathbf{z}_i^k| \end{aligned} \quad (13)$$

where  $h$  is an index of the output from an intermediate layer  $q$  with size  $H$ .

#### 4.2.2.2 Feature Consistency

It has been shown in Akcay, Atapour-Abarghouei, and Breckon [16] that adding an additional encoder in the training path of the autoencoder improves performance. This paradigm is referred to as feature consistency [131] and can be integrated in our nearest-latent-neighbours method by

$$\begin{aligned} \eta_{\text{con}} = & \frac{\alpha}{LK} \sum^L \sum^K (f_{\theta_f}(\mathbf{x}_i)[l] - f_{\text{con}}(\hat{\mathbf{x}}_i^k)[l])^2 \\ & + \frac{1-\alpha}{K} \sum^K (\mathbf{z}_i - \mathbf{z}_i^k)^2. \end{aligned} \quad (14)$$

Where  $f_{\text{con}}$  is the additional encoder that takes  $\hat{\mathbf{x}}$  as an input, with parameters,  $\theta_{f_{\text{con}}}$ . Furthermore,  $L$  is the latent space dimensionality, which is maintained between the first encoder,  $f$ , and the second encoder,  $f_{\text{con}}$  and is indexed by  $l$ . The encoder is trained jointly with the rest of the discriminative autoencoder as described in Akcay, Atapour-Abarghouei, and Breckon [16].

#### 4.2.3 The NLN algorithm

Our work concerns the integration of the NLN technique into existing autoencoding models. For this reason we explain three different modes of operation for three different novelty scores. In the first case, a vanilla autoencoding model is used with a standard reconstruction error, as shown in Equation 10. The second uses the autoencoding architecture in [16] and the feature consistency error in Equation 14. Finally, the third makes use of a discriminative autoencoding architecture and use of the residual error in Equation 13.

In all cases, an autoencoding model is first trained on a dataset with some anomalous class(es) removed. During testing, a sample is randomly chosen (which may be anomalous or not) and is input into the encoder. Then the nearest neighbours of the encoded sample are found in the latent space generated by the training data. This process is represented by the left-most half of Figure 4.3.

In the first mode of operation, the error is computed between the test sample and both the decodings and positions of its latent-neighbours in the non-anomalous latent space. Whereas when discriminative methods are used, the error is computed between the intermediate representation from the discriminator  $d_x$  of the test sample and all its decoded latent neighbours in the training data.

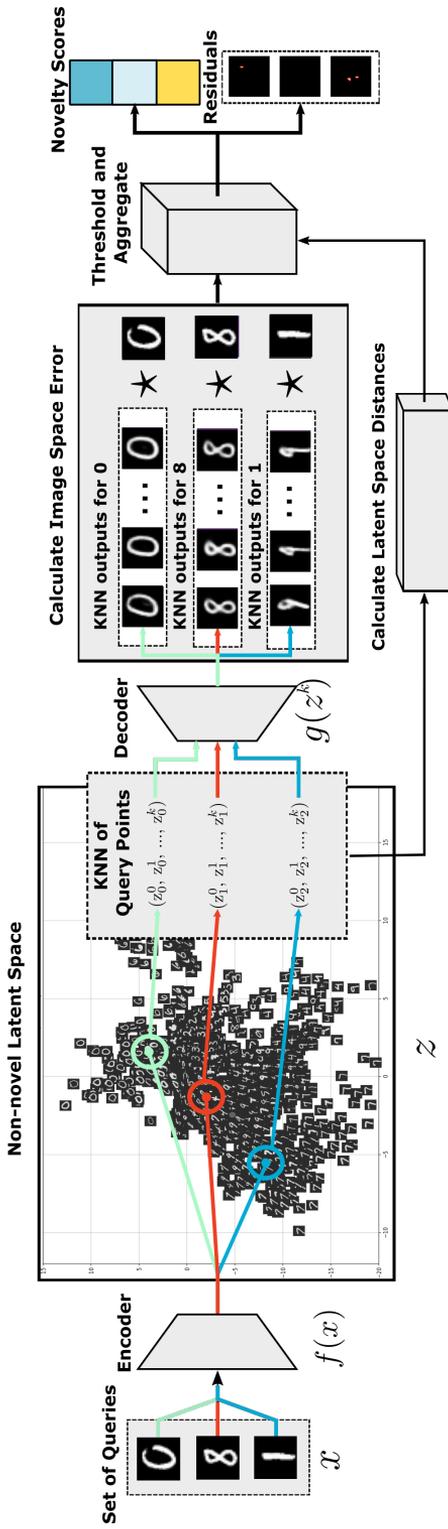


Figure 4.3: An illustration of our NLN algorithm, with three different samples from MNIST used as inputs to an autoencoding model, trained without the class of  $1$ 's (i.e.,  $1$ 's are anomalous). Each of the three input's latent representations are found and their nearest neighbours (relative to the training set) are computed. The  $\star$  represents the chosen error calculation operator. Finally, the error is thresholded and the residual-error maps and novelty scores are produced.

In the feature-consistent case, the error is computed between the encoding via  $f_{\text{con}}$  of the given sample and all its nearest-latent-neighbours in the training data. In Figure 4.3 these three operations are represented by the  $\star$  operator.

When performing anomaly detection, one of the three methods' errors are aggregated over all neighbours and normalised after which they are added to the aggregated and normalised latent-neighbour distance vector. Then they are thresholded to result in an anomaly score and a segmentation map. The threshold is determined by the AUROC method described in Section 4.3. This methodology is illustrated in the right half of Figure 4.3.

### 4.3 EXPERIMENTS

We evaluate our method experimentally in both multi-class and single-class anomaly detection contexts as outlined in [93]. Furthermore, we compare our best performing NLN-enabled autoencoder using both pixel-level and image-level anomaly detection metrics on the MVTec-AD dataset with state-of-the-art autoencoders.

#### 4.3.1 Evaluation methodology

To measure the performance of the NLN-enabled models, they are trained multiple times on a specific dataset, each time removing a different class or classes from the training set, thereby testing the anomaly detection performance on every class present in a given dataset. We do this according to Burlina, Joshi, and Wang [93], such that both the single-class or Single-Inlier-Multiple-Outlier (SIMO) and the multi-class or Multiple-Inliers-Single-Outlier (MISO) performance are evaluated.

We use the Area Under the Receiver Operating Characteristic (AUROC) score to evaluate and compare the performance of the NLN-algorithm. The AUROC metric measures the area under the ROC curve of true positive rates and false positive rates for different threshold values. Furthermore, we evaluate the per-pixel detection performance of our NLN-enabled models using Intersection over Union (IoU) score. The IoU metric is a measure of the overlap between the predicted regions and their corresponding ground-truth.

We limit our evaluation to only autoencoders as we find comparison with methods that rely on SSL [90, 129, 139], pretrained feature extractors [92, 129, 139, 140] or computationally expensive inference [90] are not easily

comparable on AUCROC alone across multiple datasets. It has been well documented that using pretrained feature extractors and SSL losses result in improved performance. However, they typically require orders of magnitude more parameters [137], and are not easily applicable across datasets or evaluation strategies. Furthermore, we regard the simplicity of AEs a crucial attribute. This is in contrast with the significant augmentation found in Li et al. [91] and the challenge of applying patch-dependant methods Yi and Yoon [90] to different datasets of varying resolutions and anomaly types.

#### 4.3.2 Datasets

We evaluate our work on five different datasets, namely MNIST [134], CIFAR-10 [141], Fashion-MNIST [142], MVTec-AD [23] and the HERA dataset [143]. MNIST is a dataset consisting of  $28 \times 28 \times 1$  handwritten digits between 0 and 9. The complexity of the dataset is low and therefore our method performs best on it. Similarly, Fashion-MNIST is composed of  $28 \times 28 \times 1$  images of different types of articles of clothing. This dataset is used as an intermediary difficulty, between MNIST and CIFAR-10. CIFAR-10 is an object recognition dataset consisting of  $32 \times 32 \times 3$  images of 10 different classes. It is the most challenging dataset for anomaly detection as each of the semantic classes may appear at different scales, viewing angles and have changing backgrounds [125].

The MVTec-AD dataset is an industrial anomaly detection dataset consisting of 15 different classes in 2 categories - objects and textures. The 10 object classes contain regularly positioned objects photographed in high resolution from the same viewing angle and the 5 texture classes contain repetitive patterns. For training on the MVTec-AD dataset we follow the augmentation scheme proposed in Bergmann et al. [23], where random rotations and crops are applied to the dataset that is broken into  $128 \times 128$  patches. For more details about the dataset's composition and the augmentation performed see Bergmann et al. [23].

Finally, we extend the evaluation to the simulated astronomical dataset used in Chapter 3. It consists of 5 different non-compounded classes corresponding to several types of radio frequency interference (RFI), astronomical sources as well as instrumentation effects. This dataset acts as a preliminary evaluation of the NLN algorithm in the radio astronomical context. Here, all samples contain diffuse background sources, noise and bandpass effects on both cross- and autocorrelation based spectrograms. The spectro-

grams are  $512 \times 512$  in time and frequency and we only use the magnitude component of the XX polarisation, and perform the preprocessing according to Chapter 3. For more information on the dataset see Kerrigan et al. [112]. We evaluate the models only in the MISO setting as this is a more realistic setting for radio telescope operations.

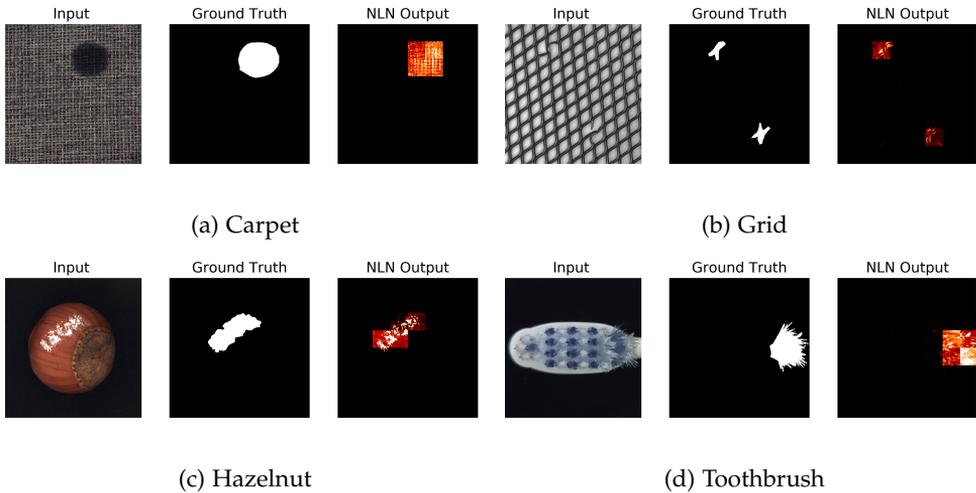


Figure 4.4: Pixel-level anomaly detection using NLN for four different MVTEC-AD classes in the textures (top) and objects (bottom) categories.

### 4.3.3 Model and parameter selection

In order to evaluate our work across a number of different datasets we adapt our models accordingly. We adopt autoencoding the architecture specified in [23] for the evaluation of the NLN algorithm on the MVTEC-AD dataset. For MNIST, CIFAR-10, F-MNIST and HERA we modify a LeNet [144] based autoencoding architecture. The encoder consists of 3 convolutional layers and the decoder has 3 transposed-convolutional layers. A base number of filters of 32 is used for the AE and is increased or decreased on each subsequent layer by a factor of 2. We use *ReLU* activations for all models and they are trained for 50 epochs using ADAM [145] with a learning rate of  $1 \times 10^{-4}$ . The image-based discriminators  $d_x$  use the same architecture as the encoder, except the final layer, which is a dense layer with a *sigmoid* activation. The latent discriminator for the AAE consists of 3 dense layers with *Leaky ReLU* activations and a dropout rate of 0.3. The base layer size is 64 and is increased by a factor of 2 for each subsequent

	AE	AAE	VAE	AE-res	AE-con
<b>MNIST</b>	9.80%	17.65%	11.65%	10.13%	14.18%
<b>CIFAR-10</b>	6.92%	7.41%	1.29%	7.30%	6.68%
<b>F-MNIST</b>	11.52%	9.53%	10.31%	11.95%	11.52%

Table 4.1: Mean MISO AUROC percentage increase using NLN

	AE	AAE	VAE	AE-res	AE-con
<b>MNIST</b>	3.45%	4.92%	3.99%	3.47%	4.21%
<b>CIFAR-10</b>	7.65%	8.81%	6.36%	6.89%	8.02%
<b>F-MNIST</b>	5.66%	5.12%	5.31%	5.65%	3.49%

Table 4.2: Mean SIMO AUROC percentage increase using NLN

layer. Furthermore, we treat the maximum number of neighbours,  $K$ , the latent dimensionality,  $L$ , and the NLN contribution,  $\alpha$ , as hyper-parameters of our algorithm.

#### 4.3.4 Results

We evaluate the performance increase of the NLN algorithm for a variety of autoencoding models across a number of different datasets in both the MISO-context in Table 4.1 and SIMO-context in Table 4.2. Here the best performing reconstruction error-based AUROC is compared with the best performing NLN-enabled model for each architecture. The NLN-based AEs achieve a performance increase between 17% and 1% across the three MISO-datasets and 8% and 3% for the SIMO-case. We suspect the low performance gains in the SIMO-case of the NLN-enabled AEs are due there being fewer latent neighbours to select from, thereby reducing performance.

In Table 4.3 we present the MISO-based class-averaged AUROC comparison of autoencoding models. For MNIST, the optimal configuration is a feature consistent AE with  $K = 2$ ,  $L = 32$  and  $\alpha = 1.0$ , for CIFAR-10 we use the discriminative AE when  $K = 1$ ,  $L = 32$  and  $\alpha = 0.5$ . Finally for F-MNIST, we use a discriminative AE when  $K = 1$ ,  $LD = 64$  and  $\alpha = 0.9$ . Here we see that the NLN-algorithm gives significant performance increases for MNIST and F-MNIST, even above the pretrained ResNet-50 proposed by Bergman, Cohen, and Hoshen [30]. Furthermore, we see that OCGAN Perera, Nallapati, and Xiang [19] is not performant in a MISO context, this

Model	MNIST	CIFAR-10	F-MNIST
GANomaly [16]	0.753	0.532	0.679
Skip-GAN [17]	0.492	0.629	0.515
OC-GAN [19]	0.683	0.510	0.678
VAE [18]	0.515	0.497	0.521
AnoGAN [27]	0.632	0.434	0.510
EGBAD [28]	0.656	0.496	0.500
DKNN [30]	0.791	<b>0.714</b>	0.746
<b>Ours</b>	<b>0.921</b>	0.560	<b>0.763</b>

Table 4.3: Mean MISO anomaly detection AUROC, bold is best.

indicates that our NLN algorithm may offer a more robust solution to the generalisability problem in AEs. We show that AEs do not perform particularly well on CIFAR-10. This is expected, as images from the same class in contain substantially different pixel-level information. For example the *aeroplane* class contains images of both the cockpit of a grounded Boeing 747 as well a fighter-jet photographed from the side-view in mid-flight. In effect, the MSE between non-anomalous images in the same class, can be greater than anomalous images thereby reducing the efficacy of MSE based anomaly detectors on CIFAR-10.

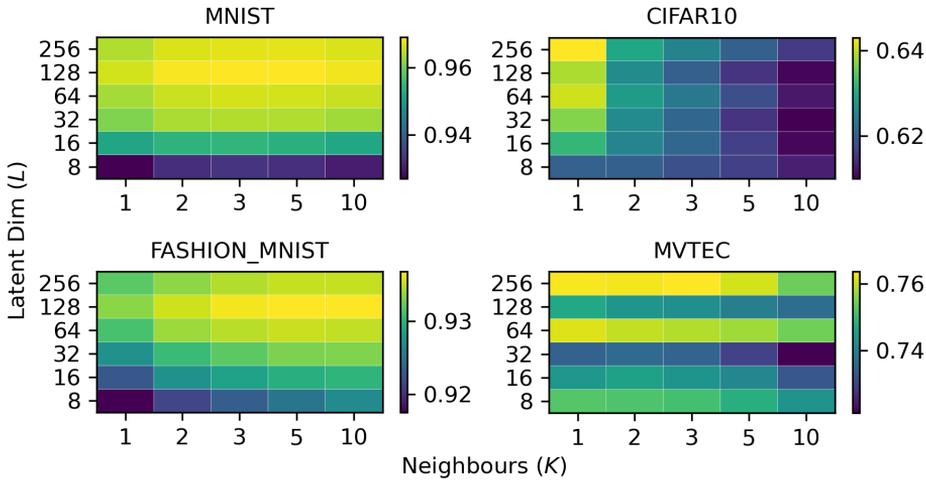
We present the class-averaged AUROC scores for the SIMO-based evaluation in Table 4.4. Here the optimal method for MNIST is a discriminative AE, with  $LD = 128$ ,  $K = 3$  and  $\alpha = 1.0$  and for CIFAR-10 we find the optimal method to be a vanilla AE with  $LD = 256$ ,  $K = 1$  and  $\alpha = 0.75$ . Furthermore, we find the best performing method on F-MNIST to be a VAE with  $LD = 32$ ,  $K = 3$  and  $\alpha = 0.9$ . For the MVTEC-AD dataset we use a discriminative AE with  $LD = 128$ ,  $K = 1$  and  $\alpha = 0.8$ . It is clear that the attention guided VAE (CAVAGA) [146] method performs best on MNIST whereas DKNN [30] on CIFAR-10. However, it is evident that the NLN-enabled autoencoding models offer increased performance over existing autoencoding and ResNet-based architectures for both the F-MNIST and MVTEC-AD datasets in the SIMO context.

In Figure 4.5 we show the effect of varying  $L$  and  $K$  on AUROC scores for vanilla AE in the SIMO context when  $\alpha = 0.8$ . For F-MNIST and MNIST a maximum AUROC score is found for  $L = 128$  and  $K > 3$ , whereas for CIFAR-10 the optimal is found when  $L = 256$  and  $K = 1$ . Finally it is shown

Model	MNIST	CIFAR-10	F-MNIST	MVTec-AD
GANomaly [16]	0.965	0.695	0.906	0.762
OC-GAN [19]	0.975	0.657	0.924	0.756
AnoGAN [27]	0.912	0.618	0.817	0.600
LFD [132]	0.977	-	0.927	0.777
CBiGAN [29]	-	-	-	0.770
CAVGA-D <sub>u</sub> [146]	<b>0.986</b>	0.737	0.885	-
DKNN <sup>1</sup> [30]	0.917	<b>0.890</b>	0.938	0.750
<b>Ours</b>	0.974	0.658	<b>0.941</b>	<b>0.783</b>

Table 4.4: Mean SIMO anomaly detection AUROC, bold is best.

that the vanilla AE offers best image-based AUROC performance when  $L = 256$  and  $K = 3$ .

Figure 4.5: Vanilla Autoencoder AUROC sensitivity to number of neighbours and latent dimensions in SIMO-context for  $\alpha = 0.8$ .

We evaluate the pixel-level anomaly detection performance in Table 4.5, and illustrate the model outputs in Figure 4.4 of both texture and object classes. In all cases we use a vanilla AE with  $K = 1$ ,  $L = 128$  and  $\alpha = 0.6$ . It is clear that the NLN-enabled AE demonstrates performance increases in

<sup>1</sup> We use the authors implementation for all datasets other than MVTEC-AD, here we use our own Tensorflow-based implementation

	Class	AE-L2 [23]	AE-SSIM [23]	SMAI L2 [22]	VE-VAE[24]	Ours
Textures	carpet	0.59	0.87	<b>0.88</b>	0.78	0.82
	grid	0.90	0.94	<b>0.97</b>	0.73	0.86
	leather	0.75	0.78	0.86	<b>0.95</b>	0.85
	tile	0.51	0.59	0.62	<b>0.80</b>	0.51
	wood	0.73	0.73	<b>0.80</b>	0.77	0.72
	mean	0.70	0.78	<b>0.83</b>	0.81	0.75
Objects	bottle	0.86	0.93	0.86	0.87	<b>0.95</b>
	cable	0.86	0.82	<b>0.92</b>	0.90	0.90
	capsule	0.88	<b>0.94</b>	0.93	0.74	<b>0.94</b>
	hazelnut	0.95	0.97	0.97	<b>0.98</b>	<b>0.98</b>
	metal nut	0.86	0.89	0.92	<b>0.94</b>	0.88
	pill	0.85	0.91	<b>0.92</b>	0.83	<b>0.92</b>
	screw	0.96	0.96	0.96	<b>0.97</b>	<b>0.97</b>
	toothbrush	0.93	0.92	0.96	0.94	<b>0.97</b>
	transistor	0.86	0.90	0.85	<b>0.93</b>	0.85
	zipper	0.77	0.88	0.9	0.78	<b>0.96</b>
	mean	0.88	0.91	0.92	0.89	<b>0.93</b>

Table 4.5: Pixel-based anomaly detection (Segmentation) AUROC score for auto-encoding models, where bold is best.

the object classes of MVTec-AD. However, this is not the case for the texture classes. We suspect that this is due to our NLN-enabled AE not being able to distinguish between different texture-patches. This behaviour is similarly demonstrated in [23], and we believe that this is an inherent weakness of standard autoencoding architectures.

In Figure 4.7 we illustrate the effect on varying alpha for the NLN-enabled autoencoding models used for the MVTec-AD dataset. Here it is demonstrated, that the NLN-based model obtain optimal AUROC segmentation-performance when  $0.25 < \alpha < 0.8$ , whereas to optimal AUROC detection-performance occurs when  $\alpha > 0.6$ . Finally we illustrate that the optimal IoU value is obtained at  $\alpha = 0.8$ , thus demonstrating the benefit of including the reconstructions of nearest-neighbours in the calculation of the anomaly score.

We evaluate the MISO AUROC performance of the autoencoder models on the HERA dataset as shown by Table 4.6. Here we find that the AAE performs best in this context when using a patch size of  $64 \times 64$ , a latent dimensionality of 32 and  $\alpha = 0.1$ . We note that the performance of the

Model	Point Source	Narrow band RFI	Broadband transient RFI	Broadband continuous RFI	Blips RFI	Mean
AE	0.604	0.727	0.711	0.654	0.672	0.674
AE-res	0.486	0.728	0.661	0.676	0.726	0.656
AE-con	<b>0.612</b>	<b>0.776</b>	0.713	<b>0.846</b>	0.799	0.749
VAE	<b>0.612</b>	0.74	0.782	0.772	0.909	0.763
AAE	0.592	0.69	<b>0.805</b>	0.838	<b>1.0</b>	<b>0.785</b>

Table 4.6: MISO AUROC of the autoencoder models evaluated on the HERA dataset, bold is best

model is slightly less than the machine learning specific datasets, especially for the point source feature. We attribute this loss in performance to the point source having significantly lower power than the other features in spectrograms. We present the associated hyper-parameter search over  $\alpha$  and patch size in Figure 4.6.

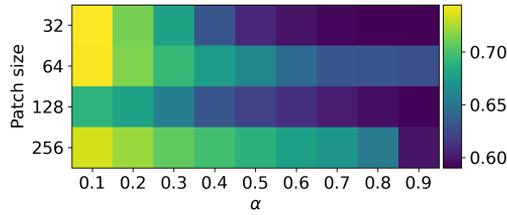


Figure 4.6: AAE AUROC sensitivity when varying the patch size and  $\alpha$  on the HERA dataset when latent dimensionality is fixed at 32.

#### 4.3.5 Time and memory efficiency

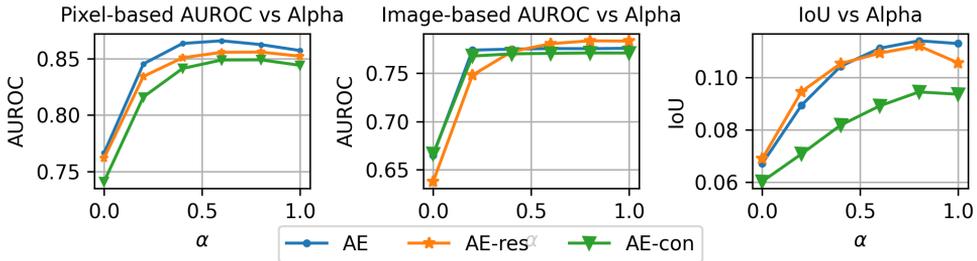


Figure 4.7: AUROC and IoU sensitivity to varying  $\alpha$  of the NLN-enabled autoencoding models applied to the MVTEC-AD dataset.

The NLN-algorithm requires a forward pass through an encoder, a KNN search of the latent-space generated by the training samples, and a forward pass of a given point’s nearest neighbours through a decoder. We evaluate the models on a Nvidia T4, where a forward pass of a single image from the MVTEC-AD dataset takes 7.41 ms for the encoder and 9.63 ms for the decoder. In comparison, a ResNet50 used in [30, 147, 148] requires 43.3ms for a forward pass of a single image. This means that our method is between  $1.3\times$  and  $2.5\times$  more efficient for a forward pass, depending on the architecture used.

For the KNN search we use a k-d tree implementation of the KNN search, which has a inference time complexity of  $\mathcal{O}(KL \log N)$ . Where K is the number of neighbours, L is the latent dimensionality and N is the number of points in the training set. In the case of the NLN-enabled models presented in this work, we find a latent dimensionality of 128 sufficient, whereas the ResNet50 in [30] uses 2048 dimensional latent space. This means that our work offers a  $16\times$  reduction in KNN search inference time in comparison with [30].

Finally, our method has comparable storage requirements as other AE based models [137] in terms of number of trainable parameters. For comparison, the AE-con model used for MVTEC-AD has 1.79 million parameters, whereas the ResNet-50 from [30] has 25.58 million parameters. The only storage-based overhead of the NLN-algorithm is the requirement of amortising the embeddings of the training set as suggested in [30]. In the case of the bottle-class of the MVTEC-AD dataset, there is an additional storage requirement of 6.85 MB<sup>2</sup>

#### 4.4 ABLATION STUDY

The AUROC performance of the NLN-algorithm is demonstrated in Table 4.7 when the loss function varied. The term in the first column,  $\mathcal{L}_{\text{recon}}$ , represents the standard reconstruction error given by Equation 9 and  $\mathcal{L}_{\text{NLN}}$  shows the NLN-based reconstruction loss given in the first half of Equation 10.  $\mathcal{L}_{\text{con}}$  represents the feature consistent adaption given by the first half of Equation 14 and  $\mathcal{L}_{\text{total}}$  is equivalent to the score obtained from Equation 14. It can be seen that through the utilisation of all terms in NLN-loss formulation we obtain optimal performance.

---

<sup>2</sup> 209 images  $\times$  16  $\times$  4 augmented patches  $\times$  128 latent dimensions  $\times$  32 bits = 6.85MB of additional memory

Dataset	$\mathcal{L}_{\text{recon}}$	$\mathcal{L}_{\text{NLN}}$	$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{total}}$
MNIST	0.778	0.822	0.913	0.921
FMNIST	0.669	0.702	0.719	0.738
CIFAR-10	0.511	0.513	0.551	0.553

Table 4.7: MISO AUROC performance of AE-con for different losses terms when  $K = 2$ ,  $L = 32$  and  $\alpha = 0.9$

#### 4.5 DISCUSSION AND CONCLUSIONS

Autoencoders learn to generalise to unseen classes which is a problem when they are used for anomaly detection. In this work, we demonstrate that when the reconstructions of a model’s nearest-latent-neighbours are harnessed we can more effectively and efficiently mitigate this problem in comparison with the state-of-the-art. This is achieved through a fairly simple algorithm that is agnostic to both the AE’s architecture and its error method. We experimentally demonstrate that the addition of the NLN algorithm consistently yields performance increases for various autoencoding architectures and various datasets and is competitive with the state-of-the-art autoencoding models. This is achieved without complex augmentation, using pretrained networks or computationally expensive inference. We note that the complexity of CIFAR-10 and the texture classes of MVTec-AD result in modest performance, but we expect this can be solved using more robust error functions or using SSL to obtain even better latent representations.

Furthermore we illustrate the NLN algorithm’s potential in the radio observatory anomaly detection setting. We find that the data agnostic approach offers performance in line with other machine learning specific datasets. We note however that due to similar feature morphologies, high dynamic range and potential feature compounding, autoencoders need radio astronomy specific refinements to make them suitable in the observatory setting.



# 5

## RADIO FREQUENCY INTERFERENCE DETECTION

---

*In this chapter we apply the Nearest Latent Neighbours algorithm to Radio Frequency Interference (RFI) detection. We consider RFI detection one of the many aspects of system health management at the radio observatories. This chapter illustrates one of the many places where an anomaly detection approach can be beneficial to telescope operations. We describe the current state-of-the-art techniques in machine learning applied to RFI detection and show how our approach has several benefits over them. This chapter clarifies research questions two and three by showing a novel application of machine learning based anomaly detection to RFI detection at radio telescopes.*

This Chapter is based on:

- **Michael Mesarcik**, Albert-Jan Boonstra, Elena Ranguelova, and Rob V. van Nieuwpoort, "Learning to detect radio frequency interference in radio astronomy without seeing " [149], in *In: Monthly Notices of the Royal Astronomical Society* 516.4 (September 2022)
- **Michael Mesarcik**, Albert-Jan Boonstra, Elena Ranguelova, and Rob V. van Nieuwpoort, "Dataset for Learning to detect RFI in radio astronomy without seeing it." [150], in *Dataset on Zenodo*

### 5.1 INTRODUCTION

Radio Frequency Interference (RFI) is a growing concern for radio astronomy due to the proliferation of electronic equipment that depends on electromagnetic emissions. Radio frequency radiation from astronomical sources is extremely faint relative to emissions from man-made systems such as

radars, telecommunication devices, large satellite constellations [151] and more. Despite international regulation to ensure radio-quiet zones and limit transmission power of emitters, there are still concerns about RFI hampering radio astronomy.

For this reason, approaches for RFI detection and mitigation have become a necessity in modern radio observatories. RFI pipelines are commonly deployed at telescopes performing RFI detection and mitigation in a post correlation setting. Traditionally, algorithms such as CUMSUM [152], Singular Value Decomposition (SVD) [41], Wavelet-based methods [153] and AOFlogger [154] have been used. These RFI detection algorithms are widely implemented for real-time RFI detection at observatories around the world [8, 108, 155, 156]. In effect all archived data from these instruments contain automatically generated RFI-masks which can be accessed with no additional cost.

Recent advances in machine learning have made data-driven approaches unprecedentedly suitable for RFI-detection. Most machine learning methods for RFI detection have been based on supervised learning using U-net [101] and its derivatives [69–71, 112]. Research has shown that these are promising methods, significantly outperforming classical approaches. However, in reality supervised methods require significant amounts of expertly labelled time/frequency data, that is not available in practice due to the related cost.

As a result, recent models are trained and evaluated using simulators or flags generated by classical methods, with limited experimentation on real expertly labelled datasets. This is problematic as the effectiveness of these methods on unseen data is difficult to measure and predict. Furthermore, recent machine learning-based methods have not been well integrated into telescope-pipelines as the cost of labelling is prohibitive to many instrument operators.

To solve these problems we propose an unsupervised learning method based on the Nearest Latent Neighbours (NLN) algorithm [86]. This approach leverages anomaly detection to perform RFI detection. This is achieved using a generative model trained on uncontaminated (RFI free) data to detect *novel* RFI contaminated emissions. Interestingly, this formulation is effectively the inverse of how existing deep learning-based methods are trained.

In this work, RFI is detected by measuring the difference between small sub-regions (patches) of spectrograms that are known to *not* contain RFI and the patches being evaluated. To select RFI-free data we break each

spectrogram into a number of equally sized patches and use their associated AOFlagger-based flags to locate the instances which contain RFI. In doing so we do not have to incur the cost of extensive labelling as the AOFlagger masks are readily available. It must be noted that NLN can tolerate high false positive rates in the training masks due to the use of the inverse problem. However, under-flagging can cause undesired effects.

We show that our method outperforms existing supervised models on several benchmarks, using less data for training. Furthermore, we demonstrate that if supervised state-of-the-art methods are trained with weak-labels they typically over-fit to the training data and do not generalise to unseen examples. Our approach does not suffer from these problems.

Additionally, as the landscape of RFI emissions changes over time we expect supervised methods to be continually retrained as future emitters occupy newer frequency bands. Conversely, the emissions from celestial bodies will remain fairly consistent over the same time-scale, effectively meaning that that our method will not have to undergo retraining.

We make the following contributions in this chapter: (1) a novel unsupervised learning-based approach to RFI detection in radio astronomy; (2) an evaluation of the effectiveness of using AOFlagger generated ground-truth for training of machine learning-based RFI detection algorithms, and (3) an expert labelled dataset that can be used for comparison and development of novel RFI detection algorithms.

This chapter begins with describing the NLN-method in Section 5.2 and explains how the NLN algorithm was adapted to work for RFI detection. In Section 5.3 we explain our data selection strategy and outline the expert-labelled dataset used for evaluation of this work. Finally, we present our results and conclusions in Section 5.4 and Section 5.5.

## 5.2 METHOD

In this work we use NLN for RFI detection. This is motivated by several factors: (1) obtaining sufficient labels for supervised segmentation of RFI has a significant overhead; (2) existing supervised techniques over-fit to flags from classical methods such as AOFlagger, leading to sub-optimal performance on unseen data; (3) the ever-changing landscape of RFI requires continual labelling and training efforts to enable supervised approaches to capture new temporal and spectral RFI structures.

We show that if the traditional RFI detection problem is inverted, we can effectively address the supervised RFI detection issues.

### 5.2.1 Model definition and training

For some complex visibility  $V(\nu, \tau, b)$  and the corresponding ground truth mask for the interference  $G(\nu, \tau, b)$ , the training objective for supervised RFI detection can be formulated as follows

$$\mathcal{L}_{\text{sup}} = \min_{\theta_m} \mathcal{H}(m_{\theta_m}(V(\nu, \tau, b)), G(\nu, \tau, b)). \quad (15)$$

Here  $m$  is a function with learnable parameters  $\theta_m$  and  $\mathcal{H}$  is the entropy-based similarity between the model prediction and the ground truth. This problem is well-posed and has been used across multiple domains to effectively train classifiers. However it relies on learning a model of the RFI using ground truth labels, which are in practice hard to obtain.

In this work, we train a model of everything other than the interference, so that we can perform RFI detection as a downstream task. This is done by first training a discriminative autoencoder, on  $n \times n$ -sized uncontaminated regions (also known as patches) of the visibility,  $V_{n \times n}(\nu, \tau, b)$ . We select these regions using the *weak-labels* generated by a classical method such as AOFlagger.

First we define the encoder  $f$  that maps from the visibility space  $\mathbb{R}^2$  to a latent space  $\mathbb{L}$ , such that

$$\mathbf{z} = f_{\theta_f}(V_{n \times n}(\nu, \tau, b)), \quad f : \mathbb{R}^2 \rightarrow \mathbb{L}. \quad (16)$$

Here  $\mathbf{z}$  is a low-dimensional projection of the an  $n \times n$  patch that contains no interference and  $\theta_f$  are the learnable parameters of the encoder. Furthermore, we define the decoder  $g$  that maps back from the low-dimensional projection to the visibility space, such that

$$\widehat{V}_{n \times n}(\nu, \tau, b) = g_{\theta_g}(\mathbf{z}), \quad g : \mathbb{L} \rightarrow \mathbb{R}^2 \quad (17)$$

where  $\theta_g$  are the parameters of the decoder. We simultaneously train the encoder and decoder using the reconstruction loss,

$$\mathcal{L}_{\text{recon}} = \min_{\theta_f, \theta_g} \mathcal{H}(V_{n \times n}(\nu, \tau, b), \widehat{V}_{n \times n}(\nu, \tau, b)). \quad (18)$$

We use Mean-Square-Error (MSE) to train a standard autoencoder. Typically, MSE-based reconstruction losses produce blurry outputs, which may affect the quality of the predicted RFI masks. The blurriness is a result

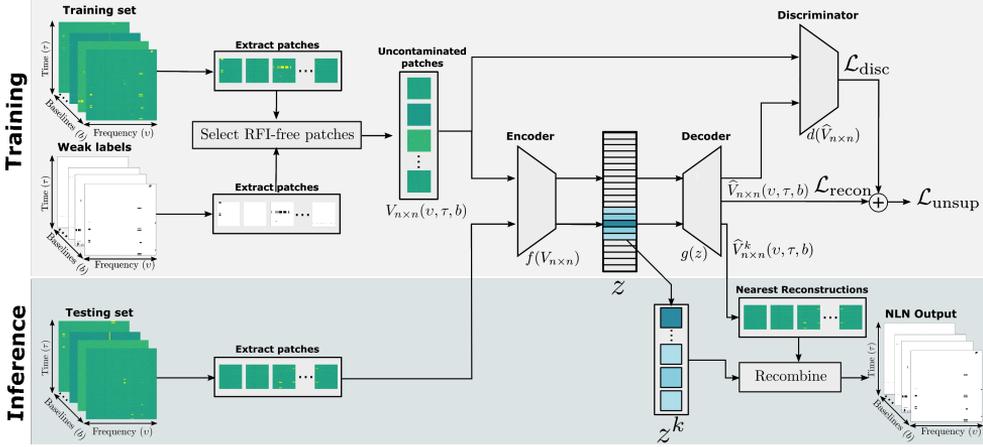


Figure 5.1: Block diagram of the training and inference procedures of NLN-based RFI detector. Here, we use a discriminative autoencoder as the backbone of the architecture. The top half of the figure shows the training procedure whereas the bottom illustrates how NLN is used for inference. The recombination of latent distances  $z^k$  and nearest-reconstructions  $\hat{V}^k(\nu, \tau, b)$  is performed according to Equation 20.

of back-propagating the gradient from the average pixel-wise error, which prohibits the autoencoder from producing high-frequency details for given inputs. In order to counteract this problem, we define a discriminator,  $d$ , that acts as a regulariser on the decoder’s output. This discriminative loss is back-propagated through the decoder, such that it learns to produce not only low frequency details but also the discriminative features [157]. Furthermore, the discriminator enables the autoencoder to be used as a generative model [21]. The discriminator maps from  $\mathbb{R}^2$  to a classification on the interval  $[0, 1]$ , effectively trying to determine if the input is generated or sampled from the original dataset. In this case, the original dataset is the uncontaminated patches selected using the weak-labels. The discriminative loss is given by

$$\mathcal{L}_{\text{disc}} = \min_{\theta_d} \mathbb{E}[\log(d_{\theta_d}((V_{n \times n}(\nu, \tau, b))))] + \mathbb{E}[\log(1 - d_{\theta_d}(\hat{V}_{n \times n}(\nu, \tau, b)))].$$
(19)

We train the discriminator simultaneously with the decoder, such that the total loss is  $\mathcal{L}_{\text{unsup}} = \alpha \mathcal{L}_{\text{recon}} + (1 - \alpha) \mathcal{L}_{\text{disc}}$ , where  $\alpha$  is a hyper-parameter between 0 and 1 that determines how much of an effect the discriminative and reconstruction losses have on the training respectively. An illustration

of the training procedure for the discriminative autoencoder is shown in the top half of Figure 5.1.

### 5.2.2 *Nearest-Latent-Neighbours for RFI detection*

Given some trained latent-variable-model such as the discriminative autoencoder,  $g(f(V))$ , we need to formulate a measure of similarity or difference between the learnt distribution of RFI-free patches and those which are unseen. In practice several options exist such as pixel-level difference [16], structured-similarity measure [84], residual measures [27], purely latent measures [30] and many more. The most important factor when selecting a measure for RFI detection is the resolution of the output. For example, using a purely latent-measure would result in the resolution of the output RFI-masks to be fixed by the resolution of each patch, as shown in Figure 5.2 (e). However using a pixel-level difference, may cause the predictions to be sensitive to noise.

To counter-act this problem we propose using a distance function that utilises both latent and pixel-wise measures of difference, namely NLN. NLN is an anomaly detection technique that works by performing a nearest-neighbour lookup in the latent space of generative models. At training time, it operates as a standard discriminative autoencoder, training on the aforementioned loss. During inference, a test-sample is given to the model, with the objective to determine which parts (if any) of the input sample are anomalous. A combination of two metrics is used, the first measures the latent distance from the given test-sample to its nearest neighbours from the distribution of in-lying data, illustrated in Figure 5.2 (e). The other is the reconstruction error between the given sample and the reconstructions of all its neighbours found in the latent space. Figures 5.2 (b) and 5.2 (c) demonstrate that when the autoencoder is trained on only RFI-free data it is capable of only reconstructing the *non-novel* astronomical signals and cannot generate RFI found in the input. In effect, the reconstruction error shown in Figure 5.2 (d) has a higher dynamic range than the input. For more details and analysis of the method see [86].

We modify the original NLN-distance function such that the latent distances are used as a coarse selection for the higher resolution pixel-based error. An illustration of this selection mechanism is shown in Figure 5.2 (f). The modified NLN measure is the reconstruction error of a test-sample's nearest neighbours multiplied by its latent-distance vector, as given by

$$D_{\text{NLN}} = \left( \frac{1}{K} \sum^K |V_{i,n \times n}(v, \tau, b) - g_{\theta_g}(\mathbf{z}_i^k)| \right) \times D_{\text{latent}} \quad (20)$$

where  $k$  is the nearest-neighbour of the  $i^{\text{th}}$  sample in the latent space given by  $\mathbf{z}$ . The nearest-neighbours are selected through the K-Nearest-Neighbours (KNN) algorithm using the default implementation of FAISS [158]. Furthermore,  $D_{\text{latent}}$  is the thresholded mean latent-distance vector of the  $i^{\text{th}}$  query patch and its  $k$  RFI-free neighbours, as given by

$$D_{\text{latent}} = \begin{cases} 1, & \frac{1}{K} \sum^K |\mathbf{z}_i - \mathbf{z}_i^k| \geq T \\ 0, & \text{otherwise} \end{cases} \quad (21)$$

We treat both  $K$  and  $T$  as hyper-parameters of our algorithm and determine them experimentally across our datasets. In effect, the latent distance function offers a coarse resolution view of the RFI, and the reconstruction error offers a finer grained resolution. It must be noted, that the only additional overhead of NLN is that it requires the representations of the training set to be stored. An illustration of the inference mode of the NLN algorithm is shown in the bottom half of Figure 5.1.

### 5.2.3 Architectural considerations

We use a strided convolutional architecture for the encoder, decoder and discriminator. Both the encoder and decoder have the same architecture except that the decoder uses transposed convolutions in place of the encoder’s convolutional layers. Furthermore, the discriminator uses the same architecture as the encoder, except for the final layer, which is a linear layer with a sigmoid activation for the discriminator.

Several parameters of the architecture are constrained by the chosen patch size and stride width. We find that a patch size of  $32 \times 32$  generally exhibits the best performance, as shown in Figure 5.8, this limits both the depth and latent dimensionality of our networks. For this reason, the three networks have 2 convolutional layers with  $3 \times 3$  filters and a stride of 2. Each convolutional layer is followed by a batch normalisation layer, and a dropout layer with a rate of 5% to regularise the network. Lastly, the convolutional output is projected to a specified latent dimensionality by a linear layer.

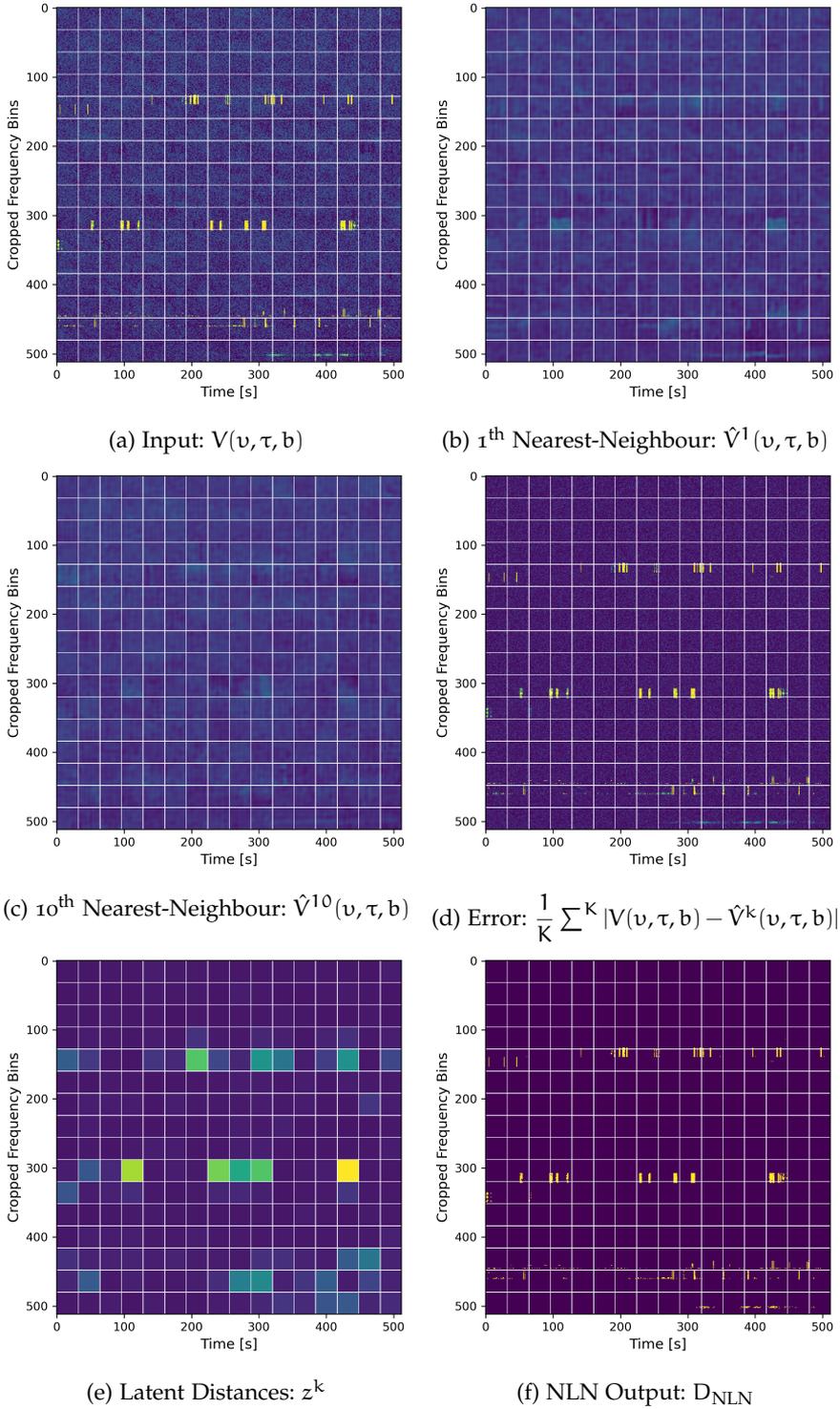


Figure 5.2: Stages of NLN-based RFI detection on the 30th sample of the LOFAR dataset. The white grid illustrates the re-composition of the  $32 \times 32$  patches to their corresponding locations in the original spectrogram. Each subplot reflects a part of the modified NLN algorithm from Equation 20.

A base number of filters of 32 is used for the AE and is increased or decreased on each subsequent layer by a factor of 2. We use *ReLU* activations for all models and they are trained for 100 epochs using ADAM [145] with a learning rate of  $1 \times 10^{-4}$ .

### 5.3 DATA SELECTION AND PREPROCESSING

Existing machine learning based approaches rely on significant amounts of labelled data for training and evaluating the models. By inverting the RFI detection problem, we do not need explicit training labels, but rather rely on the weak labels that typically come without additional cost from data archives such as the LOFAR Long Term Archive (LTA) [8]. This means that we only need very few expert-labelled examples for evaluation of our model while training on a large dataset as shown in Table 5.1.

We use two different datasets from two different telescopes to evaluate our work: simulated data from HERA [118] and calibration data from the LOFAR LTA. We use AOFlagger-based weak-labels for training all models shown in this work on both the HERA and LOFAR datasets. For evaluation with HERA we use the ground truth supplied by the simulator, whereas for the LOFAR we hand-annotate a selection of baselines obtained from the archive.

#### 5.3.1 *Simulated HERA dataset*

The HERA simulator <sup>1</sup> generates complex spectrograms from a simulated radio telescope. It uses models of diffuse sources, point sources, multiple types of RFI emissions and systematic models of the HERA telescope for parameters such as antenna cross-coupling, band-pass effects and more. Importantly for this work, the simulator gives operators a fine-grained control over the generated RFI types as well as their pixel-precise ground truth maps. Thanks to these properties, we can use the HERA data for the validation of our approach.

In this work, we simulate a hexagonal array with 14.6 m between each station, as specified by the 19 element array [118]. To synthesise our dataset, we create a 30 minute observation with an integration time of 3.52 s and bandwidth of 90 MHz from 105 MHz to 195 MHz (with 512 frequency channels). The specific integration interval is used to ensure that the resulting spectrograms are square, this is done to simplify the arithmetic of creating

<sup>1</sup> [https://github.com/HERA-Team/hera\\_sim](https://github.com/HERA-Team/hera_sim)

Dataset	# Baselines	# Training samples	# Test Samples	% RFI
HERA	28	420	140	2.76
LOFAR	2775	7500	109	1.26

Table 5.1: Attributes of each dataset used for training and evaluation. The low test-train-ratio is due to the use of weak-labels generated from classical methods, that are not directly used in the NLN training.

and reconstructing the spectrograms from patches. Furthermore, we use the default number of diffuse galactic emissions and 10 point sources as specified in [112] for the H1C observation season from 2017 to 2018. Finally, we include additive thermal noise 180 K at 180 MHz.

In the generation of our training set we individually synthesise multiple RFI emissions based on the models specified in [112]. These being narrow-band continuous emissions modelled satellite communication such as *ORB-COM*, broad-band transient emissions that imitate events such as lightning, as well as narrow-band burst RFI based on ground communication. Additionally, we include random single time-frequency *blips*.

Using the hexagonal array we simulate 28 baselines of both auto and cross correlations. We repeat this 20 times to obtain 560 complex spectrograms to train and evaluate our models with. This approach was deemed more appropriate than simulating a single long observation as performed in [112], as there is more diversity in the RFI landscape from multiple initialisations of the simulator. The simulated data has an RFI occupancy rate of 2.76% and is split into 75% for training and 25% for testing as reported in Table 5.1.

The simulated data is preprocessed before training and evaluation. For simplicity purposes, we only use the magnitude of the complex visibilities. To deal with the high dynamic range, we first clip the data between the following range  $[\mu - \sigma, \mu + 4\sigma]$  and take the natural log of the clipped spectrograms. Finally before training we standardise our data between 0 and 1 to ensure the gradients while training do not explode or vanish.

### 5.3.2 LOFAR

As previously mentioned we use publicly available data from the LOFAR Long Term Archive (LTA) [8]. Five measurements were randomly selected from 2017 to 2018 for evaluation our model. We select calibration measure-

ments that point at strong radio sources, using 51 stations in the band 120-190 MHz for 600s. The precise details of the observations are available at [94].

A common challenge in applying machine learning to radio astronomy is the amount of data generated by observations. This is especially problematic when training models with limited GPU memory. For example the 5 raw Measurement Sets (MS) used in this work are 1.7 TB in size. Therefore, in order to decrease the data-size, we use only the magnitude of the first stokes parameter and randomly sample 1500 baselines of each observation. This effectively reduces the dataset size to  $\sim 10$  GB.

For purposes of further reducing the training memory footprint of the data and simplifying the arithmetic of reconstructing the spectrograms from their respective patches, we first create approximately square spectrograms. As there are 599 time-samples per observation and 16 subbands per channel, we concatenate each 44 consecutive subbands together to form spectrograms size of  $599 \times 616$ . Additionally, we discard the first and last channel of each subband due to bandpass-effects. Finally we randomly crop the resulting spectrograms to  $512 \times 512$ . This is done, as cropping gives an equal representation of all frequency bands.

For the evaluation of the models on the LOFAR dataset, we randomly select 109 baselines for expert labelling. This is in line with the number of baselines used for evaluation in the simulator-based setting. As noted in [70, 112, 159], there are often discrepancies between the AOFlagger masks and those given by an expert, this is highlighted in Figure 5.3. For validation and evaluation of this work we treat the hand-annotated examples as the ground truth, as described in Section 5.4. Furthermore, we use the AOFlagger masks associated with the measurement sets from the LTA for training our models.

In Table 5.1 we report the percentage of RFI contamination and the dataset sizes. We ensured that the contamination is in line with what is reported in [8]. Furthermore, as our method does not rely on human-labelled examples to train, we only require a small number of expert-labelled examples to evaluate the performance of our models.

We pre-process the LOFAR data in a similar manner to the simulated data. We first clip the data between  $[\mu - \sigma, \mu + 20\sigma]$  and followed by the natural logarithm. We finally standardise the data between 0 and 1 to ensure the gradients while training do not explode or vanish.

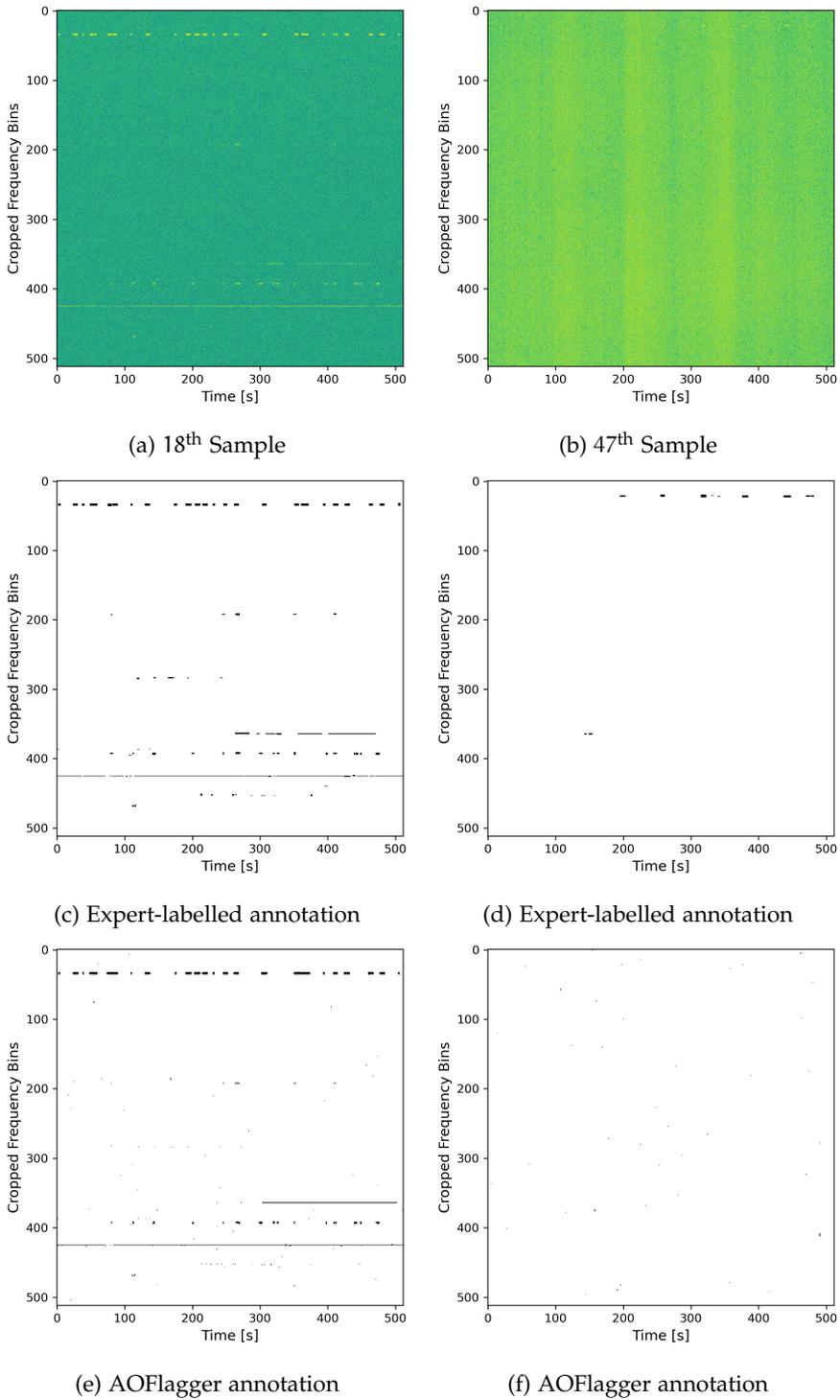


Figure 5.3: Spectrograms and their masks from the LOFAR dataset with the highest overlap between AOFlagger and expert-labelled annotations (left column) and the lowest overlap (right column), where similarity is measured using the F1-score.

## 5.4 RESULTS

We evaluate the performance of NLN applied to RFI detection experimentally on both simulated and real data from HERA and LOFAR respectively. Unlike previous works, we perform our evaluation in a two-step process, first by finding an appropriate AOFlagger threshold to generate the training annotations and then calculating the respective performance of each model on the real ground truth. To ensure the correctness of the evaluation we use the ground-truth masks from the simulator in the test-set for the HERA models and human-generated annotations for the testing of the LOFAR models.

### 5.4.1 Evaluation Methodology

Following previous efforts to evaluate the performance of RFI detection [5, 69, 70, 112], we use the three most common metrics: the Area Under the Receiver Operating Characteristic (AUROC) score, Area Under Precision Recall Curve (AUPRC) and F1-Score. The AUROC metric evaluates the ratio of True Positive Rate (TPR) and False Positive Rate (FPR) across several thresholds. In this case, the TPR is the fraction of RFI that is correctly classified as RFI and the FPR is the fraction of misclassified signals. The AUPRC metric gives the ratio of precision and recall when the model's output is evaluated across several thresholds. In this case precision refers to the fraction of correctly classified RFI across all RFI predictions, and the recall is simply the TPR. Finally, the F1 score is the harmonic mean of the precision and recall for a given threshold. For all evaluations across all models in this work the threshold is fixed to the maximum obtainable F1 score.

In the class-imbalanced scenario of RFI detection, high AUROC scores imply that a model is effective in classifying the majority class. This means that all non-RFI signals are detected as not RFI. Conversely, the AUPRC and F1 scores focus on the minority class, meaning that when AUPRC is high, the model is better at detecting RFI with a low RFI misdetection rate. Therefore, in order to maintain consistency with previous works' evaluations and to give insight into a model's performance on the both the majority and minority classes, we evaluate using both AUPRC and AUROC.

We use the AOFlagger masks to train all models in this work. In the case of LOFAR, we use the flags provided by the FLAG field of the measurement sets obtained from the LTA. For the HERA dataset, the optimal flagging strategy is determined experimentally, where each strategy is specific to

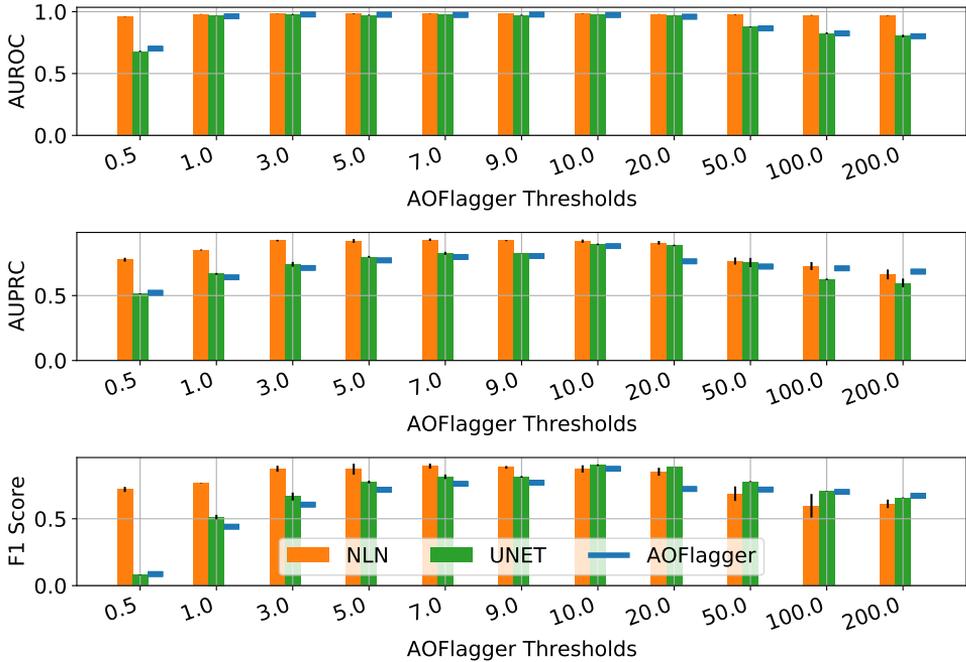


Figure 5.4: Performance of NLN, UNET and AOFlagger on the HERA data. Both U-Net and NLN are trained on the AOFlagger masks at for given starting threshold and evaluated on the simulator-ground truth. AOFlagger results are shown as only a line as a single threshold is used to evaluate the area-under-curve scores.

a particular radio telescope. As there is no pre-specified strategy for the HERA telescope, we test the HERA dataset on all available strategies for several different base-thresholds. We find that the bighorns-telescope strategy with a starting threshold of 10 to be optimal with respect to the joint-maximum of AOFlagger across AUROC, AUPRC and F1-score as shown in Figure 5.4.

For comparison with existing work we select the state-of-the-art RFI detection models. As described in Section 2, we evaluate with supervised-segmentation algorithms, based off the U-Net [101] architecture. These being, the magnitude-only U-Net for RFI detection [69, 112] as well as its residual adaptations, R-Net [70] and RFI-Net [71]. Additionally, we measure the AOFlagger on both datasets and report its performance. Furthermore to see the effect of adding some supervision to the NLN-based method, we fine tune a 2 layer CNN after the unsupervised training. This fine tuning uses the the RFI flags from the training set in the same manner as the su-

pervised architectures. We refer to this model in the subsequent sections as NLN-ft. We train and evaluate every model 3 times with a randomly initialised seed and report the mean and standard deviation for each evaluation.

For both datasets, we perform an independent coarse grid search across the hyper-parameters of NLN. These being patch size ( $n$ ), the number of latent space dimensions ( $L$ ), number of neighbours ( $K$ ), maximum latent distance threshold ( $T$ ) and discriminative training weight ( $\alpha$ ). We determine the optimal hyper-parameters based on the average maximum of AUROC and AUPRC. For LOFAR and HERA we use 16 and 20 neighbours respectively and find a patch size of  $32 \times 32$  to be optimal for both. Furthermore, the latent dimension size of 8 is used for HERA whereas we find 32 dimensions to be optimal for LOFAR. We threshold the latent distance vector at its 66<sup>th</sup> percentile for LOFAR and the 10<sup>th</sup> percentile for HERA. Finally, we find that the optimal discriminative hyper-parameter  $\alpha$  is 0.6 for both the HERA and LOFAR datasets. For all other models, we use all parameters specified by the authors other than patch size, which we fix to  $32 \times 32$  in order to keep comparison consistent.

The difference in the hyper-parameters between the datasets is due to the increased complexity of the real LOFAR data relative to the simulated HERA data. This complexity is due to stochastic instrumentation effects, ionospheric artefacts, increased dynamic range and many more. Therefore, the autoencoder requires a larger latent dimensionality to better represent this increased complexity. Similarly, an increased latent distance threshold is used to mitigate the elevated reconstruction error-based noise.

Finally, to validate the suitability of the NLN algorithm for anomaly detection-based RFI detection we compare it against three commonly used anomaly detection techniques. In this case we consider Deep K-Nearest Neighbours (DKNN) [30], autoencoding models with an L2 loss [23] and SSIM-based loss [84].

#### 5.4.2 HERA Results

In Figure 5.4 we illustrate the performance sensitivity of the magnitude-based U-Net and NLN when modifying the AOFlagger starting threshold for the bighorns-telescope strategy. It is clear NLN is less sensitive to changes in the validity of the training data, exhibiting little variation relative to the AOFlagger-based masks. This is because the model is not directly trained on the AOFlagger-based masks. However UNET's perfor-

Model	AUROC	AUPRC	F1-Score
AOFlagger [154]	<b>0.9692</b>	0.8946	0.8899
U-Net [159]	0.9544 $\pm$ 0.0097	0.9088 $\pm$ 0.0109	0.9269 $\pm$ 0.0058
RFI-Net [71]	0.9486 $\pm$ 0.0054	0.8908 $\pm$ 0.0100	0.9231 $\pm$ 0.0045
RNet [70]	0.9614 $\pm$ 0.0061	0.8541 $\pm$ 0.0282	0.7870 $\pm$ 0.0635
NLN (ours)	0.9682 $\pm$ 0.0018	<b>0.9308 <math>\pm</math> 0.0024</b>	<b>0.9350 <math>\pm</math> 0.0017</b>

Table 5.2: Performance of RFI detection models on the simulated HERA dataset when trained using the AOFlagger annotation at a threshold of 10 and evaluated on the ground truth from the simulator. We do not report standard deviation of AOFlagger as it is deterministic. Best scores in bold.

mance is shown to be more dependent on the accuracy of the AOFlagger masks, indicating that it indeed is over-fitting to the training labels.

For low thresholds ( $> 5$ ) the training data is over-flagged, meaning that all RFI is flagged but large percentage of the astronomical data is as well. Conversely, for high thresholds ( $< 10$ ) the data is under-flagged, meaning that some RFI is not flagged. From this intuition, it is clear that the NLN algorithm is less sensitive to over-flagging, but its performance deteriorates when the training data is under-flagged. For the experiments using the simulated HERA dataset in the remainder of this paper, we fix the AOFlagger threshold to 10, as it gives optimal flagging performance.

In Table 5.2 we show the performance of NLN for RFI detection relative to the current state-of-the-art machine learning models. It is clear that the NLN offers superior performance across all metrics other than AUROC demonstrating the success of our approach to use clean data for training. We consider the slight decrease in AUROC (less than 0.001%) acceptable considering that it is strongly impacted by the class imbalance prevalent in the RFI data.

In Figure 5.5 we demonstrate an interesting unintended consequence of our work in the HERA setting. Here, we train each model on the dataset but excluding a particular type of RFI, then during testing we expose the model to the unseen RFI type. This paradigm effectively evaluates how well models generalise to Out-Of-Distribution (OOD) RFI. As the training process of NLN excludes all RFI (OOD or not) from the training set, it does not learn explicit models of the RFI, in effect our method significantly

outperforms both supervised and classical methods across all metrics. This is also clearly demonstrated by the variance across each experiment.

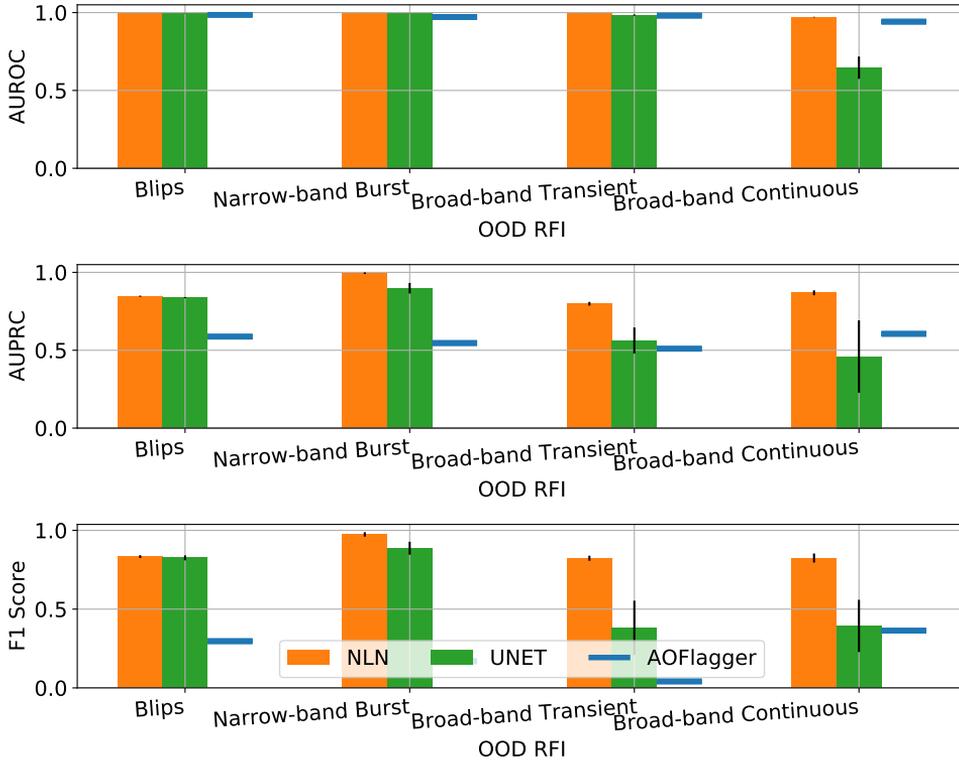


Figure 5.5: Out-Of-Distribution (OOD) RFI detection performance of the NLN, U-NET and the AOFlagger. As the RFI is excluded in the training process of NLN, the method is unaffected by OOD effects. AOFlagger results are shown as only a line as a single threshold is used to evaluate the area-under-curve scores.

### 5.4.3 LOFAR results

In Table 5.3 we show the performance of NLN relative to the state-of-the-art on the LOFAR dataset. Here we see a similar trend to the HERA-based results; NLN offers superior performance in terms of AUPRC. However in terms of AUROC and F1-score, NLN-ft offers best performance. The decrease in F1 score is due to the NLN algorithm yielding more false negatives when thresholding all predictions with a single threshold. This is attributable to the large fluctuations in power of the RFI in the LOFAR

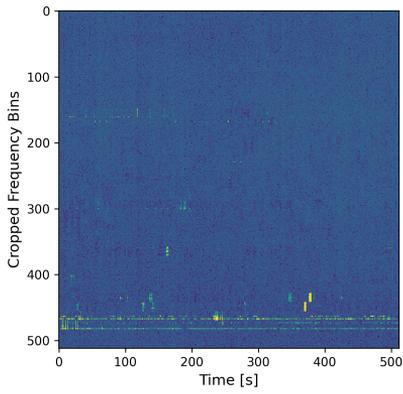
Model	AUROC	AUPRC	F1-Score
AOFlagger [154]	0.7883	0.5716	0.5698
U-Net [159]	0.8017 $\pm$ 0.0058	0.5920 $\pm$ 0.0031	0.5876 $\pm$ 0.0031
RFI-Net [71]	0.8109 $\pm$ 0.0037	0.5991 $\pm$ 0.0038	0.5979 $\pm$ 0.0012
RNet [70]	0.8301 $\pm$ 0.0084	0.5495 $\pm$ 0.0145	0.5286 $\pm$ 0.0195
AE-L2 [23]	0.8397 $\pm$ 0.0019	0.3933 $\pm$ 0.0036	0.4491 $\pm$ 0.0007
AE-SSIM [84]	0.7748 $\pm$ 0.0046	0.3913 $\pm$ 0.0186	0.4801 $\pm$ 0.0115
DKNN [30]	0.8285	0.0704	0.1528
NLN (ours)	0.8622 $\pm$ 0.0006	<b>0.6216 <math>\pm</math> 0.0005</b>	0.5114 $\pm$ 0.0004
NLN-ft (ours)	<b>0.8833 <math>\pm</math> 0.0118</b>	0.5384 $\pm$ 0.0121	<b>0.6075 <math>\pm</math> 0.0109</b>

Table 5.3: Performance of RFI detection models on the real LOFAR dataset when trained using the AOFlagger annotations from the LTA and evaluated on the expert-labelled ground truth. We do not report AOFlagger and DKNN standard deviations as they are deterministic. Best scores in bold.

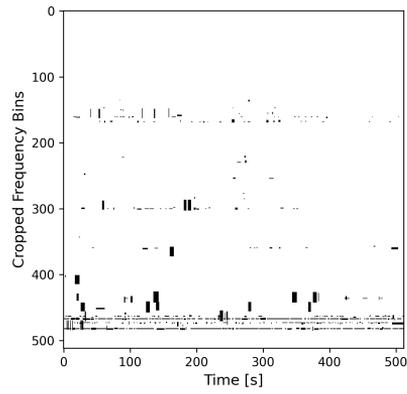
dataset in combination with the reconstruction-error term of the NLN RFI detector. When the RFI power is low, the reconstruction error will be of a low amplitude and NLN will produce predictions with low power. In effect, when these low power predictions are thresholded using the same threshold as the high power RFI (in this case one which maximises F1-score), the outputs have more false negatives. This is in contrast with the HERA dataset, that has RFI with a consistently higher power level than the astronomical and system-based signals. It can be seen that by adding a small amount of supervision to the overall method, we can improve the overall performance. An illustration of these effects and a comparison between the models is shown in Figure 5.6.

Furthermore, in the bottom half of Table 5.3 we compare NLN with commonly used anomaly detection methods. Here we see that NLN significantly outperforms these methods on all metrics. We attribute the performance improvement to the combination of reconstruction error and latent error.

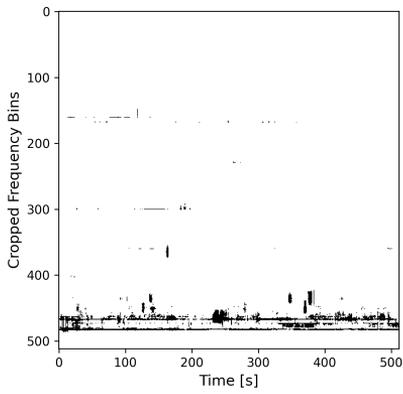
Additionally, to test the model’s reliance on dataset size, we evaluate each model on a percentage of the training data. We show in Figure 5.7 that NLN is less sensitive to reductions in training data-size, performing almost uniformly with even with large decreases in training data-size. Conversely, the supervised methods’ performance scales asymptotically with dataset size,



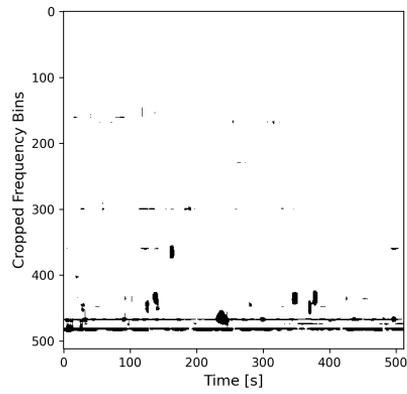
(a) Input



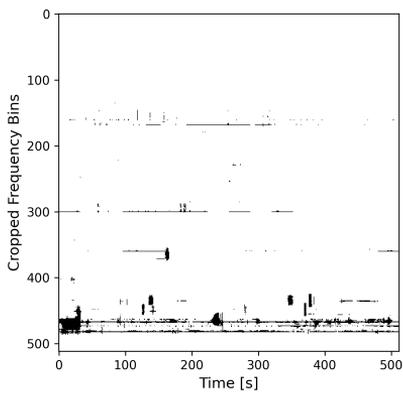
(b) Mask



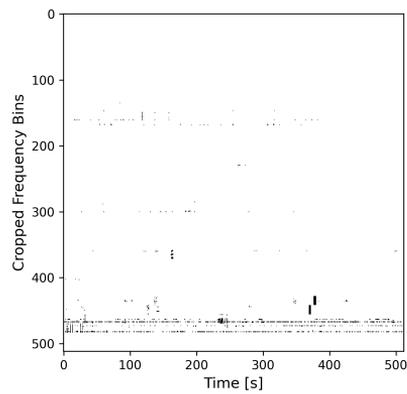
(c) U-Net Prediction



(d) RNet Prediction



(e) RFI-Net Prediction



(f) NLN Prediction

Figure 5.6: The 95<sup>th</sup> sample from the testing set, its corresponding mask and the predictions of each model.

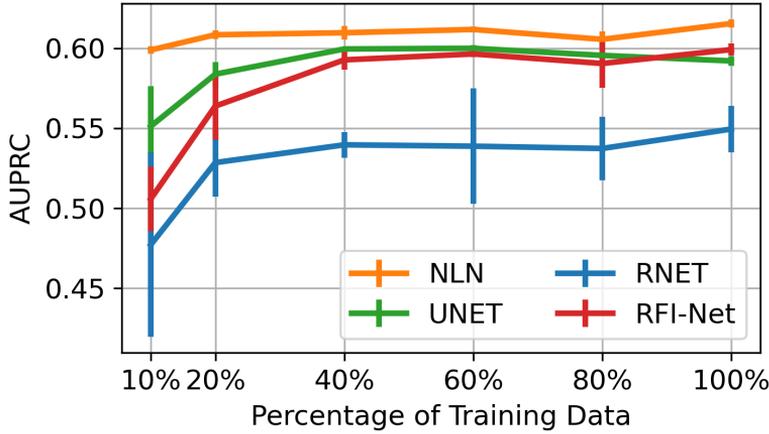


Figure 5.7: AUPRC performance of each model when training on a percentage of the original LOFAR dataset and evaluating on the original LOFAR test dataset.

exhibiting significantly higher variance in their performance with smaller dataset sizes. We associate the asymptotic-scaling and increased variance with both the supervised model’s larger capacity (due to residual and skip connections) as well as the diversity of the RFI landscape.

To determine the sensitivity of the parameters of the NLN-algorithm, we perform a course grid-search of its hyper-parameters. We search across the number of latent dimensions, patch size and number of neighbours as illustrated in Figure 5.8. In order to better visualise the 4-dimensional space, we plot cross-sections of the high-dimensional landscape. First we fix the number of neighbours to 16 as shown in Figures 5.8 (a), 5.8 (b) and 5.8 (c). It can be seen that the optimal number of latent dimensions is 32, with respect to the average maximum of AUPRC and AUROC. We then set the optimal number of latent dimensions and plot the effect of varying the number of neighbours in Figures 5.8 (d), 5.8 (e) and 5.8 (f). Through this, we determine the optimal number of neighbours to be 16 with respect to the average maximum AUROC and AUPRC. We conclude that NLN gives optimal performance when the number of latent dimensions are 32, the patch size is  $32 \times 32$  and the number of neighbours is 16.

Finally, as a consequence of our selection RFI-free selection algorithm, we find that NLN requires 66% less data in comparison to its supervised counterparts. This amount is naturally dataset dependent, however we expect that due to the reduced training data there will be less compute time, and less power consumption while training.

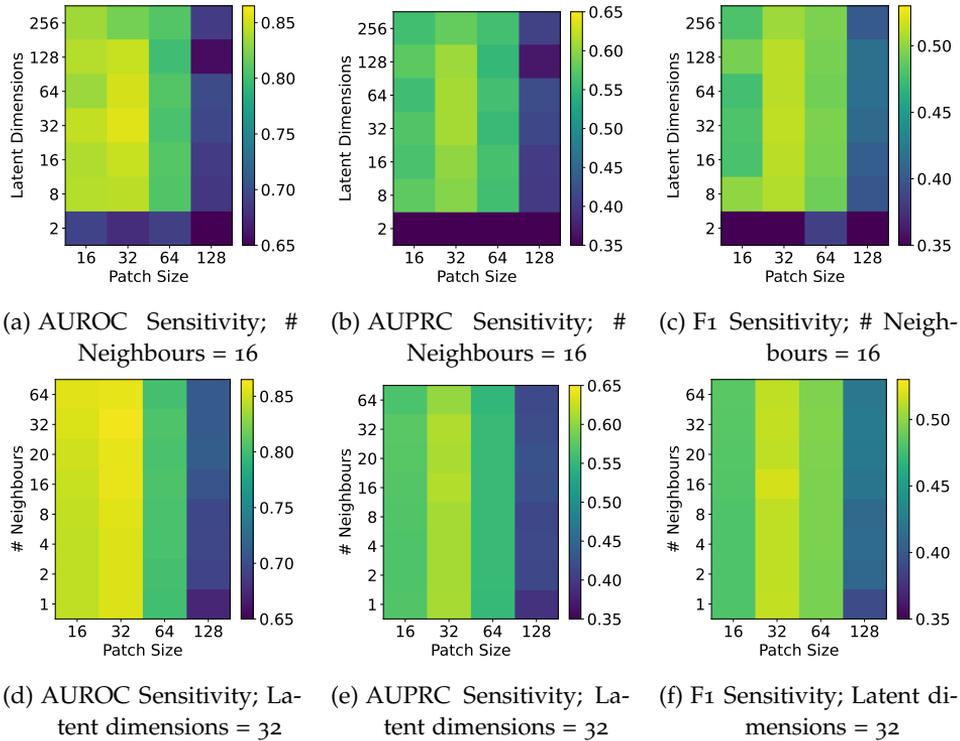


Figure 5.8: Sensitivity of the hyper-parameters on the LOFAR-based performance of NLN when varying latent dimensionality, patch size, and number of neighbours. In order to visualise the four-dimensional space, the number of neighbours is fixed to 16 in the top row, whereas in the bottom row we fix the number of latent dimensions to 32. The optimal parameters for the LOFAR dataset are a patch size of  $32 \times 32$ , 32 latent dimensions and 16 neighbours.

## 5.5 CONCLUSIONS

RFI detection is an increasingly important research topic for radio astronomy. State-of-the-art solutions to the RFI problem have been based on supervised machine learning techniques, which fail to address the prohibitive cost of labelling astronomical data. In this work we have documented how inverting the detection problem effectively addresses this issue. We have shown that NLN provides better than state-of-the-art RFI detection without incurring the cost of labelling.

Furthermore, we have demonstrated that our method better generalises to unseen RFI, whereas current supervised approaches over-fit to weak-label-based RFI masks. As a consequence, we hypothesise that our approach will better generalise to future generations of emitters, whereas existing supervised methods will have to be regularly retrained. Additionally, we find that due to our training patch selection process we need less data for training, hence decreasing both the training time and energy cost.

Overall, there is sufficient evidence in several other domains that suggests supervised segmentation algorithms will outperform unsupervised approaches given sufficient high-quality labels. However, in the current state of the RFI detection domain, where there are still few labelled datasets available and the high cost of obtaining labelled data, we propose the inverted approach as the way forward.

We plan to further improve the performance of NLN applied to RFI detection through additional training priors. Contrastive self-supervised learning is a candidate solution thanks to its ability to generate more robust latent representations that can be leveraged in the NLN-algorithm. Furthermore, in order to improve the increased false negative rate of NLN on the LOFAR dataset we suggest more research to be done into automated processing schemes to deal with the high dynamic range of astronomical data when training unsupervised models. This could be additionally improved using a hybrid approach through `SumThreshold` or trying to directly predict a threshold on a per patch-basis. Finally, we plan to extend this work to more general anomaly detection based problems within radio astronomy.

Overall we have shown compelling use-case for machine learning anomaly detection in radio astronomy. We note that RFI detection is not a holistic solution to system health management in radio telescopes. However, it does offer insight into a singular aspect of it. In subsequent chapters we use the lessons learnt propose a solution to the complete system health management problem for observatories.

# 6

## THE RADIO OBSERVATORY ANOMALY DETECTOR

---

*In this chapter we provide an end-to-end model for anomaly detection in radio observatories called the Radio Observatory Anomaly Detector (ROAD). We created this model using the building blocks and principles described in the preceding chapters. Here we also introduce a new dataset specifically created for the validation of anomaly detection models for system health management in radio telescopes. This chapter answers all four research questions culminating in a real time system for system health management for the LOFAR telescope.*

This Chapter is based on:

- **Michael Mesarcik**, Albert-Jan Boonstra, Marco Iacobelli, Elena Rangelova, Cees de Laat, and Rob V. van Nieuwpoort, "*The ROAD to discovery: machine learning-driven anomaly detection in radio astronomy spectrograms* " [160], *Accepted for publication in Astronomy & Astrophysics*
- **Michael Mesarcik**, Albert-Jan Boonstra, Marco Iacobelli, Elena Rangelova, Cees de Laat, and Rob V. van Nieuwpoort [161], "*Dataset for The Radio Observatory Anomaly Detector*", in *Dataset on Zenodo*

### 6.1 INTRODUCTION

Radio telescopes are getting bigger and are generating increasing amounts of data to improve their sensitivity and resolution [8, 51, 52, 108]. The growing system size and resulting complexity increases the likelihood of unexpected anomalies occurring thereby resulting in datasets that contain erroneous data. These anomalies include failures in instrument electronics, miscalibrated observations, environmental anomalies such as lightning, as-

tronomical effects like solar storms as well as problems in data processing systems among many others. We consider Radio Frequency Interference (RFI) unavoidable, therefore do not consider it an anomaly in this context. Currently, efforts to detect and mitigate these anomalies are performed by human operators, who manually inspect intermediate data products to determine the success or failure of a given observation. The accelerating data-rates coupled with the lack of automation, results in operator-based data quality inspection becoming increasingly infeasible [81].

Telescope operators have successfully automated many scientific data processing pipelines from calibration [40] to Radio Frequency Interference (RFI) mitigation [41], imaging [162–164] and de-dispersion [42, 43]. Additionally, continuous effort is being made to create high-performance real-time algorithms, to improve the quality and reliability of the scientific data [13, 44–47]. As of yet, there have been no attempts to fully automate the System Health Management (SHM) pipeline, and by virtue of the lack of work on this topic, no real-time implementations exist. This is in part due to the complexity of the challenge as well as the unavailability of SHM specific datasets. Furthermore, the successes of SHM-based anomaly detection systems have been extremely impactful in fields ranging from industrial manufacturing [23] to space craft system health [165, 166] thereby motivating this study.

The exponential growth of data production from modern instruments have made data-driven techniques and machine learning appealing to astronomers and telescope operators. However efforts in machine learning based anomaly detection are concentrated in scientific discovery rather than SHM, with approaches ranging from detecting unusual galaxy morphologies [98, 99] to identifying new transients [82, 95, 97]. Unfortunately, these techniques are not directly applicable to the multi-station autocorrelation-based spectrographic data obtained from radio observatories, due to increased data complexity, high dynamic range due to RFI, varying observation durations and frequency ranges as well as the feature compounding problem [81]. It must be noted this work makes use of up-stream data products in the form of spectrograms, that are produced by all radio telescopes thereby enabling its applicability to other instruments such as the Square Kilometre Array (SKA) [9].

The SHM anomaly detection problem differs from existing work for several reasons. Firstly, the data inspection performed by telescope operators involves analysing both known and unknown anomalies; where known anomalies should be classified into their respective classes and unknown

anomalies should be differentiated from all other existing classes. This is in contrast with typical anomaly detection which is normally posed as one-class-classification problem. Furthermore, we find that class imbalance not only exists between the normal and anomalous classes (which is common for anomaly detection), but there is also strong imbalance between the anomalous classes. For these reasons, we propose a new framework for detecting and classifying SHM-based anomalies, that is capable of distinguishing both regularly occurring and rare events.

We find the multi-class classification approach more appropriate as it gives more flexibility to telescope operators. This is because the anomalousness of particular events entirely depend on the context of the science-goals of an observation. For example, in observations relating to the Epoch of Reionisation (EoR) [167], Signal-to-Noise Ratio (SNR) is a huge concern, as such any high power *anomalies* such as solar storms should be identified and removed from the observation. In contrast, for solar physics-based observations [168] the high-power solar events should implicitly be kept within the data and should not be flagged as anomalous. Therefore by including a classification step within the anomaly detector system we offer greater flexibility to telescope operators in data quality inspection.

In this chapter we make the following contributions: (1) a new dataset consisting of 6708 manually labelled autocorrelation based spectrograms consisting of 10 different feature classes; (2) a self-supervised learning (SSL) framework that is effective in learning representations of time-frequency data with a high-dynamic range; (3) an anomaly detection framework that can classify both commonly occurring known anomalies and detect unknown anomalies with a high precision, and (4) we show that our implementation achieves real-time performance for LOFAR.

In the remaining part of the chapter we document our data selection strategy and outline the labelling process used for evaluation of this work in Section 6.2. In Section 6.3 we show the proposed SSL and anomaly detection frameworks. Finally, our results and conclusions are documented in Sections 6.4 and 6.5.

## 6.2 DATASET

We created a new dataset for anomaly detection in radio observatories and document the data selection, preprocessing and labelling strategy used in this section. Applying machine learning to radio astronomical datasets poses a significant challenge, particularly when using time-frequency data.

Methods for data preprocessing and selection need to be carefully considered, due to issues such as high-dynamic range (from RFI among other events), combining thousands of stations for a single observation, having complex-valued data with multiple polarisations, feature compounding and many more. An additional challenge with applying machine learning to radio astronomy is the lack of labelled time-frequency datasets from radio telescopes as well as the availability of expert knowledge and the cost associated with creating a dataset.

### 6.2.1 *Observation selection and preprocessing*

The ROAD dataset is made up of observations from the Low Frequency Array (LOFAR) telescope [8]. For more information on the LOFAR telescope and its data processing systems refer to Chapter 2.1.

Deep learning architectures typically require equally-sized inputs, however LOFAR observations can have a varying number of time samples and/or frequency bands. Therefore, additional resizing of the intermediate visibilities is done by resizing all observations to (256, 256) bins in time and frequency. This means that observations with fewer than 256 time samples are interpolated and those with more are down-sampled. Furthermore, as the autocorrelations contain no phase information, we use only the magnitude component of each spectrogram.

It must be noted that this processing does modify the morphologies of certain features, particularly those present with a low time resolution. However as this preprocessing step is consistent across all spectrograms, the overall effects on the anomaly detector and classifier are negligible. In future work, we plan to associate the labels with the full resolution LOFAR data from the Long Term Archive (LTA)<sup>1</sup> and apply it to (256, 256) crops of the full resolution spectrograms.

We selected 110 observations from the LOFAR LTA comprising of a broad set of science use cases and the corresponding observing setups. Of the selected observations, we use the autocorrelations from 2431 LBA stations and 4277 HBA stations from an observation period between 2019 and 2022.

### 6.2.2 *Labelling methodology*

The ROAD dataset contains 10 classes which describe various system-wide phenomena and anomalies from data obtained by the LOFAR telescope.

---

<sup>1</sup> <https://lta.lofar.eu/>

Category	Description	Band	Polarisation	Occurrence rate	# Samples
Normal	All non-characterised effects	Both	All	-	4687
<b>Data processing</b>					
First order data loss	Data loss from consecutive time and/or frequency channels	Both	All	0.02	146
Second order data loss	Data loss from single frequency and/or single time channels	Both	All	0.04	283
<b>Electronic systems</b>					
High noise element	High power disturbances caused by miscellaneous events	Both	All	0.01	88
Oscillating tile	Amplifier going into oscillation	High	All	0.01	56
<b>Astronomical events</b>					
Source in side-lobes	A-team source passing through side-lobes	High	All	0.06	446
Galactic plane	Galactic plane passing through the main lobe of the antenna	Both	Cross	0.08	550
Solar storm	Strong emissions from the sun	Low	All	0.02	147
<b>Environmental effects</b>					
Lightning	Lightning storm	Both	All	0.06	389
Ionospheric RFI reflections	RFI reflected from the ionosphere	Low	All	0.04	261

Table 6.1: Categorisation of data processing, electronic, astronomical and environmental anomalies in the ROAD dataset. Where A-team sources refer to the four brightest persistent radio sources in the northern sky. Note each spectrogram may contain multiple anomalies, hence the number of samples stated is greater than the overall dataset size.

These classes are categorised into 4 groups: data processing system failures, electronic anomalies, environmental effects, and unwanted astronomical events. Table 6.1 shows the classes used as well as the description of events, their band and polarisation in which they occur. We note that the term *anomaly* is used liberally in this context, while low power effects (that are only present in the cross polarisations) such as the galactic plane passing through an observation are somewhat unavoidable. Nonetheless, for observations with extremely low SNR such as The Epoch of Re-ionisation of the Universe (EoR) [167], the galactic foreground signals need to be identified and removed. For this reason, we include such events in the ROAD dataset. Furthermore, as the ROAD dataset was created using data mainly from the period around the minimum of the past Solar cycle, the high-band data are marginally impacted while the low-band data tracks these events due to the frequency dependence of the related emissions. This also explains why the dataset does not contain classes consisting of ionospheric scintillation effects. In future work, we plan to extend the dataset to consist of classes relating to more ionospheric disturbances.

Our labelling approach took into consideration anomalies which occurred at both the station- and observation-levels. For example, events such as lightning storms and high-noise events can look fairly similar, especially in the down-sampled context. However, lightning storms are geographically-bound to affect all stations in a certain region therefore only occurring at the station-level. Additionally, lightning is highly correlated across stations in time, with minimal delay between the recorded events in each station. Whereas high-noise events usually affect only a single antenna at a time with no time dependency between antennas and stations. By this logic, all stations bound to the same geographic location with broadband high power events across all polarisations that are correlated in time were considered to be corrupted by lightning storms, whereas individually affected stations were labelled as high-noise events.

We make a distinction between first and second order events; for example the first order data-loss event corresponds to dropped information from consecutive time samples and/or frequency bands and second order is for a single time sample/frequency band. We find this a useful distinction as the root cause of these events is different. In the case of first order data loss events, the problem can be traced to the correlator pipeline, whereas the second order events are most-likely from to type conversion overflows due to strong RFI. Additionally, we note some overlap between class labels, for example it is common for a high power noise events to trigger instability in

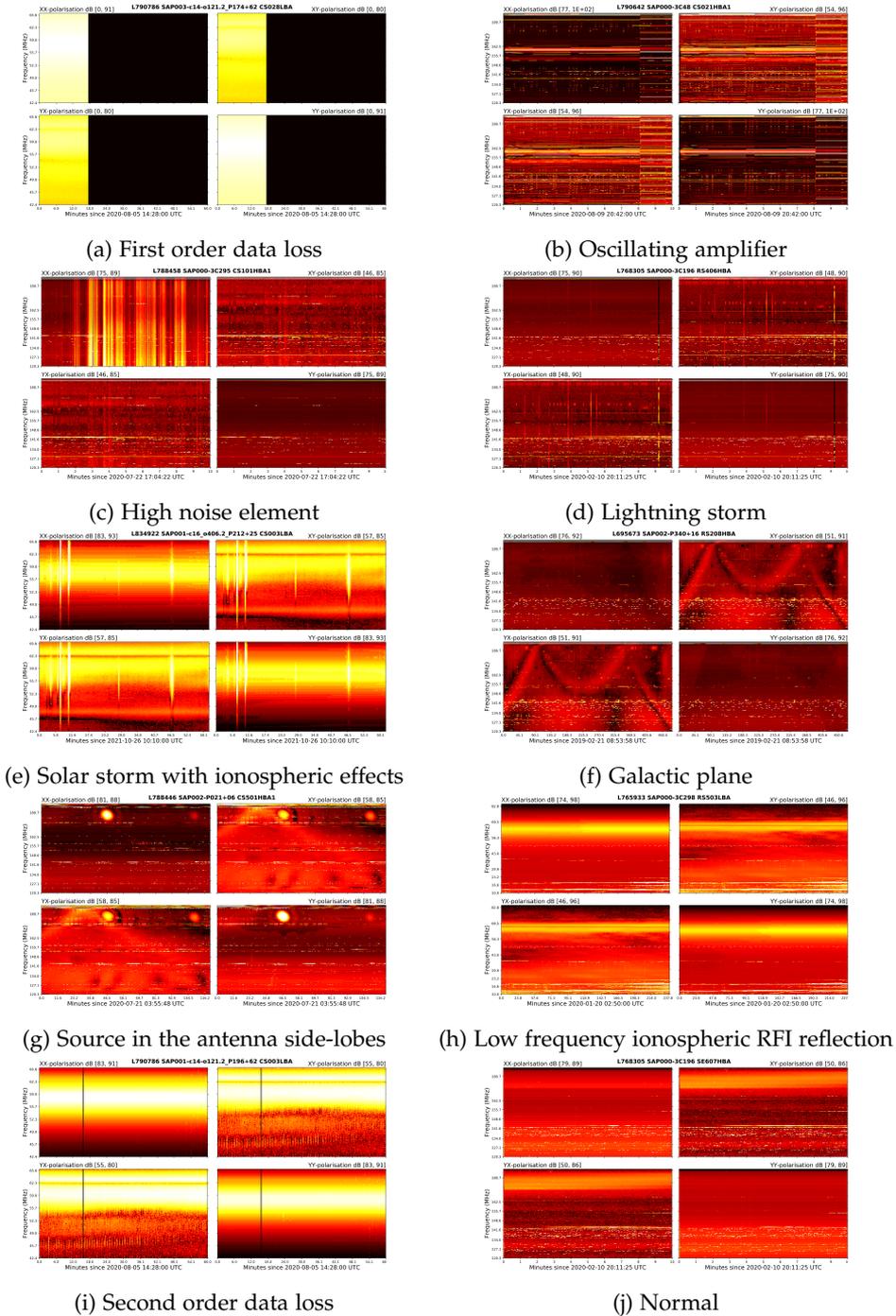


Figure 6.1: Illustration of 10 examples from the ROAD dataset.

an amplifier causing it to oscillate, however the point at which a high noise event becomes an oscillating tile is often unclear, therefore making it hard to distinguish these events from each other.

We labelled the dataset using LOFAR observations that were down-sampled and preprocessed as described in Section 6.2.1. We made multiple train-test splits during experimentation to ensure consistent performance across models. Furthermore, the ROAD dataset is publicly available<sup>2</sup>. The file is in the hdf5 format and consists of fields corresponding to the raw data, labels, frequency bands information, station name and source observation. Figure 6.1 illustrates 8 of the 10 classes labelled in the available dataset for brevity.

### 6.2.3 *Class imbalance*

Due to the nature of anomaly detection, the number of normal samples greatly outnumbers anomalous ones. In the case of the ROAD dataset, and the LOFAR telescope more generally, we find there is not only a class imbalance between normal and anomalous classes but also among the anomalous classes. For example, commonly occurring astronomical signals, such as the galactic plane, are far better represented in the observations than unlikely events like the amplifiers oscillating. Practically, this means that when we separate the samples into testing and training sets we need to also maintain the same occurrence rates with respect to the rates in the original dataset.

In effect, we down-sample the testing data such that the occurrence rate (shown in the second last column of Table 6.1) is maintained for evaluation. This means that each model needs to be tested multiple times with new samples taken from the testing pool of anomalous samples to effectively evaluate its performance. We evaluate each model 10 times with different random seeds to ensure accurate reporting.

### 6.2.4 *Radio frequency interference considerations*

As previously mentioned, we consider Radio Frequency Interference (RFI) to be unavoidable thus deeming it a normal class. A key problem with using the RFI-masks associated with the spectrograms from LOFAR (as done in Mesarcik et al. [94]) is that they are generated using A0FLagger. A0FLagger is a SumThreshold-based algorithm which indiscriminately flags all high power events as interference. In the context of this work, such an approach would result in many of the high power anomalies such as lightning, solar

---

<sup>2</sup> <https://zenodo.org/record/8028045>

storms, high-noise events and oscillating tiles being flagged as RFI. In effect, if we were to blank the RFI before training our models, we would likely remove many of these high-power features, thereby decreasing the efficacy of the model. In turn making our end goal of mitigating the anomalies more difficult, since different classes should lead to different actions by the telescope operators. For this reason, we do not use the RFI masks associated with the spectrograms.

### 6.3 THE RADIO OBSERVATORY ANOMALY DETECTOR

As outlined in preceding sections ROAD is designed to detect both previously unseen system behaviours as well as to classify known-anomalies observed by the LOFAR telescope. To accommodate these requirements, we find it necessary to combine two approaches; supervised classification, as well as self-supervised anomaly detection. This section outlines the motivations and design decisions made for the implementation of ROAD.

#### 6.3.1 Problem formulation

Given the  $i^{\text{th}}$  spectrogram  $V_i(\nu, \tau, b, p)$  from the dataset and model  $m$  with parameters  $\theta_m$ ; we would like to predict whether an anomaly is present and which class it belongs to, if it is a known event, such that,

$$m_{\theta_m}(V_i(\nu, \tau, b, p)) = \begin{cases} 0 & , \text{if normal} \\ [1, N] & , \text{if known anomaly} \\ N + 1 & , \text{if unknown anomaly} \end{cases} \quad (22)$$

where  $\nu$ ,  $\tau$ ,  $b$  and  $p$  are the indexes corresponding to frequency band, time sample, baseline and polarisation, respectively and  $N$  is the number of known anomaly classes. Supervised approaches assume that each class is represented in the training set and try to minimise the following loss function

$$\mathcal{L}_{\text{sup}} = \min_{\theta_m} \sum_i \mathcal{H}(m_{\theta_m}(V_i(\nu, \tau, b, p)), l_i) \quad (23)$$

where  $\mathcal{H}$  is an entropy-based measure of similarity and  $l$  is the encoded vector of labels corresponding to the contents of  $V$ . During inference, the

supervised classifier produces estimate of which classes are most probable in a given spectrogram, and the *argmax* function selects the most likely classification as shown in the bottom half of Figure 6.3. However, as illustrated by the results in Section 6.4, the performance of such a supervised classifier severely deteriorates when exposed to unseen or out of distribution (OOD) classes during testing. To remedy this, we disentangle the two model objectives, namely we use a supervised classifier to identify the known classes present in the training set and a self-supervised anomaly detector to classify unseen anomalies.

### 6.3.2 Self supervised representation learning

Self Supervised Learning (SSL) methods learn useful feature representations by training on secondary objectives called *pretext tasks*, so that once trained, the model weights can be utilised for downstream applications. We define two pretext tasks that allow the model to learn useful representations for anomaly detection in astronomical data: context-prediction and reconstruction error. Context-prediction is a pretext task that makes a model classify the positional relationship between two patches taken from the same image. The two patches are projected to some latent representations  $z_0$  and  $z_1$  using a backbone network  $f$ , while keeping track of their position label,  $c$ , on a  $3 \times 3$  grid as proposed by [76]. Then, using  $g$ , a 2-layer Multi-layer Perceptron (MLP), we classify the positional relationship from the latent representations; as given by

$$\mathcal{L}_{\text{con}} = \sum_i \sum_j \mathcal{H}(g(z_{i,j,0}, z_{i,j,1}), c_j) \quad (24)$$

where  $i$  corresponds the the index of each spectrogram,  $j$  is the index of each context-prediction pair in a single spectrogram and  $c_j$  is the positional label. Additionally, to ensure the model does not learn positional relationships based purely on the bordering values of each patch, we augment each neighbour in the training process. In the implementation, we randomly crop the patches between 100% and 75% of their original size followed by resizing them to their original dimensions. We illustrate the context prediction loss and patch selection in Figure 6.2.

Furthermore, to enforce consistency across the representations of similar looking patches we use reconstruction error. Reconstruction error maintains consistency by ensuring that two patches which share common features

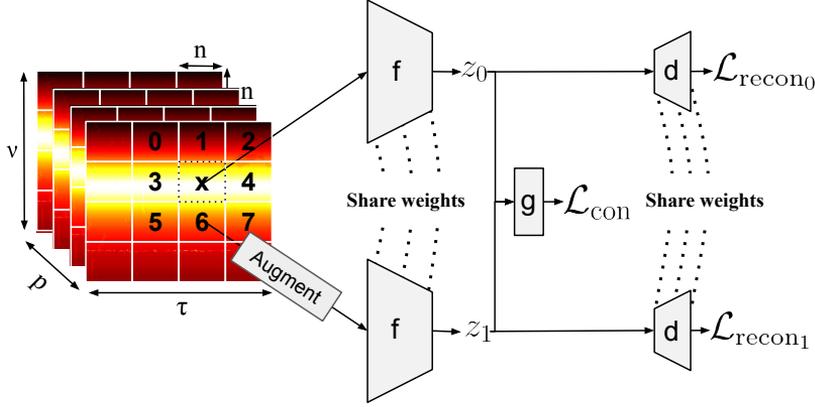


Figure 6.2: Illustration of the self-supervised training procedure used in ROAD, we use random-cropping for augmentation.

in visibility space, should occupy nearby locations in the latent space and, therefore, should be reconstructed similarly. The reconstruction loss is given by

$$\mathcal{L}_{\text{recon}} = \sum_i \sum_j |V_{i,j,0} - d(z_{i,j,0})| + |V_{i,j,1} - d(z_{i,j,1})| \quad (25)$$

where  $d$  is a de-convolutional decoder that should have significantly fewer parameters than the backbone network  $f$ . We do this to ensure that the model has more capacity to learn suitable representations instead of prioritising reconstruction. For completeness we represent the full SSL learning objective as

$$\mathcal{L}_{\text{SSL}} = \lambda \mathcal{L}_{\text{con}} + (1 - \lambda) \mathcal{L}_{\text{recon}} + \lambda_{\text{reg}} \sum_i \sum_j (z_{i,j,0}^2 + z_{i,j,1}^2) \quad (26)$$

where  $\lambda$  is a hyper-parameter which changes the influence of each component of the loss. Additionally, we use regularisation in the form of minimising the square size of the latent projections  $z$ . Regularisation is used in order to enforce the most compact representations in  $z$ . We experimentally select  $\lambda = 0.5$  and  $\lambda_{\text{reg}} = 1 \times 10^{-6}$ , and illustrate  $\lambda$ 's impact in Section 6.4.

### 6.3.3 Distinguishing normal from anomalous samples

Although we have described a method for learning representations of normal data, the model is incapable of accurately discriminating between nor-

mal and anomalous samples. Several options exist for anomaly detection when utilising the learnt representations of normal training data methods. The simplest involves measuring the distance between a given sample and the normal training data [30] using a K-Nearest-Neighbour (KNN) lookup. This assumes that larger distances correspond to more anomalous samples. However as we already make use of some of the labelled data for the supervised classifier we find it beneficial to fine-tune a shallow MLP on top of SSL representations to perform anomaly detection. As the SSL-backbone learns representations on the patch-level and ROAD dataset labels are on the spectrogram-level, we first need to concatenate the latent representations of each patch to return to the correct dimensionality before training the MLP. Notably, we propagate the gradients during fine tuning through both the MLP and the backbone network  $f$ , such that the distance between normal and anomalous representations at the spectrogram-level are consolidated. We show in Section 6.4 that fine-tuning dramatically outperforms random initialisation and KNN-based anomaly detection. Furthermore, we find that using the fine-tuned approach dramatically improves the time-complexity of the system.

Additionally, we need to determine how to threshold the anomaly scores produced by either the fine-tuned models or the KNN-distance based approach. Here we utilise the threshold from the Area-Under Precision Recall Curve (AUPRC) which results in the maximum F- $\beta$  score. A discussion on the evaluation metrics used can be found in Section 6.4 as well as the results pertaining to change of this threshold can be found in Figure 6.10.

#### 6.3.4 Combining classification with anomaly detection

The final consideration when constructing ROAD is how to effectively combine the fully supervised classifier  $y_{\text{sup}} \in [0, N]$  and the fine-tuned anomaly detector  $y_{\text{ssl}} \in [0, 1]$ . Simply, we consider normal predictions from the detector more likely to be correct, and if there is a disagreement between the two models then we flag the sample as an unknown class of anomalies that the classifier may have not seen.

The overall method is shown in Figure 6.3 and is summarised by

$$y = \begin{cases} 0 & , \text{if } y_{\text{ssl}} = 0 \\ y_{\text{sup}} & , \text{if } y_{\text{ssl}} = 1 \text{ and } y_{\text{sup}} \neq 0 \\ N + 1 & , \text{if } y_{\text{ssl}} = 1 \text{ and } y_{\text{sup}} = 0 \end{cases} \quad (27)$$

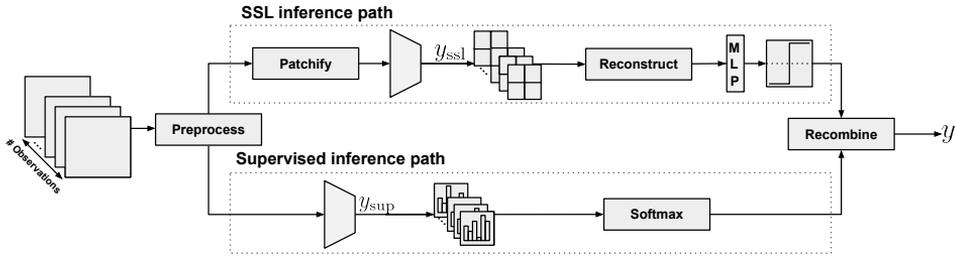


Figure 6.3: Illustration the inference pipeline of ROAD; where we combine both supervised and self supervised learning to effectively detect radio observatory based anomalies.

We validate this approach in Section 6.4 by showing that it is optimal assuming that normality is better defined by the SSL output.

## 6.4 EXPERIMENTS

We evaluate the performance of ROAD using the dataset described in Section 6.2. The evaluation considers both the computation and model performance using both the binary anomaly detection as well as the multi-class classification results. In all cases we use the  $F\text{-}\beta$  score to evaluate the model performance. The  $F\text{-}\beta$  score is the harmonic mean between precision and recall, in the context of this work precision is the anomaly detection performance that is sensitive to the number of false positives and recall is the detection performance relative to the number of false negatives. Moreover, in the context of telescope operations it is necessary to minimise the number of false negatives. In other words, it is more acceptable to classify some normal samples as anomalous than classifying anomalous samples as normal. Following this logic and work by Kerrigan et al. [112], we consider  $\beta = 2$  to be the most appropriate as it weights recall more heavily than precision. For all evaluations we use the threshold from the Area Under Precision Recall Curve (AUPRC) which maximises the  $F\text{-}2$  score.

We do not quantify the benefits of ROAD with regard to imaging. The purpose of ROAD is to provide an efficient preview of large interferometric data products to telescope operators. Thereby informing scientists how best to post-process the data in the presence of instrumental and environmental anomalies. It is anticipated that leveraging the outputs of the model would facilitate the elimination of samples containing anomalies, thereby enhancing the overall image fidelity. However, for the sake of brevity and focus of

this chapter, we leave the actual quantification of the improvements to the imaging as future work.

#### 6.4.1 *Model parameters and training*

To validate our approach we experiment with several modern machine learning architectures with various model sizes. In all cases we use the same backbone architecture for both the supervised-classifier and the SSL models, furthermore, we utilise the same 2-layer MLP for position classification. Additionally, the decoder used for the SSL-reconstruction loss is a 5-layer architecture with strided de-convolution and batch-normalisation.

For every experiment each model is trained 3 times while randomising input seeds on each run. As already mentioned in Section 6.2.3 the low occurrence rates of some anomalous features, means we need to sub-sample the anomalous classes in the test data to ensure comparable occurrences relative to normal LOFAR telescope operations. This means we run 10 separate evaluation loops for the sub-sampled test data. In effect, the results shown in this section reflect the mean and standard deviations from 30 runs of each model. The SSL and the supervised models are trained for 100 epochs (the number of times the model is exposed to the full training set) while fine-tuning using the 2-layer MLP is done for only 20 epochs to prevent over-fitting. We use a batch size, patch size and latent dimensionality of 64 across all experiments utilising the Adam optimiser with a learning rate of  $1 \times 10^{-3}$  to maintain consistency. In all cases we use the official pytorch based implementations of the various backbones, with the exception of ViT, where we utilise an open source implementation.

Furthermore, to ensure no vanishing or exploding gradients while training we clip each autocorrelation to the 1<sup>st</sup> and 99<sup>th</sup> percentiles and take its natural log. Additionally, we normalise each magnitude-based autocorrelation between 0 and 1.

#### 6.4.2 *Anomaly detection and classification*

To maximise the model performance relative to the problem specification shown in Equation 22 we find the best mean performance of several different backbones. These being different sized ResNet [169], ConvNeXt [170] and ViT [171]. Notably our method is agnostic to backbone and could easily be extended to include architectures/model sizes. In Table 6.2 we present the per-class results after applying the combination of the supervised classi-

Class	Supervised	Random init	VAE	ImageNet	ROAD-KNN	ROAD
Normal	<b>0.94 ± 0.00</b>	<b>0.94 ± 0.01</b>	<b>0.94 ± 0.00</b>	0.92 ± 0.05	0.94 ± 0.00	<b>0.94 ± 0.01</b>
First order data loss	0.98 ± 0.05	0.98 ± 0.06	<b>0.99 ± 0.03</b>	0.98 ± 0.04	<b>0.99 ± 0.02</b>	<b>0.99 ± 0.04</b>
Second order data loss	0.77 ± 0.09	0.71 ± 0.12	0.76 ± 0.10	0.78 ± 0.1	0.73 ± 0.14	<b>0.79 ± 0.07</b>
High noise element	0.78 ± 0.14	0.58 ± 0.21	0.50 ± 0.24	0.66 ± 0.26	0.72 ± 0.19	<b>0.81 ± 0.16</b>
Oscillating tile	<b>0.84 ± 0.14</b>	0.71 ± 0.20	0.75 ± 0.17	0.73 ± 0.20	<b>0.84 ± 0.14</b>	0.82 ± 0.17
Source in the side-lobes	0.78 ± 0.03	0.78 ± 0.06	0.74 ± 0.10	0.77 ± 0.08	<b>0.79 ± 0.04</b>	<b>0.79 ± 0.04</b>
Galactic plane	0.80 ± 0.08	0.78 ± 0.04	0.77 ± 0.05	<b>0.84 ± 0.05</b>	0.79 ± 0.08	0.82 ± 0.08
Solar storm	<b>0.99 ± 0.02</b>	0.98 ± 0.02	<b>0.99 ± 0.03</b>	<b>0.99 ± 0.01</b>	0.99 ± 0.02	<b>0.99 ± 0.02</b>
Lightning storm	<b>0.95 ± 0.04</b>	0.900 ± 0.06	0.92 ± 0.06	0.94 ± 0.05	<b>0.95 ± 0.05</b>	<b>0.95 ± 0.04</b>
Ionospheric RFI reflections	0.99 ± 0.01	<b>1.000 ± 0.000</b>	<b>1.000 ± 0.000</b>	<b>1.000 ± 0.000</b>	0.99 ± 0.01	0.99 ± 0.01

Table 6.2: F2-score classification performance on the ROAD dataset, where bold is the best performance per class and the ResNet34 is used for all relevant backbones.

fication and the fine tuned anomaly detector specified by Equation 27. Furthermore, we plot the mean performance of each model in Figure 6.4 for the sake of easy comparison. We note that all evaluated anomaly detection models utilised fine-tuning in order to ensure they had been exposed to the same amount of data. Additionally, ROAD-KNN utilises a KNN lookup to determine the distances in the latent space rather than using the MLP prediction.

We find that the ResNet34 exhibits overall best average performance on the classification task, giving an average increase in F-2 score of 1% relative to the purely supervised model. We note that the performance of ROAD is directly dependant on the supervised performance. We show that the SSL-pretraining is highly influential to the overall model performance as it gives a < 5% increase over the randomly initialised (random init) model without pretraining. Furthermore we find that our SSL-based approach outperforms variational autoencoder-based model with fine-tuning (VAE) by < 5% as well being as < 3% better than KNN-based anomaly detectors (ROAD-KNN). Finally we show that using pretrained weights from ImageNet classification with fine-tuning (ImageNet) results in a 2% decrease in performance relative to our SSL pretraining paradigm.

Across all experiments it is clear that the high noise element and oscillating tile classes have the highest standard deviation. We attribute this to the small number of examples present in both the testing and training set after adjusting for occurrence rates. In addition to this, the features represented in these classes can vary significantly from sample to sample and band to band.

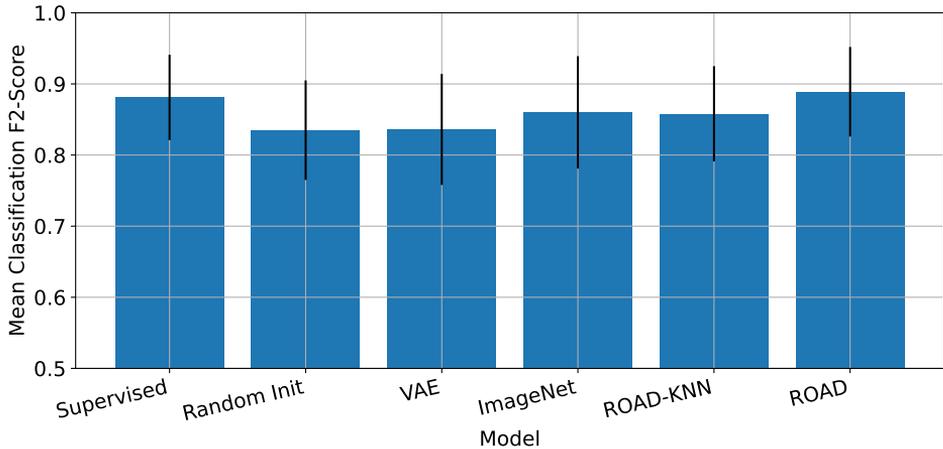


Figure 6.4: Per-class mean F-2 score based performance of each model shown in Table 6.2

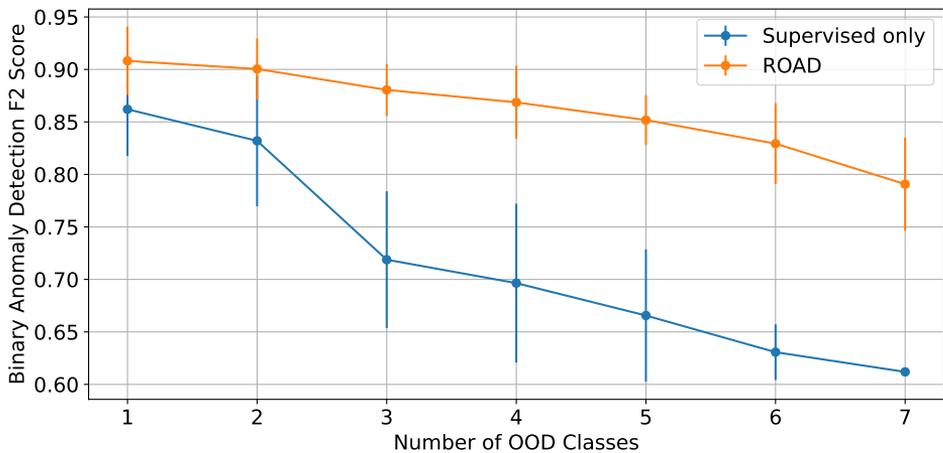


Figure 6.5: One-class anomaly detection performance for a purely supervised model and the fine-tuned SSL anomaly detector when removing a number of classes from the training set. The ResNet34 backbone is used for both training paradigms.

To simulate a real-world setting where many unknown anomalies can be present in a given observation, we remove several classes from the training set and test models' performance on the original test set. We refer these classes removed as *Out-Of-Distribution* (OOD). The objective of this experiment is to see how well the model will react to OOD anomalies and whether

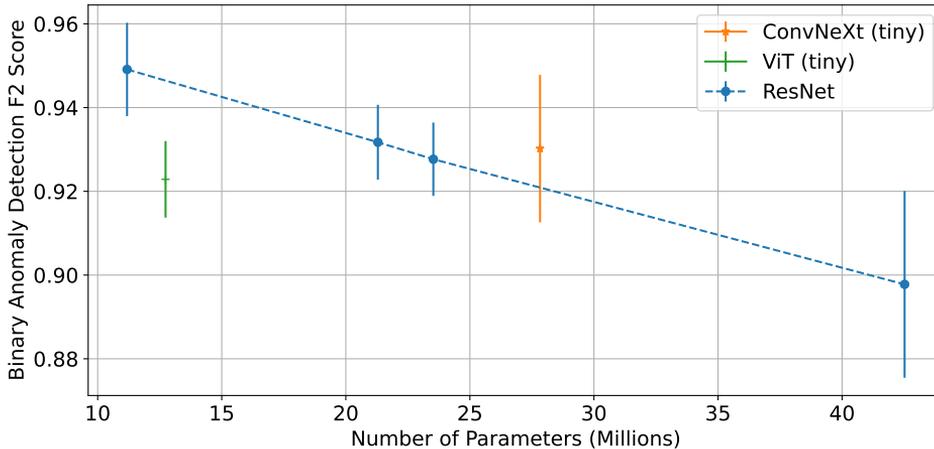


Figure 6.6: One-class anomaly detection performance after fine-tuning of various backbone networks when varying the number of available parameters.

it can correctly classify them as anomalous. To effectively simulate this scenario, we randomly remove between 1 and 7 classes and do this 10 times while training a model for each removal step. Figure 6.5 shows the average model performance from the 10 runs for both the supervised classifier as well as the fine-tuned SSL anomaly detector when removing a number of classes from the training set. Here it is clear that the supervised model suffers much more strongly from the OOD effects than the SSL-pretrained one, exhibiting a performance drop between 5% and 18%, thereby illustrating the benefit of using ROAD where both a classifier and detector are in the loop.

We illustrate the t-distributed stochastic neighbour embedding (t-SNE) projections of the latent dimensions from each models in Figure 6.7 to gain an intuition about the model performance. The same random seed and perplexity parameters are used for all plots shown, here the perplexity estimates the number of neighbours each point should have (for more information see [172]). In the topmost plot the non-fine tuned SSL-model is shown, we can see that both normal and anomalous classes are grouped closely together, with the exception of clusters pertaining to *first order data-loss*, *ionospheric RFI reflections* and *solar storms*. Furthermore, we find the normal data is distributed across two clusters, these being LBA and HBA features. It is interesting that even with no explicit training signals the SSL model without fine-tuning is still capable of distinguishing a variety of classes and phenomena. The middle plot shows the effects of fine-tuning

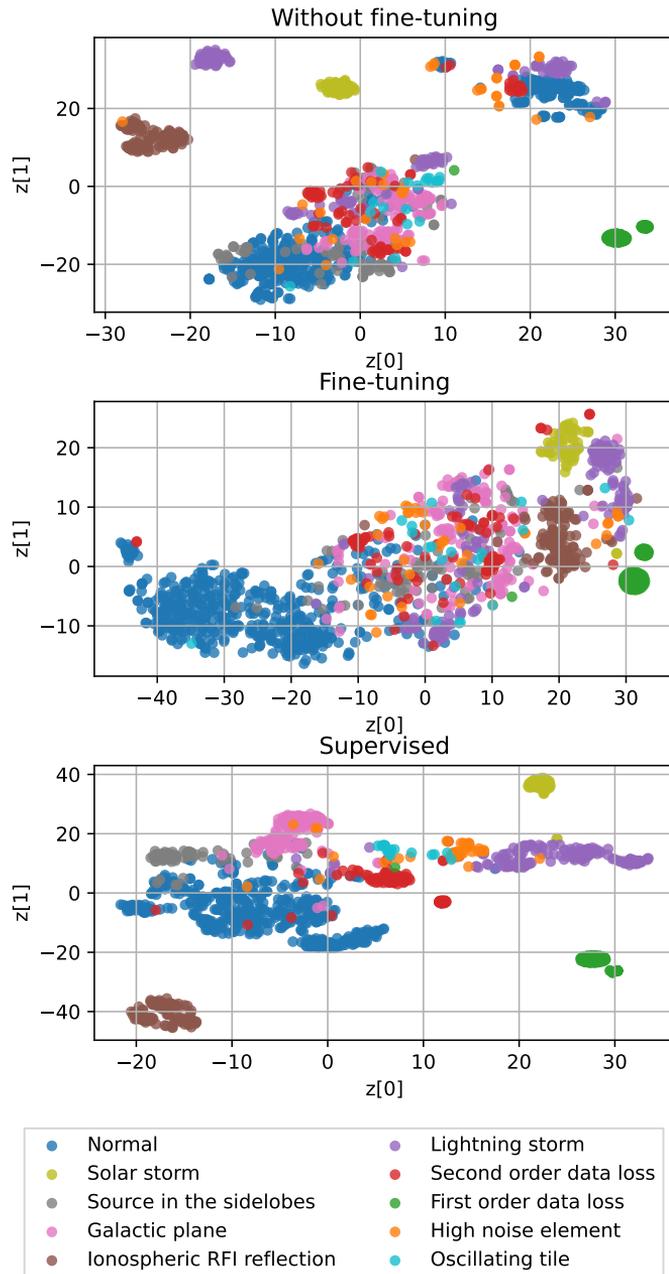


Figure 6.7: t-SNE projections of test data from the ROAD dataset using the representation from the final layer of the SSL-pretrained ResNet-34 with and without fine-tuning as well as the supervised classifier.

on the SSL representations. The fine-tuned SSL-model is significantly better at distinguishing normal from anomalous samples, with the LBA/HBA separation in the normal samples completely disappearing. Furthermore, the clusters corresponding to features that were once well separated such as *solar storm* are now better grouped with the anomalous samples. Finally in the bottom-most plot we can see the learn supervised representations of the test data. Here it is clear that the supervised model is the most capable of separating both anomalous and normal classes alike. It must be noted however that the classes relating to *galactic plane*, *source in the sidelobes* and *normal* are overlapping. Therefore by combining the boundary related to the SSL-fine-tuned embedding with the specificity of the supervised model we are able to better detect anomalies.

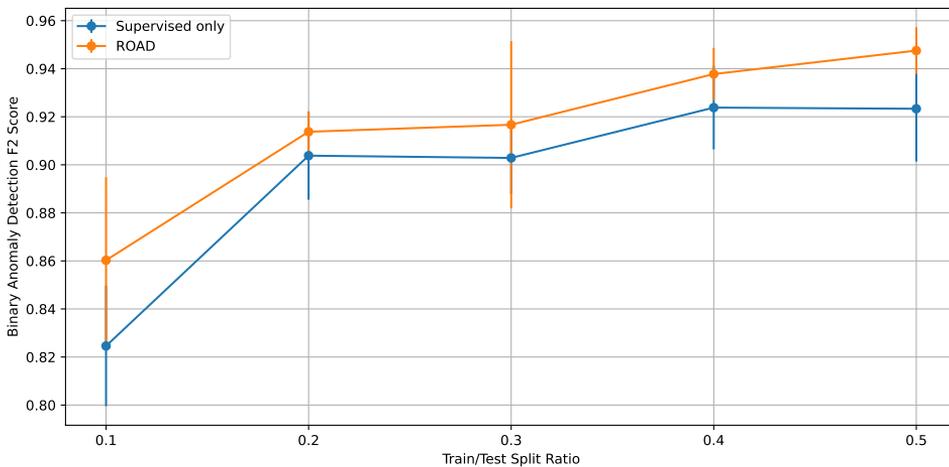


Figure 6.8: Binary anomaly detection performance when changing the amount of supervision used for training a ResNet-34 backbone for each training paradigm.

An interesting consequence of the class imbalance and the few number of samples certain events such as *oscillating tile*, is that ROAD benefits from fewer backbone parameters and does not scale with model size, as it overfits to the training data. This is illustrated in Figure 6.6, where it is also shown that ResNets offer the best performance. This being said, we expect that with more samples from the infrequent classes the model performance should scale proportionally with its number of parameters. This is further validated by Figure 6.8, where we plot the model performance relative to the amount of training data. Here it is clear that the model performance

scales linearly with training data-size. Furthermore, the fine-tuned model outperforms its purely supervised counterpart for all training set sizes.

### 6.4.3 Model ablations

To validate the correctness of the SSL-model training objective we perform several ablations. In Table 6.3 we show the effect of only using only the reconstruction term,  $\mathcal{L}_{\text{recon}}$ , or only the context prediction term,  $\mathcal{L}_{\text{con}}$ , or using the combined loss  $\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{con}}$ . We show that the combination of the two terms improves both the anomaly detection and the average classification performances by 2%, which at the scale of the LOFAR science data processing pipeline results in a significant improvement.

Performance	$\mathcal{L}_{\text{recon}}$	$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{con}}$
Anomaly detection	0.88 $\pm$ 0.03	0.92 $\pm$ 0.01	<b>0.93 <math>\pm</math> 0.01</b>
Classification	0.85 $\pm$ 0.07	0.87 $\pm$ 0.05	<b>0.89 <math>\pm</math> 0.06</b>

Table 6.3: Model performance (F2 score) after fine-tuning when varying the SSL loss function for a ResNet34 backbone.

Furthermore, in order to determine the relative contribution of each of the losses to the overall performance of ROAD we modify the  $\lambda$  hyperparameter and measure the overall model performance. Figure 6.9 shows how with  $0.3 \leq \lambda \leq 0.7$  the SSL anomaly detection obtains optimal performance.

In addition to the loss function-based ablations we also consider the effect of changing the combination function used between the supervised and SSL model shown in Equation 27. These results are shown in Figure 6.10, when we vary both the anomaly detection threshold set by the maximum  $F\text{-}\beta$  score as well as the combination function. In the plot, *combination function #1* uses the definition expressed in Equation 27, where the anomaly detector defines both normality and the unknown anomaly events. We define *combination function #2* as

$$y = \begin{cases} N + 1 & , \text{if } y_{\text{ssl}} = 1 \text{ and } y_{\text{sup}} = 0 \\ y_{\text{sup}} & , \text{otherwise} \end{cases} \quad (28)$$

such that  $y_{\text{ssl}}$  is only used to define unknown anomalous events. In the leftmost plot we can see that combination function #1 consistently offers

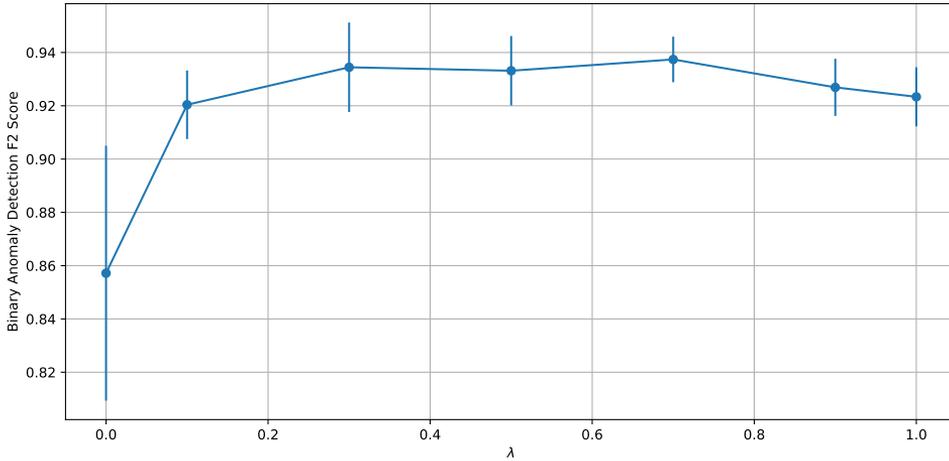


Figure 6.9: Mean anomaly detection F2-score performance of the ResNet-34 when varying the relative contribution of  $\mathcal{L}_{\text{recon}}$  and  $\mathcal{L}_{\text{con}}$ . When  $\lambda$  is minimum the loss favours  $\mathcal{L}_{\text{recon}}$  while a for greater losses  $\mathcal{L}_{\text{con}}$  dominates.

the best precision, while at the cost of a marginally decreasing the recall (<0.4%). The effect of this is that combination function #1 results in optimal F-2 score performance when the  $\beta$  is greater than 1. Furthermore, we evaluate the false positive rate using combination function #1 and find that it results in a false positive rate of approximately 2%.

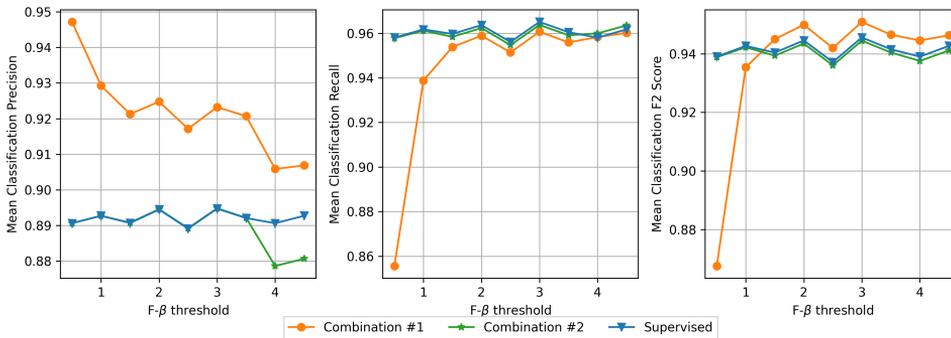


Figure 6.10: Mean classification performance of the ResNet-34 backbone after fine-tuning when changing the threshold used for anomaly detection as well as the combination function. Combinations #1 and #2 correspond to Equations 27 and 28 respectively.

#### 6.4.4 Computational performance analysis

We evaluate the computational performance of ROAD during inference on a Nvidia A10 GPU using CUDA 11.7 and using driver release 515.65.01. The KNN-based experiments utilise the GPU-based implementation of FAISS<sup>3</sup>. We use a batch size of 1024, with a patch size and a latent dimensionality of 64. Furthermore for the case of the KNN search we assume 1000 normal training samples to populate the search space. In all cases we use `bfloat16` representations of the input data such that to ensure the tensor-cores are fully utilised. In these results, we perform 1000 forward passes and measure the resulting latency, throughput in spectrograms per second as well as peak memory allocation.

The computation performance of the respective models can be seen in Table 6.4, where it is clear that the supervised model has the lowest computational overhead. We relate the difference performance between the supervised and SSL model to the dimensionality of the models inputs and required concatenation of the patches on each forward pass. As the SSL operates on the patch level, there are substantially fewer convolution operations that need to be applied (approximately 16), resulting in decreased peak memory performance. ROAD consists of both the supervised and SSL models and such the overall performance is given by the addition of the respective values, such that it takes less than 1 ms to predict the normality of a given spectrogram.

Model Type	Latency (ms)	Throughput(spec s <sup>-1</sup> )	Peak Memory(GBs)
Supervised	0.3	2844	8.11
SSL	0.4	2481	5.95
KNN	12.1	82	5.95

Table 6.4: Computational performance of anomaly detectors, where spec s<sup>-1</sup> refers to the number of spectrograms processed per second by the respective algorithm.

As described in Chapter 2, LOFAR’s correlator operates with a 80 Gbp/s output bandwidth. This implies that in order to integrate ROAD into the correlator it should be able to keep up with the incoming bandwidth without significant buffering. As previously described, ROAD has an overall latency of 0.7 ms per spectrogram, where a single spectrogram is 4 pol × 256 frequency samples × 256 time samples × 16 bits. Furthermore, ROAD

<sup>3</sup> <https://github.com/facebookresearch/faiss>

only operates on autocorrelations, hence it requires  $N_{st}/N_{tot} \approx 0.03$  of the total number of spectrograms. This means that when using all 64 antenna fields, only  $80 \text{ Gbp/s} \times 0.03 = 2.4 \text{ Gbp/s}$  need to be processed. Note, that this is a pessimistic estimate, as the correlator outputs far higher resolution data than what ROAD can effectively handle.

Given ROAD's output bandwidth of  $(4 \times 256 \times 256 \times 16)/0.7 \times 10^{-3} = 5.99 \text{ Gbp/s}$  we can conclude that ROAD is real-time in the context of the LOFAR telescope. As this estimate considers no parallelisation on a single Nvidia A10 GPU and the ROAD algorithm is embarrassingly parallel, we can expect our method to scale proportionally with number of GPUs. Additionally, we assume that the downsampling and normalisation operations required for this work are cheap and their cost would be eclipsed by the forward passes use for inference. Furthermore, we note that the KNN-based model performs significantly worse, suggesting that density based KNN anomaly detectors are less suitable for real-time applications at observatories.

## 6.5 CONCLUSIONS AND FUTURE WORK

In this work we have presented the first real time anomaly detector for system-wide anomalies in spectrographic data from radio telescopes. We produced a freely available dataset that contains 6708 autocorrelation-based spectrograms from the LOFAR telescope with labels relating to both commonly occurring anomalies as well as rare events. This work provides a formulation of anomaly detection in the SHM-context of telescope operations and illustrates how purely supervised models are ill-suited to the problem. Furthermore, we propose a new Self-Supervised Learning (SSL) paradigm for learning normal representations of spectrographic data. We combine both the SSL and supervised models and demonstrate how it remedies the shortcomings of supervised methods. We demonstrated that even with limited examples of anomalous data our fine tuned SSL model can significantly outperform its supervised counterpart, while keeping the false positive rate at an acceptable level. The radio observatory anomaly detector (ROAD) and dataset are the first major effort to address the system health management problem in radio telescopes and its potential benefit to all radio observatories is very promising.

We expect through providing open source access to both our models and dataset, continued effort by the larger community will increase the amount of training data from scarce events. Thereby enabling other train-

ing paradigms such as contrastive learning with larger models that are currently unsuited to the highly imbalanced problem. Furthermore, we identify several directions for future work in the area of radio observatory anomaly detection. Namely using the cross correlations to enhance training by using radio interferometer specific losses. Another interesting direction would be to use Bayesian deep learning to give uncertainty estimates from the classifier such that samples with low confidence would rely on the detector output. Finally, we would like to propagate the labels from the down-sampled data to the full resolution data from LOFAR Long Term Archive such that the performance could be better evaluated in the context of the full LOFAR data processing pipeline.

In future work we would like to see ROAD tested with data from different radio telescopes. We expect that instruments with roughly the same operating bands and time resolution would be good candidates. In previous work, [94] we show that unsupervised machine learning-based methods for Radio Frequency Interference (RFI) detection are directly transferable between the simulated data from the HERA telescope and real data from LOFAR. One potential problem is that there may be a domain shift between the ROAD dataset and data produced by another instrument. This could be addressed by labelling a few examples of anomalies in other instruments' spectrograms and fine tuning the ROAD model, using the supplied weights and with the new small dataset. In this manner, the overhead of extensive labelling would be avoided. However, in principle ROAD can be applied to any radio telescope provided that a new labelled data set is produced for the specific instrument. We expect that the anomaly categorisation used for the ROAD dataset is generic-enough to be directly transferred to other instruments. However we note that features such as *oscillating tile* are LOFAR specific.

Additionally, we propose investigating how best to integrate RFI detection and self-supervised anomaly detection for radio telescopes. Foundation models Bommasani et al. [173] offer a promising direction. Here a single self-supervised model could be trained on the normal data and then fine-tuned on both RFI-segmentation and anomaly detection tasks. In this manner, a model would be able to learn both representations of anomalous samples as well as RFI-contaminated data, which may improve model performance, generalisability and false-positive rates. To this end we would avoid the problem of potentially classifying RFI as anomalies and vice-versa.

# 7

## CONCLUSIONS

---

Radio telescopes will continue to increase in sensitivity and system complexity, resulting in steady growth of their data rates. This necessitates automatic system health monitoring and diagnostic tools that can report on the telescope health and inform operators of any potential problems. The ubiquity of machine learning-driven solutions in a variety of industries [27, 122, 123] and scientific disciplines [174] has motivated their application to radio astronomy-based system health management. However the complexity of the data and the expert domain specific knowledge required to interpret it, has meant that very little work has been done at the intersection of these two fields. This thesis has described an anomaly detection system for spectrogram-based data from radio telescopes that effectively offers a solution to the system health management problem. This was done by creating a data-inspection framework that can be integrated into the system monitoring and control pipeline, a representation learning method for projecting the spectrographic data to a low dimensional space as well as a discriminative measure for detecting anomalies present in spectrograms.

### 7.1 CONTRIBUTIONS

The main contributions of this thesis are as follows.

#### 7.1.1 *Representation learning for radio astronomy spectrograms*

We have proposed two novel applications of representation learning using radio astronomy spectrograms. In Chapter 3, we introduced a variational-autoencoder that can effectively integrate both the magnitude and phase information present in radio astronomy spectrograms in order produce a low dimensional prescriptive space. Chapter 6 presented a self-supervised

learning paradigm that utilised both the positional patch information as well as the reconstructions of spectrograms from the LOFAR telescope.

#### 7.1.2 *Nearest neighbour-based anomaly detection*

In Chapter 4 we proposed a novel anomaly detection algorithm that integrates the reconstructions of nearest neighbours in autoencoders. We showed its flexibility when applied to a variety of latent-variable models as well as its applicability to several different datasets. Furthermore, we demonstrated that through the simple adaption of reconstruction-error based losses we can consistently improve models' anomaly detection performance.

#### 7.1.3 *Expert labelled datasets for anomaly detection in radio astronomy*

In this work we provide three datasets to further research on both system health monitoring and radio frequency interference detection in radio telescopes. As interest grows from the radio astronomical community in machine learning so does the necessity for easily accessible annotated datasets. The datasets we provide will continue to encourage new solutions to system health management related problems in radio astronomy and telescope operations. The details and specifications of the datasets can be found in Chapters 5 and 6.

#### 7.1.4 *Data inspection framework for radio observatories*

This thesis provides a graphical interface for telescope operators to analyse data from the LOFAR telescope. The open source web framework enables representation learning models to be easily integrated into telescope operations. The interface incorporates several telescope specific controls to inspect particular aspects of the radio telescope; details regarding the data inspection framework can be found in Chapter 3.

#### 7.1.5 *Unsupervised RFI detection algorithm*

We presented a novel unsupervised RFI detection algorithm based on machine learning anomaly detection fundamentals in Chapter 5. This work documented how inverting the detection problem effectively addressed the over-fitting issue of supervised RFI detection algorithms. We have shown

that NLN provides better than state-of-the-art RFI detection without incurring the cost of labelling.

#### 7.1.6 *Real time anomaly detector for LOFAR*

In Chapter 6 we introduce the Radio Observatory Anomaly Detector, the first real time anomaly detector for system-wide anomalies in spectrographic data from radio telescopes. Here it is shown to be effective in detecting radio telescope specific anomalies from atmospheric, instrumentation, data processing and astronomical anomalies. This work provides a formulation of anomaly detection in the SHM-context of telescope operations and illustrates how purely supervised models are ill-suited to the problem. Furthermore, we propose a new Self-Supervised Learning (SSL) paradigm for learning normal representations of spectrographic data. We combine both the SSL and supervised models and demonstrate how it remedies the shortcomings of supervised methods. ROAD is the first major effort to address the system health management problem in radio telescopes and its potential benefit to all radio observatories is very promising.

## 7.2 ANSWERS TO RESEARCH QUESTIONS

The research questions defined in Chapter 1 are answered accordingly.

**Research Question 1:** *How can machine learning-based anomaly detection techniques be harnessed to improve system health management in radio telescopes?*

In this thesis we have shown two different approaches for automating system health management in radio telescopes. Firstly, in Chapter 3 we described an expert-in-the-loop data inspection framework that utilises a unsupervised representation learning model to project radio astronomy spectrograms to a prescriptive space. It enables telescope operators to easily filter observations based on several criteria using the provided web-based graphical user interface. Furthermore, we have proposed the Radio Observatory Anomaly Detector (ROAD) in Chapter 6, a model that can detect 9 different classes anomalies even when the model has not been exposed to particular classes. We show that ROAD offers suitably low false positive

rates while maintaining its detection capabilities. To this end, we find that system health management can be successfully automated.

**Research Question 2:** *What is the most effective way to learn robust representations of normal data from radio telescopes?*

We describe two novel methods for learning representations of spectrograms from radio telescopes in this thesis. In Chapter 3 we demonstrate how learning representations of spectrograms from radio telescopes is a uniquely challenging problem. This is due to high dynamic range from the presence of radio frequency interference, large data volumes and noisy data as well as the possible combinatorial effects present in these data products. To address these problems we propose a variational autoencoder model which learns joint representations from both the magnitude and phase components of spectrograms from radio telescopes. Furthermore, we evaluate these representations using a linear classifier and illustrate how the model's performance scales as we increase the number of possible features present in given observation using both simulated and real data.

In Chapter 6 we propose a self-supervised learning framework for learning both the relative positions of patches in a given spectrogram as well as their reconstruction from a latent encoding. We demonstrate that this model is capable of learning suitable representations of normal behaviour in radio telescopes, both with and without fine-tuning. We anticipate with access to larger datasets consisting of many more telescope operations-specific annotations further representation learning models could be developed and evaluated. To this end, we find self-supervised pretraining and with fine-tuning is the most effective way to learn representations of radio astronomy spectrograms.

**Research Question 3:** *Which are the most appropriate ways to discriminate between normal and anomalous samples using the learnt representations of normal data?*

Anomaly detection in high dimensional data is fundamentally dependant on learning suitable representations of normal data. With the previous

research question in mind, Chapter 6 demonstrates that fine-tuning the unsupervised and self-supervised models on small amounts of labelled data greatly improves the discriminative abilities of models to distinguish between normal and anomalous samples. This being said, with no access to labels, Chapter 4 shows that using the reconstructions of nearest neighbours in autoencoding and generative models gives consistent improvements over purely reconstruction based counterparts. This is illustrated using both synthetic data from the HERA telescope in an anomaly detection context as well as using Radio Frequency Interference specific use cases. We find that, normality is best discriminated using a fine-tuned linear classifier when labelled data is available.

**Research Question 4:** *To what extent is it possible to make system health monitoring techniques real-time in the context of LOFAR?*

The increasing data rates from modern radio telescopes necessitate real time computational performance for automated system health monitoring. As described in Chapter 2, LOFAR's correlator operates with a 80 Gbp/s output bandwidth. Furthermore, in Chapter 6 showed that ROAD only uses a small fraction of the required data produced by LOFAR, therefore our system health management system is only required to process 2.4 Gbp/s to be effectively real-time when using all 64 LOFAR antenna fields. We empirically estimated ROAD's output bandwidth to be 6 Gbp/s, demonstrating that it is real time in the context of the LOFAR telescope. Therefore we can conclude that with current data rates of the LOFAR telescope, it is indeed possible to make machine learning-based system health monitoring real-time. Furthermore, as our estimate considers no parallelisation on a single GPU and the ROAD algorithm is embarrassingly parallel, we can expect our method to scale proportionally with number of GPUs. Thus giving it headroom to be implemented with higher data-rates at potentially bigger telescopes.

### 7.3 FUTURE WORK

As radio telescopes continue to generate more data, so does the need for automated system health management workflows to ensure nominal telescope operations. The research presented in this thesis has culminated in ROAD,

a model that can effectively help to automate the tasks performed by the LOFAR telescope operators. In the creation of this model and through annotating the ROAD dataset we have enabled several more research directions in the field of system health management in radio telescopes.

An effective method to continue generating labelled spectrograms would be to integrate ROAD into the data inspection web-framework. This integration would enable an active learning loop, where misclassifications of samples near the decision boundary between normal and anomalous samples could be annotated by operators. An additional Bayesian machine learning model [175] could also be used in this context to give explicit estimates of uncertainty of samples near the decision boundary. These additional annotations could be used to retrain the model on a regular basis, until such time the model is capable of distinguishing all classes relevant to telescope operations. Furthermore, the integrated data inspection framework would enable improved error localisation, where telescope operators could immediately diagnose and shutdown stations that are experiencing anomalies during an observation.

The downsampling operation used for preprocessing the LOFAR spectrograms could potentially remove some anomalies present in the full resolution data. For example, narrowband anomalies such as lightning could be removed entirely from an observation by downsampling. Therefore, in future work we recommend an analysis of the consequences of this operation. Given the loss of consequential anomalies, we propose associating the annotations given to the downsampled spectrograms with the data from the LOFAR Long Term Archive (LTA) and retraining the models. However, as the full resolution data is in the order of terabytes per observation, we anticipate complications in training these models when using the full resolution data. In particular the supervised training on the full resolution spectrograms and the fine-tuning of the self-supervised model described in Section 6.3 will probably not be possible in its general form. Therefore, we also propose an investigation in pipelined training for radio astronomy models on the full resolution data.

The system health management system described in this work could be integrated directly into COBALT [13]. As COBALT uses a GPU cluster and ROAD makes use of a GPU to perform inference, we find it appropriate to connect these two pipelines. The only foreseeable additions that would be necessary would be to include downsampling and normalisation operations after inter-station correlation (which are computationally cheap). We believe that this would enable automatic reporting during observations.

Moreover, this would hide the memory load costs currently associated with our system.

Radio telescopes produce data in various modalities, for example system logs of the correlator, the temperature readings of each antenna processing board, the Radio Frequency Interference (RFI) flags generated by AOFlagger and the radio images used by astronomers. We believe that multi-modal machine learning based solution may offer improved representations of the radio astronomy data. Another potential option would be to perform multi-task learning, where objectives from anomaly detection, RFI mitigation and calibration could jointly be optimised.

Furthermore, the spectrograms alone contain meaningful geographical/geometric dependency between stations that could be harnessed in phase-interferometry specific self-supervised learning objective. Additionally, we see a clear research direction in associating the autocorrelation-based predictions done in ROAD to cross correlations so to give a prognosis to operators about the error localisation. This would also enable detection of inter-station errors such as clock drift, as they are not present in the autocorrelations.

## 7.4 VISION AND OUTLOOK

Scientific instruments are becoming more complex and are generating increasing amounts of data [9, 176, 177]. As such experts will probably not be able to analyse and guarantee nominal instrument operations due to the amount of data being generated. This necessitates the automation of system health management processes. Furthermore, we expect machine learning methods to continue advance and generate new compelling use-cases through all scientific disciplines. For this reason, we anticipate machine learning based anomaly detection systems to become essential in all modern instruments such as the Large Hadron Collider [176], KM3Net [177], LIGO [178] and others.

The significant improvements we have seen in machine-learning centric fields such as computer vision and natural language processing are in part due to the amount of high quality annotated datasets. Given the need for instrument specific machine learning system health management methods, we suggest scientists and instrument operators release more annotated system specific datasets on these topics. Furthermore, we anticipate active learning becoming intrinsic to system health management operations as it will allow continual model improvement.

As the data generation rates from scientific instruments increase so does the computational efficiency of the machine learning algorithms. For this reason we expect instrument specific integration of real-time machine learning models into existing scientific pipelines. Additionally, we see an additional research direction into making machine learning models in scientific disciplines more efficient through using techniques like pruning and quantisation [179].

Considering the development of foundation models [173] in machine learning, we see a similar future for system health management systems in radio telescopes [180]. Here a single self-supervised representation learning model could be trained using the multiple data modalities and fine-tuned on context specific tasks. In radio telescopes for example, a single self-supervised model could be trained on spectrograms with training objectives related to the different scientific data processing pipelines. Through this multiple objectives could be simultaneously addressed such as calibration, imaging, RFI detection and system health monitoring.

# BIBLIOGRAPHY

---

- [1] H. I. Ewen and E. M. Purcell. "Observation of a Line in the Galactic Radio Spectrum: Radiation from Galactic Hydrogen at 1,420 Mc./sec." In: *Nature* 168.4270 (Sept. 1951), pp. 356–356. ISSN: 0028-0836. DOI: [10.1038/168356a0](https://doi.org/10.1038/168356a0).
- [2] R. H. Dicke, P. J. E. Peebles, P. G. Roll, and D. T. Wilkinson. "Cosmic Black-Body Radiation." In: *The Astrophysical Journal* 142 (July 1965), p. 414. ISSN: 0004-637X. DOI: [10.1086/148306](https://doi.org/10.1086/148306).
- [3] A. HEWISH, S. J. BELL, J. D. H. PILKINGTON, P. F. SCOTT, and R. A. COLLINS. "Observation of a Rapidly Pulsating Radio Source." In: *Nature* 217.5130 (Feb. 1968), pp. 709–713. ISSN: 0028-0836. DOI: [10.1038/217709a0](https://doi.org/10.1038/217709a0).
- [4] Albert-Jan Boonstra. "Radio Frequency Interference Mitigation in Radio Astronomy." PhD thesis. TU Delft, 2005.
- [5] André R. Offringa. *Algorithms for Radio Interference Detection and Removal*. 2012, p. 212. ISBN: 9789036755450.
- [6] Kazunori Akiyama et al. "First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole." In: *The Astrophysical Journal* 875.1 (Apr. 2019), p. L1. ISSN: 2041-8213. DOI: [10.3847/2041-8213/ab0ec7](https://doi.org/10.3847/2041-8213/ab0ec7).
- [7] Katherine L. Bouman, Michael D. Johnson, Daniel Zoran, Vincent L. Fish, Sheperd S. Doeleman, and William T. Freeman. "Computational Imaging for VLBI Image Reconstruction." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016, pp. 913–922. ISBN: 978-1-4673-8851-1. DOI: [10.1109/CVPR.2016.105](https://doi.org/10.1109/CVPR.2016.105).
- [8] M. P. van Haarlem, M. W. Wise, A. W. Gunst, G. Heald, et al. "LOFAR: The LOW-Frequency ARray." In: *Astronomy & Astrophysics* 556 (Aug. 2013), A2. ISSN: 0004-6361. DOI: [10.1051/0004-6361/201220873](https://doi.org/10.1051/0004-6361/201220873). URL: <http://www.aanda.org/10.1051/0004-6361/201220873>.
- [9] P.E. Dewdney, P.J. Hall, R.T. Schilizzi, and T.J.L.W. Lazio. "The Square Kilometre Array." In: *Proceedings of the IEEE* 97.8 (Aug. 2009), pp. 1482–1496. ISSN: 0018-9219. DOI: [10.1109/JPROC.2009.2021005](https://doi.org/10.1109/JPROC.2009.2021005). URL: <http://link.springer.com/10.1007/s10686-005-4190-6><http://ieeexplore.ieee.org/document/5136190/>.
- [10] L. E. H. Godfrey, H. Bignall, S. Tingay, L. Harvey-Smith, M. Kramer, S. Burke-Spolaor, J. C. A. Miller-Jones, M. Johnston-Hollitt, R. Ekers, and S. Gulyaev. "Science at Very High Angular Resolution with the Square Kilometre Array." In: *Publications of the Astronomical Society of Australia* 29.1 (Jan. 2012), pp. 42–53. ISSN: 1323-3580. DOI: [10.1071/AS11050](https://doi.org/10.1071/AS11050).

- [11] Rik Jongerius, Stefan Wijnholds, Ronald Nijboer, and Henk Corporaal. “An End-to-End Computing Model for the Square Kilometre Array.” In: *Computer* 47.9 (Sept. 2014), pp. 48–54. ISSN: 0018-9162. DOI: [10.1109/MC.2014.235](https://doi.org/10.1109/MC.2014.235). URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6898703>.
- [12] Daniel Czech. “Time domain classification of transient radio frequency interference.” PhD thesis. Cape Town: University of Cape Town, 2019.
- [13] P. Chris Broekema, J. Jan David Mol, R. Nijboer, A. S. van Amesfoort, M. A. Brentjens, G. Marcel Loose, W. F.A. Klijn, and J. W. Romein. “Cobalt: A GPU-based correlator and beamformer for LOFAR.” In: *Astronomy and Computing* 23 (2018), pp. 180–192. ISSN: 22131337. DOI: [10.1016/j.ascom.2018.04.006](https://doi.org/10.1016/j.ascom.2018.04.006). URL: <https://doi.org/10.1016/j.ascom.2018.04.006>.
- [14] Varun Chandola, Arindam Banerjee, and Vipin Kumar. “Anomaly detection.” In: *ACM Computing Surveys* 41.3 (July 2009), pp. 1–58. ISSN: 0360-0300. DOI: [10.1145/1541880.1541882](https://doi.org/10.1145/1541880.1541882). URL: <http://cucis.ece.northwestern.edu/projects/DMS/publications/AnomalyDetection.pdf><https://dl.acm.org/doi/10.1145/1541880.1541882>.
- [15] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. “Deep Learning for Anomaly Detection.” In: *ACM Computing Surveys* 54.2 (Mar. 2022), pp. 1–38. ISSN: 0360-0300. DOI: [10.1145/3439950](https://doi.org/10.1145/3439950).
- [16] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. “GANomaly: Semi-supervised Anomaly Detection via Adversarial Training.” In: *Computer Vision – ACCV 2018*. Springer International Publishing, May 2019, pp. 622–637. DOI: [10.1007/978-3-030-20893-6\\_39](https://doi.org/10.1007/978-3-030-20893-6_39). URL: [http://link.springer.com/10.1007/978-3-030-20893-6\\_39](http://link.springer.com/10.1007/978-3-030-20893-6_39).
- [17] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. “Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection.” In: *Proceedings of the International Joint Conference on Neural Networks* 2019-July (2019). DOI: [10.1109/IJCNN.2019.8851808](https://doi.org/10.1109/IJCNN.2019.8851808).
- [18] Jinwon An and Cho Sungzoon. “Variational Autoencoder based Anomaly Detection using Reconstruction Probability.” In: *Special Lecture on IE* 20.6 (2015). ISSN: 15739031.
- [19] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. “OCGAN: One-class novelty detection using gans with constrained latent representations.” In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2019), pp. 2893–2901. ISSN: 10636919. DOI: [10.1109/CVPR.2019.00301](https://doi.org/10.1109/CVPR.2019.00301).
- [20] Stanislav Pidhorskyi, Ranya Almohsen, Donald A. Adjeroh, and Gianfranco Doretto. “Generative probabilistic novelty detection with adversarial autoencoders.” In: *Advances in Neural Information Processing Systems* NeurIPS (2018), pp. 6822–6833. ISSN: 10495258.
- [21] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. “Autoencoding beyond pixels using a learned similarity metric.” In: *33rd International Conference on Machine Learning, ICML 2016* 4 (2016), pp. 2341–2349.
- [22] Zhenyu Li. “Superpixel Masking and Inpainting for Self-Supervised Anomaly Detection.” In: *Bmvc* (2020).

- [23] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. “MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection.” In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019, pp. 9584–9592. ISBN: 978-1-7281-3293-8. DOI: [10.1109/CVPR.2019.00982](https://doi.org/10.1109/CVPR.2019.00982). URL: <https://ieeexplore.ieee.org/document/8954181/>.
- [24] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyang Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. “Towards Visually Explaining Variational Autoencoders.” In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 8639–8648. ISBN: 978-1-7281-7168-5. DOI: [10.1109/CVPR42600.2020.00867](https://doi.org/10.1109/CVPR42600.2020.00867). URL: <https://ieeexplore.ieee.org/document/9156323/>.
- [25] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. “Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows.” 2021.
- [26] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. “AnoD-DPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise.” In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2022, pp. 649–655. ISBN: 978-1-6654-8739-9. DOI: [10.1109/CVPRW56347.2022.00080](https://doi.org/10.1109/CVPRW56347.2022.00080).
- [27] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery.” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10265 LNCS (2017), pp. 146–147. ISSN: 16113349. DOI: [10.1007/978-3-319-59050-9](https://doi.org/10.1007/978-3-319-59050-9).
- [28] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. “Efficient GAN-Based Anomaly Detection.” In: (2018). URL: <http://arxiv.org/abs/1802.06222>.
- [29] Fabio Carrara, Giuseppe Amato, Luca Brombin, Fabrizio Falchi, and Claudio Genaro. “Combining GANs and AutoEncoders for efficient anomaly detection.” In: (2021), pp. 3939–3946. DOI: [10.1109/icpr48806.2021.9412253](https://doi.org/10.1109/icpr48806.2021.9412253).
- [30] Liron Bergman, Niv Cohen, and Yedid Hoshen. “Deep Nearest Neighbor Anomaly Detection.” In: *arXiv* 2002.10445 (Feb. 2020). ISSN: 2002.10445. DOI: [2002.10445](https://doi.org/2002.10445). URL: <http://arxiv.org/abs/2002.10445>.
- [31] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation Forest.” In: *ICDM* (2008). URL: <https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf%0Ahttps://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest>.
- [32] Alle-Jan van der Veen, Stefan J. Wijnholds, and Ahmad Mouri Sardarabadi. “Signal Processing for Radio Astronomy.” In: *Handbook of Signal Processing Systems*. Cham: Springer International Publishing, 2019, pp. 311–360. DOI: [10.1007/978-3-319-91734-4\\_9](https://doi.org/10.1007/978-3-319-91734-4_9).
- [33] C. Hollitt and M. Johnston-Hollitt. “Feature Detection in Radio Astronomy using the Circle Hough Transform.” In: *Publications of the Astronomical Society of Australia* 29.3 (Jan. 2012), pp. 309–317. ISSN: 1323-3580. DOI: [10.1071/AS11051](https://doi.org/10.1071/AS11051).

- [34] Ryan A. Loomis, Karin I. Öberg, Sean M. Andrews, Catherine Walsh, Ian Czekala, Jane Huang, and Katherine A. Rosenfeld. “Detecting Weak Spectral Lines in Interferometric Data through Matched Filtering.” In: *The Astronomical Journal* 155.4 (Apr. 2018), p. 182. ISSN: 1538-3881. DOI: [10.3847/1538-3881/aab604](https://doi.org/10.3847/1538-3881/aab604).
- [35] Mildred Morales-Xicohtencatl, Leticia Flores-Pulido, Carolina Rocío Sánchez-Pérez, and Juan José Córdova-Zamorano. “RASCNA: Radio Astronomy Signal Classification through Neighborhood Assemblies.” In: 2014, pp. 222–230. DOI: [10.1007/978-3-319-12568-8\\_{\\\_}28](https://doi.org/10.1007/978-3-319-12568-8_{\_}28).
- [36] P. Schellart et al. “Detecting cosmic rays with the LOFAR radio telescope.” In: *Astronomy & Astrophysics* 560 (Dec. 2013), A98. ISSN: 0004-6361. DOI: [10.1051/0004-6361/201322683](https://doi.org/10.1051/0004-6361/201322683).
- [37] F. de Gasperin et al. “Systematic effects in LOFAR data: A unified calibration strategy.” In: *Astronomy & Astrophysics* 622 (Feb. 2019), A5. ISSN: 0004-6361. DOI: [10.1051/0004-6361/201833867](https://doi.org/10.1051/0004-6361/201833867).
- [38] R. J. van Weeren et al. “LOFAR FACET CALIBRATION.” In: *The Astrophysical Journal Supplement Series* 223.1 (Mar. 2016), p. 2. ISSN: 1538-4365. DOI: [10.3847/0067-0049/223/1/2](https://doi.org/10.3847/0067-0049/223/1/2).
- [39] C. Tasse, B. Hugo, M. Mirmont, O. Smirnov, M. Atemkeng, L. Bester, M. J. Hardcastle, R. Lakhoo, S. Perkins, and T. Shimwell. “Faceting for direction-dependent spectral deconvolution.” In: *Astronomy & Astrophysics* 611 (Mar. 2018), A87. ISSN: 0004-6361. DOI: [10.1051/0004-6361/201731474](https://doi.org/10.1051/0004-6361/201731474).
- [40] Stefan J. Wijnholds, Sebastiaan Van Der Tol, Ronald Nijboer, and Alle Jan Van Der Veen. “Calibration challenges for future radio telescopes.” In: *IEEE Signal Processing Magazine* 27.1 (2010), pp. 30–42. ISSN: 10535888. DOI: [10.1109/MSP.2009.934853](https://doi.org/10.1109/MSP.2009.934853).
- [41] A. R. Offringa, A. G. de Bruyn, M. Biehl, S. Zaroubi, G. Bernardi, and V. N. Pandey. “Post-correlation radio frequency interference classification methods.” In: *Monthly Notices of the Royal Astronomical Society* 405.1 (2010), pp. 155–167. ISSN: 13652966. DOI: [10.1111/j.1365-2966.2010.16471.x](https://doi.org/10.1111/j.1365-2966.2010.16471.x).
- [42] B. R. Barsdell, M. Bailes, D. G. Barnes, and C. J. Fluke. “Accelerating incoherent dedispersion.” In: *Monthly Notices of the Royal Astronomical Society* 422.1 (May 2012), pp. 379–392. ISSN: 00358711. DOI: [10.1111/j.1365-2966.2012.20622.x](https://doi.org/10.1111/j.1365-2966.2012.20622.x).
- [43] C. G. Bassa, J. W. Romein, B. Veenboer, S. van der Vlugt, and S. J. Wijnholds. “Fourier-domain dedispersion.” In: *Astronomy & Astrophysics* 657 (Jan. 2022), A46. ISSN: 0004-6361. DOI: [10.1051/0004-6361/202142099](https://doi.org/10.1051/0004-6361/202142099).
- [44] Alessio Sclocco, Dany Vohl, and Rob V. Van Nieuwpoort. “Real-time RFI mitigation for the apertif radio transient system.” In: *RFI 2019 - Proceedings of 2019 Radio Frequency Interference: Coexisting with Radio Frequency Interference* (2019). DOI: [10.23919/RFI48793.2019.9111826](https://doi.org/10.23919/RFI48793.2019.9111826).
- [45] A. Sclocco, J. van Leeuwen, H.E. Bal, and R.V. van Nieuwpoort. “Real-time dedispersion for fast radio transient surveys, using auto tuning on many-core accelerators.” In: *Astronomy and Computing* 14 (Jan. 2016), pp. 1–7. ISSN: 22131337. DOI: [10.1016/j.ascom.2016.01.001](https://doi.org/10.1016/j.ascom.2016.01.001).
- [46] Rob V. van Nieuwpoort and John W. Romein. “Correlating Radio Astronomy Signals with Many-Core Hardware.” In: *International Journal of Parallel Programming* 39.1 (Feb. 2011), pp. 88–114. ISSN: 0885-7458. DOI: [10.1007/s10766-010-0144-3](https://doi.org/10.1007/s10766-010-0144-3).

- [47] P. La Plante et al. "A Real Time Processing system for big data in astronomy: Applications to HERA." In: *Astronomy and Computing* 36 (July 2021), p. 100489. ISSN: 22131337. DOI: [10.1016/j.ascom.2021.100489](https://doi.org/10.1016/j.ascom.2021.100489).
- [48] John W. Romein, P. Chris Broekema, Jan David Mol, and Rob V. van Nieuwpoort. "The LOFAR correlator." In: *ACM SIGPLAN Notices* 45.5 (May 2010), pp. 169–178. ISSN: 0362-1340. DOI: [10.1145/1837853.1693477](https://doi.org/10.1145/1837853.1693477).
- [49] T. J. Pearson and A. C. S. Readhead. "Image Formation by Self-Calibration in Radio Astronomy." In: *Annual Review of Astronomy and Astrophysics* 22.1 (Sept. 1984), pp. 97–130. ISSN: 0066-4146. DOI: [10.1146/annurev.aa.22.090184.000525](https://doi.org/10.1146/annurev.aa.22.090184.000525).
- [50] Amir Leshem and Alle Jan Der Van Veen. "Radio-astronomical imaging in the presence of strong radio interference." In: *IEEE Transactions on Information Theory* 46.5 (2000), pp. 1730–1747. ISSN: 00189448. DOI: [10.1109/18.857787](https://doi.org/10.1109/18.857787).
- [51] Ray P. Norris. "Data challenges for next-generation radio telescopes." In: *Proceedings - 6th IEEE International Conference on e-Science Workshops, e-ScienceW 2010* (2010), pp. 21–24. DOI: [10.1109/eScienceW.2010.13](https://doi.org/10.1109/eScienceW.2010.13).
- [52] Rendong Nan, Di Li, Chengjin Jin, Qiming Wang, Lichun Zhu, Wenbai Zhu, Haiyan Zhang, Youling Yue, and Lei Qian. "The five-hundred-meter Aperture Spherical radio Telescope (FAST) project." In: *International Journal of Modern Physics D* 20.6 (2011), pp. 989–1024. ISSN: 02182718. DOI: [10.1142/S0218271811019335](https://doi.org/10.1142/S0218271811019335).
- [53] V Morello, K M Rajwade, and B W Stappers. "IQRM: real-time adaptive RFI masking for radio transient and pulsar searches." In: *Monthly Notices of the Royal Astronomical Society* 510.1 (Dec. 2021), pp. 1393–1403. ISSN: 0035-8711. DOI: [10.1093/mnras/stab3493](https://doi.org/10.1093/mnras/stab3493).
- [54] P. Prasad, S. J. Wijnholds, F. Huizinga, and R. A. M. J. Wijers. "Real-time calibration of the AARTFAAC array for transient detection." In: *Astronomy & Astrophysics* 568 (Aug. 2014), A48. ISSN: 0004-6361. DOI: [10.1051/0004-6361/201423668](https://doi.org/10.1051/0004-6361/201423668).
- [55] Liam Connor and Joeri van Leeuwen. "Applying Deep Learning to Fast Radio Burst Classification." In: *The Astronomical Journal* 156.6 (2018), p. 256. ISSN: 1538-3881. DOI: [10.3847/1538-3881/aae649](https://doi.org/10.3847/1538-3881/aae649). URL: <http://dx.doi.org/10.3847/1538-3881/aae649>.
- [56] Daniel George and E.A. Huerta. "Deep neural networks to enable real-time multimessenger astrophysics." In: *Physical Review D* 97.4 (Feb. 2018), p. 044039. ISSN: 2470-0010. DOI: [10.1103/PhysRevD.97.044039](https://doi.org/10.1103/PhysRevD.97.044039).
- [57] Daniel Muthukrishna, Kaisey S Mandel, Michelle Lochner, Sara Webb, and Gautham Narayan. "Real-time detection of anomalies in large-scale transient surveys." In: *Monthly Notices of the Royal Astronomical Society* 517.1 (Oct. 2022), pp. 393–419. ISSN: 0035-8711. DOI: [10.1093/mnras/stac2582](https://doi.org/10.1093/mnras/stac2582).
- [58] Walter Jansen et al. "ASTRON Computing Research and Development." In: *Astronomical Data Analysis Software and Systems XXIX. ASP Conference Series*, 2020. URL: <https://ui.adsabs.harvard.edu/abs/2020ASPC..527..755J/abstract>.
- [59] Kjeld van der Schaaf and E Lawerman. "SCADA in the LOFAR radio telescope." In: *International Conference on Accelerator and Large Experimental Physics Control Systems (ICALEPCS)*. Geneva, 2005.

- [60] V Lukic, M Brügggen, B Mingo, J H Croston, G Kasieczka, and P N Best. “Morphological classification of radio galaxies: capsule networks versus convolutional neural networks.” In: *Monthly Notices of the Royal Astronomical Society* 487.2 (Aug. 2019), pp. 1729–1744. ISSN: 0035-8711. DOI: [10.1093/mnras/stz1289](https://academic.oup.com/mnras/article/487/2/1729/5492264). URL: <https://academic.oup.com/mnras/article/487/2/1729/5492264>.
- [61] V Lukic, M Br, J K Banfield, O I Wong, L Rudnick, R P Norris, and B Simmons. “Radio Galaxy Zoo : compact and extended radio source classification with deep learning 1 I N T R O D U C T I O N.” In: 260 (2018), pp. 246–260. DOI: [10.1093/mnras/sty163](https://academic.oup.com/mnras/sty163).
- [62] Chen Wu et al. “Radio Galaxy Zoo: Claran – a deep learning classifier for radio morphologies.” In: *Monthly Notices of the Royal Astronomical Society* 482.1 (Jan. 2019), pp. 1211–1230. ISSN: 0035-8711. DOI: [10.1093/mnras/sty2646](https://academic.oup.com/mnras/article/482/1/1211/5142869). URL: <https://academic.oup.com/mnras/article/482/1/1211/5142869>.
- [63] A K Aniyani and K Thorat. “Classifying Radio Galaxies with the Convolutional Neural Network.” In: *The Astrophysical Journal Supplement Series* 230.2 (2017), p. 20. ISSN: 1538-4365. DOI: [10.3847/1538-4365/aa7333](http://dx.doi.org/10.3847/1538-4365/aa7333). URL: <http://dx.doi.org/10.3847/1538-4365/aa7333>.
- [64] N Schanche et al. “Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys.” In: *Monthly Notices of the Royal Astronomical Society* 483.4 (Mar. 2019), pp. 5534–5547. ISSN: 0035-8711. DOI: [10.1093/mnras/sty3146](https://academic.oup.com/mnras/sty3146).
- [65] Abhishek Malik, Benjamin P Moster, and Christian Obermeier. “Exoplanet detection using machine learning.” In: *Monthly Notices of the Royal Astronomical Society* (Dec. 2021). ISSN: 0035-8711. DOI: [10.1093/mnras/stab3692](https://academic.oup.com/mnras/stab3692).
- [66] P. Škoda, O. Podsztavek, and P. Tvrđík. “Active deep learning method for the discovery of objects of interest in large spectroscopic surveys.” In: *Astronomy and Astrophysics* 643.Zickgraf 2003 (2020), pp. 1–15. ISSN: 14320746. DOI: [10.1051/0004-6361/201936090](https://academic.oup.com/aas/advance-article-abstract/doi/10.1051/0004-6361/201936090).
- [67] M Lochner, L Rudnick, I Heywood, K Knowles, and S S Shabala. “A unique, ring-like radio source with quadrilateral structure detected with machine learning.” In: *Monthly Notices of the Royal Astronomical Society* 520.1 (Jan. 2023), pp. 1439–1446. ISSN: 0035-8711. DOI: [10.1093/mnras/stad074](https://academic.oup.com/mnras/stad074).
- [68] Sarod Yatawatta and Ian M Avruch. “Deep reinforcement learning for smart calibration of radio telescopes.” In: *Monthly Notices of the Royal Astronomical Society* 505.2 (June 2021), pp. 2141–2150. ISSN: 0035-8711. DOI: [10.1093/mnras/stab1401](https://academic.oup.com/mnras/stab1401).
- [69] J. Akeret, C. Chang, A. Lucchi, and A. Refregier. “Radio frequency interference mitigation using deep convolutional neural networks.” In: *Astronomy and Computing* 18 (2017), pp. 35–39. ISSN: 22131337. DOI: [10.1016/j.ascom.2017.01.002](https://academic.oup.com/ascom/advance-article-abstract/doi/10.1016/j.ascom.2017.01.002).
- [70] Alireza Vafaei Sadr, Bruce A. Bassett, Nadeem Oozeer, Yabebal Fantaye, and Chris Finlay. “Deep learning improves identification of Radio Frequency Interference.” In: *Monthly Notices of the Royal Astronomical Society* 499.1 (2020), pp. 379–390. ISSN: 13652966. DOI: [10.1093/mnras/staa2724](https://academic.oup.com/mnras/staa2724).
- [71] Zhicheng Yang, Ce Yu, Jian Xiao, and Bo Zhang. “Deep residual detection of radio frequency interference for FAST.” In: *Monthly Notices of the Royal Astronomical Society* 492.1 (2020), pp. 1421–1431. ISSN: 0035-8711. DOI: [10.1093/mnras/stz3521](https://academic.oup.com/mnras/stz3521).

- [72] Yoshua Bengio, Aaron Courville, and Pascal Vincent. *Representation Learning: A Review and New Perspectives*. Tech. rep. 2014.
- [73] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. “Masked Autoencoders Are Scalable Vision Learners.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nov. 2022.
- [74] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A Simple Framework for Contrastive Learning of Visual Representations.” In: (Feb. 2020).
- [75] Jean-Bastien Grill et al. “Bootstrap your own latent: A new approach to self-supervised Learning.” In: *Advances in neural information processing systems*. June 2020.
- [76] Carl Doersch, Abhinav Gupta, and Alexei A Efros. “Unsupervised Visual Representation Learning by Context Prediction.” In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Dec. 2015, pp. 1422–1430. ISBN: 978-1-4673-8391-2. DOI: [10.1109/ICCV.2015.167](https://doi.org/10.1109/ICCV.2015.167). URL: <http://ieeexplore.ieee.org/document/7410524/>.
- [77] Mike Walmsley et al. “Practical galaxy morphology tools from deep supervised representation learning.” In: *Monthly Notices of the Royal Astronomical Society* 513.2 (May 2022), pp. 1581–1599. ISSN: 0035-8711. DOI: [10.1093/mnras/stac525](https://doi.org/10.1093/mnras/stac525).
- [78] Mike Walmsley et al. “Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314000 galaxies.” In: *Monthly Notices of the Royal Astronomical Society* 509.3 (Dec. 2021), pp. 3966–3988. ISSN: 0035-8711. DOI: [10.1093/mnras/stab2093](https://doi.org/10.1093/mnras/stab2093).
- [79] Md Abul Hayat, George Stein, Peter Harrington, Zarija Lukić, and Mustafa Mustafa. “Self-supervised Representation Learning for Astronomical Images.” In: *The Astrophysical Journal Letters* 911.2 (2021), p. L33. ISSN: 2041-8205. DOI: [10.3847/2041-8213/abf2c7](https://doi.org/10.3847/2041-8213/abf2c7).
- [80] J. E. Gunn et al. “The Sloan Digital Sky Survey Photometric Camera.” In: *The Astronomical Journal* 116.6 (Dec. 1998), pp. 3040–3081. ISSN: 00046256. DOI: [10.1086/300645](https://doi.org/10.1086/300645).
- [81] Michael Mesarcik, Albert-Jan Boonstra, Christiaan Meijer, Walter Jansen, Elena Rangelova, and Rob V van Nieuwpoort. “Deep Learning Assisted Data Inspection for Radio Astronomy.” In: *Monthly Notices of the Royal Astronomical Society* (May 2020). ISSN: 0035-8711. DOI: [10.1093/mnras/staa1412](https://doi.org/10.1093/mnras/staa1412). URL: <https://academic.oup.com/mnras/advance-article/doi/10.1093/mnras/staa1412/5848205>.
- [82] V. Ashley Villar, Miles Cranmer, Edo Berger, Gabriella Contardo, Shirley Ho, Griffin Hosseinzadeh, and Joshua Yao-Yu Lin. “A Deep-learning Approach for Live Anomaly Detection of Extragalactic Transients.” In: *The Astrophysical Journal Supplement Series* 255.2 (Aug. 2021), p. 24. ISSN: 0067-0049. DOI: [10.3847/1538-4365/ac0893](https://doi.org/10.3847/1538-4365/ac0893). URL: <http://arxiv.org/abs/2103.12102><http://dx.doi.org/10.3847/1538-4365/ac0893><https://iopscience.iop.org/article/10.3847/1538-4365/ac0893>.
- [83] Peter Melchior, Yan Liang, ChangHoon Hahn, and Andy Goulding. “Autoencoding Galaxy Spectra I: Architecture.” In: *ArXiv* 2211.07890 (Nov. 2022).
- [84] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. “Improving unsupervised defect segmentation by applying structural similarity to autoencoders.” In: *VISIGRAPP 2019 - Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* 5 (2019), pp. 372–380. DOI: [10.5220/0007364503720380](https://doi.org/10.5220/0007364503720380).

- [85] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. "Support vector method for novelty detection." In: *Proceedings of the 12th International Conference on Neural Information Processing Systems*. Denver, CO, 1999, pp. 582–588.
- [86] Michael Mesarcik, Elena Rangelova, Albert-Jan Boonstra, and Rob V van Nieuwpoort. "Improving novelty detection using the reconstructions of nearest neighbours." In: *Array* (May 2022), p. 100182. ISSN: 25900056. DOI: [10.1016/j.array.2022.100182](https://doi.org/10.1016/j.array.2022.100182). URL: <http://arxiv.org/abs/2111.06150https://linkinghub.elsevier.com/retrieve/pii/S2590005622000388>.
- [87] Tal Reiss and Yedid Hoshen. "Mean-Shifted Contrastive Loss for Anomaly Detection." In: *ArXiv* 2106.03844 (June 2021). DOI: [2106.03844](https://doi.org/10.2106.03844).
- [88] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. "Towards Total Recall in Industrial Anomaly Detection." In: *ArXiv* 2106.08265 (June 2021).
- [89] L. Fei-Fei, J. Deng, and K. Li. "ImageNet: Constructing a large-scale image database." In: *Journal of Vision* 9.8 (2010), pp. 1037–1037. DOI: [10.1167/9.8.1037](https://doi.org/10.1167/9.8.1037).
- [90] Jihun Yi and Sungroh Yoon. "Patch SVDD: Patch-Level SVDD for Anomaly Detection and Segmentation." In: *Lecture Notes in Computer Science*. 2021, pp. 375–390. DOI: [10.1007/978-3-030-69544-6\\_{\\\_}23](https://doi.org/10.1007/978-3-030-69544-6_{\_}23). URL: [http://link.springer.com/10.1007/978-3-030-69544-6\\_23](http://link.springer.com/10.1007/978-3-030-69544-6_23).
- [91] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. "CutPaste: Self-Supervised Learning for Anomaly Detection and Localization." In: *ArXiv* 2104.04015 (2021). URL: <http://arxiv.org/abs/2104.04015>.
- [92] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. "CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances." In: *ArXiv* 2007.08176.NeurIPS (July 2020). URL: <http://arxiv.org/abs/2007.08176>.
- [93] Philippe Burlina, Neil Joshi, and I. Jeng Wang. "Where's wally now? deep generative and discriminative embeddings for novelty detection." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June (2019), pp. 11499–11508. ISSN: 10636919. DOI: [10.1109/CVPR.2019.01177](https://doi.org/10.1109/CVPR.2019.01177).
- [94] Michael Mesarcik, Albert-Jan Boonstra, Elena Rangelova, and Rob V van Nieuwpoort. "Learning to detect radio frequency interference in radio astronomy without seeing it." In: *Monthly Notices of the Royal Astronomical Society* 516.4 (Sept. 2022), pp. 5367–5378. ISSN: 0035-8711. DOI: [10.1093/mnras/stac2503](https://doi.org/10.1093/mnras/stac2503). URL: <https://zenodo.org/record/6724065https://academic.oup.com/mnras/article/516/4/5367/6692884>.
- [95] Michelle Lochner and B.A. Bassett. "Astronomy: Personalised active anomaly detection in astronomical data." In: *Astronomy and Computing* 36 (July 2021), p. 100481. ISSN: 22131337. DOI: [10.1016/j.ascom.2021.100481](https://doi.org/10.1016/j.ascom.2021.100481). URL: <http://arxiv.org/abs/2010.11202https://linkinghub.elsevier.com/retrieve/pii/S2213133721000354>.
- [96] K. L. Malanchev et al. "Anomaly detection in the Zwicky Transient Facility DR3." In: *Monthly Notices of the Royal Astronomical Society* 502.4 (Mar. 2021), pp. 5147–5175. ISSN: 0035-8711. DOI: [10.1093/mnras/stab316](https://doi.org/10.1093/mnras/stab316). URL: <https://academic.oup.com/mnras/article/502/4/5147/6129794>.

- [97] Peter Xiangyuan Ma et al. "A deep-learning search for technosignatures from 820 nearby stars." In: *Nature Astronomy* (Jan. 2023). ISSN: 2397-3366. DOI: [10.1038/s41550-022-01872-z](https://doi.org/10.1038/s41550-022-01872-z).
- [98] Berta Margalef-Bentabol, Marc Huertas-Company, Tom Charnock, Carla Margalef-Bentabol, Mariangela Bernardi, Yohan Dubois, Kate Storey-Fisher, and Lorenzo Zanisi. "Detecting outliers in astronomical images with deep generative networks." In: *Monthly Notices of the Royal Astronomical Society* 000.0000 (2020). ISSN: 0035-8711. DOI: [10.1093/mnras/staa1647](https://doi.org/10.1093/mnras/staa1647).
- [99] Kate Storey-Fisher, Marc Huertas-Company, Nesar Ramachandra, Francois Lanusse, Alexie Leauthaud, Yifei Luo, Song Huang, and J Xavier Prochaska. "Anomaly detection in Hyper Suprime-Cam galaxy images with generative adversarial networks." In: *Monthly Notices of the Royal Astronomical Society* 508.2 (Oct. 2021), pp. 2946–2963. ISSN: 0035-8711. DOI: [10.1093/mnras/stab2589](https://doi.org/10.1093/mnras/stab2589). URL: <https://academic.oup.com/mnras/article/508/2/2946/6369368>.
- [100] Yunfan Gerry Zhang, Ki Hyun Won, Seung Woo Son, Andrew Siemion, and Steve Croft. "Self-supervised anomaly detection for narrowband SETI." In: *2018 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2018 - Proceedings* (2019), pp. 1114–1118. DOI: [10.1109/GlobalSIP.2018.8646437](https://doi.org/10.1109/GlobalSIP.2018.8646437).
- [101] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9351 (2015), pp. 234–241. ISSN: 16113349. DOI: [10.1007/978-3-319-24574-4\\_{\\\_}28](https://doi.org/10.1007/978-3-319-24574-4_{\_}28).
- [102] Chihway Chang, Christian Monstein, Joel Akeret, Sebastian Seehars, Alexandre Refregier, Adam Amara, Adrian Glauser, and Bruno Stuber. "An integrated system at the Bleien Observatory for mapping the Galaxy." In: *Monthly Notices of the Royal Astronomical Society* 464.2 (2017), pp. 1727–1737. ISSN: 13652966. DOI: [10.1093/mnras/stw2472](https://doi.org/10.1093/mnras/stw2472).
- [103] J. Akeret, S. Seehars, C. Chang, C. Monstein, A. Amara, and A. Refregier. "HIDE & SEEK: End-to-end packages to simulate and process radio survey data." In: *Astronomy and Computing* 18 (2017), pp. 8–17. ISSN: 22131337. DOI: [10.1016/j.ascom.2016.11.001](https://doi.org/10.1016/j.ascom.2016.11.001). URL: <http://dx.doi.org/10.1016/j.ascom.2016.11.001>.
- [104] Kevin Vinsen, Samuel Foster, and Richard Dodson. "Using Machine Learning for the detection of Radio Frequency Interference." In: *2019 URSI Asia-Pacific Radio Science Conference (AP-RASC) March* (2019), pp. 2–5.
- [105] Etienne E. Vos, P. S. Francois Luus, Chris J. Finlay, and Bruce A. Bassett. "A Generative Machine Learning Approach to RFI Mitigation for Radio Astronomy." In: *IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2019-October* (2019), pp. 1–6. ISSN: 21610371. DOI: [10.1109/MLSP.2019.8918820](https://doi.org/10.1109/MLSP.2019.8918820).
- [106] Michael Mesarcik, Albert-Jan Boonstra, Christiaan Meijer, Walter Jansen, Elena Ranguelova, and Rob V. van Nieuwpoort. "Deep learning assisted data inspection for radio astronomy." In: *Monthly Notices of the Royal Astronomical Society* 496.2 (May 2020), pp. 1517–1529. ISSN: 0035-8711. DOI: [10.1093/mnras/staa1412](https://doi.org/10.1093/mnras/staa1412). eprint: <https://academic.oup.com/mnras/article-pdf/496/2/1517/33483423/staa1412.pdf>. URL: <https://doi.org/10.1093/mnras/staa1412>.

- [107] Michael Mesarcik, Albert-Jan Boonstra, Christiaan Meijer, Walter Jansen, Elena Ranguelova, and Rob V. van Nieuwpoort. *LOFAR dataset for deep learning assisted data Inspection for radio astronomy*. Version V1.0.0. Mar. 2020. DOI: [10.5281/zenodo.3702430](https://doi.org/10.5281/zenodo.3702430). URL: <https://doi.org/10.5281/zenodo.3702430>.
- [108] A. R. Foley et al. "Engineering and Science Highlights of the KAT-7 Radio Telescope." In: *Monthly Notices of the Royal Astronomical Society* 460.2 (June 2016), pp. 1664–1679. ISSN: 13652966. DOI: [10.1093/mnras/stw1040](https://doi.org/10.1093/mnras/stw1040). URL: <http://arxiv.org/abs/1606.02929><http://dx.doi.org/10.1093/mnras/stw1040>.
- [109] Martijn Van Veelen. "System health in locally autonomic radio-telescopes." In: *3rd International Conference on Autonomic and Autonomous Systems, ICAS'07* (2007). DOI: [10.1109/CONIELECOMP.2007.100](https://doi.org/10.1109/CONIELECOMP.2007.100).
- [110] S. Maslakovic, I.R. Linscott, M. Oslick, and J.D. Twicken. "Excising radio frequency interference using the discrete wavelet transform." In: *Proceedings of Third International Symposium on Time-Frequency and Time-Scale Analysis (TFTS-96)*. Vol. 2. IEEE, 1996, pp. 349–352. ISBN: 0-7803-3512-0. DOI: [10.1109/TFSA.1996.547485](https://doi.org/10.1109/TFSA.1996.547485). URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.1869><http://ieeexplore.ieee.org/document/547485/>.
- [111] R.T. Lord and M.R. Inggs. "Efficient RFI suppression in SAR using a LMS adaptive filter with sidelobe suppression integrated with the range-Doppler algorithm." In: *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293)*. Vol. 1. 5. IEEE, 1994, pp. 574–576. ISBN: 0-7803-5207-6. DOI: [10.1109/IGARSS.1999.773569](https://doi.org/10.1109/IGARSS.1999.773569). URL: <http://ieeexplore.ieee.org/document/773569/>.
- [112] Joshua Kerrigan, Paul La Plante, Saul Kohn, Jonathan C. Pober, James Aguirre, et al. "Optimizing Sparse RFI Prediction using Deep Learning." In: *Monthly Notices of the Royal Astronomical Society* 11.February (July 2019), pp. 1–11. ISSN: 0035-8711. DOI: [10.1093/mnras/stz1865](https://doi.org/10.1093/mnras/stz1865). URL: <https://academic.oup.com/mnras/advance-article/doi/10.1093/mnras/stz1865/5529408>.
- [113] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. The MIT Press, 2016. ISBN: 0262035618.
- [114] Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, and Jun Long. "A Survey of Clustering with Deep Learning: From the Perspective of Network Architecture." In: *IEEE Access* 6 (2018), pp. 39501–39514. ISSN: 21693536. DOI: [10.1109/ACCESS.2018.2855437](https://doi.org/10.1109/ACCESS.2018.2855437).
- [115] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. "Deep Clustering with Convolutional Autoencoders." In: *Neural Information Processing - 24th International Conference*. Springer US, 2017, pp. 373–382. DOI: [10.1007/978-3-319-70096-0\\_39](https://doi.org/10.1007/978-3-319-70096-0_39). URL: <https://dblp.org/rec/conf/iconip/GuoLZY17.bib>.
- [116] Nicholas O Ralph et al. "Radio Galaxy Zoo: Unsupervised Clustering of Convolutionally Auto-encoded Radio-astronomical Images." In: *Publications of the Astronomical Society of the Pacific* 131.1004 (Oct. 2019), p. 108011. ISSN: 0004-6280. DOI: [10.1088/1538-3873/ab213d](https://doi.org/10.1088/1538-3873/ab213d). URL: <https://iopscience.iop.org/article/10.1088/1538-3873/ab213d>.
- [117] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. "Variational autoencoder for deep learning of images, labels and captions." In: *Advances in Neural Information Processing Systems Nips* (2016), pp. 2360–2368. ISSN: 10495258.

- [118] David R DeBoer et al. "Hydrogen Epoch of Reionization Array (HERA)." In: *Publications of the Astronomical Society of the Pacific* 129.974 (2017), p. 45001. ISSN: 0004-6280. DOI: [10.1088/1538-3873/129/974/045001](https://doi.org/10.1088/1538-3873/129/974/045001). URL: <http://dx.doi.org/10.1088/1538-3873/129/974/045001>.
- [119] Michael E. Mavroforakis and Sergios Theodoridis. "A geometric approach to support vector machine (SVM) classification." In: *IEEE Transactions on Neural Networks* 17.3 (2006), pp. 671–682. ISSN: 10459227. DOI: [10.1109/TNN.2006.873281](https://doi.org/10.1109/TNN.2006.873281).
- [120] Michael Mesarcik, Elena Rangelova, Albert-Jan Boonstra, and Rob V. van Nieuwpoort. "Improving novelty detection using the reconstructions of nearest neighbours." In: *Array* 14 (2022), p. 100182. ISSN: 2590-0056. DOI: <https://doi.org/10.1016/j.array.2022.100182>. URL: <https://www.sciencedirect.com/science/article/pii/S2590005622000388>.
- [121] Dongil Kim, Pilsung Kang, Sungzoon Cho, Hyoung Joo Lee, and Seungyong Doh. "Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing." In: *Expert Systems with Applications* 39.4 (2012), pp. 4075–4083. ISSN: 09574174. DOI: [10.1016/j.eswa.2011.09.088](https://doi.org/10.1016/j.eswa.2011.09.088). URL: <http://dx.doi.org/10.1016/j.eswa.2011.09.088>.
- [122] Haowen Xu et al. "Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications." In: *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018* 2 (2018), pp. 187–196. DOI: [10.1145/3178876.3185996](https://doi.org/10.1145/3178876.3185996).
- [123] Christoph Baur et al. "Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images." In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, 2019, pp. 161–169. DOI: [10.1007/978-3-030-11723-8\\_16](https://doi.org/10.1007/978-3-030-11723-8_16). URL: [http://link.springer.com/10.1007/978-3-030-11723-8\\_16](http://link.springer.com/10.1007/978-3-030-11723-8_16).
- [124] Markos Markou and Sameer Singh. "Novelty detection: a review." In: *Signal Processing* 83.12 (2003), pp. 2499–2521. ISSN: 01651684. DOI: [10.1016/j.sigpro.2003.07.019](https://doi.org/10.1016/j.sigpro.2003.07.019).
- [125] Faruk Ahmed and Aaron Courville. "Detecting semantic anomalies." In: (2019). ISSN: 2374-3468. DOI: [10.1609/aaai.v34i04.5712](https://doi.org/10.1609/aaai.v34i04.5712). URL: <http://arxiv.org/abs/1908.04388>.
- [126] Shehroz S. Khan and Michael G. Madden. "A survey of recent trends in one class classification." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6206 LNAI (2010), pp. 188–197. ISSN: 03029743. DOI: [10.1007/978-3-642-17080-5\\_21](https://doi.org/10.1007/978-3-642-17080-5_21).
- [127] John Sipple. "Interpretable, Multidimensional, Multimodal Anomaly Detection with Negative Sampling for Detection of Device Failure." In: (2020). URL: <http://arxiv.org/abs/2007.10088>.
- [128] Raghavendra Chalapathy and Sanjay Chawla. "Deep Learning for Anomaly Detection: A Survey." In: (2019), pp. 1–50. URL: <http://arxiv.org/abs/1901.03407>.
- [129] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minhoo Jin, and Tomas Pfister. "Learning and Evaluating Representations for Deep One-class Classification." In: (2020), pp. 1–32. URL: <http://arxiv.org/abs/2011.02578>.

- [130] Ye Fei, Chaoqin Huang, Cao Jinkun, Maosen Li, Ya Zhang, and Cewu Lu. "Attribute Restoration Framework for Anomaly Detection." In: *IEEE Transactions on Multimedia* April (2020), pp. 1–14. ISSN: 19410077. DOI: [10.1109/TMM.2020.3046884](https://doi.org/10.1109/TMM.2020.3046884).
- [131] Wei-Yu Lee and Yu-Chiang Frank Wang. "Learning Disentangled Feature Representations For Anomaly Detection." In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2156–2160. ISBN: 978-1-7281-6395-6. DOI: [10.1109/ICIP40778.2020.9191201](https://doi.org/10.1109/ICIP40778.2020.9191201). URL: <https://ieeexplore.ieee.org/document/9191201/>.
- [132] Eungi Hong and Yoonsik Choe. "Latent feature decentralization loss for one-class anomaly detection." In: *IEEE Access* 8 (2020), pp. 165658–165669. ISSN: 21693536. DOI: [10.1109/ACCESS.2020.3022646](https://doi.org/10.1109/ACCESS.2020.3022646).
- [133] Jia Guo, Guannan Liu, Yuan Zuo, and Junjie Wu. "An Anomaly Detection Framework Based on Autoencoder and Nearest Neighbor." In: *2018 15th International Conference on Service Systems and Service Management, ICSSSM 2018* (2018), pp. 1–6. DOI: [10.1109/ICSSSM.2018.8464983](https://doi.org/10.1109/ICSSSM.2018.8464983).
- [134] Yann LeCun, Corinna Cortes, and CJ Burges. "MNIST handwritten digit database." In: *ATT Labs* 2 (2010). URL: <http://yann.lecun.com/exdb/mnist>.
- [135] Divya Saxena and Jiannong Cao. "Generative Adversarial Networks (GANs)." In: *ACM Computing Surveys* 54.3 (June 2021), pp. 1–42. ISSN: 0360-0300. DOI: [10.1145/3446374](https://doi.org/10.1145/3446374). URL: <http://arxiv.org/abs/2005.00065><https://dl.acm.org/doi/10.1145/3446374>.
- [136] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. "PANDA: Adapting Pre-trained Features for Anomaly Detection and Segmentation." In: (2020). DOI: [2010.05903](https://doi.org/2010.05903). URL: <http://arxiv.org/abs/2010.05903>.
- [137] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. "The MVTEC Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection." In: *International Journal of Computer Vision* 129.4 (2021), pp. 1038–1059. ISSN: 15731405. DOI: [10.1007/s11263-020-01400-4](https://doi.org/10.1007/s11263-020-01400-4). URL: <https://doi.org/10.1007/s11263-020-01400-4>.
- [138] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In: *Advances in Neural Information Processing Systems* 3. January (2014), pp. 2672–2680. ISSN: 10495258.
- [139] Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. "Unsupervised Anomaly Detection with Multi-scale Interpolated Gaussian Descriptors." In: (2021). URL: <http://arxiv.org/abs/2101.10043>.
- [140] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. "Deep Anomaly Detection with Outlier Exposure." In: (Dec. 2018), pp. 1–18. URL: <http://arxiv.org/abs/1812.04606>.
- [141] A. Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. 2009. URL: <https://www.cs.toronto.edu/~kriz/cifar.html>.
- [142] Han Xiao, Kashif Rasul, and Roland Vollgraf. "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms." In: (2017), pp. 1–6. URL: <http://arxiv.org/abs/1708.07747>.

- [143] Michael Mesarcik, Albert-Jan Boonstra, Christiaan Meijer, Walter Jansen, Elena Rangelova, and Rob van Nieuwpoort. *LOFAR dataset for deep learning assisted data inspection for radio astronomy*. Amsterdam, 2020. URL: <https://zenodo.org/record/3702430#.Xme2gJNKjRY>.
- [144] Yann Lecun, Le'on Bottou, Yoshua Bengio, and Parick Haffner. "Gradient-based learning applied to document recognition." In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. ISSN: 00189219. DOI: [10.1109/5.726791](https://doi.org/10.1109/5.726791). URL: <http://ieeexplore.ieee.org/document/726791/#full-text-section><http://ieeexplore.ieee.org/document/726791/>.
- [145] Diederik P. Kingma and Jimmy Lei Ba. "Adam: A method for stochastic optimization." In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015), pp. 1–15.
- [146] Shashanka Venkataramanan, Kuan Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. "Attention Guided Anomaly Localization in Images." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12362 LNCS (2020), pp. 485–503. ISSN: 16113349. DOI: [10.1007/978-3-030-58520-4\\_29](https://doi.org/10.1007/978-3-030-58520-4_29).
- [147] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. "Uninformed Students: Student-Teacher Anomaly Detection with Discriminative Latent Embeddings." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2020), pp. 4182–4191. ISSN: 10636919. DOI: [10.1109/CVPR42600.2020.00424](https://doi.org/10.1109/CVPR42600.2020.00424).
- [148] Pramuditha Perera and Vishal M. Patel. "Deep transfer learning for multiple class novelty detection." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2019), pp. 11536–11544. ISSN: 10636919. DOI: [10.1109/CVPR.2019.01181](https://doi.org/10.1109/CVPR.2019.01181).
- [149] Michael Mesarcik, Albert-Jan Boonstra, Elena Rangelova, and Rob V. van Nieuwpoort. "Learning to detect radio frequency interference in radio astronomy without seeing it." In: *Monthly Notices of the Royal Astronomical Society* 516.4 (Sept. 2022), pp. 5367–5378. ISSN: 0035-8711. DOI: [10.1093/mnras/stac2503](https://doi.org/10.1093/mnras/stac2503). eprint: <https://academic.oup.com/mnras/article-pdf/516/4/5367/48490907/stac2503.pdf>. URL: <https://doi.org/10.1093/mnras/stac2503>.
- [150] Michael Mesarcik, Albert-Jan Boonstra, Elena Rangelova, and Rob V. van Nieuwpoort. *Learning to detect radio frequency interference in radio astronomy without seeing it*. June 2022. DOI: [10.5281/zenodo.6724065](https://doi.org/10.5281/zenodo.6724065). URL: <https://doi.org/10.5281/zenodo.6724065>.
- [151] Olivier R. Hainaut and Andrew P. Williams. "Impact of satellite constellations on astronomical observations with ESO telescopes in the visible and infrared domains." In: *Astronomy and Astrophysics* 636 (2020), pp. 1–11. ISSN: 14320746. DOI: [10.1051/0004-6361/202037501](https://doi.org/10.1051/0004-6361/202037501).
- [152] W. A. Baan, P. A. Fridman, and R. P. Millenaar. "Radio Frequency Interference Mitigation at the Westerbork Synthesis Radio Telescope: Algorithms, Test Observations, and System Implementation." In: *The Astronomical Journal* 128.2 (2004), pp. 933–949. ISSN: 0004-6256. DOI: [10.1086/422350](https://doi.org/10.1086/422350).

- [153] S. Maslakovic, I.R. Linscott, M. Oslick, and J.D. Twicken. "A Library-Based Approach to Design of Smooth Orthonormal Wavelets." In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*. Vol. 3. IEEE, 1998, pp. 1793–1796. ISBN: 0-7803-4428-6. DOI: [10.1109/ICASSP.1998.681808](https://doi.org/10.1109/ICASSP.1998.681808). URL: <http://ieeexplore.ieee.org/document/681808/>.
- [154] A. R. Offringa, J. J. Van De Gronde, and J. B.T.M. Roerdink. "A morphological algorithm for improving radio-frequency interference detection." In: *Astronomy and Astrophysics* 539 (2012). ISSN: 00046361. DOI: [10.1051/0004-6361/201118497](https://doi.org/10.1051/0004-6361/201118497).
- [155] P.-S. Kildal, L. Baker, and T. Hagfors. "Development of a dual-reflector feed for the Arecibo radio telescope: an overview." In: *IEEE Antennas and Propagation Magazine* 33.5 (Oct. 1991), pp. 12–18. ISSN: 1045-9243. DOI: [10.1109/74.97945](https://doi.org/10.1109/74.97945). URL: <http://ieeexplore.ieee.org/document/97945/>.
- [156] Marcin Sokolowski et al. "BIGHORNS - Broadband Instrument for Global Hydrogen Reionisation Signal." In: *Publications of the Astronomical Society of Australia* 32 (2015). ISSN: 14486083. DOI: [10.1017/pasa.2015.3](https://doi.org/10.1017/pasa.2015.3).
- [157] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U. Gutmann, and Charles Sutton. "VEEGAN: Reducing mode collapse in GANs using implicit variational learning." In: *Advances in Neural Information Processing Systems 2017-Decem.Nips 2017* (2017), pp. 3309–3319. ISSN: 10495258.
- [158] Jeff Johnson, Matthijs Douze, and Herve Jegou. "Billion-Scale Similarity Search with GPUs." In: *IEEE Transactions on Big Data* 7.3 (2021), pp. 535–547. ISSN: 23327790. DOI: [10.1109/TBDATA.2019.2921572](https://doi.org/10.1109/TBDATA.2019.2921572).
- [159] J Akeret, C Chang, A Lucchi, and A Refregier. "Radio frequency interference mitigation using deep convolutional." In: *Astronomy and Computing* 18 (2017), pp. 35–39. ISSN: 2213-1337. DOI: [10.1016/j.ascom.2017.01.002](https://doi.org/10.1016/j.ascom.2017.01.002). URL: <http://dx.doi.org/10.1016/j.ascom.2017.01.002>.
- [160] Michael Mesarcik, Albert-Jan Boonstra, Marco Iacobelli, Cees de Laat, Elena Ranguelova, and Rob V. van Nieuwpoort. "The ROAD to discovery: machine learning-driven anomaly detection in radio astronomy spectrograms." In: *Astronomy & Astrophysics* 680 (Dec. 2023), A74. ISSN: 0004-6361. DOI: [10.1051/0004-6361/202347182](https://doi.org/10.1051/0004-6361/202347182). eprint: <https://www.aanda.org/articles/aa/pdf/2023/12/aa47182-23.pdf>.
- [161] Michael Mesarcik, Albert-Jan Boonstra, Marco Iacobelli, Cees de Laat, Elena Ranguelova, and Rob V. van Nieuwpoort. *The radio observatory anomaly detector dataset*. June 2023. DOI: [10.5281/zenodo.8028045](https://doi.org/10.5281/zenodo.8028045). URL: <https://doi.org/10.5281/zenodo.8028045>.
- [162] Sebastiaan van der Tol, Bram Veenboer, and André R. Offringa. "Image Domain Gridding: a fast method for convolutional resampling of visibilities." In: *Astronomy & Astrophysics* 616 (Aug. 2018), A27. ISSN: 0004-6361. DOI: [10.1051/0004-6361/201832858](https://doi.org/10.1051/0004-6361/201832858).
- [163] A. R. Offringa et al. "ws-clean: an implementation of a fast, generic wide-field imager for radio astronomy." In: *Monthly Notices of the Royal Astronomical Society* 444.1 (Oct. 2014), pp. 606–619. ISSN: 0035-8711. DOI: [10.1093/mnras/stu1368](https://doi.org/10.1093/mnras/stu1368).

- [164] Chiara Salvoni, Rob R. van den Bergh, Tammo Jan Dijkema, Jakob Maljaars, Maik Nijhuis, Andre R. Offringa, Sebastiaan van der Tol, Mark de Wever, and Stefan J. Wijnholds. "Calibration and Imaging Pipeline Processing Baseline-dependent Averaged Visibilities." In: *2022 3rd URSI Atlantic and Asia Pacific Radio Science Meeting (AT-AP-RASC)*. IEEE, May 2022, pp. 1–4. ISBN: 978-9-4639-6-8058. DOI: [10.23919/AT-AP-RASC54737.2022.9814290](https://doi.org/10.23919/AT-AP-RASC54737.2022.9814290).
- [165] Sriram Baireddy, Sundip R. Desai, James L. Mathieson, Richard H. Foster, Moses W. Chan, Mary L. Comer, and Edward J. Delp. "Spacecraft Time-Series Anomaly Detection Using Transfer Learning." In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2021, pp. 1951–1960. ISBN: 978-1-6654-4899-4. DOI: [10.1109/CVPRW53098.2021.00223](https://doi.org/10.1109/CVPRW53098.2021.00223).
- [166] Lilly Spirkovska, David Iverson, David Hall, William Taylor, Ann Patterson-Hine, Barbara Brown, Bob Ferrell, and Robert Waterman. "Anomaly Detection for Next-Generation Space Launch Ground Operations." In: *SpaceOps 2010 Conference*. Reston, Virginia: American Institute of Aeronautics and Astronautics, Apr. 2010. ISBN: 978-1-62410-164-9. DOI: [10.2514/6.2010-2182](https://doi.org/10.2514/6.2010-2182).
- [167] S. Yatawatta et al. "Initial deep LOFAR observations of epoch of reionization windows." In: *Astronomy & Astrophysics* 550 (Feb. 2013), A136. ISSN: 0004-6361. DOI: [10.1051/0004-6361/201220874](https://doi.org/10.1051/0004-6361/201220874).
- [168] C. Vocks et al. "LOFAR observations of the quiet solar corona." In: *Astronomy & Astrophysics* 614 (June 2018), A54. ISSN: 0004-6361. DOI: [10.1051/0004-6361/201630067](https://doi.org/10.1051/0004-6361/201630067).
- [169] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778. ISSN: 10636919. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [170] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A ConvNet for the 2020s." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Jan. 2022.
- [171] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In: *International Conference on Learning Representations*. Oct. 2021.
- [172] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. "How to Use t-SNE Effectively." In: *Distill* 1.10 (Oct. 2016). DOI: [10.23915/distill.00002](https://doi.org/10.23915/distill.00002).
- [173] Rishi Bommasani et al. "On the Opportunities and Risks of Foundation Models." In: *ArXiv* 2108.07258 (Aug. 2021).
- [174] Jeyan Thiyaalingam, Mallikarjun Shankar, Geoffrey Fox, and Tony Hey. "Scientific machine learning benchmarks." In: *Nature Reviews Physics* 4.6 (Apr. 2022), pp. 413–420. ISSN: 2522-5820. DOI: [10.1038/s42254-022-00441-7](https://doi.org/10.1038/s42254-022-00441-7).
- [175] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In: *Proceedings of The 33rd International Conference on Machine Learning, PMLR*. June 2016.
- [176] Lyndon Evans and Philip Bryant. "LHC Machine." In: *Journal of Instrumentation* 3.08 (Aug. 2008), So8001–So8001. ISSN: 1748-0221. DOI: [10.1088/1748-0221/3/08/S08001](https://doi.org/10.1088/1748-0221/3/08/S08001).

- [177] Annarita Margiotta. "The KM<sub>3</sub>NeT deep-sea neutrino telescope." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 766 (Dec. 2014), pp. 83–87. ISSN: 01689002. DOI: [10.1016/j.nima.2014.05.090](https://doi.org/10.1016/j.nima.2014.05.090).
- [178] B.P. Abbott et al. "Observation of Gravitational Waves from a Binary Black Hole Merger." In: *Physical Review Letters* 116.6 (Feb. 2016), p. 061102. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.116.061102](https://doi.org/10.1103/PhysRevLett.116.061102).
- [179] Tailin Liang, John Glossner, Lei Wang, Shaobo Shi, and Xiaotong Zhang. "Pruning and quantization for deep neural network acceleration: A survey." In: *Neurocomputing* 461 (2021), pp. 370–403. ISSN: 18728286. DOI: [10.1016/j.neucom.2021.07.045](https://doi.org/10.1016/j.neucom.2021.07.045).
- [180] Jonathan Carter, John Feddema, Doug Kothe, Rob Neely, Jason Pruet, and Rick Stevens. *ADVANCED RESEARCH DIRECTIONS ON AI FOR SCIENCE, ENERGY, AND SECURITY*. Tech. rep. ARGONNE NATIONAL LABORATORY, May 2023. URL: <https://www.anl.gov/sites/www/files/2023-05/AI4SESReport-2023.pdf>.

## ACKNOWLEDGMENTS

---

I want to express my gratitude to everyone who supported me throughout my PhD journey. Firstly, a special thanks to Tammy for her unwavering support and commitment, even relocating across the world to be with me during this challenging process.

Rob, I appreciate your excellent supervision, patience, and positive attitude that played a crucial role in helping me navigate through my PhD. Your kindness, open-mindedness, and encouragement to explore beyond our familiar territory were invaluable. Despite minor setbacks, your ability to guide me back on track was truly impactful.

Albert-Jan, your profound knowledge of radio astronomy and your optimistic outlook significantly contributed to my research. Thank you for generously allowing me to use your desk at Astron and facilitating knowledge transfer moments at the observatory.

Elena, thank you for your consistent encouragement, suggesting conferences to attend, and dedicated guidance were instrumental in my success. Your patience and grounded approach provided the support I needed to reach the finish line.

Cees, I'm grateful for your guidance and directness. Your challenging questions made me think on my feet, and defending my research to you prepared me well for handling reviewer feedback.

My heartfelt thanks extend to my friends, both near and far, for providing much-needed distractions from daily challenges. The visits, vacations, and moments of escape offered valuable perspective on my work.

To the SNE cluster members, including Angelos, Ben, Christopher, Cyril, Damian, Daphnee, Dolly, Floris-Jan, Gabriel, Henk, Hongyun, Jelle, Joe, Julius, Leonardo, Lukas, Lukasz, Marco, Marius, Na, Peter, Pooya, Ralph, Reggie, Saba, Spiros, Xin, Yixian, and Yuandou, thanks for being there to listen my cries for help. Also thank you to the MNS staff; Adam, Chrysa and Paola for giving invaluable feedback during our weekly meetings.

Lastly, my deepest appreciation goes to my family for their unwavering support and efforts to try understand what I have been doing the last few years. Thank you, Bob, David, Paul, and Marie!



# SUMMARY

---

Radio telescopes are getting bigger and are generating increasing amounts of data to improve their sensitivity and resolution. The growing system size and resulting complexity increases the likelihood of unexpected events occurring thereby producing datasets containing anomalies. These events include failures in instrument electronics, miscalibrated observations, environmental and astronomical effects such as lightning and solar storms as well as problems in data processing systems among many more. Currently, efforts to diagnose and mitigate these events are performed by human operators, who manually inspect intermediate data products to determine the success or failure of a given observation. The accelerating data-rates coupled with the lack of automation results in operator-based data quality inspection becoming increasingly infeasible.

The exponential growth of data production from modern instruments have made data-driven techniques and machine learning appealing to astronomers and telescope operators. However efforts in machine learning-based anomaly detection are concentrated in scientific discovery rather than System Health Management (SHM), with approaches ranging from detecting unusual galaxy morphologies to identifying new transients.

Telescope operators have successfully automated most scientific data processing pipelines from calibration to Radio Frequency Interference (RFI) mitigation and dedispersion. Furthermore, continuous efforts are being made to create high-performance real-time algorithms, to improve the quality and reliability of the scientific data. As of yet, there have been no attempts to fully automate the SHM pipeline, and by virtue of the lack of work on this topic, no real-time implementations exist.

This thesis focuses on applying machine learning-based anomaly detection to spectrograms obtained from the LOFAR telescope for the purpose of SHM. It does this across several chapters, with each chapter focusing on a different aspect of SHM in radio telescopes. We provide an overview of the data processing systems in LOFAR so to create a workflow for SHM that could effectively be integrated into the scientific data processing pipeline.

First we investigate how representation learning methods can be used to create low dimensional projections of radio astronomy spectrograms. In doing so, we provide an operator-in the loop tool for data inspection, for managing data quality in radio telescopes. We propose an autoencoding architecture that separately processes both the magnitude and phase components of the radio astronomy spectrograms. In this chapter we also introduce the combinatorial feature-compounding problem in radio astronomy spectrograms as well as a simple method to evaluate performance using synthesised data.

We find that existing anomaly detection problem definitions are not well suited to the radio observatory use-case. As typical methods rely on the assumption that there is a single outlying class, whereas we want to detect both commonly occurring anomalies as well as those that are rare. Using this modified problem definition, we use the learnt representations from the previous chapter to perform anomaly detection. This is done using density-based and reconstruction-error methods for anomaly detection in latent-variable models. Here we also look at combining supervised and unsupervised approaches in order to detect both known and unknown events.

Next we consider how to apply these anomaly detection strategies to RFI detection in radio astronomy. In so doing we propose a novel formulation of the RFI detection problem and obtain state-of-the-art performance. Furthermore, we demonstrate that our method better generalises to unseen RFI, whereas current supervised approaches over-fit to weak-label-based RFI masks. As a consequence, we hypothesise that our approach will better generalise to future generations of emitters, whereas existing supervised methods will have to be regularly retrained.

Lastly, we propose the Radio Observatory Anomaly Detector (ROAD), a self-supervised learning anomaly detection method. We show that it is effective in both detecting both seen and unseen anomalies with a suitably low false positive rate. Additionally, we introduce a new dataset specifically created for the validation of anomaly detection models for system health management in radio telescopes. We show that ROAD is effectively real-time in the context of the LOFAR telescope and how with some additional training could be adapted to work with other telescopes and future developments in radio astronomy. The work we present in this thesis demonstrates that machine learning indeed is a valuable tool to reliably perform anomaly detection and SHM for radio telescopes.

## SAMENVATTING

---

Radiotelescopen worden groter en genereren steeds grotere hoeveelheden gegevens om hun gevoeligheid en resolutie te verbeteren. De groeiende omvang van het systeem en de resulterende complexiteit vergroten de kans op onverwachte verstoringen, waardoor datasets met anomalieën ontstaan. Voorbeelden van zulke verstoringen zijn storingen in de elektronica, verkeerd gekalibreerde waarnemingen, omgevings- en astronomische effecten zoals bliksem en zonnevlammen, evenals problemen in dataverwerkingssystemen en nog veel meer. Momenteel wordt de diagnose en het verhelpen van zulke onverwachte verstoringen uitgevoerd door menselijke operators, die handmatig tussenproducten van gegevens inspecteren om het succes of falen van een astronomische waarneming te bepalen. De steeds maar groeiende gegevenssnelheden in combinatie met het gebrek aan automatisering maken de inspectie van gegevenskwaliteit steeds minder haalbaar.

De exponentiële groei van gegevensproductie uit moderne instrumenten heeft data-gestuurde technieken en machine learning aantrekkelijk gemaakt voor astronomen en telescoopoperators. Op dit moment is onderzoek naar machine learning gebaseerde anomaliedetectie echter voornamelijk gericht op wetenschappelijke ontdekking in plaats van op systeemdiagnose, variërend van het detecteren van ongebruikelijke morfologie van sterrenstelsels tot het identificeren van variabele bronnen.

Telescoopoperators hebben met succes de meeste wetenschappelijke dataverwerking geautomatiseerd, van kalibratie tot Radio Frequency Interference (RFI) mitigatie en dedispersie. Bovendien worden voortdurend inspanningen geleverd om real-time algoritmes met hoge prestaties te creëren om de kwaliteit en betrouwbaarheid van wetenschappelijke gegevens te verbeteren. Tot nu toe zijn er echter geen pogingen ondernomen om systeemdiagnose volledig te automatiseren, en door het gebrek aan ontwikkelingen op dit gebied bestaan er geen real-time implementaties.

Dit proefschrift richt zich op het toepassen van op machine learning gebaseerde anomaliedetectie op spectrogrammen van de LOFAR-telescoop voor systeemdiagnose. Dit gebeurt in de verschillende hoofdstukken, waarbij elk hoofdstuk zich richt op een ander aspect van systeemdiagnose in

radiotelescopen. We geven een overzicht van de gegevensverwerkingssystemen in LOFAR om zo een workflow voor systeemdiagnose te creëren die effectief geïntegreerd kan worden in het instrument.

Eerst onderzoeken we hoe methoden voor representatieleren kunnen worden gebruikt om laag-dimensionale projecties van radioastronomie spectrogrammen te creëren. Hierbij bieden we een hulpmiddel voor data inspectie waar de operator aan het stuurwiel zit voor het beheren van de gegevenskwaliteit in radiotelescopen. We stellen een auto-encoder architectuur voor die afzonderlijk de magnitude- en fasecomponenten van de spectrogrammen verwerkt. In dit hoofdstuk beschrijven we ook de uitdagingen als er meerdere anomalieën per spectrogram zijn, evenals een eenvoudige methode om de prestaties te evalueren met gesynthetiseerde gegevens.

We constateren dat bestaande oplossingen voor anomaliedetectie niet erg geschikt zijn voor het gebruik in radiotelescopen. Bestaande methoden veronderstellen dat er een enkele afwijkende klasse is, terwijl we juist zowel veelvoorkomende anomalieën als zeldzame en volledig nieuwe willen detecteren. Met deze aangepaste probleemdefinitie gebruiken we de geleerde representaties uit het vorige hoofdstuk om anomaliedetectie uit te voeren. Dit gebeurt met behulp van dichtheid gebaseerde en reconstructiefout methoden. Hier kijken we ook naar het combineren van leren onder toezicht en onbegeleide benaderingen om zowel bekende als onbekende verstoringen te detecteren.

Vervolgens overwegen we hoe we deze anomaliedetectie strategieën kunnen toepassen op RFI-detectie in de radioastronomie. Hierbij stellen we een nieuwe formulering voor van het RFI-detectieprobleem en behalen we hogere prestaties dan bestaande methoden. Bovendien tonen we aan dat onze methode beter generaliseert naar nog niet eerder geziene RFI, terwijl bestaande methodes dat niet kunnen. Als gevolg hiervan veronderstellen we dat onze aanpak beter zal generaliseren naar toekomstige generaties (versturende ) zenders, terwijl bestaande modellen regelmatig opnieuw moeten worden getraind.

Tot slot stellen we de Radio Observatory Anomaly Detector (ROAD) voor, een zelf-lerende anomaliedetectie methode. We tonen aan dat ROAD effectief is in het detecteren van zowel bekende als onbekende anomalieën met een laag aantal vals-positieven. Daarnaast introduceren we een nieuwe dataset die speciaal is gemaakt voor de validatie van anomaliedetectie modellen voor systeembeheer in radiotelescopen. We laten zien dat ROAD real-time kan opereren in de context van de LOFAR-telescoop en ook hoe het met een beperkte hoeveelheid extra training eenvoudig kan

worden aangepast aan andere telescopen en toekomstige ontwikkelingen in de radioastronomie. Het werk dat we presenteren in dit proefschrift demonstreert dat machine learning inderdaad een waardevolle techniek is om op een betrouwbare manier anomaliedetectie en systeembeheer te implementeren voor radiotelescopen.