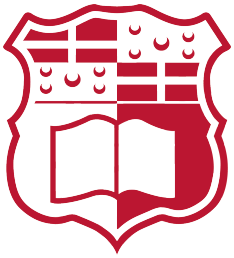


Enhancing the Northern Cross Radio Telescope for Fast Radio Burst Searching

Hayley Camilleri

Supervised by Prof A. Magro

Co-supervised by Dr K. Bugeja

Institute of Space Sciences and Astronomy

University of Malta

September, 2025

A dissertation submitted in partial fulfilment of the requirements for the degree of M.Sc. in Space Sciences and Astronomy.

L-Università ta' Malta

University of Malta Library – Electronic Thesis & Dissertations (ETD) Repository

The copyright of this thesis/dissertation belongs to the author. The author's rights in respect of this work are as defined by the Copyright Act (Chapter 415) of the Laws of Malta or as modified by any successive legislation.

Users may access this full-text thesis/dissertation and can make use of the information contained in accordance with the Copyright Act provided that the author must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the prior permission of the copyright holder.

Copyright ©2025 University of Malta

WWW.UM.EDU.MT

First edition, November 20, 2025

L-Università ta' Malta

Declaration by Postgraduate Students

(a) Authenticity of Dissertation

I hereby declare that I am the legitimate author of this Dissertation and that it is my original work. No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education. I hold the University of Malta harmless against any third party claims with regard to copyright violation, breach of confidentiality, defamation and any other third party right infringement.

(b) Research Code of Practice and Ethics Review Procedures

I declare that I have abided by the University's Research Ethics Review Procedures. Research Ethics & Data Protection form code SCI-2025-00074.

As a Master's student, as per Regulation 58 of the General Regulations for University Postgraduate Awards, I accept that should my dissertation be awarded a Grade A, it will be made publicly available on the University of Malta Institutional Repository.

Faculty/Institute/Centre/School	Institute of Space Sciences and Astronomy
Degree	M.Sc. in Space Sciences and Astronomy
Title	Enhancing the Northern Cross Radio Telescope for Fast Radio Burst Searching
Candidate (Id.)	Hayley Camilleri ([redacted])
Signature of Student	[redacted]
Date	November 20, 2025

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof Alessio Magro and my co-supervisor, Dr Keith Bugeja, for their continuous guidance, support, patience and encouragement throughout the course of this dissertation. Their expertise and thoughtful feedback have been invaluable in shaping this work and in developing my skills as a researcher.

I am also grateful to the research group at the Northern Cross in Medicina for their collaboration and for providing such a welcoming academic environment. Their insights and discussions during online meetings and visits greatly enriched the progress of this project.

I would also like to express my heartfelt appreciation to my girlfriend Giulia, whose constant support, fruitful discussions, and guidance have been invaluable throughout this journey. Her encouragement and understanding have provided me with strength and motivation during challenging times, and her presence has been a source of balance and inspiration both within and beyond my academic work.

Special thanks go to my fellow postgrads and researchers of the ultra wide-band laboratory, whose companionship, assistance, and shared experiences made this journey both productive and enjoyable.

Finally, I wish to extend my heartfelt appreciation to my friends and family. Their support and patience helps to keep me grounded and motivate me to keep pursuing my goals.

Abstract

Fast Radio Bursts (FRBs) are short, bright radio transients whose real-time detection is critical for understanding their origins. This project evaluates and optimises the Heimdall-based detection pipeline of the Northern Cross Radio Telescope for real-time deployment. Using data generated to represent data gathered from the Northern Cross, a structured grid search was performed over key hyperparameters. Performance was quantified using detection accuracy for features including the Signal to Noise Ratio (SNR), Dispersion Measure (DM), and Time of Arrival (ToA), alongside end-to-end runtime. The optimal parameter set achieved high recovery accuracy across all metrics while maintaining processing speeds compatible with real-time operations. These results demonstrate a measurable improvement in the Northern Cross pipeline's sensitivity–performance balance and strengthen its capability for ongoing FRB searches.

Contents

1 Introduction **1**

1.1 Motivation . . . 1

1.1.1 Radio Astronomy and Transient Detection . . . 2

1.1.2 Importance of Time Domain Radio Astronomy . . . 6

1.1.3 Significance of real-time detection pipelines . . . 8

1.2 Aims and Objectives . . . 10

1.3 Proposed Solution . . . 11

1.4 Outline of Dissertation . . . 12

2 Real-Time Radio Astronomy Pipelines **14**

2.1 Introduction . . . 14

2.2 Overview of FRBs . . . 15

2.2.1 Physical and Observational Properties . . . 15

2.2.2 Dispersion and Propagation Effects . . . 17

2.2.3 Repeating Versus Apparent Non Repeating FRBs . . . 18

2.2.4 Theoretical Models and Progenitor Scenarios . . . 19

2.2.5 Host Galaxies and Localisation Efforts . . . 19

2.2.6 FRBs as Cosmological and Physical Probes . . . 20

2.2.7 Relevance to Real-Time Radio Pipelines . . . 20

2.3 FRB Search Pipelines . . . 21

2.4 The Northern Cross Radio Telescope . . . 24

2.4.1 Technical Specifications . . . 24

2.4.2 Past and Current Surveys using the Northern Cross . . . 26

2.4.3 Ongoing Upgrades to the Northern Cross . . . 28

2.4.4 Comparison with Parkes, CHIME and LOFAR Telescopes . . . 29

2.5 Challenges in Real-Time Radio Pipelines . . . 32

2.5.1 Data Volumes and Processing Requirements . . . 33

2.5.2 Key Processing Steps . . . 34

2.5.3 Computational Bottlenecks and Solutions . . . 38

2.6 Software for Transient Detection . . . 39

2.6.1 Heimdall: GPU Accelerated Transient Detection . . . 39

2.6.2 Other Tools for Transient Detection . . . 40

2.7 Summary . . . 43

3 Methodology **44**
3.1 Introduction . . . 44
3.2 Data Specifications and Setup . . . 45
3.2.1 The Telescope . . . 45
3.2.2 Data Characteristics . . . 47
3.3 Computational Steps . . . 50
3.3.1 Data Selection . . . 50
3.3.2 Heimdall Processing Steps . . . 52
3.4 Forming Preliminary Results . . . 54
3.4.1 Merging and Storing Candidate Results . . . 54
3.4.2 Visualising Preliminary Results . . . 56
3.5 Statistical Tests . . . 59
3.5.1 t-SNE . . . 59
3.5.2 Cluster Analysis . . . 62
3.5.3 Friedman Test . . . 64
3.5.4 Nemenyi post-hoc Test . . . 66
3.6 Summary . . . 68

4 Results & Discussion **70**
4.1 Introduction . . . 70
4.2 Results . . . 71
4.2.1 Statistical Tests Results . . . 71
4.2.2 Final Decision . . . 79
4.3 Impact of Gulp Size on Processing Time . . . 80
4.4 Discussion . . . 81
4.5 Summary . . . 83

5 Conclusion **85**
5.1 Achieved Aims and Objectives . . . 85
5.2 Limitations . . . 86
5.3 Future Work . . . 87
5.4 Final Remarks . . . 88

Appendix A Python Packages **89**

References **92**

List of Figures

1.1 Incoherently dedispersed dynamic spectrum (bottom) and frequency-averaged time-series (top) of a burst from FRB 20240619D detected with the Westerbork RT1 25-m telescope at a central frequency of 1.27 GHz. The data are dedispersed using DM = 464.87 pc cm^{-3}. The time and frequency resolutions are labelled in the figure. The arrival time of the burst at infinite frequency at the Solar System barycentre is MJD 60511.994415889 (TDB). Figure extracted from Ould-Boukattine et al. (2024). 4

2.1 Aitoff projection showing the currently known sample of FRBs in Galactic coordinates. The FRBs shown here are predominantly those found by CHIME/FRB, which is known to have significant variations in sensitivity over the sky and only views the northern celestial hemisphere. When corrected for non-uniform sky coverage, it has been shown that the underlying distribution is isotropic, as expected for a cosmological population. Figure extracted from Petroff et al. (2022). 16

2.2 Pulse profiles for PSR B1831-03 observed at five different frequencies with the Lovell telescope and the GMRT, clearly showing the increasing effect of scattering at lower frequencies. The solid lines show exponential fits to the data. Figure extracted from Lorimer and Kramer (2005). 18

2.3 Flow chart of the key processing operations in the pipeline. Heimdall is the name of the main GPU-based pipeline implementation. Figure extracted from Barsdell (2012). 22

2.4 Aerial view of the Medicina Radio Astronomical Station, showing the N-S arm and the E-W arm. Figure extracted from Locatelli et al. (2020). 25

2.5 Scheme showing the refurbished section of the N-S arm. The red circles indicate the position of the four receivers located on the focal line of each cylinder. Figure extracted from Locatelli et al. (2020). 26

3.1 Visual plots of an FRB signal from a filterbank file from the sample data. . . . 48

3.2 Histograms of unique data points in the filterbank files which were randomly plotted to form the sample size. 52

3.3 Scatter plot representing the measured data points by Heimdall against the expected data points from the generated FRB signals. The axes represent the values of DM and ToA, while the marker size of each point is a ratio which is proportional to the representation of the SNR value. Close overlap indicates accurate recovery of both timing and dispersion measure. Horizontal offsets reflect timing errors or smearing; vertical offsets indicate dedispersion bias or DM misestimation; large separations imply poorer performance for that event setting. Axes use linear scales. It is noted that some of the measured points have not matched with their respective expected point. 58
3.4 Visualisation of how t-SNE uses a Gaussian distribution to reduce dimensionality. Figure extracted from Otten (2024). 60
3.5 A concept of HDBSCAN clustering algorithm. (a) Mutual reachability distance computation from three clusters bearing varied densities. (b) Agglomerate dendrogram showing significant clusters that can be extracted. Figure extracted from Onyango et al. (2022). 63

4.1 2-D projection of t-SNE dimensionality reduction on the data. Each point represents a parameter-file outcome embedded into two dimensions by t-SNE (axes are unitless and not directly interpretable). High-density regions indicate many outcomes with very similar feature profiles (locally preserved neighbourhoods), i.e., parameter combinations that obtained similar performance and results. Tight, well-separated clusters suggest distinct performance regimes; sparse gaps reflect low-density transitions between regimes; and small "islands" likely correspond to atypical or noisy cases. Because t-SNE prioritises local structure, distances are most meaningful within neighbourhoods, while large-scale between-cluster spacing should be interpreted qualitatively rather than metrically. 72
4.2 The same t-SNE 2-D projection from Figure 4.1 with the heatmap representing percentage accuracy for SNR (cool colours = lower accuracy; warm colours = higher accuracy, up to 100%). Dense, well-separated clusters indicate locally similar behaviour; within a cluster, smooth colour gradients suggest systematic performance trends, while mixed colours indicate parameter sensitivity or instability. Regions dominated by warm tones correspond to parameter combinations that recover SNR accurately; clusters with cooler tones highlight regimes with systematic under/over-estimation. Interpretation should focus on local neighbourhoods (t-SNE preserves local structure), while the global spacing between clusters is qualitative. 73

4.3 The same t-SNE 2-D projection from Figure 4.1 with HDBSCAN clustering. Colours indicate cluster labels returned by HDBSCAN; label –1 marks points classified as noise/outliers. Compact, well-separated colour patches denote stable density clusters, i.e. groups of outcomes with very similar behaviour, while mixed-colour boundaries reflect transitional regimes. Larger clusters represent similar performance results; smaller islands capture niche parameter settings; isolated –1 points suggest atypical or inconsistent detections. Because t-SNE preserves local neighbourhoods, interpretation should emphasise intra-cluster cohesion and boundary structure, while global inter-cluster distances are qualitative. These clusters form the basis for subsequent statistical comparisons and for selecting candidate parameter combinations. 75
4.4 Boxplot of Friedman test results for DM. 76
4.5 Boxplot of Friedman test results for SNR. 77
4.6 Critical Difference diagrams for DM, SNR and ToA. 78
4.7 Gulp duration plotted against averaged total runtime. Blue markers show the time to process a single filterbank file for a given gulp configuration (total pipeline runtime per file). Error bars denote variability across repeated runs at each gulp. The solid line is the linear fit; the dashed red line marks the real-time boundary. The green diamond highlights the operating gulp used for subsequent Heimdall tests. 81

List of Tables

3.1 Characteristics of the current acquisition system. Table extracted from Locatelli et al. (2020). . . . 46
3.2 Parameter Combination Values. . . . 53
3.3 Sample of results of performance timings (in seconds) obtained from Heimdall. This table is obtained from a file from the 400 dataset; the results show the timing metrics that contributed mostly to performance, for all parameter combinations. . . . 55
3.4 Normalised version of the total timings displayed in Table 3.3. The layout is also kept similar that the first row represents the DM Tolerance values while the first column represents the corresponding Boxcar Width values. . . . 56
3.5 Sample of results of statistical metrics obtained from analyses. In this case, this table is obtained from a file from the 400 dataset; the results show the SNR analysis. . . . 57
3.6 Critical values for the two-tailed Nemenyi test after the Friedman test. Table extracted from Demšar (2006). . . . 67

4.1 Parameter combination values which appeared in the clusters chosen. . . . 75
4.2 Table of results for Friedman test. . . . 76

List of Abbreviations

FRB Fast Radio Burst . . . 1
EM Electromagnetic . . . 1
GW Gravitational Wave . . . 1
DM Dispersion Measure . . . 1
RFI Radio Frequency Interference . . . 1
RF Radio Frequency . . . 2
SNR Signal to Noise Ratio . . . 3
FWHM Full Width Half Maximum . . . 3
PRESTO PulsaR Exploration and Search TOolkit . . . 4
CHIME Canadian Hydrogen Intensity Mapping Experiment . . . 6
LOFAR LOw-Frequency ARray . . . 6
ASKAP Australian Square Kilometre Array Pathfinder . . . 6
ML Machine Learning . . . 8
HPC High Performance Computing . . . 8
VOEvent Virtual Observatory Event . . . 8
ISM Interstellar Medium . . . 9
FETCH Fast Extragalactic Transient Candidate Hunter . . . 9
GPU Graphics Processing Unit . . . 14
DSA Deep Synoptic Array . . . 15
BPSR Berkeley-Parkes-Swinburne Recorder . . . 21
HTRU High Time Resolution Universe . . . 21
CRAFT Commensal Real-time ASKAP Fast Transient . . . 22
FREDDA Fast Real-time Engine for Dedispersing Amplitudes . . . 23
CPU Central Processing Unit . . . 23
E-W East-West . . . 24
N-S North-South . . . 24
SSA Space Situational Awareness . . . 26
BIRALES BIstatic RAdar for LEo Surveillance . . . 27
LEO Low Earth Orbit . . . 27
FPGAs Field Programmable Gate Arrays . . . 28

I/O Input/Output 34
CNN Convolutional Neural Network 35
CUDA Compute Unified Device Architecture 36
FFT Fast Fourier Transformation 39
ToA Time of Arrival 44
t-SNE t-distributed Stochastic Neighbour Embedding 44
HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise 44
IF Intermediate Frequency 45
DBSCAN Density-Based Spatial Clustering of Applications with Noise 62
CD Critical Difference 66

1

Introduction

1.1 | Motivation

In recent years, the field of radio astronomy has undergone a transformative shift, driven by advances in instrumentation, signal processing, and data acquisition. Radio astronomy has traditionally focused on the study of continuous and persistent sources that include quasars, galaxies, and pulsars; but nowadays it is increasingly focusing on transient phenomena; short-duration, often unpredictable events that provide critical insights into some of the universe's most extreme and energetic processes (Cordes and McLaughlin, 2003).

These events include Fast Radio Burst (FRB) and single pulse pulsar emissions. These brief signals, which often last only milliseconds, provide a challenge to traditional observation approaches due to their erratic nature and the large quantity of data necessary to identify them. However, the study of such events has great scientific value, ranging from investigating the intergalactic medium to determining the astrophysical origins of compact objects (Lorimer et al., 2007; Petroff et al., 2019).

The rise of time-domain radio astronomy, which focuses on the dynamic radio sky, has driven the development of new pipelines capable of processing streaming data in real-time (Law et al., 2018). Unlike standard offline searches, real-time detection systems capture transient events as they occur, allowing for fast follow-up observations across many wavelengths. These qualities are especially important for multi-messenger astronomy, which combines Electromagnetic (EM) detections with Gravitational Wave (GW) or neutrino signals (Abbott et al., 2017).

However, developing real-time detection pipelines poses a number of computational and algorithmic challenges. These include the requirement to efficiently dedisperse signals across a wide range of Dispersion Measure (DM), reduce latency, and identify true astrophysical events from Radio Frequency Interference (RFI), all while maintaining sensitivity. As data quantities increase, particularly with next generation telescopes, optimising these processes becomes not just beneficial but also essential (Macquart et al., 2015).

This dissertation focuses on one of these challenges: enhancing the process of real-time detection of the Northern Cross, a radio telescope in Medicina, Italy. By carefully examining the performance and sensitivity trade-offs of critical pipeline parameters, the objective is to enhance transient detection efficiency while retaining high accuracy. This work intends to contribute to a more robust and scalable detection approach appropriate for current time-domain astronomy by using synthetic signal injections, benchmarking, and parameter adjustment.

1.1.1 | Radio Astronomy and Transient Detection

The study of celestial objects and events by detecting Radio Frequency (RF) emissions is known as radio astronomy. Radio astronomy enables researchers to look into the universe at longer wavelengths than optical astronomy, which makes use of visible light to explore the universe. The field of radio astronomy emerged in the early 1930s, when Karl Jansky, an engineer at Bell Telephone Laboratories, unknowingly discovered radio waves originating from the centre of the Milky Way galaxy while examining causes of static that interfered with transatlantic radio communications (Jansky, 2002). His discovery launched a new discipline of astronomy, which was later spurred forward by Grote Reber's creation of the first purpose-built parabolic radio telescope in 1937 (Cohen, 2010).

Unlike optical astronomy, which detects high-energy photons in the visible range, radio astronomy excels in observing low-energy processes and cold, diffuse environments that are invisible elsewhere in the EM spectrum (McLean, 2008). This has enabled the identification of a wide range of astrophysical sources and processes, including the cosmic microwave background, synchrotron emission from supernova remnants, jets from active galactic nuclei, and thermal and non-thermal radiation from pulsars and molecular clouds (Burke et al., 2019). The ability to detect such emissions allows radio astronomers to study the structure and history of the universe at both vast and tiny scales, from galaxies to individual neutron stars. The observation and characterisation of radio transients, which are short, intense events that take place on durations ranging from microseconds to days, is one of the most attractive areas in this field.

Radio transients are of significant interest because they frequently emerge from major astrophysical settings such as neutron star magnetospheres, massive star collapses, or black hole interactions. A particularly compelling class of radio transients is FRBs,

which are millisecond-duration pulses of extragalactic origin, which were first discovered in 2007 (Lorimer et al., 2007). Despite intensive research, the physical principles underlying FRBs are still not completely understood, making them one of the most puzzling occurrences in modern astronomy. Some transitory sources are known to repeat, but others appear to be one-time phenomena. This unpredictability demands continuous sky monitoring, frequently at high temporal precision and over wide frequency bands.

In the search for transient radio signals such as FRBs and single-pulse pulsars, two key variables are often used to analyse and define a potential signal; the Signal to Noise Ratio (SNR) and the pulse's Full Width Half Maximum (FWHM). These metrics provide essential data about the signals' relevance and shape in comparison to the background noise, as can be seen in Figure 1.1, and they are used extensively in both the candidate identification and classification phases of transient detection pipelines.

SNR quantifies the strength of a signal in comparison to the underlying noise. In single-pulse searches, SNR measures the degree to which a pulse is above the mean noise level in a time series. A higher SNR often suggests a more noticeable and statistically significant event. For a time series containing a potential transient, this can be represented as,

$$x(t) = s(t) + n(t), \tag{1.1}$$

where $s(t)$ is the signal coming from the astrophysical source and $n(t)$ is the noise component, typically assumed to be stationary, have zero-mean, and Gaussian with variance σ^2_{noise} (Kay, 1998; Van Trees, 2001). For a pulse with short duration, a straightforward detection statistic is the peak amplitude of the signal in the time series, S_{peak}, which represents the maximum value attained during the event. The SNR is then calculated as,

$$\text{SNR} = \frac{S_{\text{peak}} - \mu_{\text{noise}}}{\sigma_{\text{noise}}}, \tag{1.2}$$

where S_{peak} is the peak amplitude of the signal, μ_{noise} is the mean of the background noise, and σ_{noise} is the standard deviation of the background noise (Cordes and McLaughlin, 2003; Lorimer and Kramer, 2005). This formulation is equivalent to a z-score in statistics, expressing the number of standard deviations the peak lies above the noise baseline. Under Gaussian noise assumptions, this value is directly related to the false-alarm probability, which is the probability that a noise fluctuation of equal or greater magnitude could occur by chance.

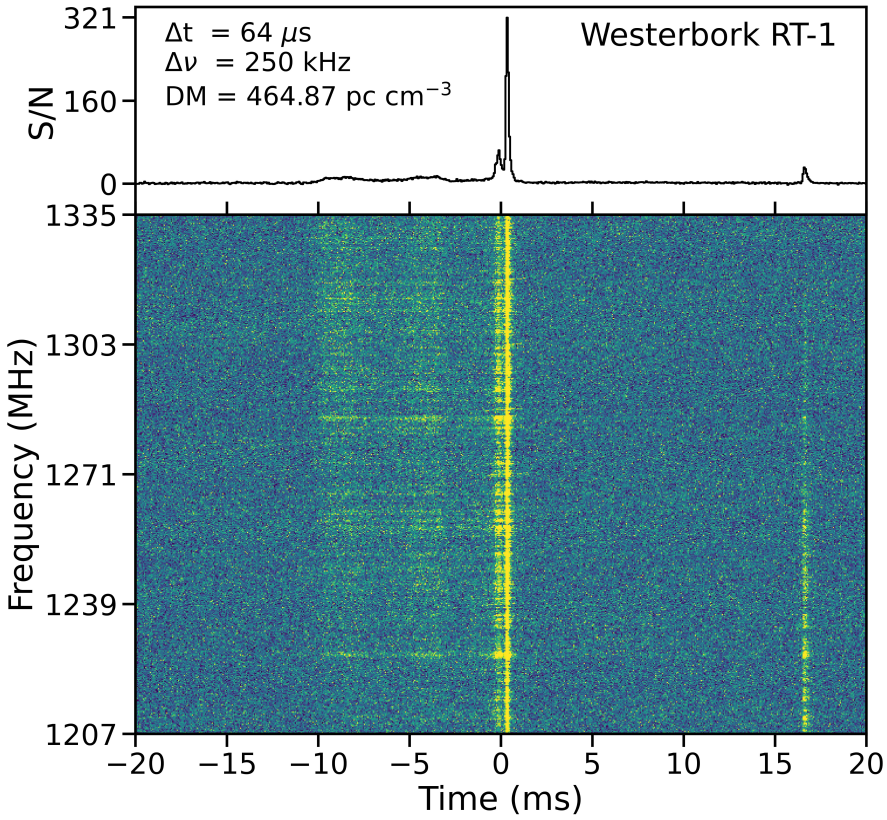


Figure 1.1: Incoherently dedispersed dynamic spectrum (bottom) and frequency-averaged time-series (top) of a burst from FRB 20240619D detected with the Westerbork RT1 25-m telescope at a central frequency of 1.27 GHz. The data are dedispersed using DM = 464.87 pc cm^{-3}. The time and frequency resolutions are labelled in the figure. The arrival time of the burst at infinite frequency at the Solar System barycentre is MJD 60511.994415889 (TDB). Figure extracted from Ould-Boukattine et al. (2024).

In practice, several transient detection software packages, which include Heimdall and PulsaR Exploration and Search TOolkit (PRESTO), calculate SNR after applying matched filtering to the time series. Matched filtering is achieved by convolving the data with an assortment of template kernels that reflect the predicted pulse forms, maximising the detection statistic when the template matches the underlying signal (Kay, 1998; Van Trees, 2001). These kernels frequently include basic boxcar functions with varying widths, which resemble the widened shape of scattered single pulses, as well as Gaussian templates that model ideal or scattering-modified pulse profiles. In fundamental terms, a signal is said to "match" the kernel when the convolution output yields a statistically significant peak: the matched filter successfully aligns with the signal's temporal structure, resulting in constructive integration of the pulse energy and max-

imum SNR relative to noise. In contrast, when the kernel does not reflect the genuine pulse shape or broadness, the convolution output is weaker, and the SNR decreases. Detection pipelines thus depend on a bank of kernels of various widths or shapes to enable sensitivity to a wide range of pulse patterns and scattering parameters. Following filtering, an SNR threshold is often used to identify true astrophysical events from random noise or RFI, while false positives can still occur, particularly in interference-rich observing conditions.

The FWHM measures a pulse's temporal duration. It describes the breadth of the signal at the point where its amplitude is half of its greatest value. This metric can be used for measuring the inherent timing of a transient event, as well as evaluating the impact of dispersive smearing, scattering, and instrumental effects.

A transient radio pulse can often be approximated by a Gaussian function in the time domain, particularly after coherent or incoherent dedispersion has removed most of the dispersion smearing (Lorimer and Kramer, 2005). For a signal modelled by a Gaussian pulse,

$$S(t) = A \exp\left(-\frac{(t - t_0)^2}{2\sigma^2}\right), \tag{1.3}$$

where $S(t)$ represents the signal amplitude, A is the peak amplitude of the pulse, t_0 is the central time of the pulse and σ is the standard deviation of the Gaussian in time. The exponential term governs how the amplitude falls from the peak symmetrically, while the denominator controls the rate of the fall-off.

The FWHM is a commonly used method for determining the width of a Gaussian pulse. This is defined as the time difference between the two points on each side of the peak where the amplitude is reduced to half of its maximum. For a Gaussian, this relationship can be derived analytically by solving;

$$\frac{S(t)}{A} = \frac{1}{2}, \tag{1.4}$$

this yields the relation,

$$\text{FWHM} = 2\sqrt{2\ln 2} \cdot \sigma \approx 2.335 \cdot \sigma. \tag{1.5}$$

In transient detection, FWHM helps define the ideal boxcar width for matched filtering, influences SNR recovery, and can help distinguish between astrophysical and non-astrophysical signals (Lorimer and Kramer, 2005). Shorter FWHMs frequently result in

better SNRs when combined with an appropriate filter, but filters that are too long or too short might suppress the signal.

The FWHM is typically used to determine pulse width since it provides a straightforward, model-independent measure of the burst's temporal width. Unlike the standard deviation (σ), which is only useful if the pulse is accurately represented by a Gaussian, FWHM can be calculated successfully even when real signals have been altered by dispersion, scattering, instrumental smearing, or asymmetry (Lorimer and Kramer, 2005). Furthermore, FWHM is naturally aligned with practical detection methods: most pipelines use banks of boxcar or Gaussian filters with widths that correspond to the observed FWHM, maximising SNR when the filter width is equivalent to the pulse duration (Cordes and McLaughlin, 2003; Kay, 1998). SNR and FWHM are important for understanding the morphology of observed pulses. An accurate estimate of both volumes is key for optimising detection pipelines and performing downstream classification or follow-up.

Detecting such short-lived signals demands refined processing pipelines capable of separating weak bursts of radio emission from massive volumes of background noise. Given the immense data rates generated by modern radio telescopes, the difficulty is not only in identifying these signals but also in doing so effectively. With a daily production of petabytes of data by facilities like Canadian Hydrogen Intensity Mapping Experiment (CHIME), LOw-Frequency ARray (LOFAR), MeerKAT, and Australian Square Kilometre Array Pathfinder (ASKAP), real-time or near-real-time detection is becoming more and more necessary. To maximise scientific gain in this instance, detection algorithm performance must be optimised (Keane and Petroff, 2015).

For instruments like the Northern Cross radio telescope, which has a long history of pulsar and transient searches, updating and fine-tuning its detection system is pivotal to maintaining scientific competitiveness (Bianchi et al., 2023). This entails re-evaluating parameter selection, processing algorithms, and detection thresholds, especially in the context of rapid, low-latency pipelines that work continuously on streaming data.

1.1.2 | Importance of Time Domain Radio Astronomy

Time domain radio astronomy examines celestial phenomena that shift over time, particularly on periods spanning from milliseconds to hours (Law et al., 2018). While tra-

ditional radio astronomy generally focuses on static or slowly evolving sources (e.g., galaxies, quasars, and H II regions), the time domain approach provides a glimpse into a dynamic universe in which rapid, often unpredictable events reveal the behaviour of some of the most extreme astrophysical systems.

This observational paradigm has gained appeal in recent years, owing partly to the uncovering of FRBs and intermittent pulsars. These events show significant temporal variability and, in certain situations, may be sporadic or one-time occurrences. The scientific relevance of these transients stems from their ability to answer fundamental astrophysical problems such as the nature of compact objects, the distribution of matter in the intergalactic medium, and the mechanics of magnetars and neutron stars (Macquart et al., 2020; McLaughlin et al., 2006).

For instance, FRBs are a strong tool for measuring cosmic baryon abundance given that their DM directly tracks the overall column density of free electrons along the line of sight. Each DM measurement includes contributions from the Milky Way, the host galaxy, and, most significantly, the intergalactic medium. By accurately localising a FRB to its host galaxy and measuring its redshift, the intergalactic medium contribution to the dispersion can be determined. Since this includes the majority of the universe's ionised baryons, the observed relationship between DM and redshift represents the cosmic distribution of baryons integrated along the path to the burst. Macquart et al. (2020) demonstrated that the average DM-z relation expected by cosmology corresponds to data from localised FRBs, allowing for direct inference of the cosmic baryon density (Ω_b). Their findings show that FRBs effectively "weigh" the universe's ionised baryons, confirming that the missing baryons are in the diffuse intergalactic medium.

Beyond astrophysical discoveries, time domain observations are important in multi-messenger astronomy, which integrates data from several "messenger" channels such as EM radiation, GWs, neutrinos, and cosmic rays (Abbott et al., 2017). Real-time detection of radio transients enables astronomers to organise follow up investigations using a variety of instruments and wavelengths. For example, detecting a radio transient in conjunction with a GW event might provide a more comprehensive picture of a binary neutron star merger or black hole accretion event.

The increasing value of time-domain radio astronomy is mirrored in the design priorities of next-generation telescopes. CHIME, MeerKAT and ASKAP are instruments designed with vast fields of view, high temporal resolution, and massive bandwidths

to allow continuous monitoring of the radio sky. These features significantly boost the chance of detecting unusual occurrences, paving the path for systematic transient surveys.

However, this possibility presents considerable computational and practical obstacles. The requirement to handle massive amounts of data in real-time, separate events from noise, and provide follow-up alerts within seconds or minutes necessitates pipelines that are both efficient and reliable. As such, time domain radio astronomy represents both a scientific and a technical frontier, with breakthroughs in Machine Learning (ML), High Performance Computing (HPC), and data systems becoming critical components of astronomical discoveries.

1.1.3 | Significance of real-time detection pipelines

The shift to real-time data processing has become an important feature of present day time domain radio astronomy. As observational abilities improve with broader fields of view, higher bandwidths, and better time precision, so does the amount and speed of data to be processed. This involves the setting up of real-time detection pipelines; which are computer systems that can recognise astrophysical transients within seconds of their occurrence.

Real-time pipelines perform several essential functions. First, they offer low latency transient detection, which is critical for time sensitive follow up observations. Many transitory phenomena, including FRBs, are not only quick but may also be accompanied by counterparts in different wavelengths (e.g., X-rays, optical) or messengers (e.g., GWs). Capturing these counterparts necessitates immediate alerts (Bannister et al., 2017), which are frequently delivered through systems such as Virtual Observatory Event (VOEvent) [1] and are only feasible when the detection pipeline runs in real-time.

Second, real-time pipelines are essential to automating the data filtering process. Given the high number of false positives, particularly in RFI contaminated areas, pipelines must efficiently separate actual astrophysical signals from noise and interference with minimal human intervention (Agarwal et al., 2020). This necessitates strong detection techniques, which frequently employ a combination of statistical filtering, ML classi-

[1]The International Virtual Observatory Alliance (IVOA) formally adopted VOEvent, a defined language for reporting observations of astronomical events (Petroff et al., 2017).

fiers, and heuristic thresholds.

A typical real-time detection pipeline includes the following stages:

- RFI excision, which involves removing known terrestrial and instrumental interference,
- Dedispersion, which corrects for frequency-dependent signal delays caused by the Interstellar Medium (ISM),
- Matched filtering, which enhances the detectability of short-duration pulses using boxcar convolution or more advanced techniques,
- Candidate identification where events are selected based on SNR, DM, and pulse width criteria,
- Candidate classification and ranking, which involves sorting and prioritising events for further analysis or alert generation.

Several software tools have been established to help with these stages, including PRESTO (Ransom, 2001), Heimdall (Barsdell and Jameson, 2024) and Fast Extragalactic Transient Candidate Hunter (FETCH) (Agarwal et al., 2020). These methods are frequently modular, allowing for incorporation into bespoke pipelines customised to the requirements of individual telescopes or investigations. For example, Heimdall does high-performance dedispersion and candidate identification, whereas FETCH classifies candidates as astrophysical or non astrophysical using deep learning.

Despite these improvements, major difficulties persist. Balancing sensitivity and speed is critical; increasing the number of dispersion attempts or applying the target pulse widths being searched for enhances the possibility of detecting weak signals while increasing computing time (Foster et al., 2018). Furthermore, pipeline tuning, the process of optimising parameter settings such as SNR thresholds, DM ranges, and boxcar widths, is frequently done manually or with minimal exploration of the parameter space.

As radio telescopes move toward continuous, real-time sky monitoring, it is critical to establish systematic ways for evaluating pipeline performance in various configurations. This involves testing on simulated data with known signal attributes, comparing run durations, and weighing false positives and missed detections.

In the case of the Northern Cross radio telescope, boosting the performance of its real-time detection process has the potential to considerably boost research output. By adjusting conditions to achieve a better balance of processing speed and detection accuracy, the system can identify transitory events in real-time, paving the path for high-impact discoveries and rapid multi-wavelength follow ups.

1.2 | Aims and Objectives

The main objective of this dissertation is to optimise the Northern Cross radio telescope's real-time transient detection pipeline in order to improve detection efficiency and flexibility while maintaining sensitivity and latency. As time domain radio astronomy develops, there is an increasing need to guarantee that older telescopes, such as the Northern Cross, are equipped with pipelines capable of meeting current data and performance requirements (Bianchi et al., 2023).

The Northern Cross has a long history of detecting pulsars and transients, but its current pipeline architecture lacks a systematic framework for targeted optimisation. Many detection settings, such as the DM range, temporal resolution, boxcar widths, and SNR thresholds, are either fixed or manually adjusted, frequently based on broad heuristics rather than a data driven examination of performance trade-offs. This reduces the system's adaptability to various observational aims and data quality conditions.

To address this, this dissertation presents a systematic examination of the impact of pipeline parameters on detection sensitivity and processing performance. The production of synthetic transients, such as FRBs, is an important component of this effort. This enables controlled testing of pipeline performance across a wide range of settings and noise levels.

The study also examines the adaptation of the Heimdall detection software. By including performance input such as candidate rates and computing load, the pipeline may adjust its own settings over time. While full closed loop automation is beyond the scope of this research, the goal is to establish the framework for smart pipeline behaviour.

The objectives of this work includes:

- To systematically evaluate the effect of key pipeline parameters on detection accuracy, data feature accuracy, and computational time.
- To develop a methodology based on synthetic FRB injections and benchmark datasets, which can be followed and adapted to replicate this methodology across other telescopes or transient detection software.
- To identify optimal parameter configurations for the Northern Cross telescope, based on realistic performance constraints.
- To provide guidelines and recommendations for future upgrades of the Northern Cross detection system and similar instruments.

By attaining these goals, the project intends to improve the potential of the Northern Cross radio telescope and contribute to the construction of robust, adaptive transient detection pipelines in radio astronomy.

1.3 | Proposed Solution

This study suggests a systematic, performance-driven technique to enhance the Northern Cross radio telescope's real-time transient detection process. The main premise is to use a controlled experimental framework based on synthetic transient injections to assess how various parameters impact detection performance as well as computing efficiency. The goal is to determine parameter combinations that offer the best balance of sensitivity and processing speed.

A key component of this approach is the use of simulated FRB injections to resemble data which is usually gathered with the Northern Cross. To replicate true astrophysical conditions, these synthetic signals are produced with known DMs, pulse lengths, and SNRs before being embedded in background noise. This technique enables precise pipeline benchmarking, since the injected signals serve as ground truth for testing detection accuracy, timing precision, and false positive rates across a range of configurations.

The transient detection pipeline used in this investigation is based on Heimdall, which is a high performance tool widely used in radio astronomy for single pulse searches which

was developed by Barsdell and Jameson (2024). While Heimdall is real-time capable and performs crucial functions like dedispersion, matched filtering, and thresholding, its default settings are not modified to the Northern Cross's hardware architecture or observing environment. As a result, there is a clear potential to modify and extend Heimdall to make it more suitable for usage with the Northern Cross.

To achieve this, the research project starts by identifying a number of key pipeline characteristics that impact detection performance. These include the DM tolerance and range, the boxcar width specified for matched filtering, and the minimum SNR threshold for candidate selection. By systematically adjusting these settings and running the pipeline on datasets with known injected signals, it is feasible to obtain a comprehensive performance map for the detection system.

This benchmarking process is supported by the development of tailored evaluation tools which quantify detection rates, false positives, and execution durations for each configuration. The findings are evaluated to determine which settings provide the optimal trade-off between computational load and performance. For example, certain settings may increase sensitivity to faint transients at the expense of longer processing times, whilst others may provide quicker throughput but lower completeness. Understanding these trade-offs is critical for building the pipeline to achieve specific scientific objectives, such as low latency FRB detection or deep pulsar searches.

This effort is planned to produce a final parameter configuration for the Northern Cross pipeline, as well as practical suggestions for future observation initiatives. Furthermore, the findings will guide future enhancements to the Northern Cross, contributing to the larger development of adaptive, real-time transient detection systems in radio astronomy.

1.4 | Outline of Dissertation

The dissertation is structured into the following chapters, where each elaborates on the motivations and objectives introduced in this chapter. Each chapter builds on the previous, progressing from the scientific background through to methodology, results, and concluding analysis.

Chapter 2 establishes the context for this research. It opens with an introduction to real-time radio astronomy and an overview of FRBs, which play a role in current time domain research. The chapter then examines typical FRB search pipelines, concentrating on architectural components and operational needs. A specific section discusses the Northern Cross radio telescope, its observational capabilities, and its significance in transient detection. The chapter also looks at the main challenges encountered in real-time radio pipelines, including latency, RFI mitigation, and resource limits. Finally, it reviews the primary software tools used for transient detection, such as Heimdall and similar frameworks.

Chapter 3 presents the experimental framework and methodologies employed in this study. It begins by describing the data specifications, which include the structure of the observational data and the properties of the synthetic FRB injections. The computing steps are then discussed, such as parameter selection, pipeline execution, and performance monitoring. A portion is dedicated to generating preliminary data that will guide further evaluation, followed by a description of the statistical tests used to measure detection performance and significance. The chapter finishes with an overview of the methodological approach and its justification.

Chapter 4 presents the empirical results and their interpretation. It begins with a brief introduction before going over the findings of the statistical tests, which include detection rates, timing accuracy, SNR recovery, and computational load. Then the focus switches to the amount of time samples which are supplied in a batch and its impact on performance. The discussion section interprets these findings in light of the project's goals, weighing the trade-offs between sensitivity and performance and considering the implications for real-time pipeline design.

Chapter 5 outlines the project's accomplishments, limits, and potential for additional research. It begins with an overview of the results and considers how the study objectives were met. The study's shortcomings are then pointed out, such as potential sources of bias and methodological constraints. The final segment discusses future development, such as establishing completely dynamic pipeline setups and investigating more advanced classification algorithms for candidate filtering.

This structure allows for the dissertation to flow logically from theoretical background to practical implementation, empirical analysis, and final conclusions, all while keeping a clear focus on the overarching goal of optimising real-time transient detection for the Northern Cross radio telescope.

2

Real-Time Radio Astronomy Pipelines

2.1 | Introduction

Modern radio astronomy has become a data intensive science that depends primarily on real-time computational processing to detect and investigate rapid, transitory astronomical events like pulsars and FRBs. As telescopes become more sensitive and bandwidths expand, the volume and velocity of data collected has increased significantly, resulting in bottlenecks that pose new problems to signal processing pipelines. Real-time detection systems must strike a careful balance between performance; the ability to handle massive data streams rapidly, sensitivity; the ability to identify weak and scattered signals with minimal information loss, and cost; since increasing hardware resources can improve both performance and sensitivity, but at the expense of higher initial capital investment, ongoing maintenance, and administrative overhead.

This chapter presents a complete review of the technological and computational landscape that supports real-time transient detection in radio astronomy. It begins with an overview of FRBs and FRB search pipelines, which have so far contributed to the current knowledge there is on the matter. This is followed by an outline of the Northern Cross Radio Telescope, the project's primary observational platform, its architecture, capabilities, and its function within the larger framework of worldwide efforts, comparing it to instruments such as CHIME, LOFAR, and Parkes. Subsequently, a discussion of the primary computing issues involved with real-time radio data processing is presented, concentrating on data throughput, dedispersion, and the need for swift, reliable detection.

The chapter also delves into the software tools and frameworks typically used in the field, focusing on Heimdall; a Graphics Processing Unit (GPU) accelerated pipeline tool for transient detection, as well as other tools used in radio astronomy pipelines like PRESTO and FETCH. These tools are evaluated not just in terms of functionality, but also in terms of performance optimisation strategies such as algorithmic decisions, hardware acceleration, and real-time limitations. Throughout, the trade-offs between computing efficiency and astrophysical sensitivity are highlighted, an important issue for the optimisation work performed in this project.

2.2 | Overview of FRBs

FRBs are very intense radio pulses that last only a few milliseconds. Lorimer et al. (2007) discovered them in 2007, when they reported a single brilliant flash in Parkes Radio Telescope archival data. This signal, subsequently termed the "Lorimer Burst," had a unique frequency sweep, with lower frequency components arriving later, which was compatible with cold plasma dispersion. The DM of around 375 $\mathrm{pc\,cm^{-3}}$ exceeds the predicted galactic contribution along that line of sight, indicating an extragalactic or potentially cosmic origin.

Following this initial discovery, the number of known FRBs has skyrocketed due to breakthroughs in high cadence surveys and real-time detection systems. Telescopes such as CHIME, ASKAP, Deep Synoptic Array (DSA)-110, and LOFAR, as well as newer additions like the Northern Cross, have all served important roles in this endeavour. As stated by Lorimer et al. (2024), as of 2024, almost a thousand FRBs have been documented. This comprises both non repeated occurrences and an increasing number of repeating FRBs. Public resources like the FRB Catalogue (Petroff et al., 2016), and the CHIME/FRB archive (Amiri et al., 2021) have proven critical for tracking burst characteristics and informing theoretical research. These databases provide essential metrics such as DMs, fluence, bandwidth, pulse widths, polarisation properties, and location information. The increasing diversity of observable burst characteristics; such as frequency structure, repetition patterns, and polarisation, indicates that FRBs may not be the result of a single progenitor type, but rather reflect a range of astrophysical processes.

2.2.1 | Physical and Observational Properties

FRBs feature multiple distinct observational traces that set them apart from other transient phenomena. Their pulse durations typically range between 0.1 to 10 ms, while certain bursts, particularly those identified at low frequencies, show substructures on µs timeframe (Snelders et al., 2023). FRBs possess dispersion measures ranging from 100 to 3000 $\mathrm{pc\,cm^{-3}}$, which typically exceed the galactic foreground contribution. This strongly supports a cosmic distance scale, with redshifts predicted to be between $z = 0.1$ and 2.5 for most sources.

Bursts are detected at a broad range of frequencies. In particular, CHIME operates at

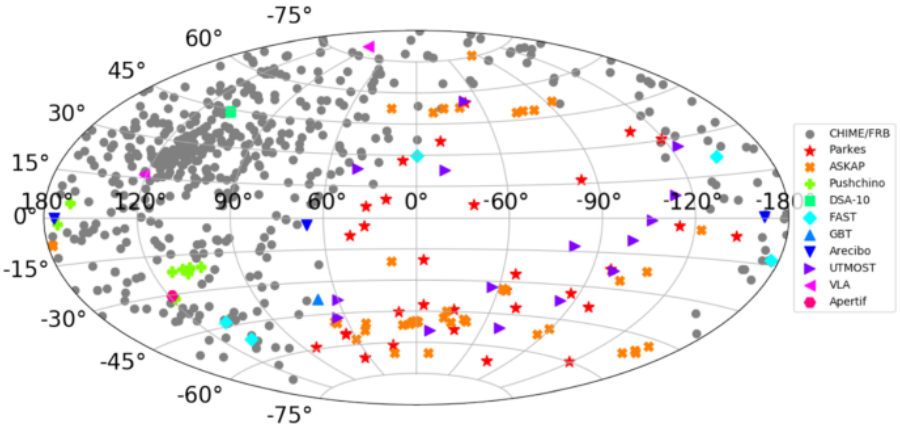


Figure 2.1: Aitoff projection showing the currently known sample of FRBs in Galactic coordinates. The FRBs shown here are predominantly those found by CHIME/FRB, which is known to have significant variations in sensitivity over the sky and only views the northern celestial hemisphere. When corrected for non-uniform sky coverage, it has been shown that the underlying distribution is isotropic, as expected for a cosmological population. Figure extracted from Petroff et al. (2022).

400 − 800 MHz, Parkes generally observes at 1.4 GHz, and other telescopes have detected bursts as high as 8 GHz. The FRBs detected with these telescopes and many others are displayed in Figure 2.1. The polarisation features are also quite varied; many bursts exhibit significant linear or circular polarisation, which is frequently accompanied by significant rotation measures. These rotation measures indicate propagation by means of magnetised plasma, implying that certain FRBs originate from settings with high magnetic fields (Michilli et al., 2018). FRBs frequently exhibit patchy or drifting structures within their burst envelopes. These characteristics may be intrinsic to the emission process or the result of propagation through intervening media. FRBs are distinguished for their exceptional brightness temperature, which can surpass 10^{35} K. These values rule out incoherent emission mechanisms while suggesting coherent processes like curvature radiation or maser amplification.

When seeking to link reported FRB features to their underlying physical origins, numerous diagnostic findings are explored. The polarisation fraction and rotation measure contribute to determining the magnetic field strength and orientation in the burst environment. The DM can be used to calculate the distance and host galaxy contribution. The spectral index and frequency structure offer information about the emission process, as well as any propagation effects like scattering or scintillation. Meanwhile, the temporal profile, which includes sub-burst drift rates and fine-time structure, reveals in-

formation on the source geometry and plasma conditions close to the emission region. By combining these multidimensional qualities, researchers may distinguish between competing hypotheses such as magnetar flares, pulsar-like magnetospheric emission, and compact binary interactions, resulting in a consistent physical description of the FRB source population.

2.2.2 | Dispersion and Propagation Effects

The presence of free electrons in the ISM and intergalactic medium causes signal dispersion, which is a distinguishing property of FRBs. The frequency dependent delay caused by this plasma dispersion is a crucial characteristic used in detection. The temporal delay between two frequencies is determined by the equation,

$$\Delta t = 4.15 \times 10^3 \text{ms} \left(\frac{1}{\nu_1^2} - \frac{1}{\nu_2^2} \right) \times \left(\frac{\text{DM}}{\text{pc cm}^{-3}} \right), \tag{2.1}$$

where ν_1 and ν_2 are frequencies in MHz and DM is the integrated column density of free electrons along the line of sight (Lorimer et al., 2024). The dispersion measure itself is given by,

$$\text{DM} = \int_0^d n_e(l) dl, \tag{2.2}$$

where $n_e(l)$ is the electron density at a distance l (Draine, 2011).

In addition to dispersion, FRBs often show propagation effects such as scattering, which is depcited in Figure 2.2, resulting in asymmetric pulse broadening, especially at lower frequencies. The scattering measure is the path integral of C_n^2 (Cordes and Lazio, 2002),

$$\text{SM} = \int_0^D ds C_n^2 \tag{2.3}$$

where D represents the independent distance measurements and C_n^2 is the spectral coefficient (the "level of turbulence").

Scintillation, generated by small scale turbulence in the plasma, introduces variation in signal intensity. Faraday rotation, which rotates the polarisation angle with frequency, offers information on the magnetic field intensity and electron density of the intervening medium. Some bursts exhibit indications of plasma lensing, which occurs when inhomogeneities in the plasma cause the signal to focus or defocus. These effects limit detection while providing a lot of information about the cosmic environment (Cordes and Wasserman, 2016).

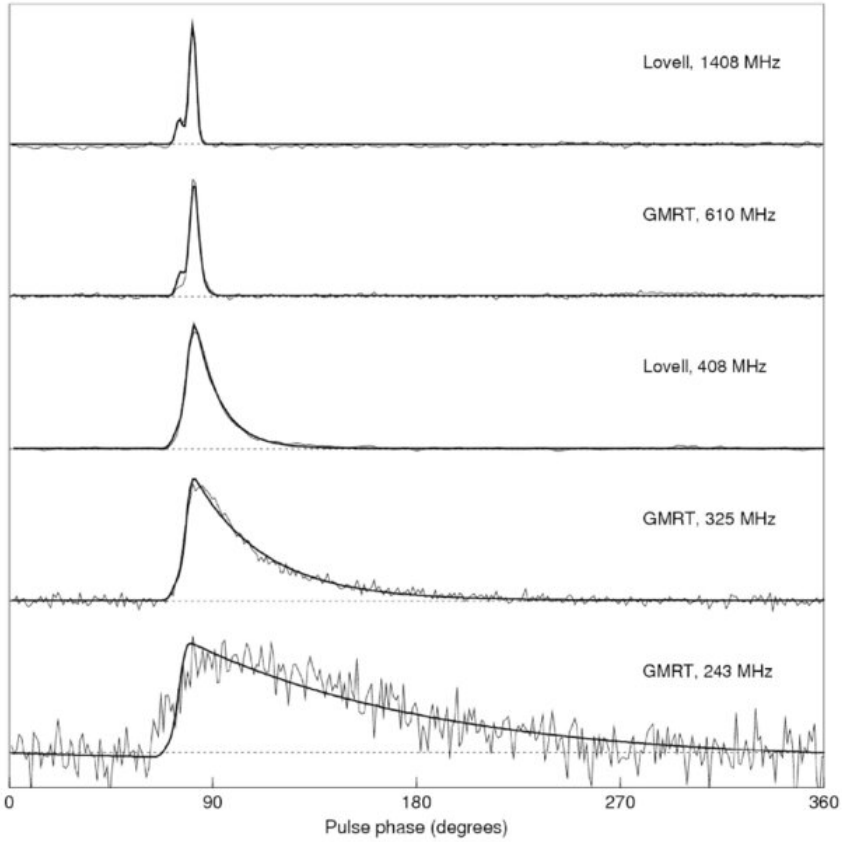


Figure 2.2: Pulse profiles for PSR B1831-03 observed at five different frequencies with the Lovell telescope and the GMRT, clearly showing the increasing effect of scattering at lower frequencies. The solid lines show exponential fits to the data. Figure extracted from Lorimer and Kramer (2005).

2.2.3 | Repeating Versus Apparent Non Repeating FRBs

Spitler et al. (2016) discovered the repeating source FRB 121102, representing a turning point in FRB research. This repeater, located in a low metallicity dwarf galaxy, demonstrated not just repeated bursts over several years, but also extraordinarily high Faraday rotation, signifying a dense, magnetised local environment. Following this finding, a number of further repeaters were found. FRB 180916.J0158+65 showed quasi-periodic activity on a $\sim$ 16-day cycle, indicating orbital motion or periodic emission.

Observationally, repeaters have broader pulse widths and more detailed substructures than non repeaters. They may also have lower average energies and downward drifting frequency patterns inside each burst. It is unknown whether these traits indicate a

fundamentally different population or are the result of observational bias. The growing perspective is that there may be a continuum of FRB behaviours, with repetition determined by both innate and environmental variables.

2.2.4 | Theoretical Models and Progenitor Scenarios

The various observational features of FRBs have resulted in a wide range of hypothesised progenitor theories. These models are roughly classified into two categories: cataclysmic or non-cataclysmic. Cataclysmic models feature one time explosive events such as neutron star mergers, supermassive neutron star collapses, or black hole accretion. These are commonly used to explain non-repeating FRBs.

Non-cataclysmic models however, account for periodic behaviour and include instances such as magnetar flares, coherent emission from pulsar like magnetospheres, and interactions with orbital material such as asteroid belts. The observation of FRB 200428, a millisecond duration burst from the galactic magnetar SGR 1935+2154 that coincided with an X-ray flare in 2020, strengthened the argument for magnetars being FRB origins. While its intensity was many orders of magnitude lower than that of cosmic FRBs, the event established the first direct connection between FRBs and known neutron star systems (Bochenek et al., 2020).

2.2.5 | Host Galaxies and Localisation Efforts

Localising FRBs to their host galaxies has proven critical in determining their extragalactic origin and understanding what surrounds them. Arrays such as ASKAP, CHIME/FRB, and the European VLBI Network (EVN) have achieved sub arcsecond localisation of numerous FRBs. These efforts have yielded a diversified host population, which includes dwarf irregulars, spiral galaxies, and even massive ellipticals. Some FRBs appear in star forming regions, whereas others are located on the galactic outskirts or in areas with minimal recent star formation (Bannister et al., 2019; Chatterjee et al., 2017; Marcote et al., 2020).

Researchers have used the redshift distribution of localised FRBs to investigate the universe's baryonic composition. Macquart et al. (2020) discovered that the measured dispersion measurements of localised FRBs scale with redshift in a way that is compatible

with theoretical expectations for baryon content, therefore resolving the long standing "missing baryon" dilemma in cosmology.

2.2.6 | FRBs as Cosmological and Physical Probes

Aside from their astrophysical appeal, FRBs are developing as instruments for precision cosmology and fundamental physics. Their DMs offer line of sight observations of free electron content, allowing for restrictions on cosmic baryon density. FRBs, when combined with redshift data from host galaxies, provide a way of studying the intergalactic medium as well as the universe's large scale structure (Zhou et al., 2014).

FRBs may also be used to study magnetic fields using rotational measurements, and they may restrict the features of dark matter through lensing, dispersion anomalies, or burst multiplicity (Ravi, 2019). Future studies including hundreds of localised FRBs could look into large scale homogeneity, cosmic reionisation history, and baryon distribution beyond galaxies (Macquart et al., 2020).

2.2.7 | Relevance to Real-Time Radio Pipelines

FRBs are not only a scientific target for instruments like the Northern Cross, but also a driving factor for technological advancement. Detecting unknown FRBs in real-time necessitates pipelines that can handle large data streams with minimal latency (Rajwade et al., 2022). Dedispersion must be conducted across hundreds or thousands of DM trials, frequently using GPU acceleration like Heimdall or AstroAccelerate (Armour et al., 2019; Barsdell and Jameson, 2024). Matched filtering across several pulse widths improves sensitivity to larger or dispersed bursts, although RFI mitigation remains critical for data quality.

Candidate classification is increasingly reliant on ML methods like FETCH, which help decrease false positives and prioritise genuine astrophysical events. Finally, real-time pipelines have to generate alert packets (e.g., VOEvents) within seconds to allow for timely multi-wavelength follow up, especially for uncommon or recurrent FRBs (Agarwal et al., 2020).

2.3 | FRB Search Pipelines

The study of FRBs has accelerated the development of real-time, high throughput data pipelines in radio astronomy. These pipelines are designed to detect millisecond long transients amidst large amounts of noisy, RFI contaminated data. While many FRB search pipelines developed from older pulsar surveys, the requirements of FRB detection, particularly real-time classification and rapid alerting, have resulted in considerable changes toward low latency, high throughput processing. Several observatories established pipelines with increasing complexity and efficiency. The main processing steps in a typical detection pipeline are represented in Figure 2.3. The steps which Heimdall carries out will be discussed in more detail in Section 2.5.2. This literature study investigates the primary FRB detection pipelines in use, examines their components and techniques, and makes comparisons that are significant to future development of instruments like that of the Northern Cross.

The Parkes Radio Telescope in Australia pioneered real-time FRB detection. The Berkeley-Parkes-Swinburne Recorder (BPSR) technology established the foundation for current pipelines by combining high resolution time series data capture and real-time processing. BPSR used Heimdall software, shown in Figure 2.3, which is a GPU accelerated tool for incoherent dedispersion and single pulse detection that can analyse thousands of DM trials at once (Barsdell, 2012). This method was essential in identifying the first repeating FRB, FRB 121102, as part of the High Time Resolution Universe (HTRU) survey, as well as creating practical dedispersion and matched filtering algorithms (Spitler et al., 2016). Although the Parkes pipeline was not originally intended for FRBs, its integration with Heimdall revealed that brute force dedispersion[1] mixed with GPU parallelism was a feasible solution for real-time detection.

The introduction of the CHIME/FRB pipeline marked an important milestone. CHIME operates at $400 - 800$ MHz and is uniquely designed with cylindrical reflectors and no moving parts. Its huge instantaneous field of view as well as excellent time-frequency resolution make it ideal for transient detection. However, CHIME generates massive amounts of data—on the order of 1.5 TB/s—which necessitates a complex, multi-stage data processing pipeline. The CHIME/FRB backend employs GPU accelerated channelisation, aggressive real-time RFI mitigation, and a bespoke tree dedispersion[2] al-

[1] Brute-force dedispersion is a method that shifts and sums frequency channels for every trial DM, providing maximum accuracy but with high computational cost.

[2] Tree dedispersion is an algorithm that reuses intermediate sums across DM trials to reduce computa-

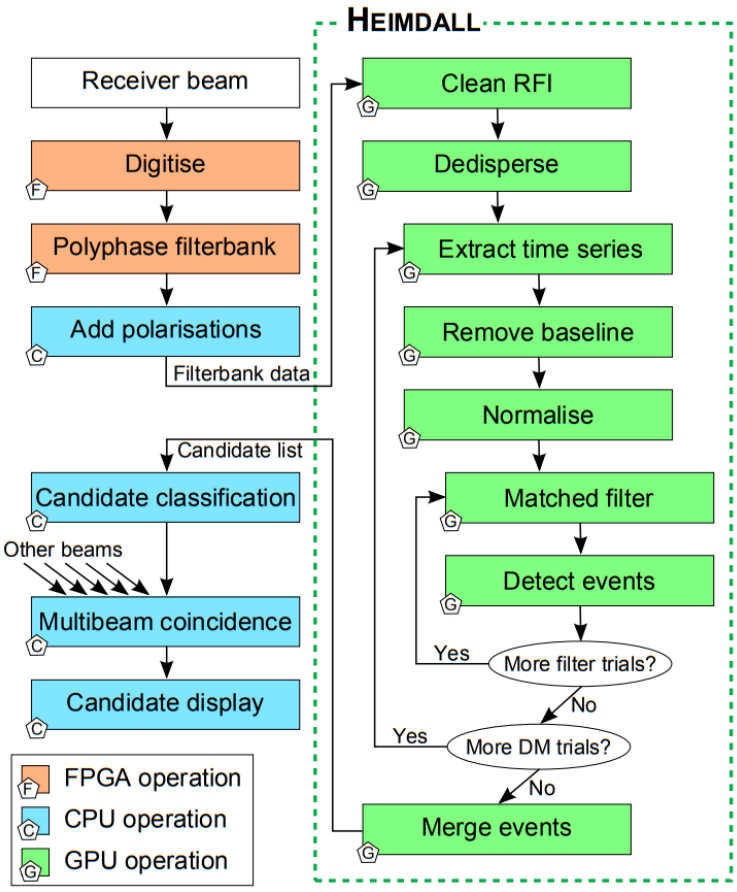


Figure 2.3: Flow chart of the key processing operations in the pipeline. Heimdall is the name of the main GPU-based pipeline implementation. Figure extracted from Barsdell (2012).

gorithm called bonsai, which produces SNR estimates for a grid of DMs and times for a discrete set of spectral indices, scattering times, and intrinsic width parameters (Amiri et al., 2018). Single pulse searches are carried out via boxcar matched filtering, while downstream candidate grouping and scoring are handled using statistical classifiers. The pipeline enables VOEvent triggering in seconds, allowing for quick multi-wavelength follow up. CHIME has subsequently become the most prolific FRB discovery instrument, reporting over 1000 FRBs, including dozens of repeaters (Andersen et al., 2020).

Meanwhile, in Australia, the ASKAP Commensal Real-time ASKAP Fast Transient (CRAFT) pipeline implemented a different technique. ASKAP runs at $\sim 1.3\,\mathrm{GHz}$,

tion, improving speed but imposing constraints on DM spacing and channelisation.

with phased array feeds that may produce up to 36 simultaneous beams. The FRB pipeline is based on Fast Real-time Engine for Dedispersing Amplitudes (FREDDA), a GPU implementation of the tree dedispersion method that is more computationally efficient than brute force alternatives. FREDDA allows for real-time processing of over 10,000 DM trials and incorporates candidate extraction and classification directly on the GPU (Bannister et al., 2017). This design enables high sensitivity searches across large fields of view. The ASKAP pipeline successfully localised FRB 180924 to a galaxy at $z \approx 0.3$ with sub arcsecond accuracy, a crucial step in connecting FRBs to their host galaxies (Bannister et al., 2019).

The DSA is another milestone in FRB detection, with a focus on instantaneous localisation. The DSA-110, which operates at 1.4 GHz with 110 small dishes, uses real-time detection with Heimdall-like dedispersion kernels and voltage capture for interferometric localisation. Its architecture is optimised for low latency baseband buffering and rapid imaging pipelines, allowing it to locate FRBs within $3''$ in real-time (Ravi et al., 2023). The future DSA-2000 will build on this concept with over 2000 antennae, allowing for daily FRB localisations throughout the whole sky.

In South Africa, the MeerTRAP project on the MeerKAT telescope adds another layer of innovation by combining coherent beamforming with GPU-based transient detection. MeerTRAP employs AstroAccelerate, a GPU accelerated C++ toolkit designed for rapid Fourier transforms, dedispersion, and matched filtering (Armour et al., 2019). Real-time data from tied array beams undergoes dedispersion over $\sim$ 3000 DMs, followed by matched filtering and candidate scoring. FETCH, a ML classifier, is incorporated for final candidate validation, and triggers can launch VOEvent messages or baseband recording buffers (Bezuidenhout et al., 2022). MeerTRAP's pipeline has already showed success in identifying low frequency FRBs, including CHIME discovered repeaters, and it provides an advanced reference model for mid-frequency FRB search designs.

Historically, pipelines at UTMOST (The Molonglo Observatory Synthesis Telescope) and the HTRU study (Parkes telescope) have provided useful architecture concepts. UTMOST, which runs at 843 MHz, created a real-time pipeline with Central Processing Unit (CPU)-GPU hybrid nodes for parallelised dedispersion and single pulse searching (Caleb et al., 2017). While not as sensitive or wide field as CHIME or ASKAP, UTMOST proved the viability of FRB detection with smaller arrays and stressed candidate grouping and sorting approaches, many of which have been incorporated into present systems.

Despite differences in hardware, frequency coverage, and software components, certain key design concepts emerge. First, most systems rely on GPU acceleration for real-time dedispersion and filtering, with a progressive shift toward tree-based techniques to improve computational efficiency. Second, matched filtering, particularly a sequence of boxcar smoothings, remains the most effective strategy for improving transient signal detection. Third, ML has become an emerging field for post detection candidate classification, with models such as FETCH being incorporated into pipelines to reduce false positives in real-time (Agarwal et al., 2020). Fourth, latency is a typical design restriction, especially in pipelines that enable multi-wavelength or multi-messenger triggers, where VOEvent systems are now universal.

These systems establish the standard for future Northern Cross development, which may take use of contemporary FRB pipelines' modular, GPU-driven, and ML augmented design. While the Northern Cross works at lower frequencies and smaller bandwidths than CHIME or ASKAP, it provides high cadence and continuous sky coverage, making it excellent for long term FRB monitoring and pipeline testing. Integrating real-time dedispersion tools like Heimdall, improving candidate scoring, and implementing adaptive DM trials might help the Northern Cross line with worldwide best practices.

2.4 | The Northern Cross Radio Telescope

2.4.1 | Technical Specifications

The Northern Cross Radio Telescope, located at the Medicina Radio Observatory in Italy, was one of the first large scale interferometric radio arrays to go into operation. It is a transit instrument that can only be mechanically steered in declination, therefore it can only examine celestial sources that are crossing the local meridian at the given time. The telescope's primary purpose is to receive radio waves centered at 408 MHz, which corresponds to a wavelength of 73.5 cm (Locatelli et al., 2020).

Initially, the telescope was operating with a RF bandwidth of 2.7 MHz, but due to the features of the East-West (E-W) arm's focal line, this could be increased to 16 MHz by simply altering the amplifiers and filters. To obtain the same capacity on the North-South (N-S) arm, the focal line was modified with an impedance adaptor to accommodate operations up to 16 MHz. These enhancements were prompted primarily

by the growing interest in pulsar research, where a larger band allows for greater sensitivity and temporal precision (Bianchi et al., 2023).

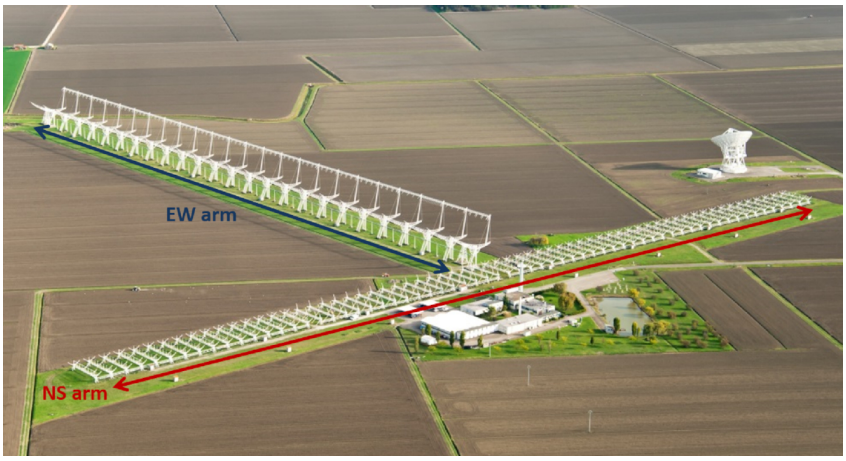


Figure 2.4: Aerial view of the Medicina Radio Astronomical Station, showing the N-S arm and the E-W arm. Figure extracted from Locatelli et al. (2020).

The telescope's trademark T-shaped configuration, which can be seen in Figure 2.4, consists of two major components:

- The E-W arm is a continuous parabolic cylinder reflector that is 564 m long and 35 m wide. 1,536 half wavelength dipoles are spaced 36 cm apart along its focal axis, which is half the operational wavelength. The focus axis runs parallel to Earth's rotation axis and is 20 m above it. Each dipole converts receiving radio waves into electrical impulses, which are then amplified and processed.

- The N-S arm, shown in Figure 2.5, is made up of 64 separate parabolic cylinder antennas, each 23.5 m long and 7.5 m wide and positioned 10 m apart. Like the E-W arm, each N-S antenna has 64 dipoles arranged along its focal axis. These reflectors comprise a sparse, steerable phased array that augments the E-W arm by improving angular resolution in the orthogonal direction.

The Northern Cross has an extensive collecting area of almost 27 000 m^2, making it ideal for detecting weak extragalactic radio sources. The system has an angular resolution of approximately 5′ in the N-S direction and 3′ in the E-W direction, providing a reasonable balance of sensitivity and spatial resolution for time domain radio surveys. These

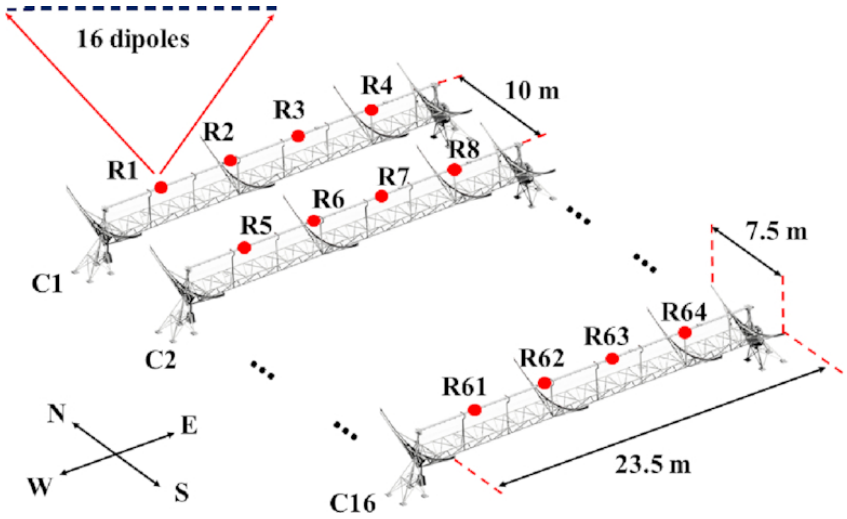


Figure 2.5: Scheme showing the refurbished section of the N-S arm. The red circles indicate the position of the four receivers located on the focal line of each cylinder. Figure extracted from Locatelli et al. (2020).

technological characteristics make the Northern Cross an important tool for pulsar timing, transient detection, and survey research, particularly when integrated with current digital backends and real-time data processing pipelines (Bianchi et al., 2023).

2.4.2 | Past and Current Surveys using the Northern Cross

The Northern Cross Radio Telescope has been instrumental in the advancement of European radio astronomy since it was put into service in the 1960s. Originally intended for detecting extragalactic radio sources, the telescope gained prominence by making early contributions to pulsar research, all sky surveys, and more recently, Space Situational Awareness (SSA) and time domain astrophysics.

The telescope's early operation was focused on continuum surveys at 408 MHz, which resulted in substantial datasets that culminated in the Bologna B2 and B3 catalogs. The B2 survey, released in 1970, included approximately 3,200 extragalactic radio sources, whereas the B3 catalog included over 13,000 entries, considerably enhancing the accuracy of source identification in the northern celestial hemisphere (Ficarra et al., 1985). These surveys were extensively utilised to investigate radio galaxy populations, spectral indices, and source distributions at low frequencies. The telescope's immense collecting area, steady gain, and long E-W baseline permitted it to identify weak extended sources that other devices operating at higher frequencies or with lower apertures would generally miss.

In the decades that followed, the Northern Cross was modified for pulsar observations, utilising its transit arrangement and coherent beamforming capabilities. Upgrades to its backend electronics and beam forming technologies enabled the telescope to focus on high cadence surveillance of known pulsars and participation in coordinated international observation operations (D'Amico et al., 1996). Its capacity to create several tied array beams along the declination axis allowed for a primitive drift scan survey, which provided vital time domain data across large areas of the sky. While the telescope's meridian only sky access limited its pointing flexibility, its regular observation cadence and beam stability made it ideal for long term pulsar timing studies, such as neutron star spin down and emission variability testing (Lower et al., 2025).

The Northern Cross has undergone major upgrading in recent years, making it suitable for detecting FRBs. The telescope was converted into a time domain transient detector by upgrading signal chains, expanding bandwidth receivers, and including real-time dedispersion software. In 2022, it reached a significant milestone by detecting FRB 20220912A, the first verified FRB observation at 408 MHz with the Northern Cross (Pelliciari et al., 2024). This identification was made possible using a custom real-time processing pipeline, showcasing the telescope's ability to handle massive amounts of high resolution data while applying on the fly dispersion adjustments. The findings validated the telescope's ability to contribute to worldwide transient detection efforts, prompting more expenditures in real-time software and GPU-based accelerations.

Beyond astrophysical research, the Northern Cross has been repurposed for SSA under the BIstatic RAdar for LEo Surveillance (BIRALES) project. This bistatic radar system detects and characterises space debris in Low Earth Orbit (LEO) by using the Northern Cross as a passive receiver in conjunction with an independent transmitter. In a 2018 demonstration, the BIRALES system spotted numerous LEO objects using real-time signal processing and Doppler tracking methods. The telescope's stable position and high resolution sampling made it ideal for the passive radar function in this design. The study revealed the Northern Cross's versatility beyond astrophysics, emphasising its practical uses in military, aerospace tracking, and orbital debris reduction measures (Cutajar et al., 2018). This reapplication emphasises the telescope's multi-mission capability and the importance of legacy infrastructure in solving scientific and applied technological difficulties.

Currently, the Northern Cross serves as both a research observatory and a development platform for real-time signal processing. It is used to test pipelines like Heimdall

and tailored implementations of GPU accelerated dedispersion algorithms such as FREDDA, as well as to train and validate ML-based classification tools like FETCH (Locatelli et al., 2020; Trudu et al., 2022). Collaboration with institutions such as INAF, the University of Bologna, and colleagues in Malta and Germany has expanded its scope. Its significance in both basic research and applied surveillance demonstrates a renewed scientific adaptability based on ongoing technological progress.

2.4.3 | Ongoing Upgrades to the Northern Cross

The Northern Cross is currently undergoing extensive renovation, which is being funded by the NextGenerationEU initiative. A significant component of this project is the complete restoration of the E-W arm, which has been out of service since 2012 due to prohibitively expensive maintenance and repair expenditures. The renovation entails extensive mechanical work and the replacement of old components with modern technology. The original focal line, which consisted of dipoles and a corner reflector, will be replaced with an array of 416 Yagi-Uda antennas, each with its own low-noise amplifier positioned near the antenna element. This construction, along with the already enhanced N-S arm, will restore the telescope's entire T-shaped structure resulting in a total collecting area of nearly $30\,000\,\mathrm{m}^2$, considerably increasing its sensitivity (Naldi et al., 2025).

In addition to hardware maintenance, the project includes funding for a new digital collection and processing system. The analogue signals from 672 receivers (256 from the N-S arm and 416 from the E-W arm) will be digitised using 21 Field Programmable Gate Arrays (FPGAs)-based backend boards, each with 32 inputs. These boards, based on designs created for the SKA-Low telescope, enable high-speed digitisation and real-time channelisation using polyphase filter banks, with data packetised and delivered via 40 GbE interfaces. A transient buffer on each FPGAs board preserves several seconds of raw voltage data, which may be recorded on command when an astrophysical trigger is identified (Naldi et al., 2025).

The digitised data streams will be handled by a new HPC cluster built for massive parallelisation. The cluster design consists of 18 real-time processing nodes, each with two CPUs, NVIDIA L40S GPUs, 2 TB of RAM, and high-bandwidth network connections. These nodes handle the computationally difficult functions of beamforming, correlation, and transient detection. Given the large number of input channels and

beams, the dedispersion search for FRBs continues to be the most resource demanding operation. GPU acceleration is thus required to maintain real-time performance. The system also includes object storage nodes with a total disk capacity of 8 PB, which can accommodate aggregate read/write speeds of 30-45 GB/s, allowing raw and reduced data to be buffered, archived, and re-analysed (Naldi et al., 2025).

The magnitude of this digital infrastructure demonstrates why extensive parallelisation is required. Serial processing is impractical with hundreds of input streams flowing continually, dedispersion spanning thousands of trial DMs, and real-time beamforming to cover the telescope's large field of view. Instead, computing is divided over CPUs, GPUs, and FPGAs, with operations like filtering, dedispersion, and candidate identification running in parallel. This architecture assures that the updated Northern Cross can function as both a strong FRB detection instrument and a radar system for monitoring LEO objects, with the ability to adapt to future advances in real-time radio astronomy.

2.4.4 | Comparison with Parkes, CHIME and LOFAR Telescopes

The Northern Cross Radio Telescope has a distinct architecture compared to instruments such as CHIME, LOFAR, and the Parkes Radio Telescope. While it differs in design, frequency range, and data processing capabilities, it remains relevant for transient detection and time domain studies. Comparing its performance with other facilities helps clarify its operational strengths and constraints, and informs opportunities for targeted optimisation.

2.4.4.1 | Parkes

The Parkes Radio Telescope, operated by CSIRO in Australia, is one of the most widely recognised single dish radio observatories. Parkes, which has a 64-meter parabolic dish, performs across a wide range of frequencies, primarily centered around 1.4 GHz but encompassing 700 MHz to several GHz with specialised receivers (Manchester et al., 2001).

Many of the known pulsars were first detected by Parkes, making substantial contributions to FRB research, most notably with the initial Lorimer Burst discovery which was previously mentioned at the beginning of Section 2.2. In terms of sensitivity and sky coverage, Parkes' fully steerable mount allows it to target any source observable from its southern position.

In comparison, the Northern Cross lacks Parkes' point source sensitivity and sky flexibility. However, the Northern Cross compensates by providing more instantaneous sky coverage in its transit strip and longer stay periods, which are required for random transient detection without the requirement for repositioning. Furthermore, the running expenses of a large steerable dish like Parkes are far greater than those of a semi-static array like the Northern Cross.

Parkes has historically used backend systems such as BPSR and newer GPU-based processing pipelines for real-time transient identification. The Northern Cross is pursuing comparable ideas, notably through the use of Heimdall and GPU accelerated dedispersion algorithms, but on a smaller computing scale.

2.4.4.2 | CHIME

Based in British Columbia, CHIME is a radio telescope developed primarily for cosmic surveys and transient searches, particularly FRBs. The interferometer is made up of four cylindrical reflectors, each 100 m long and 20 m wide, with over a thousand dual polarisation radio receivers installed along the focal lines (Bandura et al., 2014).

CHIME's frequency range, 400 MHz to 800 MHz, overlaps considerably with that of the Northern Cross, but with more extensive spectral coverage. What distinguishes CHIME is its advanced digital beamforming capability. Unlike the mechanically steerable N-S arm of the Northern Cross, CHIME electronically guides many beams across the sky at the same time, producing up to thousands of independent beams using real-time GPU processing (Amiri et al., 2018).

Consequently, this enables CHIME to survey an extensive region of the sky without using mechanical components, providing an advantage for high cadence FRB detection. In contrast, the Northern Cross is a transit only instrument with mechanically restricted declination, but its fast observation cadence and large E-W baseline enable outstanding sensitivity in small meridian slices.

While CHIME exceeds the Northern Cross in terms of immediate sky coverage and beamforming flexibility, conversely the Northern Cross provides a simpler infrastructure that is ideal for targeted experiments and technique development. Furthermore, although CHIME's frequent usage of high performance GPU clusters adds to significant

energy and maintenance requirements, the Northern Cross requires a smaller GPU system due to its smaller bandwidth and fewer simultaneous beams. This reduces total running expenses, making it ideal for ongoing surveys and long term monitoring.

2.4.4.3 | LOFAR

LOFAR, which is primarily located in the Netherlands but with stations across Europe, is a generation of radio interferometers that operate at low frequencies ranging from 10 MHz to 240 MHz (van Haarlem et al., 2013). To provide a range of resolution and sensitivity, LOFAR makes use of a dispersed array of basic dipole antennas which are coupled into phased arrays.

While LOFAR's operating frequency is substantially lower than the Northern Cross's 408 MHz focus, the two instruments share certain design principles, particularly their use of broad fields of vision and digital beamforming. LOFAR's flexible configuration enables it to generate high resolution images using long baseline interferometry, but its transient search capabilities, particularly in the LOFAR Transients Key Science Project (TKSP), detect radio bursts using fast imaging modes and time domain pipelines (Carbone et al., 2016).

Compared to LOFAR, the Northern Cross operates at a higher frequency, resulting in higher angular resolution per baseline length, meaning that the system can resolve finer angular detail on the sky and can distinguish between sources that are closer together in angle. However, because of its higher digital flexibility and networked architecture, LOFAR may dynamically distribute resources between imaging, transient searches, and monitoring, whereas the Northern Cross is more targeted and restricted to a fixed array configuration.

In addition, LOFAR's modular architecture allows for future upgrades and reconfiguration, whereas the Northern Cross's permanent physical construction makes this less likely. Nonetheless, for focused long term monitoring projects, particularly those that do not require wideband imaging, the Northern Cross is still an operationally straightforward tool.

In conclusion, the Northern Cross continues to be a powerful and adaptable instrument in the context of real-time transient detection, regardless of how CHIME, LOFAR,

and Parkes each have enhanced capabilities in areas like dynamic beamforming, wide band frequency coverage, and point source sensitivity. Although its adaptability in sky coverage is limited, its transit based architecture enables reliable and repeatable observations with minimal mechanical complexity. Strong sensitivity along the meridian is supported by the telescope's large effective collecting area and extended E-W baseline, which enable dependable detection of dispersed, short duration signals like FRBs.

Recent backend upgrades have developed the system to support GPU accelerated processing and multiple tied array beams, enabling the development and testing of real-time pipelines under controlled, reproducible conditions. The Northern Cross's straightforward infrastructure makes it a great candidate for long duration surveys, high cadence monitoring, and continuous pipeline refinement; without the scheduling pressures typical of more oversubscribed facilities.

In light of these features, the Northern Cross acts as an invaluable platform for improving operational strategies and detection algorithms in time domain radio astronomy. The optimisation techniques examined in this research, which seek to maximise both performance and observational data within the telescope's architectural bounds, depend on an understanding of the telescope's relative advantages and limitations.

2.5 | Challenges in Real-Time Radio Pipelines

As radio astronomy transitions into the era of large scale data, one of the most significant challenges is to develop and execute efficient, real-time data processing pipelines that fully use the capabilities offered by modern telescopes. Instruments like the Northern Cross, CHIME, LOFAR, and MeerKAT demonstrate this change by producing data streams at speeds ranging from several gigabits to tens of terabits per second (Price, 2020). These telescopes, with their high observation cadence and high resolution, narrow band outputs, highlight the demand for pipelines that can process large data quantities efficiently and securely.

The computational complexity is brought on not only by the large number of observational datasets, but also by the need to analyse, dedisperse, categorise, and detect transitory astronomical phenomena in real-time. Rapid detection and characterisation of transitory events, such as pulsars and FRBs, sometimes necessitates subsecond reac-

tion times. Meeting these rigorous requirements necessitates tackling complex issues in algorithm creation, appropriate hardware selection, and smooth pipeline integration. Effective handling of these constraints is critical to realising the full scientific potential of present and future radio telescopes.

2.5.1 | Data Volumes and Processing Requirements

Radio telescopes that carry out time domain surveys, such as the Northern Cross, must have high temporal and spectral resolutions in order to discover fast transients such as pulsars and FRBs. As a result, even very narrow band systems can provide high raw data rates. This is consistent with the system level data flow limits reported in Malenta et al. (2020), where transient pipelines such as MeerTRAP are intended to accommodate hundreds of Gbit/s across several beams via GPU accelerated processing. Similarly, De and Gupta (2016) explore a 16–32 MHz coherent dedispersion backend at the GMRT, citing data rates of 120–240 GB/h per beam, demonstrating the high throughput nature of time domain radio systems.

Instruments such as the Northern Cross demonstrate the data processing issues that come with real-time FRB searches. Despite being limited to a single polarisation, scanning meridian strips using multiple tied array beams produces large amounts of data. Such massive amounts are impossible to store, especially for lengthy observation periods, necessitating real-time processing systems (Price, 2020). These solutions entail instantaneous data processing, dedispersion, and compression into manageable candidate events while retaining only relevant outputs. Hence, these systems frequently use distributed computing clusters, which generally combine GPUs and CPUs, to divide the workload between nodes and handle numerous repetitive tasks through parallelism.

Real-time dedispersion with brute force is computationally demanding. The level of complexity scales poorly with higher resolution and DM range, requiring algorithmic modifications or hardware acceleration, such as GPUs, for real-time performance. Modern systems can search up to $2000 - 3000$ DM trials with fine spacing, making dedispersion the most significant expense in the pipeline. According to Barsdell (2012), dedispersion is essentially memory bound, and performance is restricted by memory bandwidth rather than computing speed. It is shown that GPUs can greatly accelerate dedispersion algorithms, with speedups of up to $9\times$ over optimised CPU versions. This acceleration is essential for real-time processing in large scale surveys.

Furthermore, real-time pipelines must operate with low latency and high reliability, needing not just raw computational power but also adaptive buffering, parallel tasking, and fault tolerant stream management. The system must ensure that no transient signal is lost while maintaining consistent throughput, even during Input/Output (I/O) bottlenecks or temporary processing interruptions.

2.5.2 | Key Processing Steps

Real-time transient detection pipelines typically consist of a number of linked and computationally expensive stages, each of which is crucial to providing fast and accurate detection. These stages have been developed to effectively handle large data streams, decrease latency, maximise sensitivity to transient events, and maintain robust performance in the face of operational and technical challenges. The entire architecture combines high throughput data handling with complex algorithms and parallel processing techniques to successfully manage the difficulties of real-time astronomical signal processing.

2.5.2.1 | Pre-processing and RFI Mitigation

The pre-processing stage of a pipeline is important for ensuring data integrity before performing more demanding operations like dedispersion and candidate detection. This phase often comprises data intake, synchronisation, reformatting, and calibration, but the most important aspect is RFI mitigation, which is essential for time domain radio astronomy.

RFI from both terrestrial and satellite sources may significantly undermine the performance of transient detection systems by simulating astrophysical signals or masking true events. Effective mitigation solutions must be computationally efficient and scalable to thousands of frequency channels and time samples. Common techniques include:

- Simple thresholding, which eliminates samples over a given amplitude level.
- Median filtering, which is an approach to suppress outliers and short duration spikes.

- Time-frequency masking, which includes the well known SumThreshold method, which dynamically finds and flags polluted zones based on local signal characteristics (Offringa et al., 2010).

In addition to these classic methods, advanced pipelines are also using higher order statistical techniques, such as kurtosis based filtering, to improve RFI detection. Kurtosis, which measures the "tailedness" or peakness of a signal distribution, may identify Gaussian thermal noise (with a kurtosis of 3) from non-Gaussian signals such as RFI. When computed over frequency channels or time samples, high kurtosis values might indicate interference corrupted areas, allowing for more precise masking (Nita et al., 2007). These approaches are especially helpful for detecting low-level or intermittent RFI that escapes simpler thresholding techniques, and they are becoming more widespread in real-time pipelines due to their versatility.

In addition to traditional approaches like thresholding and median filtering, present radio astronomy is increasingly making use of ML for robust and automated RFI detection and excision. Akeret et al. (2017) developed a pioneering technique that used U-Net Convolutional Neural Network (CNN) to discriminate RFI-contaminated time-frequency data from clean signals, with accuracy comparable to conventional algorithms such as SumThreshold. Similarly, Saliwanchik and Slosar (2022) used a self-learning neural network based on variational autoencoders to identify and eliminate RFI without labelled training data, indicating a possible path for automated cleaning in contexts where ground truth is lacking. More recently, Pritchard et al. (2025) applied Liquid State Machines, a type of spiking neural network, to perform fine-grained semantic segmentation on RFI-affected spectrograms with an impressive 98% per-pixel accuracy, demonstrating the potential of event-driven architectures in real-time RFI mitigation. These ML algorithms improve flexibility and efficiency in detecting complex and developing interference patterns in real-time pipelines.

Despite their simplicity, using these algorithms repeatedly over numerous beams, high temporal resolutions, and large bandwidths adds a significant processing load, especially on systems with GPU accelerated or multicore pipelines. To satisfy such requirements, projects like MeerTRAP use pre-processing subsystems like PANDA (for data buffering) and CHEETAH (for memory alignment and stream reshaping) to sustain throughput while retaining sensitivity to small signals (Rajwade et al., 2022). These components include the input layer of real-time processing chains, ensuring that future stages obtain clean, well structured data.

2.5.2.2 | Dedispersion and Detection

Dedispersion is at the heart of transient search pipelines. It accounts for the frequency-dependent temporal delay that radio pulses suffer as they travel through the ionised ISM. This delay, known as the DM, is proportional to the integrated column density of free electrons along the line of sight.

In real-time FRB searches, pipelines often use incoherent dedispersion, which corrects for temporal delays across frequency channels but does not recover phase information within each channel. Although coherent dedispersion, which uses raw voltage data to correct intra-channel dispersion, is more precise, it is computationally far more costly and often not feasible for wide field, high-throughput surveys. As a result of its balance of precision and efficiency, incoherent dedispersion is still used in most real-time pipelines today (Barsdell et al., 2012).

The direct (brute force) dedispersion technique computes the delay for each trial DM over all frequency channels and adjusts each channel appropriately before summing. Its computational complexity is $\mathcal{O}(N_t \cdot N_\nu \cdot N_{\mathrm{DM}})$ [3], where N_t is the number of time samples, N_ν the number of frequency channels, and N_{DM} the number of DM trials. This simplistic method scales poorly with increasing resolution and DM range, making it computationally costly for high throughput pipelines (Barsdell, 2012).

To address this, several optimised algorithms have been developed, which include Tree dedispersion and Sub-Band dedispersion. Tree dedispersion achieves $\mathcal{O}(N_t \cdot N_\nu \cdot \log N_\nu)$ scaling with normal channel spacing (Taylor, 1974). Sub-band Dedispersion is a technique that balances performance and accuracy by separating the frequency band into sub-bands, dedispersing them individually, and then recombining them at the end (Magro et al., 2011).

The growing demand for real-time responsiveness has resulted in widespread use of GPU accelerated dedispersion software. These include: Heimdall, which is intended for transitory searches using GPU parallelism, FREDDA which is a Compute Unified Device Architecture (CUDA) implementation of tree dedispersion and AstroAccelerate, which is a versatile GPU framework used in MeerTRAP to do quick single pulse searches (Adámek and Armour, 2016).

[3] $\mathcal{O}()$ represents order of magnitude.

Modern systems generally run $1000 - 3000$ DM trials in real-time, testing a wide range of dispersion events. Matched filtering using pulse templates of varying widths (boxcar functions) is frequently used immediately following dedispersion to improve SNR ratios for widened or dispersed signals.

2.5.2.3 | Smoothing and SNR Optimisation

Pulses are frequently broadened by multipath scattering and intra-channel dispersion, particularly at lower RF. However, a more fundamental consideration in transient search pipelines is the fact that the intrinsic pulse widths of FRBs can vary significantly, from µs to several ms. To ensure sensitivity across this range, real-time pipelines implement matched filtering using boxcar functions of varying widths; typically spanning from the native sampling interval (e.g., ∼64 µs) up to several ms. This process helps maximise the SNR for bursts of different temporal duration, improving detection efficiency across the diverse FRB population.

This entails convolving the dedispersed time series with boxcar kernels with increasing width and then measuring the peak SNR. The method is repeated across all trial DMs. Although theoretically straightforward, this technique is done thousands of times per second across hundreds of beams and DM trials, necessitating careful parallelisation to prevent exhausting GPU or CPU resources. Matched filtering is critical for improving detectability, especially for weak, dispersed signals like FRBs seen at low DMs or with broad temporal structure. This step's computing effort is frequently second only to that of dedispersion.

2.5.2.4 | Candidate Detection and Thresholding

Following dedispersion and smoothing, potential signals are retrieved from the processed data stream using thresholding methods. Each sample is compared to a baseline noise level (often derived from a running mean or variance) and alerted if the SNR exceeds a user defined value, usually about $6 - 8\sigma$ depending on the system's false positive tolerance.

Although less computationally intensive than dedispersion or matched filtering, this stage must be carefully tuned to minimise bottlenecks. Efficient memory access is critical to maintaining performance at this point. Some systems use dynamic thresholds or clustering of close peaks to decrease redundancy.

2.5.2.5 | Post Processing and Candidate Classification

In multi-beam systems, a significant number of candidate events are generated, many of which are noise or RFI related. This requires a post processing step in which candidates are sorted, grouped, and categorised.

Sophisticated heuristic filters (for example, rejecting events detected in all beams at the same time) are often employed, while increasingly powerful pipelines are relying on ML. FETCH is an excellent example of a CNN trained on huge labelled datasets for classifying FRB and pulsar candidates.

FETCH has achieved >99.5% accuracy in FRB classification and can run in real-time on GPU systems, making it appropriate for integration with current pipelines (Agarwal et al., 2020). These classifiers limit the number of false positives and help triage candidates for follow up or archiving.

2.5.2.6 | Inspection, Output, Storage, and Alerting

Despite automation, final candidate evaluation frequently requires visual analysis of plots such as dynamic spectra and DM-time diagrams. Real-time systems must therefore deliver both candidate files and visual outputs as quickly as possible. CUDA accelerated plotting tools (such as MeerTRAP) provide near instantaneous depiction of thousands of candidates for human assessment (Rajwade et al., 2022).

Real-time alerting systems, such as VOEvent brokers or email triggers, are also utilised to notify follow up instruments. Latency is essential here: delays of more than a few seconds might result in missed opportunities for multi-wavelength observation, especially for short duration or repeated FRBs. Efficient storage solutions, such as writing only the buffered window around each candidate rather than complete time streams, are critical for keeping data manageable while maintaining scientifically relevant signals.

2.5.3 | Computational Bottlenecks and Solutions

Some of the critical recurrent difficulties for these pipelines include processing latency, computational cost and GPU memory restrictions. Computationally expensive ap-

proaches, such as Fast Fourier Transformation (FFT)-based coherent dedispersion, scale poorly with bandwidth (complexity squared), demanding strategies such as sub-band dedispersion to properly manage GPU workloads (Barsdell, 2012).

In transient detection pipelines, maintaining real-time performance requires diligent resource management in addition to efficient algorithms. In multi-CPU systems, where NUMA-aware programming enables each processor to access its own local memory, lowering latency and increasing throughput, efficient memory handling becomes particularly important. To prevent bandwidth bottlenecks, it's also essential to make use of the PCIe bus, which transports data between CPUs, GPUs, and storage, as efficiently as possible. Further reducing unneeded overhead is achieved by using direct I/O techniques, which move data straight between memory and disk or devices, avoiding the operating system's cache (Barsdell, 2012).

Addressing the computational challenges of real-time pipelines requires a complex interaction of high throughput data processing, latency optimised algorithms, and parallel computing architectures. The Northern Cross telescope, with its consistent pointing and set shape, is an excellent platform for prototyping and optimising such real-time pipelines under realistic, repeatable conditions. Tools like Heimdall, which incorporates GPU-based dedispersion and improved candidate filtering, demonstrate plausible routes to scalable, real-time transient discovery (Adámek and Armour, 2016).

2.6 | Software for Transient Detection

Real-time radio transient search pipelines depend greatly on efficient software design and execution to be successful. As data quantities from instruments like the Northern Cross progress, software applications must be scalable, low latency, and high throughput for signal detection. In this section, the major software tools for transient detection are assessed, with a particular focus on Heimdall, and are compared to other frequently used systems.

2.6.1 | Heimdall: GPU Accelerated Transient Detection

Heimdall is a widely used open source software program for conducting real-time, single pulse transient searches in radio astronomy data. Originally designed for use

in the HTRU survey and later adapted by several FRB focused projects (e.g. SUPERB, MeerTRAP, CHIME/FRB), Heimdall is based on GPU acceleration, with NVIDIA's CUDA platform handling the computationally expensive dedispersion and matched filtering stages (Barsdell, 2012; Bhandari et al., 2017).

At its core, Heimdall executes incoherent dedispersion across thousands of trial DMs to account for the ISM's frequency dependent delays. Following dedispersion, the program uses boxcar matched filtering over a user defined range of pulse widths to detect signals that have been widened due to scattering or instrumental effects. Once potential signals have been detected, Heimdall follows with candidate extraction, awarding a score to each detection based on SNR and other criteria before logging the findings for future evaluation or grouping (Trudu et al., 2022).

One of Heimdall's advantages is its adaptability. It is compatible with a wide range of front end systems (e.g., ring buffers, packetised data streams) and has been applied in both offline and streaming real-time environments. It also allows concurrent multi-beam processing, making it ideal for telescopes such as the Northern Cross, which tile their field of view with numerous tied array beams.

Heimdall's dedispersion procedure takes a brute force approach, but by making use of GPUs' fast throughput, it is still feasible for usage with up to several thousand DM trials and several boxcar widths. The program employs advanced memory management and overlapping kernel execution algorithms to enhance GPU occupancy while minimising idle time. Depending on the arrangement, Heimdall can process time series in real-time at speeds faster than 1 GB/s per GPU. Heimdall has been successfully implemented in; the Parkes radio telescope, the ASKAP CRAFT Survey, MeerTRAP on MeerKAT (as a comparative baseline) and in recent installations of the Northern Cross for low frequency FRB detection. It is currently developed and actively utilised in the FRB research community, and it can interface with ML-based candidate classifiers like FETCH (Barsdell, 2012; Rajwade et al., 2022; Trudu et al., 2022).

2.6.2 | Other Tools for Transient Detection

While Heimdall remains a standard in many real-time pipelines, a variety of additional software solutions have developed, each with a distinct constraint or use case.

- FREDDA

  FREDDA is a CUDA-based version of the tree dedispersion algorithm designed primarily for real-time use with instruments such as ASKAP. FREDDA, unlike Heimdall's brute force dedispersion, decreases complexity to about $\mathcal{O}(N_t \cdot \log N_\nu)$ each DM trial with specific frequency spacing assumptions. This makes it ideal for wide field, high beam count systems in which brute force is impractical. FREDDA also allows matched filtering and returns candidate events in Heimdall compatible formats (Bannister et al., 2017).

- AstroAccelerate

  Carels et al. (2019) created AstroAccelerate, a high performance computing toolkit for radio astronomy that includes GPU accelerated modules for dedispersion, FFT, convolution, and single pulse detection. It is applied to MeerTRAP for both real-time and offline single pulse searches and is compatible with the PANDA-CHEETAH pre-processing suite for data intake and buffering. AstroAccelerate focuses on portability and speed, delivering real-time performance with sub-millisecond latency across several processing nodes. It has been tested to handle up to 16 beams in real-time on GPU clusters.

- PRESTO

  PRESTO is an assortment of tools created specifically for pulsar searches in archival data. While not designed for real-time use, it remains a key component in candidate folding, periodicity searches, and RFI elimination. It has been adapted for use in near real-time pipelines, particularly in combination with other tools (such as Heimdall for detection and PRESTO for refining). It remains popular due to its robustness and extensive documentation (Ransom, 2001).

- FETCH

  FETCH is a deep learning based classifier trained on actual and simulated FRB spectrograms, rather than a detector in its own right. FETCH, which is based on CNNs, can separate astrophysical signals from noise or RFI with an accuracy of more than 99.5%. It is frequently incorporated into pipelines downstream of Heimdall or FREDDA and runs in real or near real-time throughout the candidate

evaluation stage. The model is trained on big datasets from several backends, making it adaptable to various observation methods (Agarwal et al., 2020).

While each of these software packages are developed to aid in transient detection, their architecture, purpose, and computational procedure vary considerably. FREDDA and AstroAccelerate are two GPU-accelerated real-time dedispersion engines developed for large-scale surveys like ASKAP and MeerTRAP. FREDDA uses a tree dedispersion technique to provide high processing speed and scalability across thousands of DM trials, whereas AstroAccelerate uses a brute-force dedispersion approach with considerable kernel optimisation to maintain near-real-time performance on multicore computers. PRESTO, on the other hand, is a more general-purpose CPU-based pulsar and transient search suite that has traditionally been employed in offline analysis. While it is highly accurate and adaptable, its lack of native GPU acceleration limits its application in modern real-time pipelines. FETCH, on the other hand, operates beyond these detection steps; it is a ML classifier that uses CNNs to distinguish between astronomical bursts and RFI.

While FREDDA, AstroAccelerate, and PRESTO all focus on signal processing and candidate production, FETCH addresses the next challenge: candidate verification. These programs represent complementary stages of a modern transient detection pipeline: dedispersion and filtering (FREDDA, AstroAccelerate, PRESTO), followed by automated categorisation (FETCH). Their combined use at facilities like ASKAP, MeerKAT, and Parkes illustrates that performance optimisation is increasingly reliant on GPU parallelism and ML integration rather than algorithmic design alone.

The collection of transient detection software in use today reflects the variety of telescope designs and research objectives. Heimdall is the go-to tool for GPU accelerated brute force dedispersion, while FREDDA and AstroAccelerate provide algorithmic and performance benefits for niche use cases. Integration with ML-based classification systems such as FETCH represents the next step toward completely autonomous, real-time transient detection pipelines.

Heimdall and FETCH offer a well balanced solution for telescopes like the Northern Cross, which favour high cadence, low frequency, and small bandwidths. However, when the system is updated, the use of more effective dedispersion techniques or hybrid pipelines may be advantageous.

2.7 | Summary

This chapter presented a review of the scientific background and technological foundations essential to the detection of FRBs, with a particular emphasis on real-time search pipelines. It started with a review of FRBs, highlighting their astrophysical relevance, and then looked at the progress of FRB detection pipelines created at key observatories. These systems, despite their differences in implementation, have shared aims of speed, accuracy, and scalability; elements required for the construction of new pipelines at instruments such as the Northern Cross.

The chapter also presented the Northern Cross Radio Telescope, outlining its structure, capabilities, and recent developments that make it suitable for current transient surveys. The computational problems of real-time radio pipelines were discussed, with an emphasis on effective dedispersion, RFI mitigation, and candidate classification. Finally, the software tools most relevant to this task were examined, including Heimdall and other GPU accelerated frameworks often used in the area.

Together, these components provide the foundation for the project's application. The next chapter will describe how this information contributed to develop and evaluate a part of the real-time FRB search pipeline for the Northern Cross. It will describe the tools chosen, the data implemented in the study, the parameter testing approach, and performance assessment methodologies used during the development process.

3

Methodology

3.1 | Introduction

The detection and interpretation of FRBs requires precise and computationally expensive processes, as was previously detailed in Section 2.3. The procedure which was implemented in this project is discussed in this chapter.

First, a description of the synthetic data used throughout the analysis is presented. The details of the dataset, including observational parameters such as SNR, DM and Time of Arrival (ToA) are outlined to highlight their relevance. This is also important as it forms the basis for the subsequent computational and statistical analyses.

This is followed by an outline of the computational steps executed on the data, as well as an explanation of the statistical analysis approaches used to obtain the required final results. The computational approaches outlined include the systematic processing of data on a HPC cluster. This technique allows for the evaluation of different parameter combinations in the Heimdall software, an established tool for FRB signal detection. Heimdall was chosen since it was already in use at the Northern Cross, was efficient in processing huge quantities of data, and had a strong capacity to filter potential signals from noise.

Following computational analysis, candidate results from various parameter combinations were methodically combined, analysed, and visualised. Statistical tests, including t-distributed Stochastic Neighbour Embedding (t-SNE), the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) clustering technique, the Friedman test, and the Nemenyi post-hoc test, were then performed. These statistical studies were chosen particularly to thoroughly test and enhance the performance of parameter combinations, hence ensuring methodological validity.

Additionally, the chapter aims to justify the selection of specific algorithms and parameters through statistical validation, ultimately leading to the identification of optimal parameter settings for FRB detection and efficiency in real-time processing.

3.2 | Data Specifications and Setup

3.2.1 | The Telescope

The Northern Cross radio telescope, which was discussed in Section 2.4.1, went through significant modifications, notably in its digital backend system, resulting in higher quality FRB surveys. The digital backend is based on modern hardware and powerful digital signal processing methods that are designed to efficiently handle large volumes of observational data. The backend architecture relies on FPGAs and GPUs for real-time signal processing. These allow fundamental operations such as dedispersion, interference mitigation, and efficient time domain data handling, which are important for understanding transient radio signals typical of FRBs (Locatelli et al., 2020).

The digital backend converts the incoming analog signals from the antenna array. The signals are initially collected as EM waves and then transformed to digital form using Analog to Digital Converters (ADCs) with high sampling precision. A greater sampling rate leads to better temporal resolution, enhancing the capacity to precisely determine the arrival times of FRBs, which is an observational parameter known as the ToA. The digital signals are then passed through several signal processing stages, including channelisation into narrower frequency sub-bands, to enhance comprehensive analysis and decrease computing overhead (Locatelli et al., 2020).

Signal processing is performed on FPGAs, and the generated data is transmitted to a processing server via a 40 Gbit Ethernet connection. The processing chain includes the following key steps. Initially, it corrects for cable mismatch by adjusting relative delays using time domain shifts applied to each of the 32 Intermediate Frequency (IF) inputs. Each IF input is then channelised into 512 channels, each 781 kHz wide, using an oversampled polyphase filterbank. Calibration and pointing coefficients are also given to the FPGAs to correct for instrumental amplitude and phase responses, and separate pointing coefficients enable independent beam pointing. The final step involves frequency domain beamforming, which generates four simultaneous beams with a minimum integration time of roughly 70 µs or a single beam with an integration time of 1.08 µs. The previously determined coefficients are applied to calibrate and direct each beam. Table 3.1 summarises the acquisition system's main features.

The Northern Cross telescope's FRB survey aimed to improve sensitivity and detection capabilities. The survey design included a variety of observational parameters, each

N. of frequency channels	1024
Channel width	781.25 kHz
Time resolution	1.08 μs
Single beam beamformer	
N. bits	16 complex
N. channels	21
Throughput	311.11 Mb/s

Table 3.1: Characteristics of the current acquisition system. Table extracted from Locatelli et al. (2020).

designed to increase the chance of identifying transient events. The survey's principal observational frequency is 408 MHz, chosen for its balance of sensitivity, sky coverage, and minimal susceptibility to interference. This frequency range was chosen to examine areas less studied by earlier studies, offering supplementary information about the FRB population and emission processes.

Bandwidth is important to the survey's sensitivity and frequency coverage. The Northern Cross has a bandwidth of approximately 16 MHz, split into smaller channels via digital signal processing methods. This channelisation improves spectrum resolution and distinguishes true astrophysical signals from terrestrial interference, making it easier to isolate and identify possible FRB signals (Locatelli et al., 2020).

Another important observational parameter, DM, is closely related to radio signal propagation in ISM. As explained in Section 2.2.2, DM measures the total column density of free electrons between the source and the observer, resulting in frequency dependent delays in signal arrival times. This dispersion effect causes lower frequency radio pulse components to arrive later than higher frequencies, with the delay proportional to the DM value and inversely related to the square of the observed frequency. Accurate measurement and analysis of DM require both high frequency resolution and large bandwidth, which are delivered by the telescope's digital backend architecture. The backend's real-time dedispersion enhances sensitivity to heavily dispersed signals, increasing the possibility of identifying distant FRBs.

The SNR, which relates the detected signal's amplitude to the background noise's root mean square (RMS) level, is an important indicator of detection confidence. High SNR values distinguish the astrophysical signal from noise fluctuations, allowing

for more accurate detection of FRB pulses. SNR factors significantly affect operating strategy. The observational design intended to maximise SNR by optimal integration times, careful management of instrumental noise, and strict interference mitigation measures. The survey's persistent high SNR levels helped detect genuine FRBs against instrumental and ambient noise.

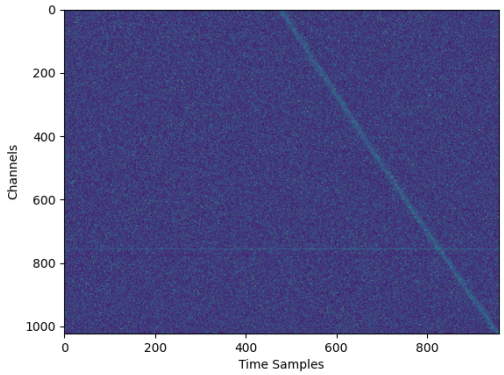
The modifications made to the Northern Cross telescope satisfy the requirements for FRB observations. These changes were intended to improve the instruments' sensitivity and computational power. Key modifications included the installation of digital signal processing technology, which improved real-time processing and allowed for faster detection and storing of transient events. This real-time capability is significant since FRBs are brief, lasting only milliseconds.

Furthermore, enhancements were also made to the software architecture, notably with the implementation of robust algorithms for RFI mitigation. These methods served for improving data quality by minimising terrestrial noise, ensuring that possible astrophysical signals were saved and correctly detected. These software changes, paired with hardware upgrades, increased the telescope's detection capability, allowing it to undertake more sensitive and reliable FRB searches (Locatelli et al., 2020).

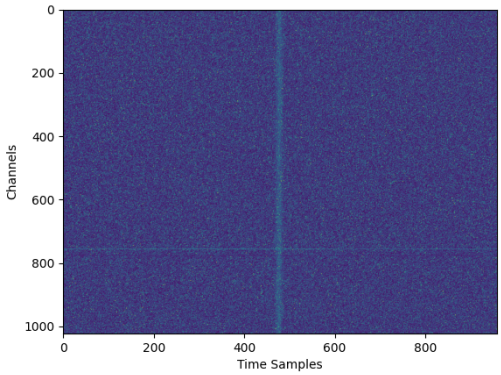
3.2.2 | Data Characteristics

The dataset generated for this investigation consisted of 950 filterbank files, each having 140 seconds of synthetic data that was developed to model observational data from the Northern Cross telescope. An example of how the data can be visualised as a signal is depicted in Figure 3.1. To maintain data integrity, the first and last 10 seconds of each file were removed from the analysis. The first 10 seconds are set aside to accommodate for startup transients and stabilisation procedures on the HPC cluster, which might produce misleading artifacts. Similarly, the last 10 seconds of the file include no injected signals and are typically used as a buffer time before data transmission or shutdown, rendering them meaningless for detection purposes. This trimming yields 120 seconds of clean, useful data per file, with injected test signals dispersed methodically.

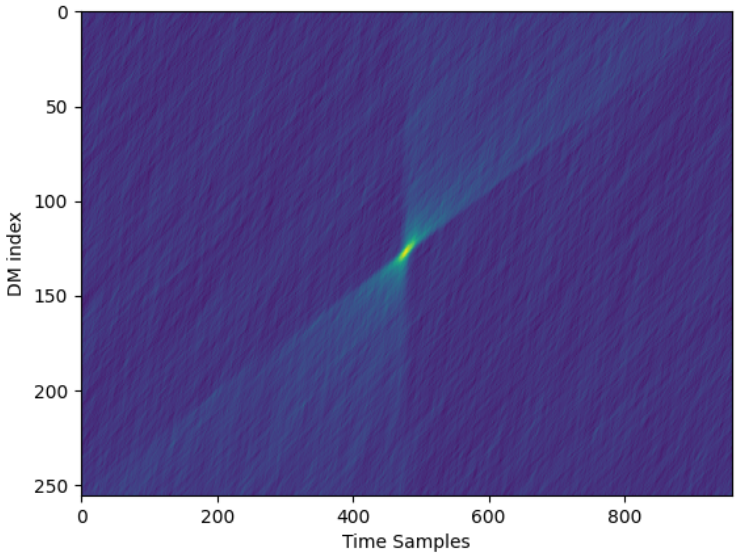
Within each 120 second segment, 13 synthetic FRBs were injected to test the Heimdall detection pipeline's performance under controlled conditions. The number of injections per file was set to strike a compromise between computational tractability and



The raw signal.



The dedispersed signal.



Plot of DM-time data. The bow-tie shape and the peak in the middle represent the FRB signal.

Figure 3.1: Visual plots of an FRB signal from a filterbank file from the sample data.

statistical robustness, ensuring that there are enough test cases without overloading the detection system. Injection timings were about every 10 seconds, allowing the pipeline's sensitivity and timing accuracy to be tested across the entire observation window.

This dataset allocated DMs based on a log-uniform distribution from 20 to 3000 pc cm^{-3}. The log-uniform selection matches astrophysical expectations; low DM events inves-

tigate nearby sources, whereas high DM events represent distant or widely scattered signals. By sampling evenly in logarithmic space, the dataset ensures a representative mix over multiple orders of magnitude, putting the pipeline to the test for its ability to handle both softly and heavily scattered bursts equally well.

The FWHM of a peaked function, often a Gaussian or other bell-shaped profile, refers to the distance between two points on the curve where the function's value falls to half of its maximum. In FRB analysis, FWHM measures the intrinsic temporal width of a detected burst, affecting time domain detection sensitivity and boxcar smoothing widths in search algorithms.

The temporal properties of the injected FRBs were represented by Gaussian profiles, a standard approximation for intrinsic burst patterns after experimental and propagation effects. The FWHM of each burst was calculated using a log-uniform distribution that ranged from 0.0005 to 0.13 seconds. This spectrum includes both very small, millisecond scale bursts and wider pulses lasting a few hundred milliseconds, which put the pipeline's resolution to the test. Narrow bursts test the detection systems' temporal precision, whereas bigger bursts test sensitivity to longer duration signals.

SNR is another essential injection parameter that influences detectability versus background noise. In this dataset, SNR values for injected bursts were drawn from a uniform distribution ranging from 3 to 13. An SNR of 3 indicates a barely detectable threshold, whereas an SNR of 13 simulates easily identifiable events. Uniform sampling over this range presents the pipeline with a number of detection challenges; the algorithm must separate barely identifiable signals at low SNR from noise variations while reliably characterising high SNR events without saturation or clipping.

Aside from these injection parameters, each filterbank file has information describing both injection and observational features. The descriptive elements for each injected burst include the pulse profile shape (`SHAPE`) and two measurements of pulse width (`FWHM_1` and `FWHM_2`). The frequency domain properties of the Gaussian pulse are also recorded, including its central frequency (`F_G_CEN`), amplitude (`F_G_AMP`), spectral index (`SP_IDX`), and drift rate (`DRIFT_R`). A native SNR value (`NATIVE_SNR`) is kept to indicate the raw detection measure before any pipeline modifications.

Along with burst specific information, each file includes observational parameters set by the digital backend. This includes the center frequency of each channel which is

about 415 MHz, the total number of frequency channels is 1024, the sampling time interval is 138 µs, and the total number of samples per file is 1,012,962. This information enables exact reconstruction and detailed analysis of each synthetic FRB, so that Heimdall pipeline's performance is properly evaluated under actual observational settings.

3.3 | Computational Steps

All computationally intensive tasks in this study were carried out on the university's cluster, which was built to handle large scale data processing and advanced signal analysis processes. The NVIDIA RTX 6000 Ada GPU is a significant part of this system, with 48 GB of high speed GDDR6 memory and powerful tensor cores for enhanced deep learning and parallel computation. This GPU, when combined with multi-core CPU nodes and high throughput interconnections, allows for the efficient execution of Heimdall parameter sweeps, dedispersion, and subsequent statistical analysis. Using the HPC cluster assured that all data intensive operations, from filterbank processing to dimensionality reduction and clustering, were completed within reasonable timescales, allowing for quick iteration and optimisation across the methodological pipeline.

The Northern Cross will be receiving further improvements to its digital backend and observation setup. All computations and methodologies described here were created and verified using the current system parameters (frequency coverage, channelisation, sampling time, beamforming, and RFI management). Because the process is modular and parameter-driven, the improved system may benefit directly from the same operations such as data selection, Heimdall parameter sweeps, candidate merging, time analysis, and statistical testing. It is important to note that the GPU employed is the same as the upgraded system.

3.3.1 | Data Selection

Due to the large volume of data which was made available, which is 950 filterbank files containing substantial information, analysing the full dataset was deemed impractical within the time constraints. To produce a representative yet manageable sample, a systematic selection approach was conducted. This approach started by identifying the core characteristics for each file, such as SNR, DM and pulse width. These charac-

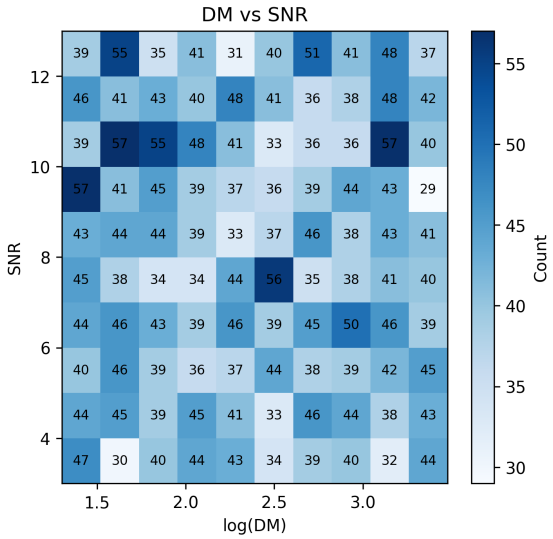
teristics were transformed using a logarithmic scale for DM and pulse width, which allowed for equal binning across orders of magnitude and ensured that both weak and strong signals, as well as narrow and broad pulses, were accurately represented.

Using these modified properties, a three-dimensional feature space was created and partitioned into a regular grid of volumetric bins. The injected data from each candidate file were assessed to identify which bins they occupied, resulting in an estimate of coverage over the feature space. Beginning with an empty sample set, files were randomly selected and their feature distributions compared to the existing sample. Files that significantly added additional occupied bins, boosting overall coverage, were kept, whereas those whose characteristics closely matched previously picked files beyond a predetermined threshold were skipped. Iterative progress was recorded, with the proportion of bins occupied and the number of files picked at each stage.

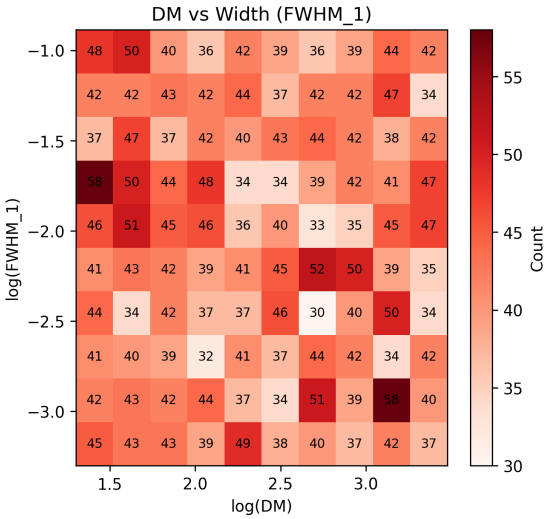
Throughout the iterative sampling process, tolerance limits for feature similarity were carefully selected to create a compromise between coverage expansion and redundancy avoidance. A tolerance parameter specified the maximum percentage of data points that might overlap in feature values before a file was considered too similar to the current sample. During this process, the tolerance parameter was varied between 1-10% to allow some similarity in the signals but overall varied signals in the sample dataset. By modifying this value accordingly and evaluating coverage improvements, the approach resulted in a subset of files whose combined feature distributions densely packed the multidimensional space, offering a reliable sample for analysis.

Visual evaluation was important in directing the selection process. After each batch of data was included, two-dimensional histograms and scatter plots of feature pairs (DM vs SNR, DM vs log FWHM, and SNR vs log FWHM), which can be seen in Figure 3.2, were generated to verify that the chosen subset maintained the necessary variety. These visual tests guaranteed that the final sample accurately represented the original dataset's variability while staying computationally feasible.

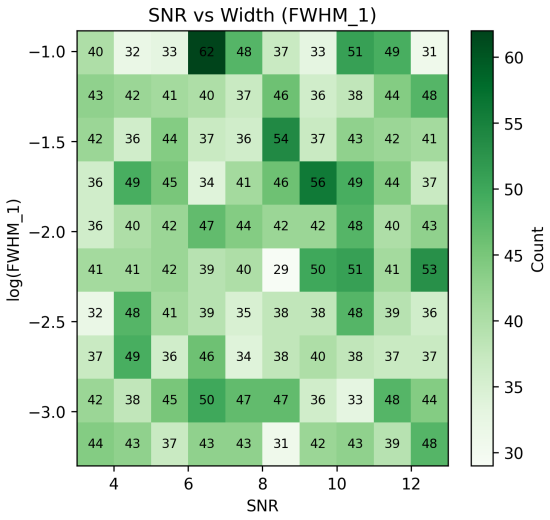
Finally, the data selection process produced a well balanced sample of 400 filterbank files that adequately covered the injection parameter space. By prioritising a variation in key signal properties and removing repetitive files, the technique increased the sample's value in evaluating the Heimdall pipeline's performance without having to evaluate the complete dataset.



Histogram of DM vs SNR.



Histogram of DM vs FWHM.



Histogram of SNR vs FWHM.

Figure 3.2: Histograms of unique data points in the filterbank files which were randomly plotted to form the sample size.

3.3.2 | Heimdall Processing Steps

Following data selection, where a total of 400 filterbank files were selected for the sample size, these files were processed with the Heimdall detection software, which was previously described in Section 2.6.1. The filterbank files were put under a systematic evaluation of 25 parameter combinations, which are shown in Table 3.2.

DM Tolerance	1.001	1.01	1.05	1.1	1.2
Boxcar Width	32	64	128	256	512

Table 3.2: Parameter Combination Values.

These combinations originated by combining five DM tolerance settings with five boxcar smoothing width samples. Preliminary trials revealing their sensitivity to fluctuations in DM alignment influenced the selection of these specific DM tolerance levels, which matched the likelihood of missing scattered signals against the possibility for increased false positives. Similarly, the chosen boxcar widths ranged in powers of two to investigate Heimdall's reaction to various time domain smoothing scales, from fine scale detection of small pulses to broader integration suited for longer duration bursts.

The execution started by reading a log file with the paths to the 400 processed filterbank files. Each file path was checked and mapped from its corresponding CSV metadata file, ensuring that only existent files were handled. Each filterbank file was assigned its own output directory inside a hierarchy based on the file name and parameter combination. This arrangement assisted downstream result gathering and analysis, making it possible to easily cross reference parameter settings and detection outcomes.

Heimdall was run for each combination of DM tolerance and boxcar values with standardised settings; a global DM search range from 0 to 3000 pc cm^{-3}, the current DM tolerance setting, the maximum boxcar width, GPU acceleration, and verbose logging. Each command execution collected both standard output and error streams, which were then routed to log files in the appropriate output folders. This extensive logging guaranteed that any problems such as missing data segments, GPU allocation failures, or runtime warnings were documented for debugging.

This automated technique allowed Heimdall jobs to be submitted in parallel across several GPU enabled cluster nodes, substantially reducing computational time. By methodically adjusting DM tolerance and boxcar width, the pipeline generated a large dataset of potential detections for each file and parameter combination. Subsequent phases combined these results, extracting parameters such as the number of discovered candidates, time performance, and accuracy measurements for SNR, DM, ToA for statistical analysis. Finally, this procedure supported the assessment and optimisation of Heimdall's parameter settings for FRB detection.

3.4 | Forming Preliminary Results

3.4.1 | Merging and Storing Candidate Results

Once Heimdall finished processing for all 400 filterbank files for each of the 25 parameter combinations, the first step in generating preliminary findings was to merge the separate candidate output files. Using Heimdall's inbuilt function, all candidate files associated with a specific parameter combination for a filterbank file were combined into a single comprehensive candidate file. This guaranteed that all detection events for a specific filterbank file and specific parameter combination were stored in a single location, allowing for more efficient downstream processing.

After generating all the candidate lists, the focus shifted to performance timings. Heimdall generates a set of internal time metrics for each filterbank file and parameter combination, each representing a different stage of the detection process. To avoid bias, start and final time measurements, which are vulnerable to cluster setup and shutdown overheads, were removed, as was the case with previous data trimming techniques. The remaining timing measurements, which represented the most significant computational effort, were averaged to obtain the mean execution time for each processing stage. A sample of the performance timings along with a normalised version of the total timings are presented in Tables 3.3 and 3.4 respectively. These mean timings offered a clear, comparative representation of how varying DM tolerances and boxcar widths affected computing performance across all files.

In addition to the time analysis, accuracy measurements for significant detection parameters were calculated. Several percentage based accuracy measurements for SNR, DM, and ToA were obtained by comparing Heimdall's candidate outputs to known injection parameters. The following were considered: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), coefficient of determination (R^2) and mean absolute percentage error (MAPE). Such measures gave an extensive view of detection performance, including both absolute deviations and relative error characteristics. A sample of these metrics is provided in Table 3.5. It is important to note that all accuracy metrics for SNR were calculated with the parameter combination of DM Tolerance of 1.001 and Boxcar Width of 512 acting as the benchmark point, since it is the finest combination.

A structured data repository helped in organising the large amount of results. An

Boxcar Width	Time Execution (s)	DM Tolerance				
		1.001	1.01	1.05	1.1	1.2
		±0.01	±0.01	±0.01	±0.01	±0.01
	0-DM Cleaning	1.78	1.65	1.81	1.67	1.79
	Dedispersion	3.75	0.85	0.47	0.35	0.22
	Baselining	2.01	0.60	0.26	0.18	0.12
32	Normalisation	1.76	0.52	0.23	0.15	0.11
	Filtering	1.75	0.52	0.22	0.15	0.10
	Find Giants	11.95	3.57	1.54	1.06	0.72
	Total	31.72	10.09	5.55	4.27	3.60
	0-DM Cleaning	1.77	1.80	1.66	1.66	1.66
	Dedispersion	3.79	0.89	0.42	0.30	0.22
	Baselining	2.01	0.60	0.26	0.18	0.12
64	Normalisation	1.76	0.52	0.23	0.16	0.11
	Filtering	2.07	0.62	0.27	0.18	0.12
	Find Giants	15.11	4.51	1.93	1.33	0.92
	Total	35.13	11.35	5.87	4.59	3.70
	0-DM Cleaning	1.80	1.64	1.83	1.67	1.80
	Dedispersion	3.95	0.89	0.44	0.32	0.22
	Baselining	2.01	0.60	0.26	0.18	0.12
128	Normalisation	1.77	0.52	0.23	0.16	0.11
	Filtering	2.38	0.70	0.31	0.21	0.14
	Find Giants	18.41	5.50	2.38	1.63	1.11
	Total	39.11	12.32	6.56	4.96	4.04
	0-DM Cleaning	1.86	1.80	1.69	1.81	1.67
	Dedispersion	3.78	0.93	0.47	0.32	0.22
	Baselining	2.02	0.60	0.26	0.18	0.12
256	Normalisation	1.77	0.52	0.23	0.16	0.11
	Filtering	2.68	0.80	0.35	0.24	0.16
	Find Giants	21.55	6.51	2.82	1.94	1.31
	Total	42.47	13.65	6.94	5.44	4.15
	0-DM Cleaning	1.77	1.66	1.83	1.66	1.81
	Dedispersion	3.93	0.93	0.40	0.35	0.24
	Baselining	1.99	0.60	0.26	0.18	0.12
512	Normalisation	1.75	0.53	0.23	0.16	0.11
	Filtering	2.98	0.90	0.39	0.27	0.18
	Find Giants	25.46	7.67	3.29	2.29	1.56
	Total	46.96	14.86	7.43	5.70	4.58

Table 3.3: Sample of results of performance timings (in seconds) obtained from Heimdall. This table is obtained from a file from the 400 dataset; the results show the timing metrics that contributed mostly to performance, for all parameter combinations.

	1.001	1.01	1.05	1.1	1.2
32	0.68	0.21	0.12	0.09	0.08
64	0.75	0.24	0.13	0.10	0.08
128	0.83	0.26	0.14	0.11	0.09
256	0.90	0.29	0.15	0.12	0.09
512	1.00	0.32	0.16	0.12	0.10

Table 3.4: Normalised version of the total timings displayed in Table 3.3. The layout is also kept similar that the first row represents the DM Tolerance values while the first column represents the corresponding Boxcar Width values.

initial workbook was constructed in Excel format, with each sheet allotted to one of the 400 filterbank files. Sections on each page documented average time results along with related parameter combinations. In another spreadsheet, accuracy measurements were similarly kept, with annotations connecting each accuracy result to its boxcar width and DM tolerance setting. Finally, a master dataframe was created by combining the timing and accuracy workbooks. This dataframe included one row per parameter file combination and columns for file identities, DM tolerance, boxcar width, average runtime, and each accuracy statistic. The resulting combined table served as the foundation for all subsequent data analysis and visual results.

3.4.2 | Visualising Preliminary Results

Prior to undertaking statistical analyses, a visual review was carried out to determine the alignment between Heimdall's detections and the known injection parameters. To identify the measured signals that should match the injected signals, the thirteen largest SNR values were chosen from the candidate files. This option focused on the recognised events while also collecting edge instances around the detection threshold.

The scatter plots were created with the ToA of each candidate as the horizontal axis and the DM as the vertical axis. Each point's marker size was scaled in direct proportion to its SNR, offering a visual clue to detection confidence; larger markers linked to stronger signals, while smaller markers showed marginal detections with more variability. Benchmark points, which indicate the true injection values, were overlaid with a separate colour. A nearest-neighbour matching technique linked each measured detection to its closest benchmark counterpart. An example of these scatter plots can be seen in Figure 3.3.

Boxcar Width	Metric	DM Tolerance				
		1.001	1.01	1.05	1.1	1.2
		±0.01	±0.01	±0.01	±0.01	±0.01
32	MAE	17.70	20.02	19.57	19.46	20.13
	MSE	660.83	736.26	729.47	736.34	749.35
	RMSE	25.71	27.13	27.01	27.14	27.37
	R^2	0.00	0.00	0.00	0.00	0.00
	MAPE	62.42	68.61	69.83	70.03	72.61
	Accuracy	37.58	31.39	30.17	29.97	27.39
64	MAE	13.41	13.80	13.90	13.67	16.77
	MSE	412.18	404.80	418.11	414.91	548.53
	RMSE	20.30	20.12	20.45	20.37	23.42
	R^2	0.00	0.00	0.00	0.00	0.00
	MAPE	36.56	36.54	37.85	37.17	47.74
	Accuracy	63.44	63.46	62.15	62.83	52.26
128	MAE	8.18	7.09	7.57	7.48	8.22
	MSE	186.40	157.77	165.80	162.03	179.29
	RMSE	13.65	12.56	12.88	12.73	13.39
	R^2	0.00	0.00	0.00	0.00	0.00
	MAPE	17.25	15.07	16.29	15.89	17.97
	Accuracy	82.75	84.93	83.71	84.11	82.03
256	MAE	2.11	2.53	2.86	2.95	5.25
	MSE	26.06	29.94	32.99	35.18	82.41
	RMSE	5.11	5.47	5.74	5.93	9.08
	R^2	0.82	0.79	0.77	0.75	0.45
	MAPE	3.55	4.32	4.99	5.15	12.38
	Accuracy	96.45	95.68	95.01	94.85	87.62
512	MAE	0.00	0.41	0.79	0.48	1.57
	MSE	0.00	0.30	1.41	0.42	6.54
	RMSE	0.00	0.55	1.19	0.65	2.56
	R^2	1.00	0.99	0.99	1.00	0.96
	MAPE	0.00	0.75	1.55	0.92	3.10
	Accuracy	100	99.25	98.45	99.08	96.90

Table 3.5: Sample of results of statistical metrics obtained from analyses. In this case, this table is obtained from a file from the 400 dataset; the results show the SNR analysis.

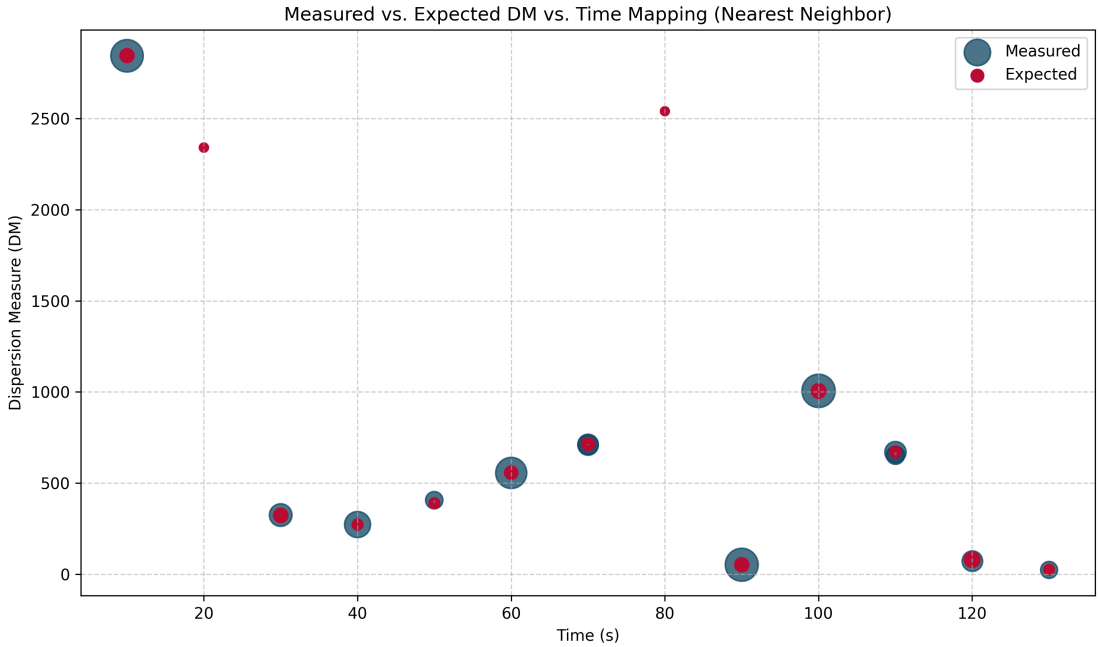


Figure 3.3: Scatter plot representing the measured data points by Heimdall against the expected data points from the generated FRB signals. The axes represent the values of DM and ToA, while the marker size of each point is a ratio which is proportional to the representation of the SNR value. Close overlap indicates accurate recovery of both timing and dispersion measure. Horizontal offsets reflect timing errors or smearing; vertical offsets indicate dedispersion bias or DM misestimation; large separations imply poorer performance for that event setting. Axes use linear scales. It is noted that some of the measured points have not matched with their respective expected point.

This representation of observed and predicted values served different purposes. First, it offered a check on detection fidelity; measured points around their benchmark equivalents indicated correct performance, whereas deviations suggested potential biases or errors caused by specific parameter choices. Second, the visual overlap between large, high SNR markers and their benchmark equivalents revealed the pipeline's most reliable performance under noise impact. Third, vertical distributions suggested dispersion related mistakes, whereas horizontal spreads indicated timing faults.

Overall, this stage played a role in establishing the initial assessment of the data, as well as displaying how the various parameter selections impacted detection accuracy.

3.5 | Statistical Tests

Several statistical tests were carried out to evaluate the performance of multiple parameter combinations employed during Heimdall processing. These methods are required to objectively determine the significance and reliability of observed variations between parameter combinations. The next section introduces and discusses each test, outlining its underlying reasoning and methodological foundation.
To investigate intrinsic structures in the data, dimensionality reduction by t-SNE was used, followed by cluster analysis through the use of HDBSCAN. Furthermore, the Friedman test, together with the Nemenyi post-hoc analysis, were used to assess the comparative performance of the investigated parameter combinations. This organised statistical technique provides a thorough and transparent examination, eventually directing the selection of the most efficient parameter values.

3.5.1 | t-SNE

Traditional analytical approaches find it difficult to identify minor but important patterns in complex datasets derived from radio astronomy observations. Dimensionality reduction approaches, such as t-SNE, have emerged as effective tools for detecting hidden patterns in high dimensional data (Van der Maaten and Hinton, 2008). The key purpose for using t-SNE in this research was to show and explain the complex correlations between the many parameter combinations explored with Heimdall. By decreasing the dataset's dimensionality to two dimensions, t-SNE allowed straightforward interpretation and presentation of underlying groups or clusters that would otherwise be hidden.

Van der Maaten and Hinton (2008) developed the t-SNE algorithm, which was created specifically for representing high dimensional data $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ by reducing it to two or three dimensional data $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$ that can be displayed in a scatter plot. Unlike linear dimensionality reduction approaches such as Principal Component Analysis (PCA), that focuses on keeping the low dimensional representations of dissimilar data points far apart, t-SNE detects non-linear patterns in data, resulting in more understandable representations of complex relationships. The objective of dimensionality reduction is to retain as much of the relevant structure of the high dimensional data that is possible in the low dimensional map. When dealing with high dimensional data on a low dimensional, non-linear manifold, it's essential to maintain

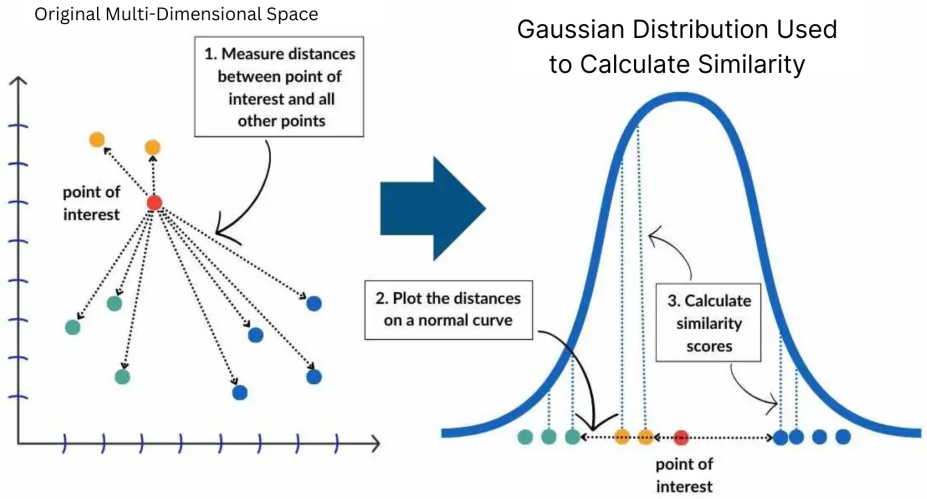


Figure 3.4: Visualisation of how t-SNE uses a Gaussian distribution to reduce dimensionality. Figure extracted from Otten (2024).

low dimensional representations of equivalent data points close together, which linear mapping might not be suitable for.

SNE as presented by Hinton and Roweis (2002), forms the basis for t-SNE. Initially, t-SNE computes probabilities for pairwise similarities between data points in the original high dimensional space. These similarities are approximated using a Gaussian distribution, which ensures that points close together in high dimensional space are more likely to be neighbours, which is represented in Figure 3.4. Specifically, SNE translates high dimensional Euclidean distances into conditional probabilities that express similarities. The conditional probability $p_{j|i}$ describes the similarity between data points x_i and x_j as,

$$p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2/2\sigma_i^2\right)}{\Sigma_{k\neq i}\exp\left(-||x_i - x_k||^2/2\sigma_i^2\right)}, \tag{3.1}$$

where σ_i is the variance of the Gaussian distribution centred at x_i, which is determined through optimisation for local neighbourhood density. It is important to note that the covariance between data points would need to be checked, as t-SNE does not check for correlation.

Next, t-SNE defines a probability distribution for points in the lower dimensional space,

using the Student's t-distribution. In the lower dimensional map points y_i and y_j, the conditional probability $q_{j|i}$ is defined as,

$$q_{j|i} = \frac{\exp\left(-||y_i - y_j||^2\right)}{\Sigma_{k \neq i} \exp\left(-||y_i - y_k||^2\right)}. \tag{3.2}$$

This decision allows t-SNE to effectively represent outliers while maintaining correct relative distances across clusters.

Finally, t-SNE iteratively reduces the Kullback-Leibler divergence, which is a measure of difference between two probability distributions, by altering points in the low dimensional space until a stable, understandable image is obtained. SNE minimises the sum of Kullback-Leibler divergences over all data points using a gradient descent method[1]. The cost function C is given by,

$$C = \sum_i KL\left(P_i||Q_i\right) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \tag{3.3}$$

in which P_i represents the conditional probability distribution over all other data points given data point x_i, and Q_i represents the conditional probability distribution over all other map points given map point y_i.

t-SNE has several parameters, including perplexity, learning rate, and number of iterations. Given the results were sensitive to these factors, a systematic grid search was performed. The perplexity parameter regulates the balance between local and global features of the data structure. A range of 5 to 55 was methodically investigated to identify an ideal perplexity that efficiently captured relevant local clusters without unnecessarily fragmenting the data.

A calculated learning rate enables stable and meaningful embeddings. This parameter was set to `"auto"`, since this option sets the learning rate to `max(N / early_exaggeration / 4, 50)` where `N` is the sample size, following Belkina et al. (2019) and Kobak and Berens (2019). Iteration values of 500 to 5000 were assessed. Sufficient iterations were required to provide consistent and reproducible visuals without increasing processing cost. The optimal parameters used for visual coherence and interpretability were a perplexity of 55 and 5000 iterations, resulting in clear, distinct, and stable clusters that were both interpretable and reproducible.

[1]Gradient descent is an iterative optimisation algorithm used to find the minimum value of a function (Boyd and Vandenberghe, 2004).

Representation of the t-SNE output involves producing two-dimensional scatter plots that clearly displayed the relative locations of data points representing various parameter combinations. These visualisations provided for fast understanding of clusters, assisting in the identification of groups of parameter combinations that performed similarly. Data points in tight, distinct clusters indicated parameter combinations that produced consistent outcomes, whereas scattered points indicated parameter combinations with less consistent or lower performance.

The interpretation of these plots was intuitive, considerably supporting further investigations such as clustering and statistical testing. Furthermore, the clear visual patterns generated by t-SNE were valuable for guiding the subsequent cluster selection and evaluating the performance consistency of parameter combinations, allowing for informed decision making on the optimal Heimdall settings.

3.5.2 | Cluster Analysis

Clustering is a widely recognised data analysis technique that groups data points so that elements within the same group, or cluster, are more similar to one another than to those in other groups. Unlike supervised learning frameworks, clustering is classified as unsupervised learning since it does not require labelled training data. Instead, clustering approaches detect intrinsic patterns in datasets based only on data attributes, making them valuable in exploratory data analysis (Jain et al., 1999).

HDBSCAN is an advanced clustering technique that is capable of finding clusters of different densities and shapes. Unlike traditional clustering algorithms such as k-means, which require a predefined number of clusters, and hierarchical clustering, which typically relies on predefined linkage criteria, HDBSCAN detects clusters without a predefined number while handling noise and outliers (McInnes et al., 2017).

By implementing hierarchical clustering approaches and producing a condensed cluster hierarchy, HDBSCAN outperforms Density-Based Spatial Clustering of Applications with Noise (DBSCAN) which is the technique which it was built upon. According to Campello et al. (2013) this method enables HDBSCAN to effectively determine clusters of varying densities, automatically detect a suitable number of clusters without predefinition, and handle and label noise points that do not fall into any cluster.

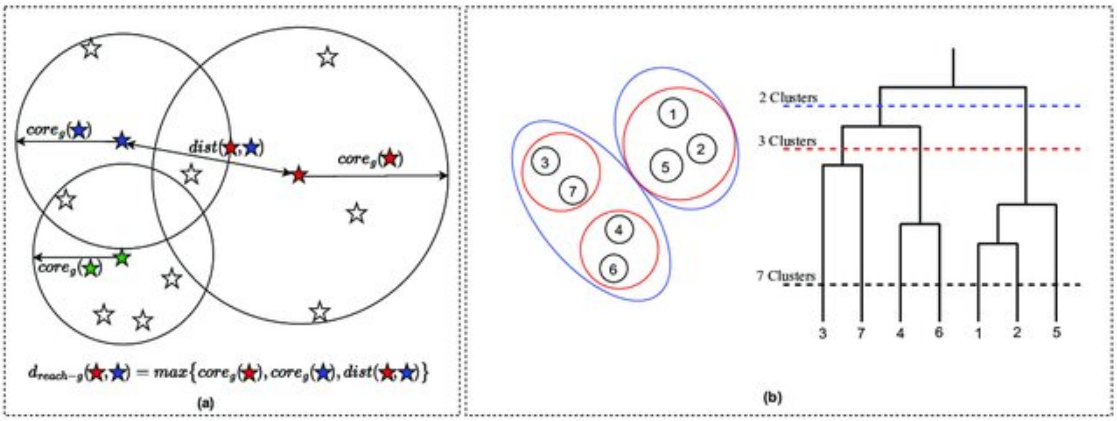


Figure 3.5: A concept of HDBSCAN clustering algorithm. (a) Mutual reachability distance computation from three clusters bearing varied densities. (b) Agglomerate dendrogram showing significant clusters that can be extracted. Figure extracted from Onyango et al. (2022).

The HDBSCAN algorithm operates through several key steps, these include; the construction of a mutual reachability graph, performance of hierarchical clustering, the condensed cluster tree formation and cluster extraction. Initially, HDBSCAN computes a mutual reachability distance by taking into account each point's core distance, which is the distance to its k^{th} nearest neighbour. In doing this, the initial dataset is converted into a weighted graph structure in which distances are represented by the edges (Campello et al., 2013). The system then undergoes hierarchical clustering, resulting in clusters with a dendrogram like structure. Figure 3.5 depicts how the mutual reachability distance is calculated and how information is extracted from dendrograms. This dendrogram depicts how clusters form and merge dependent on density, and it presents a representation of clustering at multiple scales (McInnes et al., 2017).

The dendrogram is condensed by eliminating clusters with a lower persistence threshold, thereby pruning unstable clusters and noise. Persistence is the stability of a cluster at different distance scales. Finally, stable clusters are identified based on their persistence, whereas unstable points are classified as noise. Thus, HDBSCAN produces robust, density-based clusters while also identifying noise explicitly (Campello et al., 2013).

HDBSCAN includes several hyperparameters that impact clustering results, notably the `min_cluster_size`, `min_samples`, and the distance metric. `min_cluster_size` specifies the minimal group size that makes an acceptable cluster. A grid search

strategy was used to test several values ranging from 250 to 500 in order to find one that created significant clusters without excessive fragmentation. `min_samples` determines the clusters' sensitivity to outliers. Multiple configurations were tried in a methodical manner, balancing cluster definition and noise management. Because of the nature of the t-SNE embeddings, Euclidean distance was used as the primary distance metric, however alternative metrics were examined to confirm robustness. It was found that the optimal value for `min_cluster_size` was 450.

The optimal hyperparameters were determined by evaluating cluster quality through the silhouette score, which is a metric used to evaluate the quality of clustering, and visual inspection of the clusters' coherence within the t-SNE feature space (Rousseeuw, 1987).

After identifying the ideal hyperparameters, HDBSCAN was applied directly to the two-dimensional t-SNE embeddings. The combination of these two approaches greatly improved interpretability. The t-SNE result had already reduced the data's complexity to a comprehensible visual form, then HDBSCAN revealed underlying structures by grouping related parameter combinations into useful clusters.

Visual evaluation of HDBSCAN clustering findings in the t-SNE space revealed several clusters, each reflecting an assortment of parameter combinations with comparable performance characteristics. HDBSCAN detected noise points as parameter combinations that did not consistently produce trustworthy or interpretable findings. This clear visual separation enabled the instant identification of parameter groups that required additional statistical analysis and comparison, allowing for more informed selection of optimal parameter values for Heimdall.

In summary, using HDBSCAN on the t-SNE embeddings not only revealed the underlying data structure, but also offered a solid foundation for further statistical analyses and informed decision making.

3.5.3 | Friedman Test

When analysing performance data from the Heimdall processing pipeline, assessing numerous parameter combinations requires repeated measurements and comparisons across different experimental settings. When working with such repeated measure-

ments data, it is necessary to select an appropriate statistical test that takes into account the data's intrinsic assumptions. The Friedman test, which is a non-parametric alternative to repeated measures ANOVA, appears to be suited for datasets where normality assumptions are questionable or the data is ordinal in nature [2]. This test compares differences across numerous related groups or conditions, which correspond to the parameter combinations assessed by Heimdall (Sheldon et al., 1996).

The Friedman test was selected due to its robustness to normality violations, which are typical in datasets produced from complex radio astronomy observations. Unlike parametric tests, the Friedman test does not require the data to follow a precise normal distribution, making it a more accurate framework for analysing parameter performance. Furthermore, the use of ranks in the Friedman test successfully handles data heterogeneity, making it a reliable choice for dealing with the unpredictability associated with FRB astronomical data.

Friedman (1937) developed the Friedman test, which makes use of ranked data to determine if there are significant differences between various related groups. Mathematically, the Friedman test entails rating each measurement within each subject or observation block individually. These rankings are then added together in blocks for each experimental condition or parameter combination. The test statistic is determined using these rank sums to assess how consistently one condition ranks differently from another over numerous blocks or samples.

Given a dataset $\{X_{ij}\}_{n\times k}$ in the form of a matrix with n rows and k columns and a single observation at the intersection of each row and column, the ranks within each block are calculated. If there are tied values, they are assigned the average of the rankings that would have been assigned without ties. The data is replaced with a new matrix $\{R_{ij}\}_{n\times k}$ where the entry R_{ij} is the rank of X_{ij} within block i. The values of R_j are found using $\frac{1}{n}\sum_{i=1}^{n} R_{ij}$. Hence the Friedman statistic, denoted as χ_F^2, is computed using the formula,

$$\chi_F^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1), \tag{3.4}$$

where n represents the number of subjects or blocks, k denotes the number of conditions or groups, and R_j indicates the sum of ranks for each condition j (Friedman, 1937). The resulting χ_F^2 statistic approximately follows a chi-square distribution with $k-1$ degrees of freedom, allowing the determination of statistical significance of observed

[2]Ordinal data is a type of data that has a natural, ordered relationship between its categories.

rank differences (Zimmerman and Zumbo, 1993).

The Friedman test outcomes indicate if at least one parameter combination performs substantially differently than others. A significant Friedman test result (usually tested at an alpha level of $\alpha = 0.05$) suggests that not all parameter combinations perform equally well. Hence, if the Friedman test is statistically significant, this indicates that there is evidence against the null hypothesis. However, it does not specify which parameter combinations differ or outperform from each other. This necessitates further examination using post-hoc analysis to identify particular differences (Demšar, 2006).

Therefore, the Friedman test was used for assessing the performance of different parameter combinations across the dataset and for different FRB features. This was necessary to determine whether certain parameter combinations with highly comparable performance and accuracy scores retained a statistically significant difference between them.

3.5.4 | Nemenyi post-hoc Test

Following a significant Friedman test result, the following step involves identifying which parameter combinations have statistically significant differences from one another, and which perform better or worse. Although the Friedman test effectively detects overall differences across several parameter combinations, it does not define which pairings vary substantially. Therefore, completing post-hoc analysis becomes necessary. Among the several available post-hoc tests, the Nemenyi post-hoc test, developed by Nemenyi (1963), is best suited to non-parametric conditions [3], as there is in this case. The Nemenyi test complements the Friedman test by allowing for multiple comparisons without strict distributional assumptions, ensuring robust and distribution free results (Demšar, 2006).

The Nemenyi post-hoc test is intended for being implemented after a significant Friedman test and can be used to rank data across different groups or conditions. The Nemenyi test evaluates pairwise comparisons of all parameter combinations based on rank differences. It calculates the Critical Difference (CD) between average ranks to determine if the differences are statistically significant. The Nemenyi test offers the

[3]Non-parametric conditions, in the context of statistical analysis, refer to situations where assumptions about the distribution of data are either minimal or absent. Non-parametric methods are "distribution-free" and can be applied to various data types, including those that are not normally distributed, have outliers, or are ordinal (ranked).

advantage of being distribution free, which eliminates the need for assumptions of normality or variance homogeneity, making it resilient against the irregularities that are common in real world datasets.

#classifiers	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

Table 3.6: Critical values for the two-tailed Nemenyi test after the Friedman test. Table extracted from Demšar (2006).

Mathematically, the Nemenyi test determines a CD threshold based on the number of combinations being compared, the number of observations, and the desired significance level. Two classifiers have significantly different performance when their average rankings differ by at least the CD,

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{3.5}$$

where critical values q_α are based on the Studentised range statistic divided by $\sqrt{2}$, whose values can be seen in Table 3.6; α is the chosen significance level, k is the number of parameter combinations and N is the number of data files. Comparisons between pairs of parameter combinations are considered significantly different if their difference in average ranks exceeds this CD threshold.

The implementation of the Nemenyi test consists of the following steps; beginning with calculating average ranks for each parameter combination using the Friedman test findings. The pairwise differences in average rankings are then calculated for all possible combinations. Each pairwise difference is then compared directly to the estimated CD value. When the pairwise rank difference exceeds this CD threshold, the difference between these two parameter combinations is deemed statistically significant.

Pairwise comparisons were used in this experiment to achieve robust and thorough analysis. The interpretation of the data involves creating visual representations, usually using CD diagrams, that clearly demonstrate significant differences. These diagrams improve the interpretability of the data by providing the positions of parameter combinations that perform similarly while clearly indicating outliers or significantly different parameter combinations.

Applying the Nemenyi post-hoc test to the Friedman test findings allowed for more exact identification of the Heimdall pipeline's most effective parameter combinations. These were classified based on their relative performance using evident visual and statistical distinctions. This offered strong evidence for differing between parameter combinations that resulted in greater performance and those that performed less effectively. By carefully detecting these discrepancies, the Nemenyi test improved the validity of the results.

3.6 | Summary

This chapter explained the complete methodology used to handle, analyse, and validate FRB data simulated to represent data from the Northern Cross telescope. Beginning with a description of the digital backend, the discussion centred on the telescope's signal processing architecture, operational settings, and observational parameters such as sampling rate, bandwidth, and integration times. The data characteristics section then discussed how FRBs were synthetically injected into 950 filterbank files, with distributions for DM, pulse width, and SNR, as well as the metadata fields captured for each event.

To manage the large dataset, a sample technique was used, with 400 representative files picked using a feature space binning and redundancy avoidance procedure. This provided full coverage of the DM, SNR, pulse width parameter space while remaining computationally feasible. The selected subset was processed using the Heimdall pipeline with 25 different parameter combinations.

The subsequent parts detailed the collection of the initial results; merged candidate files for each filterbank, averaged timing metrics, and computation of numerous accuracy measures (MAE, MSE, RMSE, R^2 and MAPE) for SNR, DM, and ToA. The data repository structure was described to show how the findings were structured for analysis. A first visual validation, which plotted observed against predicted detections with marker size connected to SNR, offered confirmation of detection quality and informed parameter tuning.

Finally, the statistical analyses were presented. Dimensionality reduction with t-SNE, followed by density-based clustering with HDBSCAN, showed intrinsic features in detection results. Non parametric tests including the Friedman test and Nemenyi

post-hoc comparisons were then used to evaluate parameter values throughout the full dataset without making any distribution assumptions. These statistical approaches provided the foundation for determining the best parameter combinations to potentially be implemented.

With this framework established, the following chapter will go into the results of these research efforts. Quantitative outcomes, statistical test results, and visualisations are displayed, using the approach given here to get conclusions on Heimdall's ideal setup.

4

Results & Discussion

4.1 | Introduction

Presented in this chapter are the findings from the analyses completed during the course of this research. This chapter aims to provide results obtained from various statistical tests, assess their implications, and ultimately determine the most successful parameter combination for implementation in the Northern Cross telescope.

The chapter opens with an evaluation of the results obtained from dimensionality reduction techniques and clustering algorithms, namely t-SNE and HDBSCAN. These approaches provide a clear display and logical classification of the parameter combinations examined, revealing underlying patterns and groups within the dataset.

Following that, the Friedman statistical test results are described, as well as the findings from the Nemenyi post-hoc analysis. These tests were useful in determining the statistical significance of differences across parameter combinations, allowing for logical decision making processes. The chapter reveals the parameter combination that was considered to be optimal based on these statistical calculations and the strike between sensitivity and accuracy for a real-time processing pipeline.

Furthermore, practical considerations for the computational efficiency and data processing capability of the chosen parameter combination are discussed. Another parameter of Heimdall is also explored to observe how the number of time samples supplied in a batch impacts the performance. These calculations demonstrate the practical scalability of the selected configuration.

Finally, the chapter concludes with a discussion interpreting the effects of the presented results. The findings are contextualised within the broader goals of optimising telescope performance, emphasising how these results contribute to balancing sensitivity and computational efficiency in real-time observational pipelines.

4.2 | Results

The outcomes of all conducted analyses, including dimensionality reduction, clustering, and statistical testing are presented in this section and interpreted to guide the selection of the optimal parameter combination for the Northern Cross telescope's real-time pipeline. These are followed by a short discussion to provide the information behind the choice made. A full list of Python packages and their versions used in these analyses is provided in Appendix A.

4.2.1 | Statistical Tests Results

4.2.1.1 | t-SNE

As mentioned in Section 3.5.1, t-SNE was implemented to visualise high dimensional data in two dimensions, allowing the identification of inherent patterns in the data structure. Prior to performing t-SNE, the Hopkins statistic [1] was calculated to assess the dataset's clustering tendency, resulting in a value of $H = 0.0193$. This result suggests that a strong clustering structure was identified, since the package employed, `pyclustertend`, stated that for values tending to 0, a high cluster tendency is expressed.

Two main plots depict important outcomes of this research. It is important to note that the axes of both plots are arbitrary, since the algorithm attempts to preserve distances in the high dimensional space; hence the actual point positioning is not important, only the relationships between them. The initial t-SNE plot, shown in Figure 4.1, clearly displays potential clusters created by different parameter combinations. These potential clusters successfully classify parameter sets based on fundamental similarity, exhibiting both unambiguous groupings and potential outliers.

The t-SNE projection depicts a structured landscape of parameter behaviour, with several distinct, high-density islands interspersed with more diffuse regions and isolated regions. Certain islands seem compact and tightly packed, indicating sets of parameter combinations that give very similar results; other islands are more diffuse, indicating internal variety within those parameter systems. The relative sizes of the islands are also instructive: larger clusters correlate to often occurring behaviours across the tested parameter grid, whereas smaller clusters reveal niche configurations with behaviours that differ from the main population.

[1] The Hopkins statistic is a way of measuring the cluster tendency of a data set. The method was introduced by Hopkins and Skellam (1954).

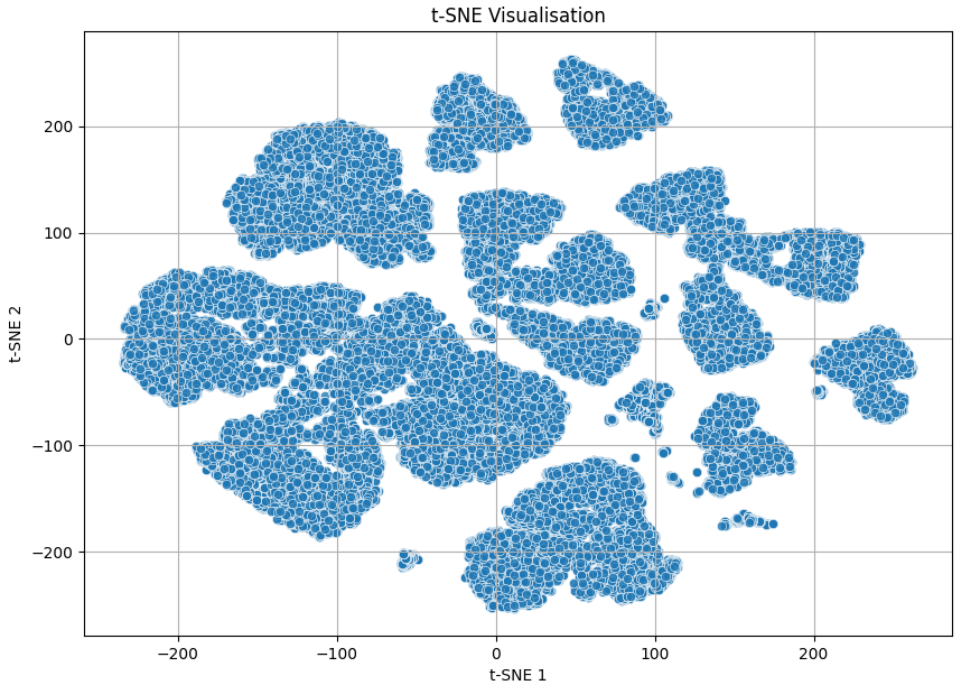


Figure 4.1: 2-D projection of t-SNE dimensionality reduction on the data. Each point represents a parameter-file outcome embedded into two dimensions by t-SNE (axes are unitless and not directly interpretable). High-density regions indicate many outcomes with very similar feature profiles (locally preserved neighbourhoods), i.e., parameter combinations that obtained similar performance and results. Tight, well-separated clusters suggest distinct performance regimes; sparse gaps reflect low-density transitions between regimes; and small "islands" likely correspond to atypical or noisy cases. Because t-SNE prioritises local structure, distances are most meaningful within neighbourhoods, while large-scale between-cluster spacing should be interpreted qualitatively rather than metrically.

Cluster separation in the embedding is mixed. Several clusters are visibly distinct, indicating qualitatively different operating conditions. In contrast, some clusters are near, touch, or are connected by narrow low-density bridges; these topologies indicate incremental transitions in performance as one or more factors change, rather than an abrupt categorical shift. A small amount of points are outside the major cluster bodies and appear as isolated or sparse points in the embedding. These non-clustered points should be viewed as potential outliers: they could represent runs tainted by RFI, measurement error, unstable SNR estimates, or pathological parameter selection.

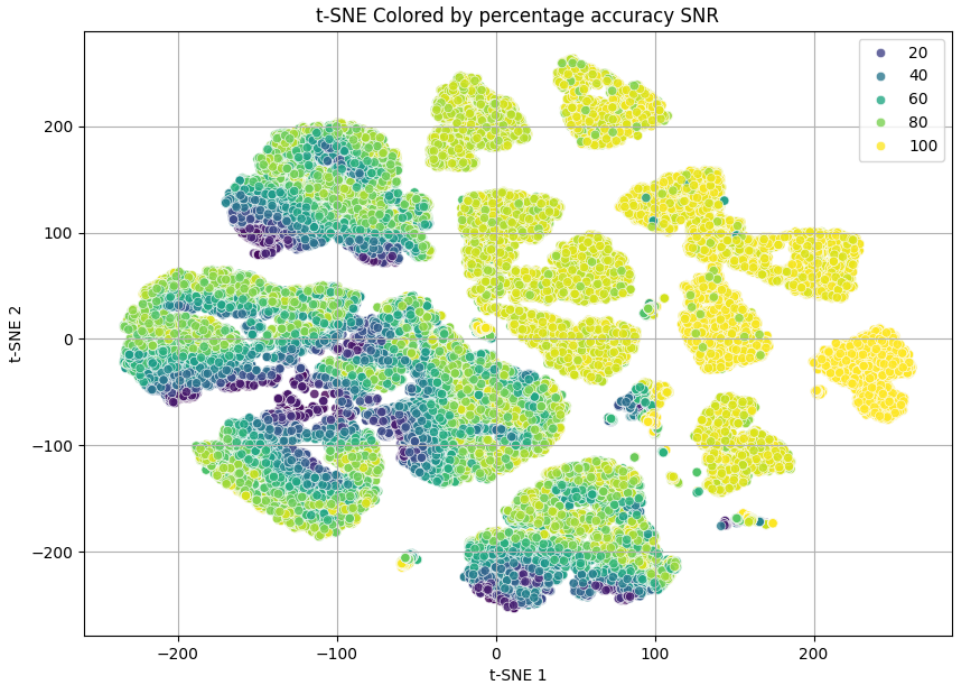


Figure 4.2: The same t-SNE 2-D projection from Figure 4.1 with the heatmap representing percentage accuracy for SNR (cool colours = lower accuracy; warm colours = higher accuracy, up to 100%). Dense, well-separated clusters indicate locally similar behaviour; within a cluster, smooth colour gradients suggest systematic performance trends, while mixed colours indicate parameter sensitivity or instability. Regions dominated by warm tones correspond to parameter combinations that recover SNR accurately; clusters with cooler tones highlight regimes with systematic under/over-estimation. Interpretation should focus on local neighbourhoods (t-SNE preserves local structure), while the global spacing between clusters is qualitative.

Figure 4.2 overlays a heatmap based on the accuracy scores of the SNR feature onto the same t-SNE projection, improving interpretability. This visualisation illustrates regions with changing signal clarity, finding potential clusters associated with greater or lower SNR accuracy values. Several of these possible clusters consistently showed higher accuracy for SNR, indicating their potential for obtaining optimal telescope performance. A similar procedure was followed to create heatmaps based on the accuracy scores of DM and ToA, but no significant effect was found in these plots.

Dense, well-separated clusters of points represent local regions of parameter space that give similar results. When a cluster has a smooth gradient of colour across its length,

it indicates a systematic performance trend within that neighbourhood, for example, a progressive improvement in SNR recovery as a certain parameter moves through a range. Such smooth colour shifts provide solid evidence that a given parameter (or a small set of parameters) causes predictable variations in SNR accuracy.

In contrast, irregular clusters or patches with an intermixing of cool and warm colours at short t-SNE distances indicate parameter sensitivity or instability. In some places, slight changes in parameter combination might result in huge swings in SNR accuracy, indicating that the pipeline response is less robust. Isolated warm-toned islands reveal parameter combinations that regularly recover SNR well and are thus promising candidates for further inquiry; cooler-toned clusters indicate regimes where the algorithm routinely under or overestimates SNR and should be investigated or excluded.

4.2.1.2 | Cluster Analysis

The data patterns identified by t-SNE were examined further using HDBSCAN. The clusters discovered by HDBSCAN, as shown in Figure 4.3, were directly compared and connected with the visual groups previously detected in the t-SNE projection. The Python packages used in this clustering stage were `Scikit-learn` for general data processing and the specific HDBSCAN library for clustering implementation. From the clusters seen in Figure 4.3, the data points labelled as -1 indicate noise/outliers while the rest are members of a respective cluster. In this case, matching with the groups that indicated the highest match percentage for SNR, the chosen clusters for further evaluation were; $0, 1, 2, 3, 6, 8, 10, 11$.

The HDBSCAN clustering technique identified discrete groups within the data, matching closely with the t-SNE visualisation, especially in parts of the t-SNE plot that suggested higher percentage accuracy for SNR values. The significant relationship enabled informed selection of clusters with consistently higher percentage accuracy SNRs. The discovered subset of data, which includes these recommended clusters, indicates ideal parameter combinations with the potential to improve performance and reliability. The findings of HDBSCAN clustering gave important assistance for statistical evaluation and practical decision making.

The subset of data included a total of 10 parameter combinations, where all the values of DM tolerance were included but only 2 values of Boxcar width were present. Hence, the following analyses were conducted exclusively on the results of the data points which

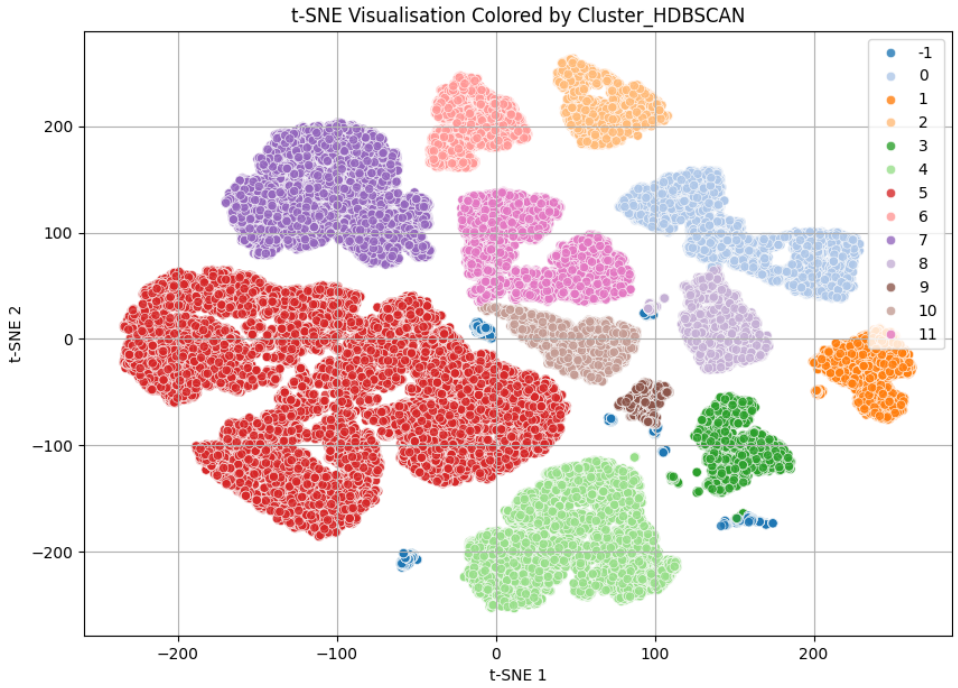


Figure 4.3: The same t-SNE 2-D projection from Figure 4.1 with HDBSCAN clustering. Colours indicate cluster labels returned by HDBSCAN; label –1 marks points classified as noise/outliers. Compact, well-separated colour patches denote stable density clusters, i.e. groups of outcomes with very similar behaviour, while mixed-colour boundaries reflect transitional regimes. Larger clusters represent similar performance results; smaller islands capture niche parameter settings; isolated –1 points suggest atypical or inconsistent detections. Because t-SNE preserves local neighbourhoods, interpretation should emphasise intra-cluster cohesion and boundary structure, while global inter-cluster distances are qualitative. These clusters form the basis for subsequent statistical comparisons and for selecting candidate parameter combinations.

fell under the parameter combinations found in Table 4.1.

DM Tolerance	1.001	1.01	1.05	1.1	1.2
Boxcar Width		256 512			

Table 4.1: Parameter combination values which appeared in the clusters chosen.

4.2.1.3 | Friedman Test

A Friedman test was carried out to determine the statistical significance of the observed differences across the chosen parameter combinations. HDBSCAN detected various parameter groups, and the test compared their most significant performance indicators. The analysis produced the following results, which are found in Table 4.2.

Feature	Friedman Statistic	p-value
SNR	2737.073	0.000
DM	234.072	2.283×10^{-45}
ToA	52.131	4.270×10^{-8}

Table 4.2: Table of results for Friedman test.

A p-value of < 0.05 reveals substantial variances across samples, indicating that at least one parameter combination performed differently than others. Because the Friedman test rejected the null hypothesis of equal medians, a post-hoc study was necessary to determine which combinations varied. This necessitated the use of the Nemenyi post-hoc test, which is discussed in the subsequent section.

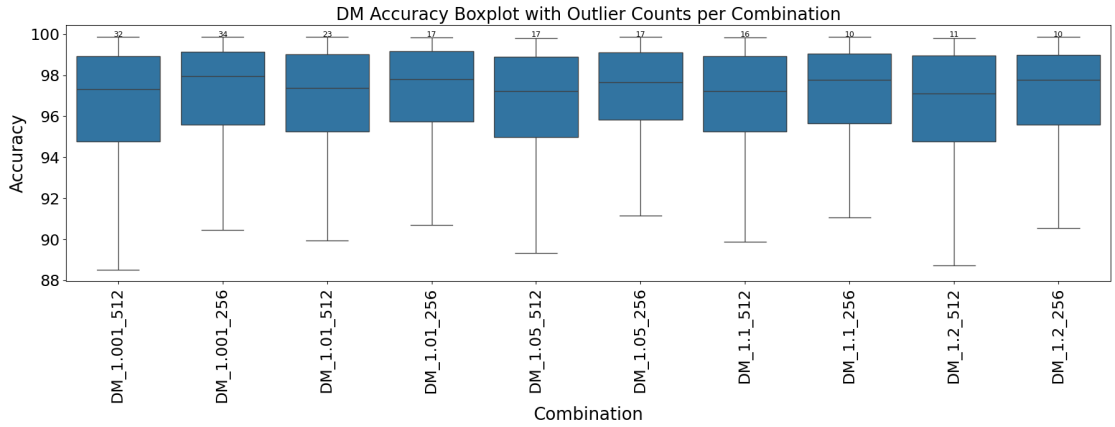


Figure 4.4: Boxplot of Friedman test results for DM.

To support the Friedman results, boxplots were generated using `Seaborn`'s `boxplot` function, displaying the distribution of the features for each parameter combination. Figures 4.4 and 4.5 depict these. The boxplots show variations in median SNR and DM accuracies, as well as likely outliers within each group. The findings of ToA were not included as no significant information was obtained. Parameter combinations with greater median features and densely clustered interquartile ranges support the Friedman test results, leading the selection of the most advantageous configuration. It

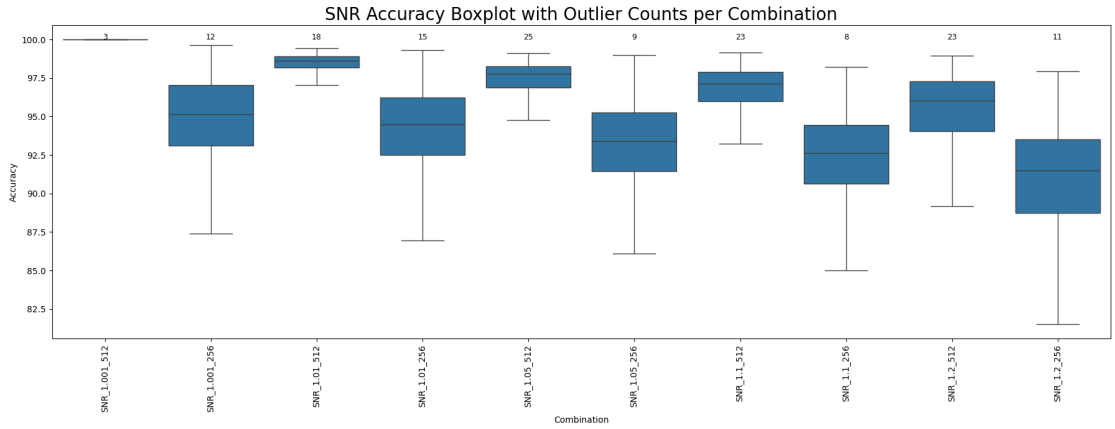


Figure 4.5: Boxplot of Friedman test results for SNR.

is observed in Figure 4.4 that the performance of each combination is more or less the same for DM accuracy, while in Figure 4.5 there are more fluctuations in results. It is also important to note that in the case for SNR, the first combination is shown to be 100% accurate, this is because as was previously mentioned, it was used as a benchmark for the other results since it is the finest parameter combination. The combinations with a Boxcar Width value of 512 displayed tighter interquartile ranges but had more outliers. The result that the parameter combinations which have a Boxcar width value of 256 show more varied interquartile ranges is expected since it is considered as a more "coarse" setting when compared to 512.

4.2.1.4 | Nemenyi post-hoc Test

Given the significant Friedman test result, a Nemenyi post-hoc analysis was performed using the `scikit-posthocs` package (`posthoc_nemenyi_friedman` function) with a significance level of $\alpha = 0.05$ to keep consistency with the Friedman test. This criterion was chosen over a more lenient $\alpha = 0.10$ to maintain a balance between Type I and Type II error rates [2] and comply with conventional practice in statistical testing. It ensures that identified differences are less likely related to random variability. The CD value was determined to be 0.55. A CD diagram was created for every performance metric, such as SNR, DM, and ToA. These are shown in Figure 4.6.

[2] In hypothesis testing, Type I and Type II errors are potential errors that can occur when drawing conclusions about a population based on sample data. A Type I error (false positive) occurs when a true null hypothesis is incorrectly rejected. Conversely, a Type II error (false negative) occurs when a false null hypothesis is not rejected.

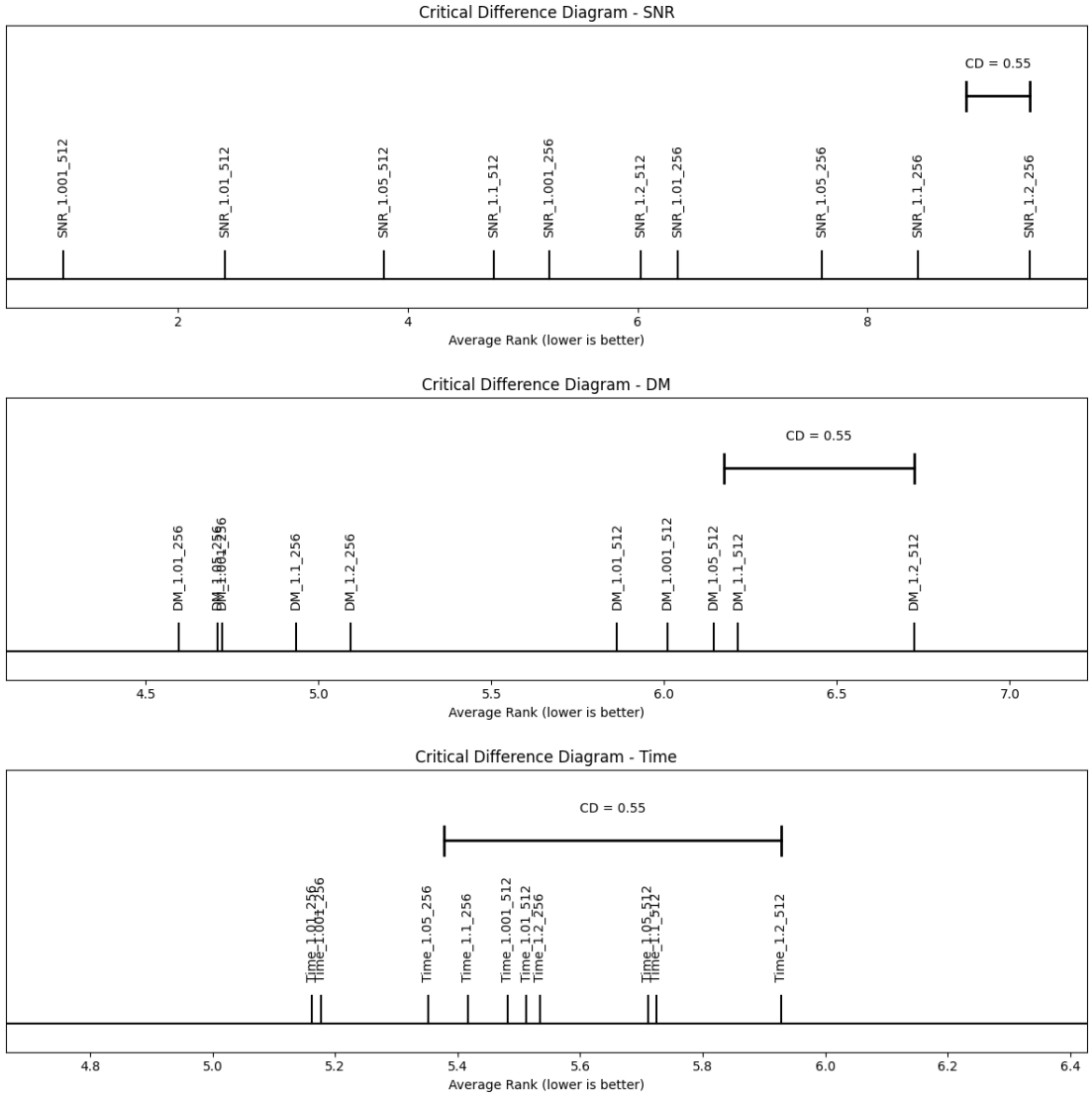


Figure 4.6: Critical Difference diagrams for DM, SNR and ToA.

These CD diagrams plot the average ranks of each parameter combination on a horizontal axis. The lower the value of CD that a combination has the better its performance was during the test, meaning that the best performing combination would be found to the left of the diagrams. Combinations which have a distance between them that is less than the CD value indicate no statistically significant difference at the calculated α level.

For the SNR feature, the CD diagram revealed that there is a critical difference between

almost all of the parameter combinations, with the parameter combination having a DM tolerance of 1.001 and Boxcar Width of 512 obtaining the best result since it was used as a benchmark for this feature. On the other hand, for the DM feature, the top 5 combinations were the ones which had a Boxcar Width of 256 and between them there is no significant difference according to the CD range. This implies that any combination with this Boxcar value will perform on the same level in the case for DM accuracy. Finally in the CD diagram for ToA accuracy, it is noted that for the majority of the parameter combinations there is no significant difference, since most of them fall within the CD range.

The results obtained in the CD diagrams match the previously obtained results in the boxplots depicted in Figures 4.4 and 4.5, where there were significant differences in results for SNR but similar performance for DM. Together, these CD diagrams provide a clear, visual summary of which parameter combinations differ significantly across each evaluation metric and support a multi-criteria decision approach for selecting the final optimal set.

4.2.2 | Final Decision

After evaluating the statistical analyses and the runtimes, the chosen parameter combination which was determined as the most suitable configuration was the one which included a DM Tolerance of 1.01 and a Boxcar Width of 256. This combination achieves detection accuracy on par with other top performing parameter sets while requiring significantly less computation time, making it ideal for real-time telescope operations. Boxplot comparisons further demonstrated that this configuration resulted in the fewest extreme SNR outliers, reflecting consistent and reliable signal detection across observations.

The Nemenyi post-hoc results reinforce this selection; the 1.01/256 combination attained an average rank advantage over alternative sets that exceeded the CD threshold ($\text{CD} = 0.55$ at $\alpha = 0.05$), confirming its performance. Beyond performance metrics, this configuration also balances sensitivity and computational cost effectively, preventing excessively long runtimes without sacrificing detection capability.

Moreover, by selecting a moderate DM tolerance and a reasonable boxcar width, the total number of trials remains manageable, reducing the processing burden on the Northern Cross telescope's hardware. Taken together, these considerations; accuracy, reliabil-

ity, statistical validation, and system resource constraints affirm that a DM tolerance of 1.01 and boxcar width of 256 offer the optimal trade-off for integration into the Northern Cross telescope's real-time pipeline.

4.3 | Impact of Gulp Size on Processing Time

The relationship between data gulp size and processing time was investigated in order to identify another significant variable impacting the pipeline's real-time performance. In this context, a "gulp" is the contiguous block of beam-formed samples that the pipeline processes in a single batch; and the term "gulp size" refers to the amount of time samples which are supplied into Heimdall in every batch. The product specifies the effective duration (in seconds) of each gulp:

$$\text{Gulp Duration}\,(s) = \text{Gulp Size (samples)} \times \text{Sampling Time}\,(s). \tag{4.1}$$

While gulp size is determined in the time-sample domain, system performance is ultimately limited by the time required to process a comparable segment of continuous time. For this study, the sample time was fixed, allowing for direct change in gulp duration by modifying the gulp size.

To investigate the effect of gulp size on performance, a filterbank file of real data acquired using the Northern Cross telescope of the pulsar PSR B0329+54, was used to obtain this relation. The filterbank file was tested under Heimdall, keeping all parameters fixed except for the `nsamps_gulp` which controls the gulp size. In this case the DM tolerance was set to 1.01 and the Boxcar width was set to 256 to see the results of gulp size on the optimal configuration which was obtained previously. The average runtime for each setting was recorded and plotted against the corresponding gulp duration. The resulting plot is shown in Figure 4.7.

Each data point represents the average runtime at a given gulp duration, and the red-dotted diagonal line denotes the real-time threshold, i.e. 1 s of compute time per 1 s of data. The data point represented by the green rhombus represents the gulp size that was fixed for the previous analysis. The linear regression fit reveals two key components; firstly a fixed initialisation overhead of around 5.04 s represents the minimum runtime, and a per-second processing cost of around 0.122 s shows the data processing speed.

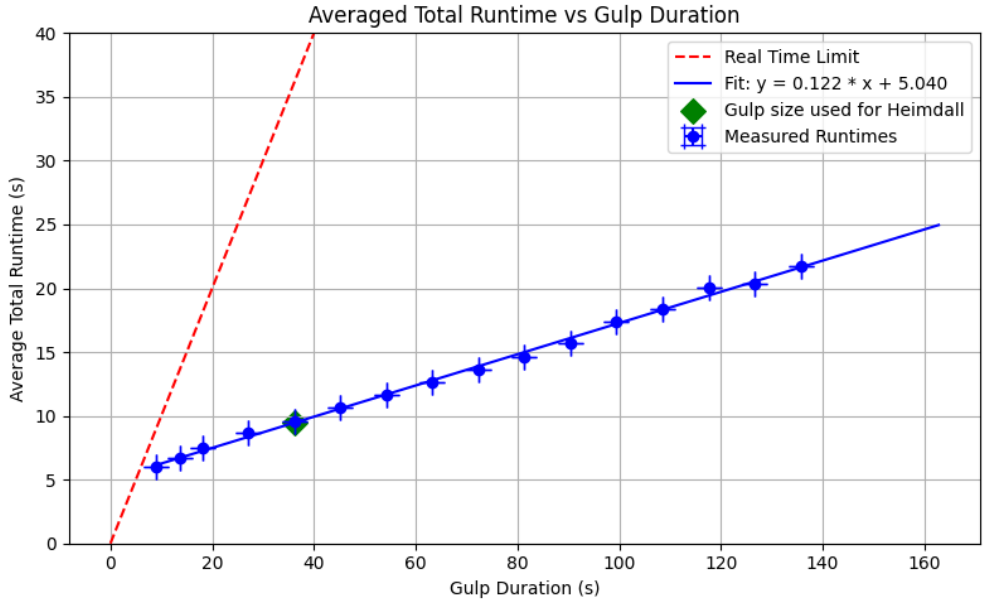


Figure 4.7: Gulp duration plotted against averaged total runtime. Blue markers show the time to process a single filterbank file for a given gulp configuration (total pipeline runtime per file). Error bars denote variability across repeated runs at each gulp. The solid line is the linear fit; the dashed red line marks the real-time boundary. The green diamond highlights the operating gulp used for subsequent Heimdall tests.

Since the value of the processing speed is less than 1, this means that the system processes data faster than real-time, meaning that it takes 0.122 s to process 1 s of data. This is beneficial for catching up during data bursts or parallel tasks. The value of the fixed overhead is insignificant for the values of the gulp durations which are typically used, but would be inefficient for gulp durations which are ≤ 10 s. The strong linearity demonstrates that runtime grows consistently with gulp duration, allowing for simple modelling of system scalability and delay budgeting using known observable factors.

4.4 | Discussion

The multi-stage methodology which was presented here produced an interconnected explanation of parameter performance. Instead of simply evaluating each test separately, the results of dimensionality reduction, clustering, and statistical hypothesis testing combine to inform the final decision. The t-SNE embeddings suggested natural groupings, and HDBSCAN provided objective boundaries around these groupings; but it is the combination of these visualisations with statistical validation that provided the confidence in the drawn conclusions.

The Friedman test demonstrated that not all parameter combinations perform similarly, and the Nemenyi post-hoc analysis identified which groupings reflect the best possible configurations. This layered technique prevents over reliance on any single measure or visualisation: high SNR clusters discovered in t-SNE are regarded as ideal if they exceed the CD level in the Nemenyi test and show strong consistency in boxplot distributions.

The analysis presented in Section 4.3 demonstrated a linear relationship between gulp duration and pipeline runtime, based on tests using a filterbank file from a Northern Cross observation of PSR B0329+54. The linear model revealed a steady processing overhead of 5.04 s and a scalable processing cost of 0.122 s per second of data. These results demonstrated that, larger gulp sizes enhance efficiency by reducing the influence of fixed overhead but the duration of the gulp is longer. The 40 s operating gulp size, which was implemented in the other obtained results, combines responsiveness and efficiency, allowing adequate duration for auxiliary operations. Overall, determining the gradient and intercept of the runtime model allows for adjusting of system parameters, providing robust, real-time performance.

While increasing gulp duration boosts computational efficiency by reducing fixed overheads, it also introduces a proportional detection delay; the pipeline only finishes processing a gulp after the entire batch has been acquired, delaying the identification of fast transients that include FRBs. If the buffered antenna voltage data must be stored at the moment of detection (for example, coherent dedispersion or high-time-resolution post-analysis), a greater gulp demands proportionally larger upstream buffers to hold the earlier samples until processing is complete. In reality, choosing a bigger gulp shifts resource needs away from CPU time and toward memory and storage capacity in the front-end architecture.

Delayed detection has an impact on downstream alerting as well. The generation of VOEvent triggers (or other real-time alerts) is dependent on rapid candidate identification; longer gulps increase the delay between the physical arrival of a transient and the emitted external alarm. To maintain scientific usefulness for multi-messenger follow-up, either buffer sizes must be expanded to allow retrieval of pre-trigger antenna data, or the operational policy should prioritise smaller gulps when low-latency warnings are critical.

Applying this model to real world operational limitations reveals that the chosen DM

tolerance of 1.01 and boxcar width of 256 provide practical advantages that go beyond statistical significance. These settings reduce computational load, preserving telescope throughput, while still providing high detection sensitivity. The absence of outliers in the boxplots indicates that these parameters will perform consistently even under changing noise conditions or unexpected RFI patterns.

Nonetheless, it is important to keep in mind the limitations of this method. The data presented here, while extensive, may not capture all conceivable temporal or spectral fluctuations experienced during live observations. Adaptability might be improved by incorporating an online feedback system that allows the pipeline to dynamically adjust DM tolerance or boxcar width in response to detect performance drift.

Overall, this integrative analysis strikes a balance between exploratory and confirmatory techniques to reach a measured decision that is also practical to implement. By basing the final parameter combination on the provided evidence, a replicable guideline for future parameter optimisation in radio astronomy applications was established.

4.5 | Summary

This chapter revealed the findings from investigations conducted to enhance the setup of the software used in the Northern Cross telescope's real-time detection pipeline. The study applied dimensionality reduction, clustering, and statistical hypothesis testing to identify parameter combinations that balance detection sensitivity with processing efficiency. Initially, t-SNE embeddings and Hopkins statistic evaluation validated the data's clustering tendency and helped to visualise the data, while the SNR heatmap indicated high-performance regions. HDBSCAN clustering then identified natural groups for further statistical analysis.

A Friedman test revealed substantial differences across candidate configurations, prompting Nemenyi post-hoc comparisons at $\alpha = 0.05$ (CD $= 0.55$). CD diagrams and boxplot distributions reliably identified the combination of DM tolerance = 1.01 and boxcar width = 256 as statistically better, with few outliers and narrow interquartile ranges. A gulp size compared to runtime investigation revealed a linear scaling relationship, guiding the selection of a 40 s gulp size to preserve real-time margins, which was the value used in the prior analyses.

The final parameter set of DM tolerance of 1.01, boxcar width of 256, and a 40-second gulp size meets the objectives of enhancing detection performance while respecting hardware limits and real-time requirements. This integrated approach presents a reproducible methodology for parameter optimisation in radio astronomy, demonstrating how exploratory visualisations and confirmatory statistical tests may be linked with system-level performance models. By basing conclusions on various evidence, the chapter reveals that the project's objectives of delivering reliable, real-time detection without consuming excessive computational resources were fulfilled.

5

Conclusion

5.1 | Achieved Aims and Objectives

This dissertation aimed to improve the performance and adaptability of the FRB detection pipeline at the Northern Cross radio telescope, which operates in a fast paced environment requiring real-time signal detection. One of the main challenges encountered in the field is striking a balance between detection sensitivity, computing efficiency, and adaptation to real-world observational restrictions. By concentrating on the systematic evaluation of important pipeline parameters, this study addressed a demand for evidence based detection optimisation.

The first objective was to determine how certain parameters such as DM step sizes, box-car widths, and SNR thresholds affect detection accuracy, data integrity, and pipeline duration. These factors were not picked at random; each has a direct impact on signal identification and processing efficiency. The investigation made use of synthetic FRB injections, which provide a controlled and consistent setting for evaluating different combinations. This enabled the evaluation of detection results independent of the random nature of live telescope data, providing clarity when considering performance trade-offs.

Through repeated testing and benchmarking, the research established a parameter setup that resulted in excellent performance across various parameters. This ideal setup is now available for real-world implementation, allowing researchers to use these results. Importantly, this optimal combination was established using a systematic process that can be replicated and modified. This repeatability is a defining feature of reliable practice, ensuring that subsequent investigations may verify, criticise, and improve on the findings.

In addition to this deliverable, the study proposed a testing methodology that can be repeated and tailored for other telescopes facing similar optimisation challenges. This system, which is based on synthetic data generation and benchmark validation, can serve as a model for future research into pipeline performance and dependability. Thus, the study not only meets the Northern Cross's immediate operational aims, but also contributes to the larger field of radio astronomy instrumentation and signal processing.

5.2 | Limitations

Notwithstanding the findings, the research had numerous drawbacks. The most pressing of these is due to hardware constraints. The trials were carried out on a single GPU, which limited the extent of the investigation, especially in terms of multi-beam parallel processing. In operational contexts, particularly those requiring large-scale sky surveys, the capacity to handle data from several beams at the same time is critical. It was not possible to recreate such scenarios in this work, due to the fact that the GPU was being overloaded when there was the attempt to process more than one beam on the GPU. Furthermore, the detection software used (Heimdall) was not intended for native multi-beam processing, complicating efforts to emulate these kinds of setups. This indicates that the reported optimum settings may require additional confirmation in multi-threaded or distributed processing systems.

Furthermore, the reliance on Heimdall as the sole detection program added another layer of restriction. While Heimdall is successful and widely implemented, it may be a bit of a black box in certain instances. Many of its core processing stages such as candidate clustering, threshold methods, and dedispersion subroutines are either difficult to modify or not visible to the user. This meant that numerous potentially relevant parameters could not be examined, while others had to be considered fixed. In addition, Heimdall does not make use of the latest GPU hardware capabilities or recent improvements in parallel computing frameworks, which limits its scalability and flexibility in high-performance contexts. As detection techniques progress, including ML and GPU-accelerated algorithms, many of these advancements are unlikely to be deployed inside Heimdall's current architecture. While this technique reflects the realities of present operations at the Northern Cross, it restricts the generalisation of the results and the capacity to investigate more complex correlations between factors.

In addition, concerns about Heimdall's long-term viability are growing. The software is no longer actively updated across all branches, and when detection pipelines and hardware configurations develop, compatibility issues may emerge. If Heimdall is deprecated or unsupported, the work done in this research may be difficult to reproduce or use without significant changes. This reliance on a single, perhaps outdated tool generates a level of technical debt that might hinder long-term progress.

Moreover, the study concentrated on a selection of variables known to impact performance, while many others; such as beam overlap settings, data reduction algorithms,

and RFI rejection filters were left unchanged. Their impact is unknown and untested. While this limiting of scope was important for practicality, it illustrates an area of incompleteness that future research should address.

5.3 | Future Work

Looking forward, the approach and findings of this study provide a solid platform for various prospective research topics. One of the most immediate and important next steps would be to expand the current experimental design to include additional pipeline parameters. Varying the beam overlap factors, for example, or implementing advanced RFI rejection systems, might provide new insights into the sensitivity-to-computational load trade-off. The same synthetic injection methodology can be used to assess these variables in a controlled and repeatable manner.

Another area of noteworthy research is the use of ML techniques into the detection pipeline. Specifically, using supervised or semi-supervised learning algorithms to classify candidate signals has the potential to significantly reduce false positives and improve detection precision. Given the amount of data produced by current radio telescopes, manual classification is not scalable. A well-trained ML model might be used as a post-processing filter or even earlier in the pipeline to make informed judgments regarding signal quality and source type.

Furthermore, there is a clear need to broaden the software tools employed in detection processes. While Heimdall provided a useful platform for this research, it is critical to investigate alternatives that provide better flexibility, transparency, or community support. Comparative assessments of open-source tools like FETCH and newer GPU-accelerated dedispersion engines could provide useful benchmarks and assist to future-proof the detection infrastructure. Ideally, the pipeline should be flexible enough to support plug-and-play components, such as those powered by ML or distributed computing systems.

Longer-term research might look at the real-time deployment of efficient pipelines on distributed systems like multi-GPU clusters or cloud-based infrastructures. These systems may provide the required scalability to handle next-generation survey data while also allowing for real-time modification based on observational priorities. Finally,

collaboration with international facilities operating under different environmental conditions or scientific goals may validate the framework's transferability and reveal new optimisation strategies.

5.4 | Final Remarks

In conclusion, this dissertation contributed practical and methodological additions to the improvement of real-time signal detection pipelines in radio astronomy. The establishment of a replicable testing methodology and the determination of optimal processing parameters has offered immediate improvements to the Northern Cross telescope's operating capability. In addition, the study highlights the need of controlled experimentation and systematic parameter analysis in increasing the efficiency and accuracy of scientific instruments.

The limitations described do not diminish the importance of the findings; but rather highlight the challenges of real-world pipeline design and execution. Addressing these challenges necessitates a combination of scientific methods, adaptable software infrastructure, and design concepts. The work presented here is an important step in that direction, providing a solid foundation on which future innovations can be built.

Ultimately, the study's conclusions are more than a series of proposals; they represent a dynamic approach to understanding and enhancing signal detection in an era of ever-increasing data and technology possibilities. As the field of astronomy continues to push beyond the boundaries of what is observable, similar research guarantees that these tools are as accurate and flexible as the science they serve.

A

Python Packages

All of the programming scripts which were created and used during the methodology section of this work can be found in the following repository: GitHub Repository Link. The repository contains the exact scripts used to generate results, including Heimdall runs, candidate collation, accuracy scoring, timing extraction, dataframe construction, visualisation, t-SNE dimensionality reduction, HDBSCAN clustering, and statistical tests.

Input benchmark CSVs and processed outputs are organised as described in the README; paths are configurable in the scripts. Randomness is controlled where applicable (`random_state=42`); any remaining nondeterminism (e.g., GPU scheduling) does not affect qualitative conclusions. All figures and tables in the manuscript can be regenerated from the outputs written by the scripts.

Here is a list of the Python packages and the specific versions used in the virtual environment, which was created for the computational steps of the methodology. Experiments were executed with Python >= 3.12. These are specified since some of the plots which were generated, and the statistical results which were calculated using these packages and their results may vary according to which package or version/s might be used. The relevant information and documentation of how each function and package work may be found online.

Package	Version
astropy	6.1.7
astropy-iers-data	0.2025.1.20.0.32.27
contourpy	1.3.1
cycler	0.12.1
et_xmlfile	2.0.0
exceptiongroup	1.3.0
fonttools	4.55.5
h5py	3.12.1

hdbscan	0.8.40
imageio	2.37.0
iniconfig	2.1.0
joblib	1.4.2
kiwisolver	1.4.8
lazy_loader	0.4
llvmlite	0.43.0
markdown-it-py	3.0.0
matplotlib	3.10.0
mdurl	0.1.2
narwhals	1.39.0
networkx	3.4.2
numba	0.60.0
numpy	2.0.2
openpyxl	3.1.5
packaging	24.2
pandas	2.2.3
pillow	11.1.0
pip	22.0.2
plotly	6.0.1
pluggy	1.5.0
pyclustertend	1.9.0
pyerfa	2.0.1.5
Pygments	2.19.1
pyparsing	3.2.1
pytest	8.3.5
python-dateutil	2.9.0.post0
pytz	2024.2
PyYAML	6.0.2
rich	13.9.4
scikit-image	0.25.0
scikit-learn	1.6.1
scipy	1.15.1
seaborn	0.13.2
setuptools	59.6.0
six	1.17.0

threadpoolctl	3.5.0
tifffile	2025.1.10
tomli	2.2.1
tqdm	4.67.1
typing_extensions	4.12.2
tzdata	2025.1
XlsxWriter	3.2.2
your	0.6.7

References

BP Abbott, R Abbott, TD Abbott, F Acernese, K Ackley, C Adams, T Adams, P Addesso, RX Adhikari, VB Adya, et al. Multi-messenger observations of a binary neutron star merger. *The Astrophysical Journal Letters*, 848(2):l12, 2017.

Karel Adámek and Wesley Armour. A real-time single pulse detection algorithm for gpus, 2016. URL `https://arxiv.org/abs/1611.09704`.

Devansh Agarwal, Kshitij Aggarwal, Sarah Burke-Spolaor, Duncan R Lorimer, and Nathaniel Garver-Daniels. Fetch: A deep-learning based classifier for fast transient classification. *Monthly Notices of the Royal Astronomical Society*, 497(2):1661–1674, 2020.

Joel Akeret, Chihway Chang, Aurelien Lucchi, and Alexandre Refregier. Radio frequency interference mitigation using deep convolutional neural networks. *Astronomy and computing*, 18:35–39, 2017.

M Amiri, K Bandura, P Berger, M Bhardwaj, MM Boyce, PJ Boyle, C Brar, M Burhanpurkar, P Chawla, J Chowdhury, et al. The chime fast radio burst project: system overview. *The Astrophysical Journal*, 863 (1):48, 2018.

Mandana Amiri, Bridget C. Andersen, Kevin Bandura, Sabrina Berger, Mohit Bhardwaj, Michelle M. Boyce, P. J. Boyle, Charanjot Brar, Daniela Breitman, Tomas Cassanelli, Pragya Chawla, Tianyue Chen, J.-F. Cliche, Amanda Cook, Davor Cubranic, Alice P. Curtin, Meiling Deng, Matt Dobbs, Fengqiu (Adam) Dong, Gwendolyn Eadie, Mateus Fandino, Emmanuel Fonseca, B. M. Gaensler, Utkarsh Giri, Deborah C. Good, Mark Halpern, Alex S. Hill, Gary Hinshaw, Alexander Josephy, Jane F. Kaczmarek, Zarif Kader, Joseph W. Kania, Victoria M. Kaspi, T. L. Landecker, Dustin Lang, Calvin Leung, Dongzi Li, Hsiu-Hsien Lin, Kiyoshi W. Masui, Ryan Mckinven, Juan Mena-Parra, Marcus Merryfield, Bradley W. Meyers, Daniele Michilli, Nikola Milutinovic, Arash Mirhosseini, Moritz Münchmeyer, Arun Naidu, Laura Newburgh, Cherry Ng, Chitrang Patel, Ue-Li Pen, Emily Petroff, Tristan Pinsonneault-Marotte, Ziggy Pleunis, Masoud Rafiei-Ravandi, Mubdi Rahman, Scott M. Ransom, Andre Renard, Pranav Sanghavi, Paul Scholz, J. Richard Shaw, Kaitlyn Shin, Seth R. Siegel, Andrew E. Sikora, Saurabh Singh, Kendrick M. Smith, Ingrid Stairs, Chia Min Tan, S. P. Tendulkar, Keith Vanderlinde, Haochen Wang, Dallas Wulf, and A. V. Zwaniga. The first chime/frb fast radio burst catalog. *The Astrophysical Journal Supplement Series*, 257(2):59, December 2021. ISSN 1538-4365. doi: 10.3847/1538-4365/ac33ab.

B. C. Andersen, K. M. Bandura, M. Bhardwaj, A. Bij, M. M. Boyce, P. J. Boyle, C. Brar, T. Cassanelli, P. Chawla, T. Chen, J.-F. Cliche, A. Cook, D. Cubranic, A. P. Curtin, N. T. Denman, M. Dobbs, F. Q. Dong, M. Fandino, E. Fonseca, B. M. Gaensler, U. Giri, D. C. Good, M. Halpern, A. S. Hill, G. F. Hinshaw, C. Höfer, A. Josephy, J. W. Kania, V. M. Kaspi, T. L. Landecker, C. Leung, D. Z. Li, H.-H. Lin, K. W. Masui, R. Mckinven, J. Mena-Parra, M. Merryfield, B. W. Meyers, D. Michilli, N. Milutinovic, A. Mirhosseini, M. Münchmeyer, A. Naidu, L. B. Newburgh, C. Ng, C. Patel, U.-L. Pen, T. Pinsonneault-Marotte, Z. Pleunis, B. M. Quine, M. Rafiei-Ravandi, M. Rahman, S. M. Ransom, A. Renard, P. Sanghavi, P. Scholz, J. R. Shaw, K. Shin, S. R. Siegel, S. Singh, R. J. Smegal, K. M. Smith, I. H. Stairs, C. M. Tan, S. P. Tendulkar, I. Tretyakov, K. Vanderlinde, H. Wang, D. Wulf, and A. V. Zwaniga. A bright millisecond-duration

radio burst from a galactic magnetar. *Nature*, 587(7832):54–58, November 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2863-y.

Wes Armour, Karel Adamek, Jan Novotny, Sofia Dimoudi, Cees Carels, Nassim Quannoughi, Mike Giles, Aris Karastergiou, Chris Williams, Steve Roberts, et al. Astroaccelerate. *Zenodo*, 2019.

Kevin Bandura, Graeme E Addison, Mandana Amiri, J Richard Bond, Duncan Campbell-Wilson, Liam Connor, Jean-François Cliche, Greg Davis, Meiling Deng, Nolan Denman, et al. Canadian hydrogen intensity mapping experiment (chime) pathfinder. In *Ground-based and Airborne Telescopes V*, volume 9145, pages 738–757. SPIE, 2014.

K. W. Bannister, A. T. Deller, C. Phillips, J.-P. Macquart, J. X. Prochaska, N. Tejos, S. D. Ryder, E. M. Sadler, R. M. Shannon, S. Simha, C. K. Day, M. McQuinn, F. O. North-Hickey, S. Bhandari, W. R. Arcus, V. N. Bennert, J. Burchett, M. Bouwhuis, R. Dodson, R. D. Ekers, W. Farah, C. Flynn, C. W. James, M. Kerr, E. Lenc, E. K. Mahony, J. O'Meara, S. Osłowski, H. Qiu, T. Treu, V. U, T. J. Bateman, D. C.-J. Bock, R. J. Bolton, A. Brown, J. D. Bunton, A. P. Chippendale, F. R. Cooray, T. Cornwell, N. Gupta, D. B. Hayman, M. Kesteven, B. S. Koribalski, A. MacLeod, N. M. McClure-Griffiths, S. Neuhold, R. P. Norris, M. A. Pilawa, R.-Y. Qiao, J. Reynolds, D. N. Roxby, T. W. Shimwell, M. A. Voronkov, and C. D. Wilson. A single fast radio burst localized to a massive galaxy at cosmological distance. *Science*, 365(6453):565–570, August 2019. ISSN 1095-9203. doi: 10.1126/science.aaw5903.

KW Bannister, RM Shannon, J-P Macquart, Chris Flynn, PG Edwards, Morgan O'Neill, Stefan Osłowski, Matthew Bailes, Barak Zackay, Nathan Clarke, et al. The detection of an extremely bright fast radio burst in a phased array feed survey. *The Astrophysical Journal Letters*, 841(1):L12, 2017.

Benjamin R. Barsdell. Advanced architectures for astrophysical supercomputing. 1 2012. doi: 10.25916/sut.26286541.v1.

Benjamin R Barsdell and Andrew Jameson. Heimdall: Gpu accelerated transient detection pipeline for radio astronomy. *Astrophysics Source Code Library*, pages ascl–2407, 2024.

Benjamin R Barsdell, Matthew Bailes, David G Barnes, and Christopher J Fluke. Accelerating incoherent dedispersion. *Monthly Notices of the Royal Astronomical Society*, 422(1):379–392, 2012.

Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications*, 10(1):5415, 2019.

M C Bezuidenhout, E Barr, M Caleb, L N Driessen, F Jankowski, M Kramer, M Malenta, V Morello, K Rajwade, S Sanidas, B W Stappers, and M Surnis. Meertrap: 12 galactic fast transients detected in a real-time, commensal meerkat survey. *Monthly Notices of the Royal Astronomical Society*, 512(1):1483–1498, March 2022. ISSN 1365-2966. doi: 10.1093/mnras/stac579.

S Bhandari, E F Keane, E D Barr, A Jameson, E Petroff, S Johnston, M Bailes, N D R Bhat, M Burgay, S Burke-Spolaor, M Caleb, R P Eatough, C Flynn, J A Green, F Jankowski, M Kramer, V Venkatraman Krishnan, V Morello, A Possenti, B Stappers, C Tiburzi, W van Straten, I Andreoni, T Butterley, P Chandra, J Cooke, A Corongiu, D M Coward, V S Dhillon, R Dodson, L K Hardy, E J Howell, P Jaroenjittichai, A Klotz, S P Littlefair, T R Marsh, M Mickaliger, T Muxlow, D Perrodin, T Pritchard, U Sawangwit, T Terai, N Tominaga, P Torne, T Totani, A Trois, D Turpin, Y Niino, R W Wilson, A Albert, M André, M Anghinolfi,

G Anton, M Ardid, J-J Aubert, T Avgitas, B Baret, J Barrios-Martí, S Basa, B Belhorma, V Bertin, S Biagi, R Bormuth, S Bourret, M C Bouwhuis, H Brânzaş, R Bruijn, J Brunner, J Busto, A Capone, L Caramete, J Carr, S Celli, R Cherkaoui El Moursli, T Chiarusi, M Circella, J A B Coelho, A Coleiro, R Coniglione, H Costantini, P Coyle, A Creusot, A F Díaz, A Deschamps, G De Bonis, C Distefano, I Di Palma, A Domi, C Donzaud, D Dornic, D Drouhin, T Eberl, I El Bojaddaini, N El Khayati, D Elsässer, A Enzenhöfer, A Ettahiri, F Fassi, I Felis, L A Fusco, P Gay, V Giordano, H Glotin, T Gregoire, R Gracia-Ruiz, K Graf, S Hallmann, H van Haren, A J Heijboer, Y Hello, J J Hernández-Rey, J Hößl, J Hofestädt, C Hugon, G Illuminati, C W James, M de Jong, M Jongen, M Kadler, O Kalekin, U Katz, D Kießling, A Kouchner, M Kreter, I Kreykenbohm, V Kulikovskiy, C Lachaud, R Lahmann, D Lefèvre, E Leonora, S Loucatos, M Marcelin, A Margiotta, A Marinelli, J A Martínez-Mora, R Mele, K Melis, T Michael, P Migliozzi, A Moussa, S Navas, E Nezri, M Organokov, G E Păvălaş, C Pellegrino, C Perrina, P Piattelli, V Popa, T Pradier, L Quinn, C Racca, G Riccobene, A Sánchez-Losa, M Saldaña, I Salvadori, D F E Samtleben, M Sanguineti, P Sapienza, F Schüssler, C Sieger, M Spurio, Th Stolarczyk, M Taiuti, Y Tayalati, A Trovato, D Turpin, C Tönnis, B Vallage, V Van Elewyck, F Versari, D Vivolo, A Vizzocca, J Wilms, J D Zornoza, and J Zúñiga. The survey for pulsars and extragalactic radio bursts – ii. new frb discoveries and their follow-up. *Monthly Notices of the Royal Astronomical Society*, 475(2):1427–1446, November 2017. ISSN 1365-2966.

Germano Bianchi, Federico Perini, and Giancarlo Setti. A brief history of the northern cross radio telescope. In *2023 8th IEEE History of Electrotechnology Conference (HISTELCON)*, pages 175–178. IEEE, 2023.

C. D. Bochenek, V. Ravi, K. V. Belov, G. Hallinan, J. Kocz, S. R. Kulkarni, and D. L. McKenna. A fast radio burst associated with a galactic magnetar. *Nature*, 587(7832):59–62, November 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2872-x.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Bernard F Burke, Francis Graham-Smith, and Peter N Wilkinson. *An introduction to radio astronomy*. Cambridge University Press, 2019.

M. Caleb, C. Flynn, M. Bailes, E. D. Barr, T. Bateman, S. Bhandari, D. Campbell-Wilson, W. Farah, A. J. Green, R. W. Hunstead, A. Jameson, F. Jankowski, E. F. Keane, A. Parthasarathy, V. Ravi, P. A. Rosado, W. van Straten, and V. Venkatraman Krishnan. The first interferometric detections of fast radio bursts. *Monthly Notices of the Royal Astronomical Society*, 468(3):3746–3756, March 2017. ISSN 1365-2966. doi: 10.1093/mnras/stx638.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

Dario Carbone, AJ van Der Horst, RAMJ Wijers, JD Swinbank, A Rowlinson, JW Broderick, YN Cendes, AJ Stewart, ME Bell, RP Breton, et al. New methods to constrain the radio transient rate: results from a survey of four fields with lofar. *Monthly Notices of the Royal Astronomical Society*, 459(3):3161–3174, 2016.

Cees Carels, Karel Adámek, Jan Novotnỳ, Wesley Armour, Nassim Ouannoughi, and Sofia Dimoudi. Astroaccelerate: Accelerated software package for processing time-domain radio astronomy data. *Astrophysics Source Code Library*, pages ascl–1912, 2019.

S. Chatterjee, C. J. Law, R. S. Wharton, S. Burke-Spolaor, J. W. T. Hessels, G. C. Bower, J. M. Cordes, S. P. Tendulkar, C. G. Bassa, P. Demorest, B. J. Butler, A. Seymour, P. Scholz, M. W. Abruzzo, S. Bogdanov, V. M. Kaspi, A. Keimpema, T. J. W. Lazio, B. Marcote, M. A. McLaughlin, Z. Paragi, S. M. Ransom, M. Rupen, L. G. Spitler, and H. J. van Langevelde. A direct localization of a fast radio burst and its host. *Nature*, 541(7635):58–61, January 2017. ISSN 1476-4687. doi: 10.1038/nature20797.

Marshall H Cohen. Cosmic noise: A history of early radio astronomy. *Physics Today*, 63(6):49–49, 2010.

J. M. Cordes and Ira Wasserman. Supergiant pulses from extragalactic neutron stars. *Monthly Notices of the Royal Astronomical Society*, 457(1):232–257, January 2016. ISSN 1365-2966. doi: 10.1093/mnras/stv2948.

James M Cordes and T Joseph W Lazio. Ne2001. i. a new model for the galactic distribution of free electrons and its fluctuations. *arXiv preprint astro-ph/0207156*, 2002.

JM Cordes and Maura A McLaughlin. Searches for fast radio transients. *The Astrophysical Journal*, 596(2): 1142, 2003.

Denis Cutajar, Alessio Magro, Josef Borg, Kristian Zarb Adami, Germano Bianchi, Claudio Bortolotti, ALESSANDRO Cattani, FRANCO Fiocchi, LUCA Lama, ANDREA Maccaferri, et al. A real-time space debris detection system for birales. In *INTERNATIONAL ASTRONAUTICAL CONGRESS: IAC PROCEEDINGS*, pages 1–9. International Astronautical Federation, IAF, 2018.

N D'Amico, G Grueff, S Montebugnoli, A Maccaferri, A Cattani, C Bortolotti, L Nicastro, F Fauci, G Tomassetti, M Roma, et al. The new northern cross pulsar system: four years of pulsar timing observations. *Astrophysical Journal Supplement v. 106, p. 611*, 106:611, 1996.

Kishalay De and Yashwant Gupta. A real-time coherent dedispersion pipeline for the giant metrewave radio telescope. *Experimental Astronomy*, 41(1):67–93, 2016.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Bruce T Draine. *Physics of the interstellar and intergalactic medium*, volume 19. Princeton University Press, 2011.

A Ficarra, G Grueff, and G Tomassetti. A new bologna sky survey at 408 mhz. *Astronomy and Astrophysics Supplement Series (ISSN 0365-0138), vol. 59, Feb. 1985, p. 255-347.*, 59:255–347, 1985.

Griffin Foster, Aris Karastergiou, Marisa Geyer, Mayuresh Surnis, Golnoosh Golpayegani, Kejia Lee, Duncan Lorimer, Danny C Price, and Kaustubh Rajwade. Verifying and reporting fast radio bursts. *Monthly Notices of the Royal Astronomical Society*, 481(2):2612–2627, 2018.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

Brian Hopkins and John Gordon Skellam. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(2):213–227, 1954.

Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

Karl G Jansky. Electrical disturbances apparently of extraterrestrial origin. *Proceedings of the IEEE*, 86(7): 1510–1515, 2002.

Steven M Kay. Statistical signal processing, volume ii, detection theory, 1998.

EF Keane and E Petroff. Fast radio bursts: search sensitivities and completeness. *Monthly Notices of the Royal Astronomical Society*, 447(3):2852–2856, 2015.

Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):5416, 2019.

CJ Law, GC Bower, S Burke-Spolaor, BJ Butler, P Demorest, A Halle, S Khudikyan, TJW Lazio, M Pokorny, J Robnett, et al. realfast: real-time, commensal fast transient surveys with the very large array. *The Astrophysical Journal Supplement Series*, 236(1):8, 2018.

Nicola T Locatelli, Gianni Bernardi, Germano Bianchi, Riccardo Chiello, Alessio Magro, Giovanni Naldi, Maura Pilia, Giuseppe Pupillo, Alessandro Ridolfi, Giancarlo Setti, et al. The northern cross fast radio burst project–i. overview and pilot observations at 408 mhz. *Monthly Notices of the Royal Astronomical Society*, 494(1):1229–1236, 2020.

D. R. Lorimer, M. Bailes, M. A. McLaughlin, D. J. Narkevic, and F. Crawford. A bright millisecond radio burst of extragalactic origin. *Science*, 318(5851):777–780, November 2007. ISSN 1095-9203. doi: 10.1126/science.1147532.

Duncan R Lorimer, Maura A McLaughlin, and Matthew Bailes. The discovery and significance of fast radio bursts. *Astrophysics and Space Science*, 369(6):59, 2024.

Duncan Ross Lorimer and Michael Kramer. *Handbook of pulsar astronomy*, volume 4. Cambridge university press, 2005.

ME Lower, A Karastergiou, S Johnston, PR Brook, S Dai, M Kerr, RN Manchester, LS Oswald, RM Shannon, C Sobey, et al. The ubiquity of variable radio emission and spin-down rates in pulsars. *Monthly Notices of the Royal Astronomical Society*, 538(4):3104–3129, 2025.

J-P Macquart, Evan Keane, Keith Grainge, Matthew McQuinn, RP Fender, Jason Hessels, Adam Deller, Ramesh Bhat, René Breton, Shami Chatterjee, et al. Fast transients at cosmological distances with the ska. *arXiv preprint arXiv:1501.07535*, 2015.

J.-P. Macquart, J. X. Prochaska, M. McQuinn, K. W. Bannister, S. Bhandari, C. K. Day, A. T. Deller, R. D. Ekers, C. W. James, L. Marnoch, S. Osłowski, C. Phillips, S. D. Ryder, D. R. Scott, R. M. Shannon, and N. Tejos. A census of baryons in the universe from localized fast radio bursts. *Nature*, 581(7809):391–395, May 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2300-2.

Alessio Magro, Aris Karastergiou, Stefano Salvini, Benjamin Mort, Fred Dulwich, and Kristian Zarb Adami. Real-time, fast radio transient searches with gpu de-dispersion. *Monthly Notices of the Royal Astronomical Society*, 417(4):2642–2650, 2011.

M Malenta, B Stappers, MC Bezuidenhout, M Caleb, L Driessen, F Jankowski, V Morello, K Rajwade, and S Sanidas. Accelerating radio astronomy with high performance computing. *Astronomical Data Analysis Software and Systems XXIX*, 527:457, 2020.

Richard N Manchester, Andrew G Lyne, Fernando Camilo, Jon F Bell, VM Kaspi, Nicolo' D'Amico, NPF McKay, F Crawford, IH Stairs, A Possenti, et al. The parkes multi-beam pulsar survey–i. observing and data analysis systems, discovery and timing of 100 pulsars. *Monthly Notices of the Royal Astronomical Society*, 328(1):17–35, 2001.

B. Marcote et al. A repeating fast radio burst source localized to a nearby spiral galaxy. *Nature*, 577(7789): 190–194, 2020. doi: 10.1038/s41586-019-1866-z.

Leland McInnes, John Healy, Steve Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

Maura A McLaughlin, AG Lyne, DR Lorimer, M Kramer, AJ Faulkner, RN Manchester, JM Cordes, F Camilo, A Possenti, IH Stairs, et al. Transient radio bursts from rotating neutron stars. *Nature*, 439 (7078):817–820, 2006.

Ian S McLean. *Electronic imaging in astronomy: detectors and instrumentation*. Springer, 2008.

D. Michilli, A. Seymour, J. W. T. Hessels, L. G. Spitler, V. Gajjar, A. M. Archibald, G. C. Bower, S. Chatterjee, J. M. Cordes, K. Gourdji, G. H. Heald, V. M. Kaspi, C. J. Law, C. Sobey, E. A. K. Adams, C. G. Bassa, S. Bogdanov, C. Brinkman, P. Demorest, F. Fernandez, G. Hellbourg, T. J. W. Lazio, R. S. Lynch, N. Maddox, B. Marcote, M. A. McLaughlin, Z. Paragi, S. M. Ransom, P. Scholz, A. P. V. Siemion, S. P. Tendulkar, P. Van Rooy, R. S. Wharton, and D. Whitlow. An extreme magneto-ionic environment associated with the fast radio burst source frb 121102. *Nature*, 553(7687):182–185, January 2018. ISSN 1476-4687. doi: 10.1038/nature25149.

G. Naldi, A. Magro, F. Fiori, A. Mattana, F. Perini, N. Ragno, M. Schiaffino, H. Camilleri, A. DeBarro, K. Bugeja, L. Beduzzi, G. Bernardi, G. Bianchi, L. Bruno, M.A. De Luca, P. Di Lizia, M. Fiorentini, A. Geminardi, M. Massari, M.F. Montaruli, A. Orlati, D. Pelliciari, M. Pilia, A. Poli, and M. Trudu. The design of a new digital signal processing system for the upgraded northern cross radio telescope. In *Proceedings of the URSI Asia-Pacific Radio Science Conference (AP-RASC)*, Sydney, Australia, August 2025. in press.

Peter Bjorn Nemenyi. *Distribution-free multiple comparisons.* Princeton University, 1963.

Gelu M Nita, Dale E Gary, Zhiwei Liu, Gordon J Hurford, and Stephen M White. Radio frequency interference excision using spectral-domain statistics. *Publications of the Astronomical Society of the Pacific*, 119 (857):805, 2007.

AR Offringa, AG De Bruyn, M Biehl, S Zaroubi, G Bernardi, and VN Pandey. Post-correlation radio frequency interference classification methods. *Monthly Notices of the Royal Astronomical Society*, 405(1):155–167, 2010.

Shem Otoi Onyango, Solomon Amoah Owiredu, Kwang-Il Kim, and Sang-Lok Yoo. A quasi-intelligent maritime route extraction from ais data. *Sensors*, 22(22):8639, 2022.

Neri Van Otten. T-sne explained: Top 6 tools practical how to get started guide [python], Aug 2024. URL `https://spotintelligence.com/2023/12/22/practical-guide-t-sne/`.

OS Ould-Boukattine, JWT Hessels, MP Snelders, F Kirsten, R Blaauw, JJ Sluman, and H Mulder. A bright burst detection from frb 20240619d at 1.3 ghz using the westerbork-rt1 25-m telescope. *The Astronomer's Telegram*, 16732:1, 2024.

D Pelliciari, G Bernardi, M Pilia, G Naldi, G Maccaferri, F Verrecchia, C Casentini, M Perri, F Kirsten, G Bianchi, et al. The northern cross fast radio burst project-iv. multi-wavelength study of the actively repeating frb 20220912a. *Astronomy & Astrophysics*, 690:A219, 2024.

E. Petroff, E. D. Barr, A. Jameson, E. F. Keane, M. Bailes, M. Kramer, V. Morello, D. Tabbara, and W. van Straten. Frbcat: The fast radio burst catalogue. *Publications of the Astronomical Society of Australia*, 33, 2016. ISSN 1448-6083. doi: 10.1017/pasa.2016.35.

Emily Petroff, Leon Houben, Keith Bannister, Sarah Burke-Spolaor, Jim Cordes, Heino Falcke, Ronald van Haren, Aris Karastergiou, Michael Kramer, Casey Law, et al. Voevent standard for fast radio bursts. *arXiv preprint arXiv:1710.08155*, 2017.

Emily Petroff, JWT Hessels, and DR Lorimer. Fast radio bursts. *The Astronomy and Astrophysics Review*, 27 (1):4, 2019.

Emily Petroff, JWT Hessels, and DR Lorimer. Fast radio bursts at the dawn of the 2020s. *The Astronomy and Astrophysics Review*, 30(1):2, 2022.

Danny C Price. Real-time stream processing in radio astronomy. In *Big Data in Astronomy*, pages 83–112. Elsevier, 2020.

Nicholas J Pritchard, Andreas Wicenec, Mohammed Bennamoun, and Richard Dodson. Advancing rfi-detection in radio astronomy with liquid state machines. *arXiv preprint arXiv:2504.09796*, 2025.

KM Rajwade, Mechiel Christiaan Bezuidenhout, Manisha Caleb, LN Driessen, F Jankowski, M Malenta, V Morello, S Sanidas, BW Stappers, MP Surnis, et al. First discoveries and localizations of fast radio bursts with meertrap: real-time, commensal meerkat survey. *Monthly Notices of the Royal Astronomical Society*, 514(2):1961–1974, 2022.

Scott Mitchell Ransom. *New search techniques for binary pulsars*. Harvard University, 2001.

Vikram Ravi. The prevalence of repeating fast radio bursts. *Nature Astronomy*, 3(10):928–931, 2019.

Vikram Ravi, Morgan Catha, Ge Chen, Liam Connor, Jakob T. Faber, James W. Lamb, Gregg Hallinan, Charlie Harnach, Greg Hellbourg, Rick Hobbs, David Hodge, Mark Hodges, Casey Law, Paul Rasmussen, Kritti Sharma, Myles B. Sherman, Jun Shi, Dana Simard, Reynier Squillace, Sander Weinreb, David P. Woody, Nitika Yadlapalli, Tomas Ahumada, Dillon Dong, Christoffer Fremling, Yuping Huang, Viraj Karambelkar, and Jessie M. Miller. Deep synoptic array science: Discovery of the host galaxy of frb 20220912a. *The Astrophysical Journal Letters*, 949(1):L3, May 2023. ISSN 2041-8213. doi: 10.3847/2041-8213/acc4b6.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Benjamin RB Saliwanchik and Anže Slosar. A self-learning neural network approach for radio frequency interference detection and removal in radio astronomy. *Publications of the Astronomical Society of the Pacific*, 134(1041):114503, 2022.

Michael R Sheldon, Michael J Fillyaw, and W Douglas Thompson. The use and interpretation of the friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiotherapy Research International*, 1(4):221–228, 1996.

MP Snelders, K Nimmo, JWT Hessels, Z Bensellam, LP Zwaan, P Chawla, OS Ould-Boukattine, F Kirsten, JT Faber, and V Gajjar. Detection of ultra-fast radio bursts from frb 20121102a. *Nature Astronomy*, 7(12): 1486–1496, 2023.

L. G. Spitler, P. Scholz, J. W. T. Hessels, S. Bogdanov, A. Brazier, F. Camilo, S. Chatterjee, J. M. Cordes, F. Crawford, J. Deneva, R. D. Ferdman, P. C. C. Freire, V. M. Kaspi, P. Lazarus, R. Lynch, E. C. Madsen, M. A. McLaughlin, C. Patel, S. M. Ransom, A. Seymour, I. H. Stairs, B. W. Stappers, J. van Leeuwen, and W. W. Zhu. A repeating fast radio burst. *Nature*, 531(7593):202–205, March 2016. ISSN 1476-4687. doi: 10.1038/nature17168.

JH Taylor. A sensitive method for detecting dispersed radio emission. *Astronomy and Astrophysics Supplement, Vol. 15, p. 367*, 15:367, 1974.

M Trudu, M Pilia, G Bernardi, A Addis, G Bianchi, A Magro, G Naldi, D Pelliciari, G Pupillo, G Setti, et al. The northern cross fast radio burst project–ii. monitoring of repeating frb 20180916b, 20181030a, 20200120e, and 20201124a. *Monthly Notices of the Royal Astronomical Society*, 513(2):1858–1866, 2022.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Michael P van Haarlem, Michael W Wise, AW Gunst, George Heald, John P McKean, Jason WT Hessels, A Ger de Bruyn, Ronald Nijboer, John Swinbank, Richard Fallows, et al. Lofar: The low-frequency array. *Astronomy & astrophysics*, 556:A2, 2013.

Harry L Van Trees. *Detection, estimation, and modulation theory, part III: radar-sonar signal processing and Gaussian signals in noise*. John Wiley & Sons, 2001.

Bei Zhou, Xiang Li, Tao Wang, Yi-Zhong Fan, and Da-Ming Wei. Fast radio bursts as a cosmic probe? *Phys. Rev. D*, 89(10):107303, 2014. doi: 10.1103/PhysRevD.89.107303.

Donald W Zimmerman and Bruno D Zumbo. Relative power of the wilcoxon test, the friedman test, and repeated-measures anova on ranks. *The Journal of Experimental Education*, 62(1):75–86, 1993.