# Analysis of archival wide-band radio spectrograph data from YAMAGAWA observatory

**A Thesis**

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Amoghavarsha A V


Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

April, 2025

Supervisor: Prof. Divya Oberoi

# Certificate

This is to certify that this dissertation entitled Analysis of archival wide-band radio spectrograph data from YAMAGAWA observatorytowards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Amoghavarsha A Vat Indian Institute of Science Education and Research under the supervision of Prof. Divya Oberoi, Associate Professor, National Centre for Radio Astrophysics - Tata Institue of Fundamental Research , during the academic year 2024-2025.

Prof. Divya Oberoi

Committee:

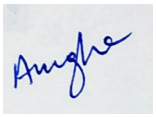
Prof. Divya Oberoi

Prasad Subramanian

This thesis is dedicated to my parents, Prof. Oberoi, and my friends.

# Declaration

I hereby declare that the matter embodied in the report entitled Analysis of archival wide-band radio spectrograph data from YAMAGAWA observatory are the results of the work carried out by me at the National Centre for Radio Astrophysics - Tata Institue of Fundamental Research, Indian Institute of Science Education and Research, Pune, under the supervision of Prof. Divya Oberoi and the same has not been submitted elsewhere for any other degree.



Amoghavarsha A V

# Acknowledgments

I would firstly like to thank my parents who have been supportive throughout my BS-MS journey at IISER. I would also like to thank Prof. Divya Oberoi for providing me an opportunity to contribute to this project, and for his kindness and support throughout its duration. I would like to Guido van Rossum for building an amazing programming language - Python, where most of the algorithms for this study has been written. I would like to thank my friends for their support, encouragement and companionship through the journey.

# Abstract

Solar radio bursts are unambiguous and sensitive tracers of non-thermal electrons in the corona. These non-thermal electrons owe their origin to episodes of solar activity and lead to a wide variety of radio emissions, making observations of solar radio bursts a very useful probe of solar activity. These emissions are usually bright enough to outshine the quiescent solar emission. The well known classes of solar radio bursts differ dramatically in their appearance in the time-frequency plane and carry useful information about solar activity, especially when combined with observations in higher frequency bands (EUV and X-ray). Spectrographs have been the most commonly used tool for observing solar radio bursts and have played a pivotal role in building our current understanding of these phenomena.

With notably few exceptions, these radio spectrographs have been operating in what is traditionally referred to as low radio frequency bands ($<\sim 500\,\mathrm{MHz}$), and the nature of these burst emissions in higher parts of the radio band are yet to be studied in similar detail. This project focuses on studying the archival data from a solar radio observatory operated by the National Institute of Information and Communications Technology (NICT), Japan, namely the YAMAGAWA radio spectrograph, which has been operating since 2016 to the present times covering the band from $70 - 9000\,\mathrm{MHz}$. The instrument cover bands well beyond the traditional low frequency bands, observe routinely from sunrise to sunset with comparatively few data gaps, provide good quality data including L and R polarizations and span a period of 10 years.

Algorithms for the detection and the classification of bursts was implemented on the data provided by the instrument. The data required extensive pre-processing before the implementation of detection algorithms. The study also explore feature detection and extraction algorithms using contours and other algorithms for a robust detection of solar signal. Further more two machine learning models were implemented on the data to detect bursts - Random Forest Classifier ann YOLO. The insights gained from this study make future investigations more practical and impactful survey of solar radio bursts in the YAMAGAWA data; the algorithms heuristics used in this study are transferrable to other telescopes.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Radio emission from Sun

### 1.1.1 Radio Window

The electromagnetic waves coming from the cosmic sources have to pass through the atmosphere, which is not transparent to all the frequencies equally; most of the radiation is absorbed or scattered by gases and other particles. Atmospheric window is defined as the range of wavelengths of electromagnetic radiation that can penetrate the Earth's atmosphere. The Earth's atmosphere absorbs radiation at Infrared, Ultraviolet and higher frequencies. The windows that are available for observation through ground based telescopes are Visible-light, and Radio windows [7]. An important atmospheric window is the Radio window which enables the observation of radio waves through ground-based telescopes.

The visible-light window is narrow and the range is defined by blackbodies emitting at $T \sim 6000\,\mathrm{K}$ to $T \sim 10000\,\mathrm{K}$. Usually the observable features in the visible-light wavelengths are the hot thermal sources. Extrapolating the blackbody spectrum, considering that stars are nearly perfect blackbodies, it was found that stars would be faint radio sources; it was also assumed that there would be no other celestial radio sources, but it was refuted when a bright radio emission was observed from the direction of the center of the Milkyway. The physical processes that are responsible for the extents of the radio window and due to the vibrational and rotational transitions

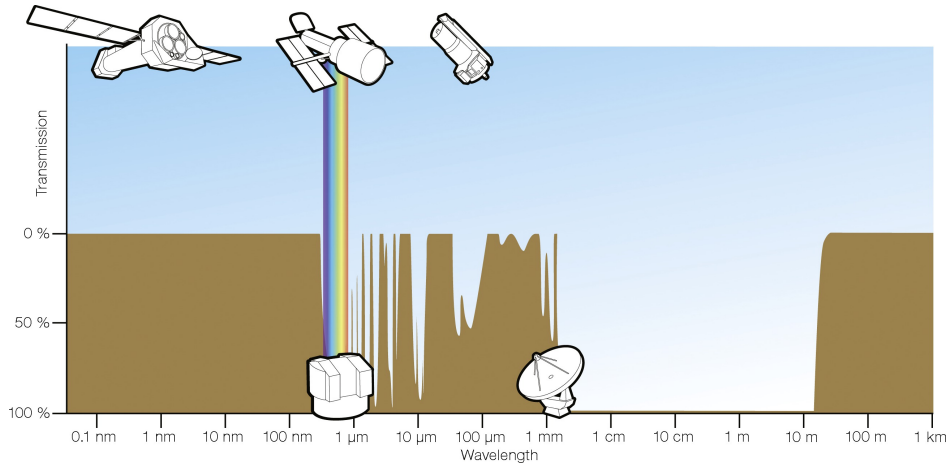



Figure 1.1: Schematic showing the opacity or the transmission percentages through the Earth's atmosphere for different wavelengths; the two broad windows are in the visible and the radio spectrum [7].

of molecules in the atmosphere. $CO_2$, $O_2$ and $H_2O$ have their vibrational energies comparable to that of the mid-infrared photons; the far-infrared photons and the upper-cutoff for the Radio window at $\nu \sim 1\,\mathrm{THz}$ high-frequency Radio waves have energies similar to the broad transitions in the molecules mentioned earlier. The lower-limit of the frequency cutoff for the Radio window is due to the Ionospheric cutoff for wavelengths beyond 30 m or frequency $< 10\,\mathrm{MHz}$; beyond this limit the electromagnetic waves are reflected back to space, or undergo total internal reflection if the waves are emitted from ground based sources.

### 1.1.2 Quiet Sun

Solar radio emission lead to the establishment of the fact that the undisturbed or *quiet* sun has two components:

i. Steady component - which is constant on scale of months to years

ii. Slowly varying component - changes day to day with a period of 27 days [17].

The Active sun is superimposed over the background quiet sun and the slowly varying component with transient events like flares in the solar atmosphere. These events are called Solar Radio Bursts. The solar background is defined as the emission of solar atmosphere when the

contribution from all discrete source of the slowly varying component have been subtracted, i.e., the background component is the emission which is stable for a period of months or years. The background component is from thermal emission of the solar atmosphere. In the centimeter, the emission originates from corona of temperatures from $T \sim 6000\,\mathrm{K}$ to $T \sim 30000\,\mathrm{K}$; for decimeter wavelength it originates from parts of the chromosphere and partly from the million degree corona. In meter wavelength, the emission originates from the regions in corona with temperatures $T \sim 10^6\,\mathrm{K}$.

Quiet Sun measurements are essential as they provide information about the kinetic temperature of the corona. In conjunction with the optical data, it also provides information about the electron densities, and we know that there exists a relation between the electron density ($n_e$ is in $\mathrm{cm}^{-3}$) and the frequency of plasma oscillations given by

$$\nu_p = 9000\sqrt{n_e}\,\mathrm{Hz} \tag{1.1}$$

Moreover, using different wavelengths in radio regime - centimeter and decimeter, it is possible to obtain temperatures and densities in the solar atmosphere - which is a challenging task as the opacity or the optical depth is small in regions of outer corona.

### 1.1.3 Active Emission

During heightened Solar activity, the sun emits through a variety of mechanisms – coherent and incoherent, which is associated with various solar phenomena. The active solar emission in the radio wavelengths generally consists of the Radio bursts and they are generally associated with solar flares. Bursts originate from all levels of the solar atmosphere between the lower chromosphere (millimeter and centimeter waves) and the outer corona (meter and decimeter waves). Unlike the background emission, the active emission achieves a very high brightness temperatures up to $10^{12}\,\mathrm{K}$ [17]. The emission originates from thermal Bremsstrahlung, Gyromagnetic, and plasma radiation; the first two emission mechanisms are incoherent and the plasma radiation is the coherent mechanism; another coherent mechanism is the Electron Cyclotron Maser Emission (ECME) which can produce intense emissions, one such study is presented by Morosan, D. E. et al. [20].

Bremsstrahlung or the Free-free emission comes from the electrons which are redirected by the Coulomb Field in regions of upper Chromosphere. Free-free scatter gets its name from the state of

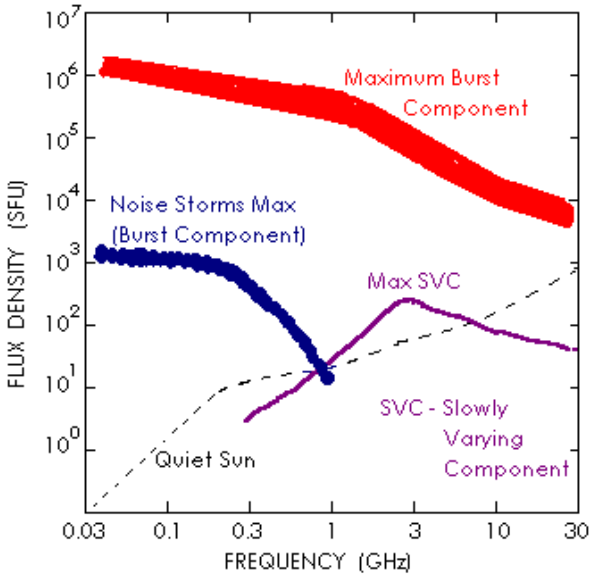



Figure 1.2: The figure shows the solar radio spectrum showing the flux densities for each of the components of emission discussed above [11]. The slowly varying components are the sunspot maxima, and quiet sun component is the sunspot minima. The noise storms and the bursts have their maximum value plotted in terms of flux density [17]. The units of flux given here is in SFU, which in SI units is given by $1\,\mathrm{SFU} = 10^{-22}\,\mathrm{Wm^{-2}Hz^{-1}}$

the electron; they are not bound before or after scattering. In such free-free scattering, the bulk of the emission comes from electrons [21], making only the electron-ion scattering relevant. This is a consequence of

$$\frac{dP}{d\Omega} = \frac{q^2 a^2}{4\pi c^2} \sin^2\theta \tag{1.2}$$

where $dP$ is the power emitted by a particle of charge $q$, mass $m$, and acceleration $a$, inside a solid angle $d\Omega$, $c$ is the speed of light, $\theta$ is the direction of the particle relative to the acceleration vector; considering the total power

$$P = \frac{2q^2 a^2}{3c^2} \tag{1.3}$$

since $a \propto 1/m$, and $a^2 \propto a/m^2$, the proton is much heavier than an electron, and radiation from the proton can be ignored. The interaction between like-charged particles can also be ignored as the radiation power is proportional to the second derivative of the change in the dipole moment, which is zero for the interaction of like-charges. Hence, eq. 1.3 supports the above statement. The emissive power for the free-free emission [22] in a optically thin region is given by

$$\eta_\nu = \frac{2^5 \pi e^6}{3 m_e c^3} \left( \frac{2\pi}{3 m_e k_B T_e} \right)^{1/2} n_e n_i Z^2 G_{ff}(T_e, \nu) \tag{1.4}$$

where the $\eta_\nu$ is the emissive power, $G_{ff}$ is the Gaunt factor [15], and the other symbols have their usual meanings. The bremsstrahlung emission is effective in regions with high electron density and temperature - active regions of the corona and the resulting spectrum is continuous and decreases with increasing frequency, making it a source of dominant radio emission below 300 MHz [9]. This follows from the optical depth ($\tau_\nu$) calculations

$$\tau_\nu \approx 0.2 \frac{\int n_e^2 dl}{\nu^2 T_e^{3/2}} \tag{1.5}$$

and the brightness temperature is given by

$$T_b = \tau_\nu T_{eff} \tag{1.6}$$

$T_{eff}$ is the effective temperature of the source. From these eq. 1.5 and 1.6, we arrive at $T_b \propto \nu^{-2}$ for optically thin corona.

The gyromagnetic emission comprises of the radiation due to the acceleration of electrons in presence of magnetic field; for relativistic electrons, the emission is referred by gyro-synchrotron emission. The gyromagnetic or gyro-resonance occurs when low-energy, non-relativistic electrons are under the influence of magnetic fields. The frequency of emission corresponds to the harmonics of the cyclotron frequency of the electron.

$$\nu_B = \frac{eB}{2\pi m_e c} \tag{1.7}$$

Gyromagnetic emission is significant in regions where the magnetic field strength is high, leading to observable radiation at harmonics of $\nu_B$.

Plasma emission is a resonant process where the electrostatic Langmuir waves are converted to propagating transverse electromagnetic waves. The frequency of the Langmuir waves $\nu_p$ is the given by the electron plasma frequency $\nu_p \approx 9000\sqrt{n_e}$, where $n_e$ is the ambient electron density; the converted electromagnetic waves are observed at $\nu_p$ or its harmonic $2\nu_p$ [27]. Plasma emissions achieve extremely high brightness temperatures $\geq 10^12\,\mathrm{K}$ [19], upholding the coherent nature of the emission. The plasma emission mechanism gives rise to fine and bright structures in the dynamic spectra. This emission mechanism is responsible for the conversion of energy in the *cold* or the non-thermal electrons to observable radio waves.

The bursts are classified based on their spectral and temporal characteristics; though this classification is slightly ambiguous, it provides valuable information about solar activity; the classification of the bursts and the types of bursts are described in section 1.1.4.

### 1.1.4 Frequency-time plane characteristics of Active emission

Bursts are classified into types based on their spectral and temporal characteristics as follows:

1. Centimeter wavelength:
   These bursts are the simplest types of bursts; usually appear as a rapid rise in intensity followed by a slow decline. The burst radiation is circularly polarized and is mostly a smooth continuum of radiative flux over the temporal course [10]. The classification of the cen-

timeter wavelength bursts have been ambiguous due to their lack of sharp peaks [10], but an attempt to classify them morphologically have been made and were classified into three distinct classes [17]: (i) impulsive bursts – lasting few minutes, represented the rapid rise and fall in intensity; (ii) post-burst – lasting up to ten minutes, with rapid increase and a slow decay phase; (iii) bursts which showed gradual rise and fall in intensity, often lasting up to a tens of minutes.

2. Decimeter wavelength:
   Unlike the bursts in the centimeter wavelength, the decimeter bursts are complex and show a variety of fluctuations super imposed on the continuum. Decimeter bursts are also associated with hard X-ray flares [9] [2]. The fluctuation are mostly in the form of fast drift in the bursts ($\sim 100\,\mathrm{MHz/s}$). Additionally, there are other transient events that can be superimposed on decimeter bursts.

3. Meter and Decameter wavelength:
   These bursts are well studied and morphologically classified into 5 types - types I to V [28], named on the chronological order of discovery. The classification is made on the basis their spectral characters.

Description of Meter and Decameter wavelength bursts:

1. Type I:
   Type I bursts are non-flare related components which consists of two components - continuum and burst. The continuum is typically referred to as noise storms and spans the ranges between $100-400\,\mathrm{MHz}$ and has variations in the order of hours. The long burst durations suggest the trapping of electrons in closed coronal magnetic lines. The burst component is usually narrow-band - channels of $10-20\,\mathrm{MHz}$ and fast - lasts a few seconds. Type I bursts are harder to study as there are no diagnostics available in other wavelengths. The bursts continue for a timescale of a few hours, which makes it intriguing to study as there are no energy release event associated with it in other wavelengths.

2. Type II:
   Type II bursts occur around the time when a peak in soft X-rays during a flare is observed. They are readily identified by slower drift rates and the presence of both the fundamental and the second-harmonic frequencies which creates two lanes in the dynamic spectra. Type II bursts are assumed to be markers of shocks in the corona - the drift rate can be converted

to velocity if the electron density $n_e$ at a particular height is known; the burst velocity is typically around $10^6$ m/s; the Alfvén speeds in the corona are smaller than the burst velocity, this provides evidence for shocks in the corona. Type II bursts are low-frequency phenomena,

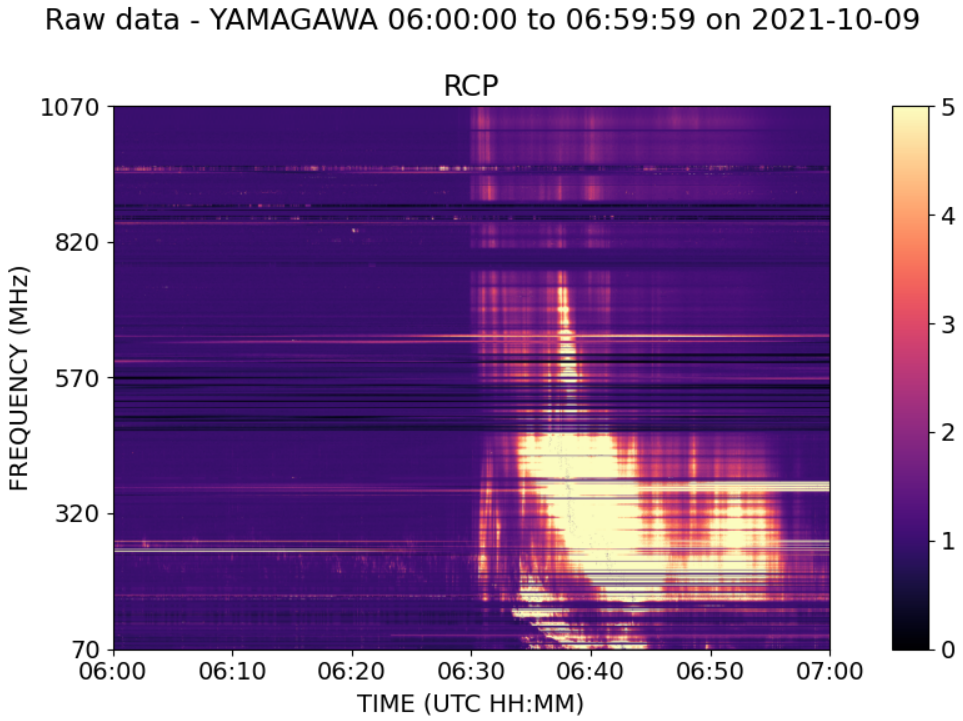



Figure 1.3: Type II burst observed by YAMAGAWA on 2021-10-09 at 06:34 to 06:58 UTC. The two lanes which show the fundamental and the harmonic frequencies can be observed in this dynamic spectrum clearly. There is also a group of type III burst preceding the type II burst around 06:30 UTC.

where most of the recorded bursts occur below 100 MHz

3. Type III:
   These bursts are characterized by their fast drift rates; as the emission is at the plasma frequency and its harmonic, the drift in frequency with time can be directly converted in terms of coronal density going from high to low density region; using the coronal density models, we can infer the velocities. The general results for the generator velocities of the types III bursts are in the order of $1/3$th of the speed of light ($\approx 0.3c$); the only possible generators of the bursts are beams of electrons with energies up to $\sim 10$ keV; such electrons are known to produce Langmuir waves [27]. Isolated bursts are found in the impulsive phase of flares; they imply connections between the accelerating regions and the and open field lines reach-

ing the solar wind make them an important source in understanding field line connectivity in flares.

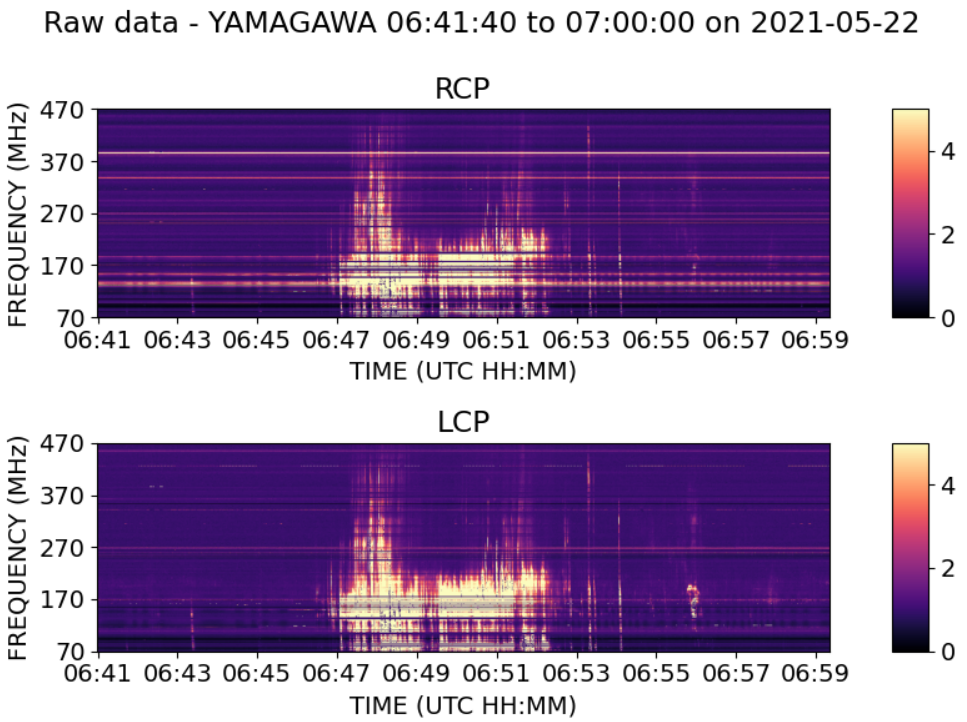



Figure 1.4: Type III bursts observed by YAMAGAWA on 2021-05-22 between 06:47 to 06:54 UTC. A study from Dulk, Suzuki, and Sheridan [8] suggests that the degree polarization of the type III burst is generally less than 0.5; difference in the intensities from both the plots can be noticed at closer inspection at the lower spectral channels between $70 - 170\,\text{MHz}$.

4. Type IV:

   These bursts are broad-band quasi-continuum features that are associated with the decay phase of a solar flare. They were first found as bursts whose radio sources were moving outwards (moving outwards in dynamic spectra with speeds similar to CMEs) from the sun [4], but later it was also found that there can exist stationary type IV bursts (bursts which are not associated with a type II burst, and are stationary in the corona). Initially, the radio sources were theorized to have been emitting through synchrotron or gyro-synchrotron emission and were trapped inside CME loops; but with the discovery of stationary type IV bursts, it was also shown plasma emission was also responsible [1]. It was shown that IV bursts are associated with large flares and those with longer durations, ranging from 1-2 hrs [5]; moreover, type IVs are also associated with type II bursts [6] they appear during the decay of a type

II burst; they are broad band and have a substructure - broad band modulations resembling fast-drift bursts. Type IVs are always associated with large scale flares.

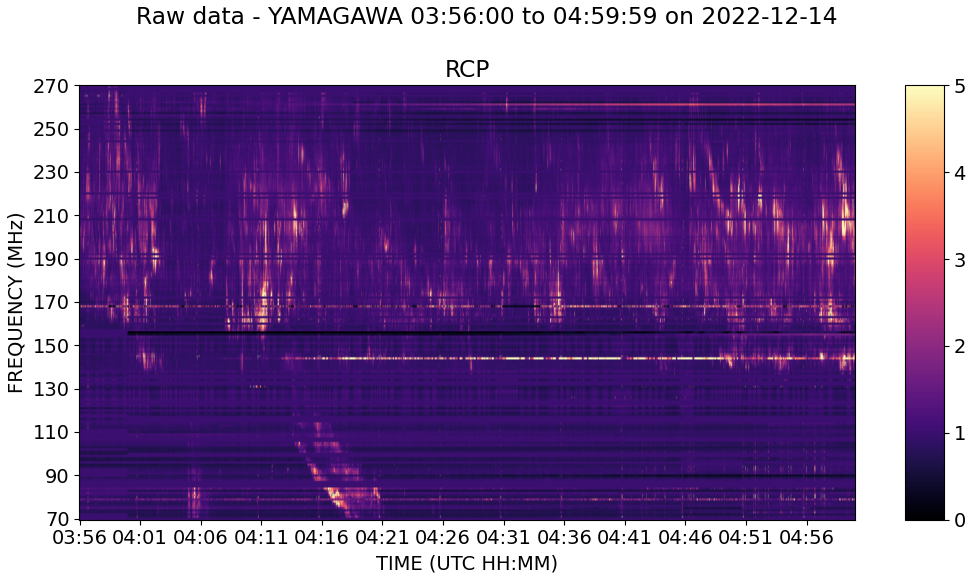



Figure 1.5: A type IV burst was identified starting from 04:04 following a type II burst at 03:58. The bursts were identified with SWPC dataset.

5. Type V:
   These are similar to continuum events like type IVs but they only last a few minutes and are limited to meter-waves [17], and often follow type III bursts or burst groups. They are also observed in lower wavelengths and typically weakly polarized. Type V are generated when the fast electrons from regions of the burst are ejected and pass through the middle corona where they interact with the plasma and setup plasma oscillation which generate the type III bursts; these electrons when captured in a magnetic loop in sufficiently high corona lead to the generation of these bursts [30].

## 1.2 Instruments used for Solar Radio observations

### 1.2.1 Interferometers

Radio Interferometry is a technique used by astronomers to study celestial objects by analyzing the interference patterns observed by combining signals from multiple radio telescopes. The need for interferometry arises from the fact that the angular resolution of a microscope or a telescope is

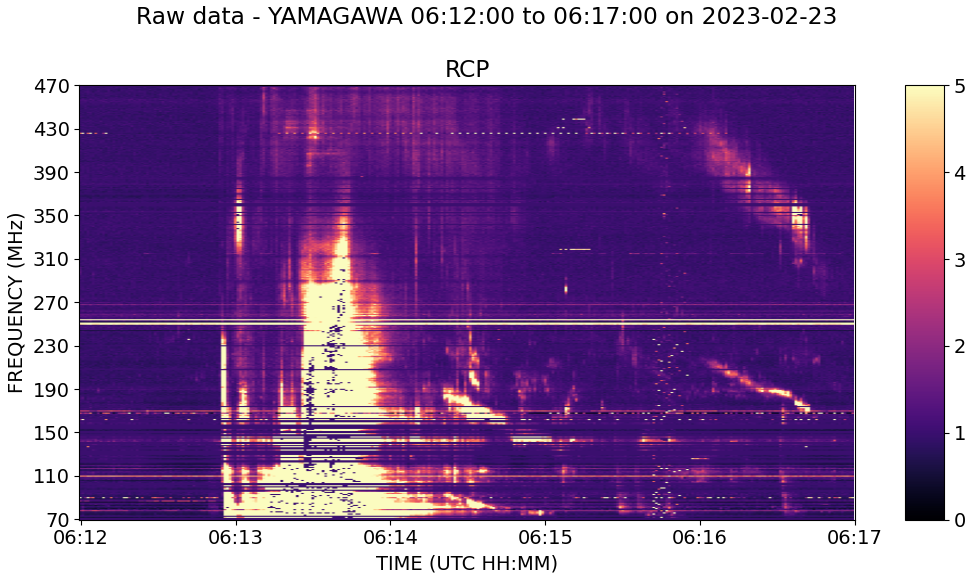



Figure 1.6: Type V burst identified between 06:13 and 06:17 UTC. The burst was follows a strong type III visible before 06:13; the burst characterized by their fast drift rates.

diffraction limited, given by

$$\theta \approx \frac{\lambda}{D} \tag{1.8}$$

where $D$ is the aperture diameter and $\lambda$ is the wavelength of light; to increase the angular resolution of radio telescope, either the aperture of the telescope has to be increased - which is impractical as the diameter of such a telescope would be in the order of thousands of kilometers, or to use other techniques to increase the angular resolution - aperture synthesis. This technique synthesizes a virtual telescope of the aperture equal to the maximum separation between the telescopes.

The Van Cittert-Zernike theorem states that the spatial coherence function $V(r_1, r_2) = \langle E(r_1)E^*(r_2)\rangle$ is related to the incoming radiation, and if all the measurements are in plane then the spatial corelation function only depends on $r_1 - r_2$.

$$V(r_1, r_2) = \mathscr{F}\{I(s)\} \tag{1.9}$$

where $\mathscr{F}$ represent the Fourier Transform operator and $I(s)$ is the intensity of the source. The problem of measuring the source intensity can be translated into measuring a quality called the

visibility ($\mathscr{V}$) given by

$$\mathscr{V}(u,v,w) = e^{-i2\pi w} \int I(l,m) e^{-i2\pi[lu+mv]} dl\ dm \tag{1.10}$$

where $(l,m,n)$ are the directional cosines considering the source to lie on a celestial sphere of radius $R$, and the source is at $P_1'(x_1', y_2', z_3')$ and the observer is at $P_1(x_1, y_2, z_3)$, and the direction cosines are $x_1' = R\cos(\theta_x) = Rl$, $y_1' = R\cos(\theta_y) = Rm$, $z_1' = R\cos(\theta_z) = Rn$. Distances in the observing plane area measured in the units of wavelength ($\lambda$) and the *baseline-coordinates* $(u,v,w)$ is defined such that

$$u = (x_2 - x_1)/\lambda \qquad v = (y_2 - y_1)/\lambda \qquad w = (z_2 - z_1)/\lambda \tag{1.11}$$

From the above equation, the spatial corelation of the electric field in the U-V plane is related to the brightness distribution of the source. Corelation of the voltages from any two radio antennas gives the measurement of a Fourier component; the visibility and the source brightness is related by the Fourier Transform. With sufficient measurements, the source brightness distribution can be obtained through the inverse Fourier Transform. When the source is tracked from rise to set, the sampling in the U-V plane is dense to allow precise reconstruction of the source brightness; the technique to use the rotation of the Earth to increase sampling in the U-V plane is called *aperture synthesis*.

The main advantage of using such a method is the high angular resolution of the telescope compared to single dish telescopes; we know that the angular resolution is given by eq. 1.8, and as the aperture increases $\theta$ decreases giving resolutions comparable to optical telescopes [24]. The sensitivity of the observing instrument increases with the collecting area of the instrument. Radio interferometers also have a high field of view, unlike optical telescopes and these instruments also allow us to study very faint source with high precision.

### 1.2.2 Spectrographs

Spectrographs are an important tool for spectroscopy which enable the study of cosmic emission. Spectrograph split the incoming light from the instruments into its constituent wavelengths for analysis. In the radio regime, the spectrographs reveal various physical and chemical properties

of the source by resolving the spectral lines. The lines also provide information about the composition, temperature, density, and the velocities of the emitting particles [12]; thus helping us understand the properties of phenomena such as Solar Radio Bursts, Pulsar emissions, etc.

Spectrographs are divided into two types – Analogue and Digital spectrographs. The two types are described below.

### 1.2.3 Scanning Spectrographs

Scanning spectrograph, also referred to as dispersive spectrographs, as they use optical dispersive methods to separate the components of incoming electromagnetic signal. The instruments working in the radio regime, are equipped with a tunable filter to spatially spread the radio signal into its constituent frequencies [16]. A mechanical tunable filter involves a frequency sensitive local oscillator which physically directs the frequency bands through an exit slit, subsequently to a detector, and usually utilize RF Micro Electro Mechanical Systems (MEMS) for such operations. Since the spectrograph is scanning a range in the frequency axis in small chunks, $\delta\nu$, and the time take for this scan, $\delta t$. Hence, the time resolution of the spectrograph now becomes the total time taken to scan all the frequency channels, this is referred to as the *sweep time*, $\Delta t = \sum_{\delta\nu} \delta t$. Spectrographs which were associated with a sweep time were also called Sweep spectrographs.

The design of the scanning spectrographs have limitations; reducing the slit width can increase the frequency resolution of the spectra, but this also reduces the flux of the light entering and can adversely effect the SNR and make the observations of faint source challenging. Moreover, the spectrographs data acquisition time is directly proportional to the sweep time, $\Delta t$. While observing transient sources, and can cause smearing in the temporal axis of the spectrograph.

### 1.2.4 Fourier Transform Spectrographs

Fourier Transform (FT) or Fast Fourier Transform (FFT) spectrographs use digital signal processing instead of mechanical parts to disperse incoming signal. The digitization of the signal happens through a high-speed Analogue to Digital Converter (ADC). A FFT algorithm is implemented is convert the time-domain signals into the frequency domains, producing a power spectrum that represents the intensity of in all the frequency components simultaneously. The several advantages to

this procedure are listed:

- Multiplex:
  The FFT spectrographs collect all the frequency information simultaneously and produce the spectrum in one go, unlike the scanning spectrographs which has an associated scan/sweep time, reducing the data acquisition time greatly. FFT spectrographs are beneficial when observing rapidly varying or transient events [3].
- High throughput:
  Since the FFT spectrographs do not need any tunable filter or slits for the dispersion of signal, the throughput or the flux of light/information is higher, improving the sensitivity and the SNR. It is also beneficial when the signals are fainter and challenging to observe.
- Digital Flexibility and Stability:
  Modern FFT spectrographs are implements using a digital Field Programmable Gate Array (FPGA) or specialize processors for stability and repeatability [14].

Some of the limitations of the FFT spectrographs include the requirement of high computational resources to handle the FFT computations and data storage in real-time. The use of FPGAs and GPUs mitigate the computational issue. Digitization of the signal imposes a Nyquist limit, and higher frequency signals will alias into the lower frequencies, if they are not filtered properly. The choice of the window function (Hanning, Blackman, etc.) also effects the spectral leakage and resolution.

YAMAGAWA uses the OCTAD-S spectrograph, its advantages, limitations and sensitivity calculation have been described in the section 1.3; it also explains the data acquisition and the other workings of the instrument.

### 1.2.5 Spectrographs vs Interferometers

Interferometers are instruments that make high fidelity maps of the sky, where as Spectrographs are instruments which help in studying the composition of the emitting body, and both of these instruments have their pros and cons. Spectrographs can also isolate multiple frequencies without the need for multiple antenna dishes. The operation is either performed physically with diffraction gratings or digitally with methods like Fourier Transforms. These instruments provide a high

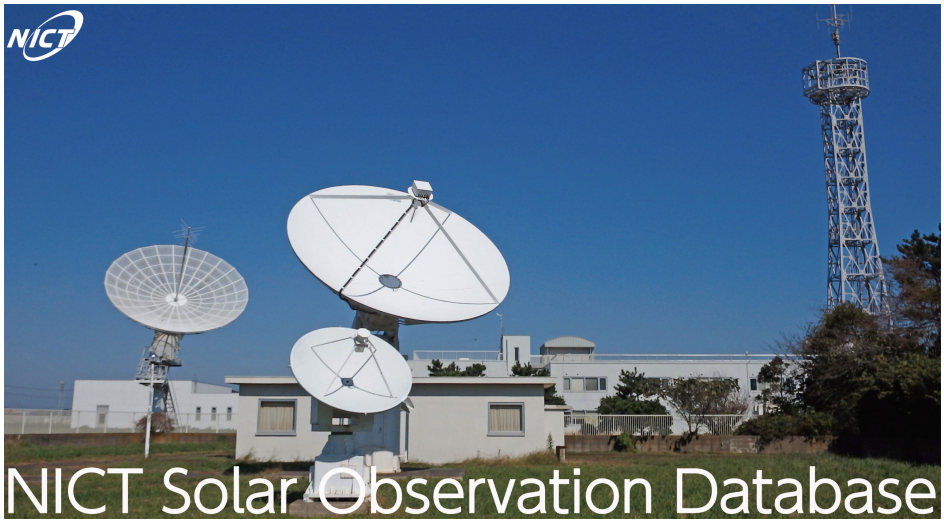



Figure 1.7: The YAMAGAWA observatory, operated by NICT, Japan. Located at $31.204^{o}N$ $130.617^{o}E$. Image taken from Information and Communications Technology website [13].

spectral resolution compared to the interferometers which can only isolate limited number, corresponding to the number of baselines, of Fourier components and have narrower spectral ranges.

Interferometers on the other hand, have high sensitivity and angular resolution compared to Spectrographs. These instruments reconstruct the voltages received from the sky precisely, allowing the study of faint sources with small angular widths by synthesizing apertures using the rotation of the Earth and consist of a array with many receiving antennas, thereby increasing the sensitivity of the images. Spectrographs produce spectra with almost no localization of the event or the emitting source; the spectra in Solar observation are called as Sun-as-a-star measurements which imply that spectrographs cannot resolve the regions of the sun where the flaring event or the burst has taken place. This is the major drawback of the Spectrograph compared to the Interferometer. The Interferometers also come with drawbacks such as the time to precess and analyze the data is computationally expensive, often requiring High Performance Computers. Arrays like the Murchinson Widefield Array (MWA) produce high resolution images of the Sun, but the analysis of one hour of MWA takes about a few days. The heavy computational cost and the limited spectral resolution of Interferometers make Spectrographs favorable for the study of Solar Radio Bursts where large amounts of data has to be processed and analyzed to flag and characterize bursts.

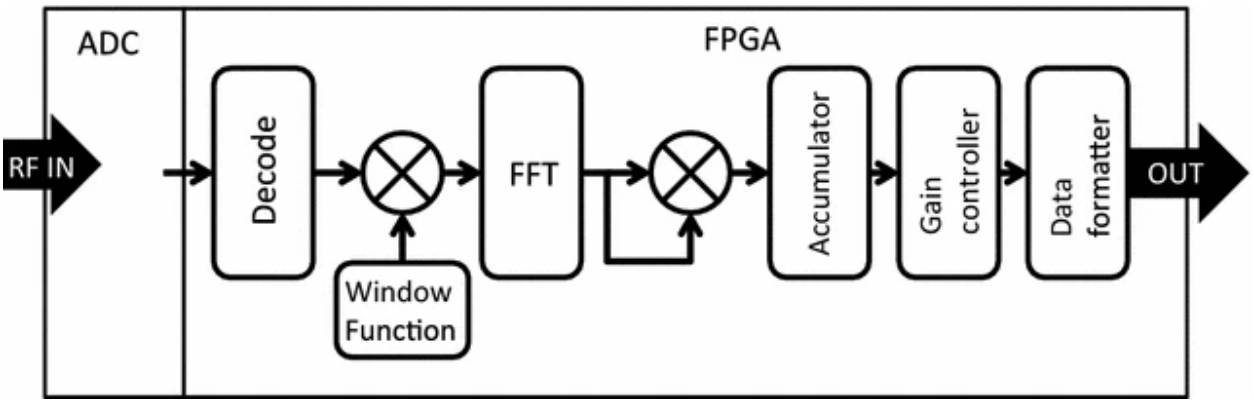



Figure 1.8: A schematic diagram of the data flow from ADC module to the FPGA module in the OCTAD-S spectrometer [14].

## 1.3 YAMAGAWA spectrograph

The YAMAGAWA spectrograph is a digital Fast Fourier Transform spectrograph, as described in section 1.2.4, designed for high resolution solar radio observations. It uses a Field-Programmable Gate Array (FPGA) and high-speed Analogue to Digital Converters (ADCs) for real-time spectral processing with high frequency and time resolutions.

The spectrograph was developed based on the digital instrument called OCTAD-S [14], or the Optically Connected Transmission system for Analogue-Digital conversion - Spectrometer. This instrument is ideal for solar observations due to the instruments high sampling speed and efficiency in signal processing; offering about 8 GHz direct sampling, the instrument completely eliminates the need for a traditional heterodyne receiver mechanism. This section further deals with details about the design, architecture, signal processing, and performance characteristics of the spectrograph.

### 1.3.1 Hardware design

The OCTAD-S digital spectrometer has three modules - Analogue to Digital Converter (ADC) Module, Digital Signal Processing (DSP) module, and the Main control module.

1. Analogue to Digital Converter (ADC) Module:
   This module converts the incoming radio signals to a digital format. The ADC module has a

track-and-hold circuit which accepts input in microwaves and a 10-bit ADC with a sampling rate of $5\,\mathrm{GS/s}$, achieved by overlapping four $1.25\,\mathrm{GS/s}$ ADC cores with 8 effective bits. This circuit is implemented in the front of the ADC and converts the single-ended input signal to a differential signal; a phase shifter is used to correct the phase of the components. The spectrometer now has a direct sampling spectroscopy from $0-8\,\mathrm{GHz}$; the ADC of the OCTAD-S spectrometer does not have a good response in the microwave frequencies.

2. Digital Signal Processing (DSP) Module:
   The DSP module hosts the FPGA chip, which executes Fast Fourier Transforms (FFT) in real-time. Traditionally, a window function is placed in front of the decoded signal before passing it to the FFT module. It has been found that using a Poly Phase Filter Bank is a better alternative than a window-function based methods. Therefore, a poly phase finite impulse response filter is place between the decoded signal and the FFT processing component, in place of the window-function. The information flux through the FFT core is measured by a FPGA clock, and is found to be smaller than the sampling rate of ADC; hence multiple such FFT cores can be carried out in a FPGA and the cores are operated in parallel. To prevent data loss, a buffer memory is used before FFT module to store data. The above implementations enable the instrument to make real-time observations without dead time. The power spectrum received from the FFT cores is stored in the accumulator memory and sent to the next step after the accumulation time, which depends on the number of bits in the accumulator; the number of buts can vary from $40-48$ and the corresponding accumulation time is in the range $0.0008-2\,\mathrm{s}$. The gain controller is used to extract the data from the spectra received the accumulator, and to reduce the data, only a 16-bit fraction is extracted from the 40-bit and the 45-bit spectrum; this is the FFT gain, and can be defined by the user.

3. Main Control Module:
   The instrument is synchronized by an external $10\,\mathrm{MHz}$ clock, and also has an internal clock which is made from a phase-locked oscillator (PLO). Hence, the instrument can function even without the presence of an external clock. The ADC and the DSP modules are independent of the control module. Control module also hosts the interfaces with which the external transfer of data can occur.

### 1.3.2 Spectrometers

The observational needs of Solar radio waves inspired the implementation of spectrometers on OCTAD-S – OCTAD-S 4G4K and OCTAD-S 2G64K. The configurations of the spectrometers are given as:

1. OCTAD-S 4G4K: This configuration of the spectrometer high-speed spectral acquisition with a moderate frequency resolution, which ideal for capturing broad-band solar emission.

   - Sampling speed : $4.096\,\mathrm{GS/s}$
   - Number of Frequency channels : 2048
   - Spectral resolution: $1\,\mathrm{MHz}$
   - Acquisition time: $1\,\mu\mathrm{s}$

   This configuration uses a single FPGA per ADC, thus making the real-time spectral acquisition fast. As the frequency resolution is not very high, they are best suited to observe solar bursts like type IIs and IIIs. This configuration also has a higher noise floor due to its broad channel width.

2. OCTAD-S 2G64K: This configuration is optimized for high-spectral resolution, hence making it ideal for detecting narrow-band and fine solar features.

   - Sampling speed : $2.048\,\mathrm{GS/s}$
   - Number of Frequency channels : 32768
   - Spectral resolution: $31.25\,\mathrm{KHz}$
   - Acquisition time: $32\,\mu\mathrm{s}$

   It uses two FPGAs per ADC, hence increasing the computations resources required to calculate a high-resolution FFT. The spectral resolution offered by this spectrograph is significantly better than 4G4K, hence making it possible to study fine structures in solar emission like radio spikes.

Both the above spectrographs operate without any dead time, and maintain a continuous data collection at an interval of 8 milliseconds.

### 1.3.3 Performance of Spectrograph

The performance metrics for the laboratory testing of spectrometers are described below; these metrics test and analyses the capabilities of the YAMAGAWA spectrograph.

1. Dynamic range and Linearity:

    - The spectrometer has a dynamic range of $\sim 80\,\text{dB}$, and the theorized dynamic range for a 10-bit ADC is 60 dB.
    - The extended range is due to spectral power distribution, where the noise power is distributed among multiple channels.

2. Ghost Free Dynamic Range:

    - Spurious Free Dynamic Range (SFDR) is defined as the ratio between the input signal and the largest spurious, which is a characteristic of the spectrometer.
    - Since the OCTAD-S uses four interleaved spectrometers, the spurious signals are not similar to the typically observed spurious signals in other spectrometers.
    - Spectrometers with interleaving creates ghost signal at the Nyquist Frequency ($f_N$) and $f_{N/2}$. Apart from the ghosts at Nyquist frequency, the interaction between the internal ghosts and the input signal also causes ghost signals ($f_R$); $f_{N/2} \pm f_R$ and $f_N - f_R$.
    - The Ghost Free Dynamic Range is defined as the ratio between the input and the largest ghost signal.
    - The GFDR of the spectrometer is $\sim 37\,\text{dB}$, which is less than the SFDR, this ensures cleaner spectral observations.
    - The GFDR can be reduced by correcting the offset, gain and phase of the interleaving ADCs.

3. Allan variance:

    - The spectrometer has an Allan variance of 1500 seconds, which informs us that the spectrometer can remain stable for that duration, making it ideal for long-duration observations.

4. Frequency Response and Aliasing

- The spectrograph achieves a direct sampling of about 8 GHz without the need for heterodyne down conversion.
- Signals above the Nyquist frequency, cause aliasing where the high frequency signals fold into lower frequencies and distort the spectrum.
- The instrument has a band-pass filter and microwave compatible ADC to reduce the effects of aliasing.

### 1.3.4 Observations

The OCTAD-S has been used in the 8-meter Solar Radio Telescope operated by the National Institute of Information and Communications Technology (NICT). The observations cover a spectral range of $70 - 9000$ MHz, with simultaneous detection of right and left circular polarization. The spectrograph was used to capture solar radio bursts with sub-second time resolution; while in comparison with HiRAS having a time resolution of 500 ms, YAMAGAWA can detect finer details from the bursts.

Traditional spectrometers require heterodyne conversion to shift the radio signals to a appropriate frequency band for processing. YAMAGAWA spectrometers have eliminated this step and can sample a larger frequency range compared to it's predecessor HiRAS.

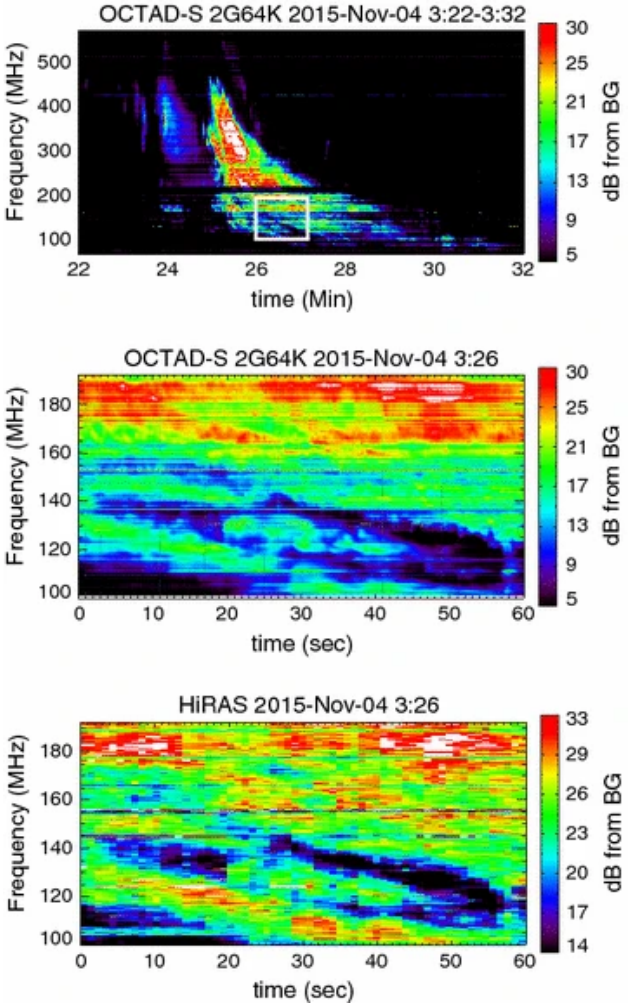



Figure 1.9: Top: Type II burst captured by OCTAD-S 2G64K spectrometer. Middle: An element of the burst zoomed. Bottom: Same burst at the same time captured by HiRAS spectrometer. The color represents the strength of the signal from the background and is given in dB units from background. Figure was taken from Iwai et al.

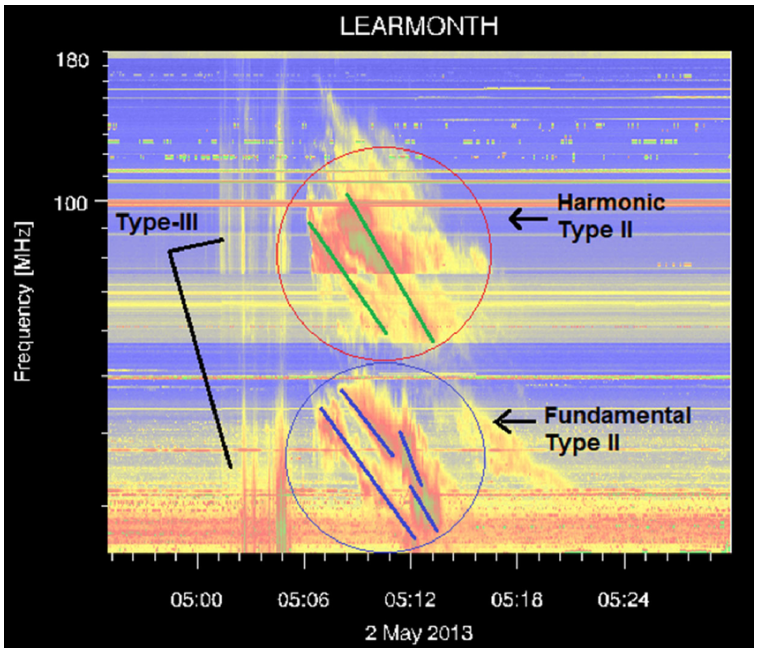



Figure 1.10: Dynamic Spectra taken from the Learmonth observatory showing the presence of type II and III in the spectra. The resolution of Fig. 1.9 much greater than the above dynamic spectra; which is produced by a Scanning Spectrograph. The image is taken from the study conducted by Soni, Ebenezer, and Yadav [23].

# Chapter 2

# Methods and Experiments

## 2.1 Image description and viewing

The images downloaded from the NICT website are provided as a FITS file [26], each containing one hour of observations. The observatory runs routinely from sunrise to sunset and has a relatively clean and continuous data collection. The resolution of the data provided by the spectrographs is 1 MHz, 1 second in the frequency and time axes, respectively; and spans the range from 70 MHz – 9000 MHz. The number of pixels in the dynamic spectra then becomes $8930 \times 3600$. Moreover, the website provides data from the observatory in both right and left circular polarization and their sum in quadrature, the Stokes I vector; the RCP and LCP files are considered to be the raw data, where the quiet sun is not subtracted from the data, but in case of the Stokes' I vector 33% quantile is subtracted from the data, and this is considered to be the quiet sun measurement. This affects the calculation of the quiet sun level in case of a long-duration burst and might suppress the other signals and bursts observed. The three files provided have the same time and frequency resolutions.

## 2.2 Instrumental response

The instrumental response function informs us about the modification of the signal by the measurement system, making it an integral part of measurements in spectroscopy. The response function includes the characteristics of the signal, such as frequency broadening, or scrunching, attenuation,

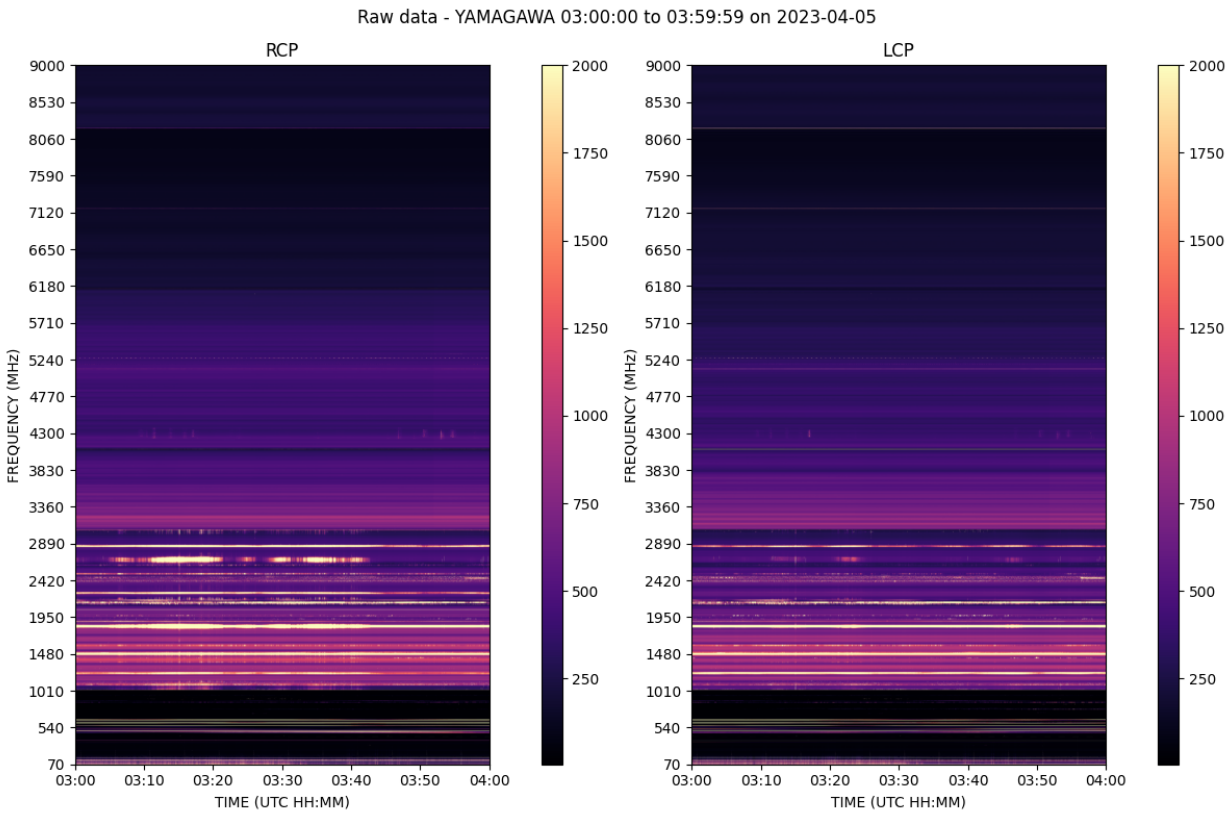



Figure 2.1: Raw right and left circularly polarized images from the YAMAGAWA spectrograph observed on 2023-04-05 between 03:00 hrs to 04:00 hrs UTC. The dynamic range provided by this colour map is insufficient to encompass the range of values from a minimum of 5 to a maximum of 65535 units. The flux measured is in arbitrary units and does not correspond to any particular unit system.

and the resolution of the instrument. The response function can lead to strong misinterpretation of the data if not accounted for during the calculations and incorrect results. Here, the response function can also be called the band shape. The band shape is a function of the frequency and time. The function is assumed to be constant for a specific observing frequency channel, as we know from the section 1.3 that the frequency resolution of the instrument is finer than the frequency presented in the final data. The function is also assumed to be constant in time for the duration of observation during the day, but in reality, the function varies slowly with time. Hence, the band shape must be calculated daily to normalize data. In the case of a spectrograph and similar experimental settings, the raw signal is a multiplication between the true signal and the band shape. The signal without the instrumental gains can be obtained by the division of the gains in their respective channels, and the characterization of the band shape determines the profile of the gains in each spectral channel.

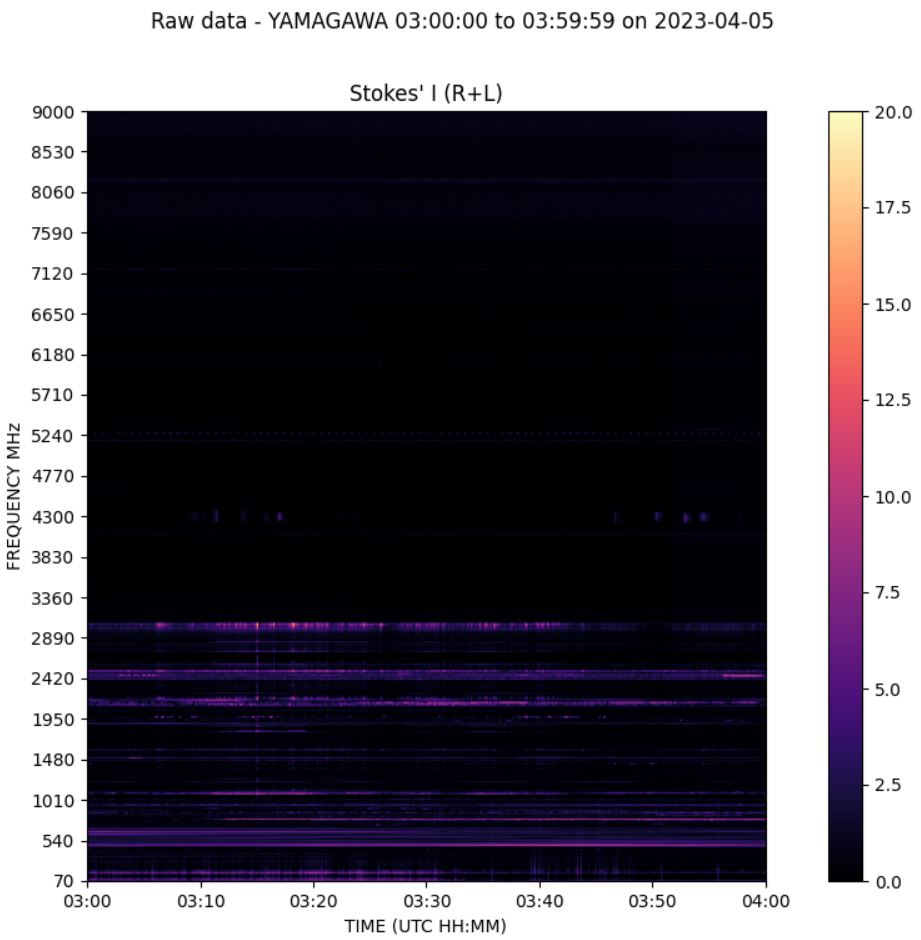



Figure 2.2: Raw R+L (Stokes I) image from YAMAGAWA spectrograph was observed on 2023-04-05 between 03:00 and 04:00 hrs UTC. The image here is saturated at 20 dB from the background or quiet Sun, and the minimum and maximum values in this spectra are 0 and 41.31 dB, respectively. The dB units are derived after the normalization and pre-processing done by YAMAGAWA.

From section 1.1.3, it is evident that transient solar signals are observed as disturbances or deviations from the quiet or the "undisturbed" sun. The median of the signal for the frequency channel is the reference source for the calibration of the signal. The median presents the center of the data and is resistant to the outliers present in the data. In cases of contaminated frequency channels, where the values of dynamic spectra are saturated, the median does not represent the quiet sun; these values can be normalized to unity. A transient signal that lies above the quiet background radiation is better thought of as an outlier in the data. On the other hand, the mean is more likely to be skewed by the outliers in the data, which might lead to misinterpretation of the data as the values of the intensities recorded by the instrument during the period of solar signals are

at least of an order of magnitude above the background. Hence, to capture the correct reference, we use the median of the frequency channel throughout the day to normalize the data against the instrumental response.

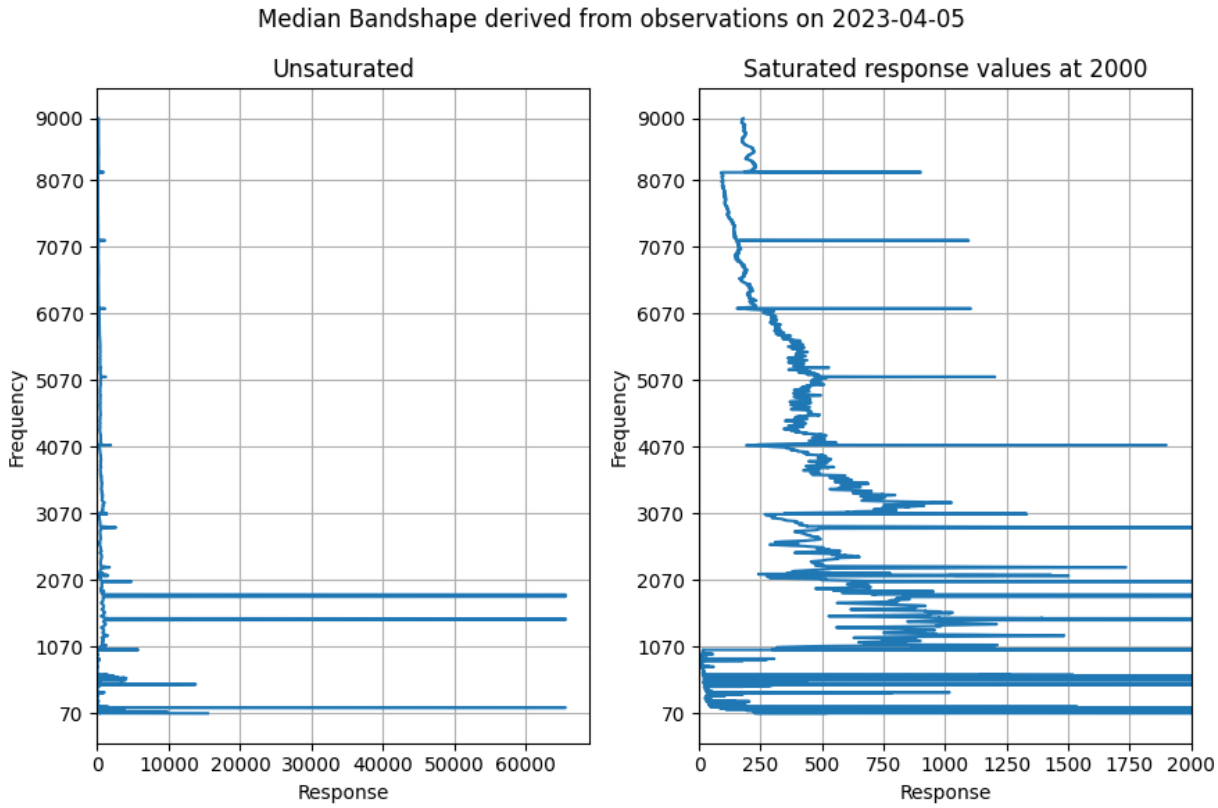



Figure 2.3: The Bandshape here is computed by taking the median of all the frequency channels from sunrise to sunset. The value of the median here is labeled as Response and is presented in arbitrary flux units. Left: Bandshape without any saturation. Right: Bandshape saturated to 2000 on the Response axis.

As discussed above, the median is obtained from the observations taken on the previous day, keeping in mind that the band shape or the intrinsic response of the instrument need not remain constant over very long periods of time. The previously derived median is used to divide the signal, ensuring that the channel median remains close to unity. This procedure is applied across all channels, effectively normalizing their medians to unity. This approach also corrects for the frequency dependence due to instrumental gains, thus simplifying the dataset and making it more amenable for downstream processes and analysis.

In the left panel of Fig. 2.3, the median of some channels reaches a maximum value of 65535,

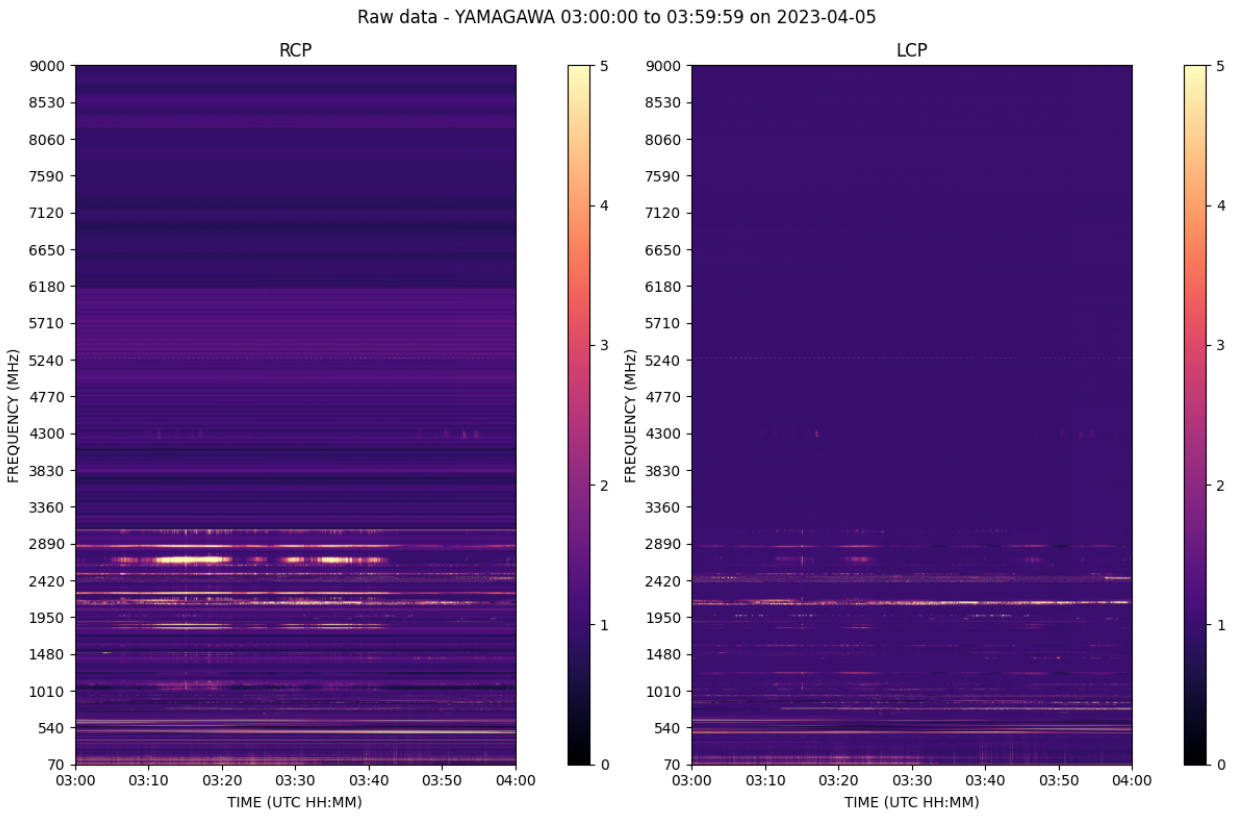



Figure 2.4: Dynamic Spectra normalized with the bandshape calculated from the previous day's observation. Right: RCP component; Left: LCP component.

the highest value allowed by the spectrograph for any observation. The essence of the quantity computed to derive the bandshape informs us that at least half the data points in the channel must be less than or equal to the value of the median, which implies that the entire channel has been affected by interference from a non-solar source. These non-solar sources, which appear as very bright channels in the dynamic spectra, can be called Radio Frequency Interference.

## 2.3 Radio Frequency Interference

In optical astronomical observations, light pollution becomes a serious issue; it makes the background sky brighter thereby making it harder or sometimes impossible to detect the emission from faint astronomical sources. Similarly, for radio telescopes, the noise is due to the radio transmitters and ground-based man-made radio signals occupying frequency channels in the observing range of

radio telescopes. Radio signals arriving from the sun and other astronomical objects are extremely weak, usually many orders of magnitude weaker than the signals produced by transmitters. These signals can completely mask the solar radio signals, or interfere with the solar signals and cause misinterpretation of the data.

In the YAMAGAWA dataset, we regularly come across radio frequency interference spreading across the entire band of observation. The RFI present themselves in various forms and morphologies ranging from transient to long-lived and narrow-band to broadband RFI. Cleaning the data from these interferences caused by man-made and ground-based signals is an important step in the interpretation and analysis of bursts and other solar features. The plots below are all saturated to three times their median values for better visualization of the solar signal and RFI in the dynamic spectrum.

Various types of RFI are described below with examples from the dynamic spectra:

1. **Long-lived narrow-band RFI:**
   This type of RFI presents itself as long-lived bright patches and is confined to a few frequency channels, typically occurring in the bandwidth of $5-10\,\mathrm{MHz}$, and can persist for hours. Potential sources of such RFI can be the nearby television or FM stations broadcasting at fixed frequencies. Due to the persistent nature of the transmission these man-made radio signals, the median Intensity is the same as the values of intensity in that channel. This negates the effect of the RFI in the normalized spectra and the reduces the data in the channel to unity after normalization.

   The systematic removal of persistent RFI from the data is implemented as follows:

   - Smoothing the data: The difference of smoothed spectra with the Savitsky-Golay (Sav-Gol) filter and the initial spectra is calculated.

     $$M'_{ij} = Y_{ij} - M_{ij}, \tag{2.1}$$

     where $M'$ is the difference, $M$ is the initial spectra and $Y$ is the matrix of smoothed spectra, an outline of the smoothing process is described in section 3.5.1.

   - Computing Signal to Noise Ratio (SNR): The difference is used to calculate the SNR.

     $$\mathrm{SNR} = 0.6745 \times \frac{M' - \tilde{M}}{median(|M - \tilde{M}|)}, \tag{2.2}$$

where $\tilde{M}$ is the median of the spectra, and the Median Absolute deviation for a univariate set is given by $\text{MAD} = median(|X - \tilde{X}|)$. The SNR is computed for every data point and its deviation from the median.

- Thresholding: A threshold is set and the values below the threshold are masked with zeros, whereas the values above it are assigned the value of unity.
- Flagging *bad* channels: The fraction of zeros and NaN values in the channel is computed. If the number exceeds 50% of the observation time, then the channel is flagged as a *bad* or a *contaminated* channel.
- Grouping flagged channels: After flagging the channels grouped ensuring that no groups contains more than a user-defined (between $5 - 10\,\text{MHz}$) number of contiguous channels.

2. **Periodic and band-limited :**
   Some types of radio interference are periodic in nature, often arising from devices that send calibrating and detecting pulses. The sort of RFI that falls into this category usually appears sprinkled in the lower parts of the spectrum with fixed periodicity. They also appear in the higher frequencies of the spectra in a band-limited fashion - where the bright pixels are confined to a single or a few channels. Unlike the persistent RFI, the periodic RFI appears and disappears with a fixed period, making it highly predictable and distinguishable from solar or other astronomical signals.

   Steps for periodic RFI excision:

   - Calculate the RMS from MAD: The relationship between the Root Mean Square (RMS) and the MAD is given by

   $$\sigma \approx \frac{MAD}{0.6745} \approx 1.483 \times \text{MAD} \tag{2.3}$$

   where $\sigma$ represents the RMS, is used to calculate the RMS of a one-dimensional time-series array, or the array of Intensities for a given channel over time.
   - Find peaks in time-series data: Peaks in time-series data, or one-dimensional array at a particular frequency of the dynamic spectra, can indicate the presence of Radio Frequency Interference (RFI) if they occur periodically. To identify such peaks, a peak-finding algorithm is implemented based on specific parameters. If the periodicity of the peaks is known, the peak-finding algorithm can use the spacing between them as

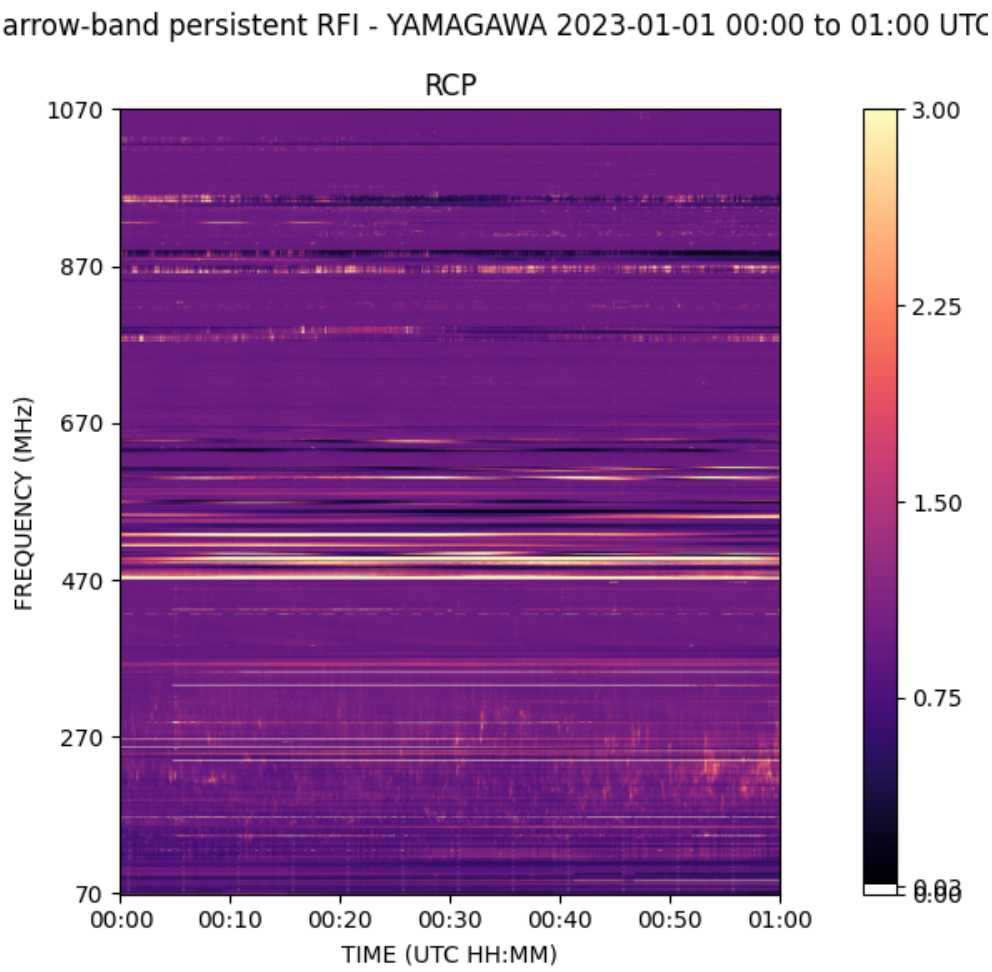



Figure 2.5: Narrow-band persistent RFI can be identified with the long bright lines spanning across the dynamic spectrum in the time-axis. A group of RFI contaminated channels exist in between $470-570\,\mathrm{MHz}$, similar set of channels can be observed around $270\,\mathrm{MHz}$.

a criterion to distinguish RFI from solar or other astronomical signals. Additionally, a threshold for the minimum peak height is set by user to describe the sensitivity of the detection, using the formula $\tilde{X}+k\sigma$, where $\tilde{X}$ is the median of the time-series input, $\sigma$ is the root mean square of the time-series array as defined above, and $k$ is a user-defined constant that determines sensitivity. To further refine the detection, peak prominence is set to $k/100$. As the peak prominence here is small quantity, this ensures that anything that is continually flat is not flagged.

- Masking: Finally, the peaks returned from the above algorithm are flagged as RFI and masked with zeros, and the other values are left unchanged.

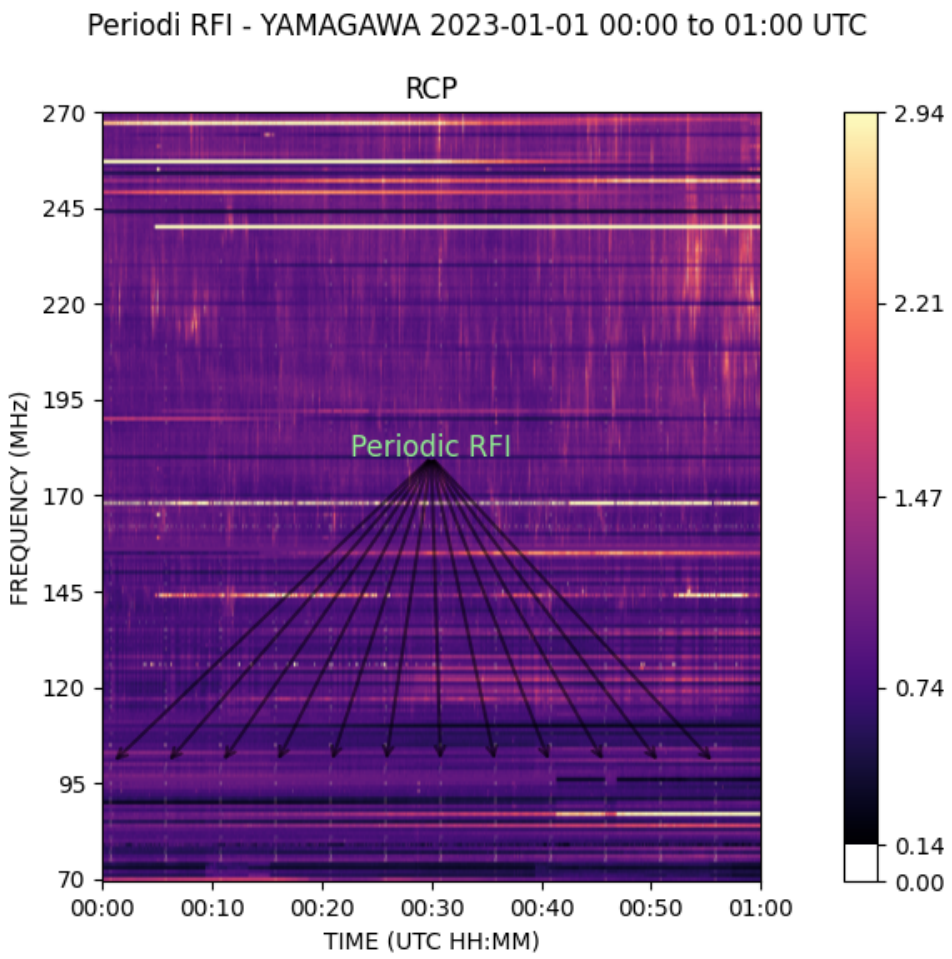



Figure 2.6: Faint vertical structures at intervals of $\sim 5$ mins can be see throughout the lower part of the spectrum ranging from a little over 70 Mhz to 100 MHz. These often appear in

3. **Broad-band transient RFI:**
   The broad-band transient RFI is the other class of RFIs encountered in the dynamic spectra. This class of RFI mimics transient solar signals like type III bursts, as seen in section 1.1.3. The normal operation of some electrical devices, like electrical motors, switches, and relays, might produce low-frequency radio waves capable of contaminating the spectra. The challenges in identifying this type of RFI are due to their broad-band and impulsive nature. They are also prone to being misidentified as legitimate solar signals, as mentioned above.

   Steps for Broad-band RFI excision

   - Preparing the data: The data is scrunched in the frequency axis, i.e., intensities across

all frequency channels is averaged for every time-stamp. The spectra is converted to a one dimensional array of the same length as the observing time in seconds.

- Computing SNR: Frequency scrunched data is then passed to a median filtering algorithm with a user-defined kernel of size. Then, the SNR and RMS are calculated for the difference between the median filtered and unfiltered arrays, from eq 2.2 and eq 2.3 respectively.
- Find peaks: Peaks in the difference array are found with the peak-finding algorithm. The distance between the identified peaks is set to be integer value of the half the length of the difference array. Heights and prominences of the peaks are given as $\tilde{X} + k\sigma$ and $k$ respectively. The higher value of prominence compared to the periodic RFI excision algorithm ensures that only strong and well-separated peaks are flagged.
- Masking: The peaks are masked with a threshold, and all the values in the frequency scrunched data are set to the median of the spectra. Similarly, the time-stamps affected with RFI are replaced with the median value of the spectra.

4. **Median de-noising**

   Median de-noising is a robust technique to suppress noise without altering the signal in a dynamic spectrum. This process, unlike median filtering, is suited to handle impulsive noise patterns and RFI due to its resistance to outliers or extreme values in the data. The algorithm calculates the running median over a defined window. The normalization of the current point is done by dividing the median instead of subtraction to preserve the unit median. The process allows each point to be scaled with the long-term effects of the spectrum. The ratio between the median and the current data point is computed; a threshold value of 2 is selected, where the correction to the current point happens only when the ratio calculated earlier is less than the threshold. This selective normalization preserves the underlying signal while reducing the effect of impulse or transient noise, typically the types seen with RFI.

   The median de-noising algorithm is not included in the pre-processing stage due to its high computational cost, which outweighs the benefits from the improved level of data cleaning it provides.

5. **Pre-processing data:**

   The algorithms for RFI excision discussed above are included in the pre-processing of the data. This step prepares the data for the downstream processes of burst detection and analysis. First, the data is bandshape normalized by division to remove the instrumental response

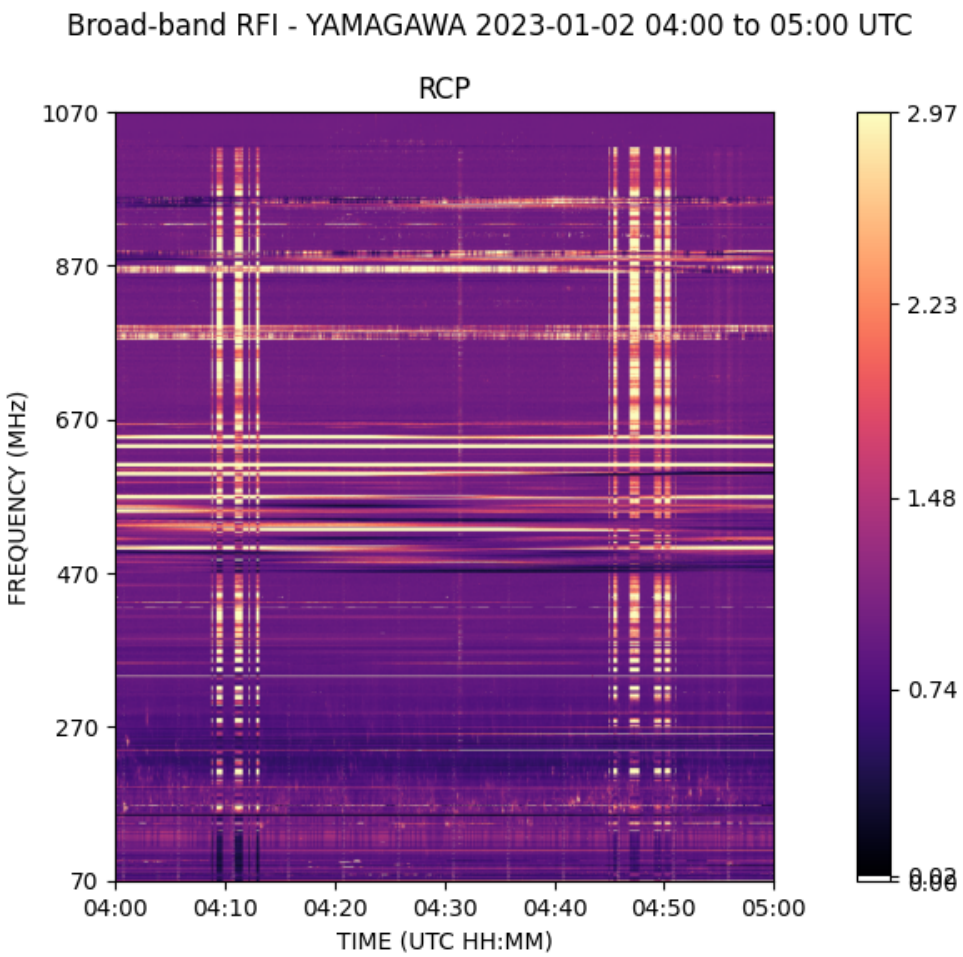



Figure 2.7: Broad-band RFI covering the bandwidth from $70-1070\,\text{MHz}$ and around $04:10$ and $04:50$ UTC. In the lower parts of the RFI can be confused for type III bursts; the characteristic feature of solar signals is the drift, since the RFI does not show any drifting and presents itself as straight lines, this can be easily classified as broad-band transient RFI. Multiple types of RFI can occur concurrently; we can see the presence of narrow-band RFI between $470-670\,\text{MHz}$.

and to bring the median to unity. If the bandshape is not provided then it is computed from provided spectra, which might cause flagging of long-duration emission like type I noise storms as persistent RFI. From section 1, it can be noted that they usually last for hours, making the use of the bandshape computed from the previous day effective for normalization.

The algorithms for RFI excision are implemented on the normalized spectra starting with the periodic RFI removal. In the dynamic spectra from YAMAGAWA spectrograph, we see prominent periodic RFI at the lower frequency channels, at a fixed period of 300 seconds.

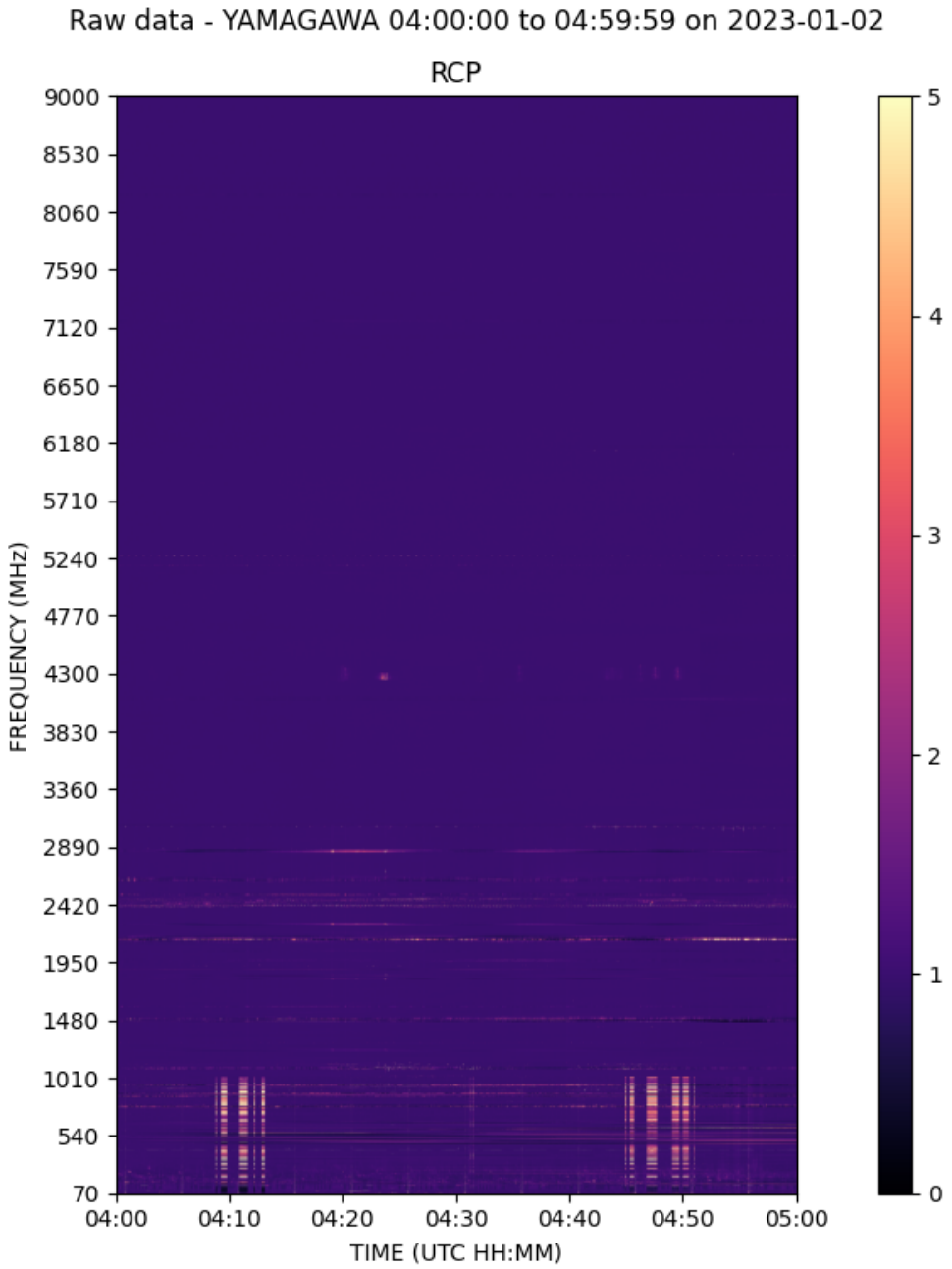



Figure 2.8: De-nosing algorithm is implemented on the normalized spectra with any RFI excision or removal. A more efficient algorithm using sorted arrays to calculate the running median instead of the conventional method described above. Both these methods do not change the required statistics of the spectra and the median is still close the unity, similar to the normalized spectra.

Along with the period, the thresholds for peak height is set at $100, 50, 25$, and these values are used iteratively. The above steps are implemented on the $1^{st}$ discrete difference along the

frequency axis in the original spectra. The discrete difference of sequences is defined by the difference of the consecutive terms of the sequence, in the description above, the $1^{st}$ discrete difference refers to $(x_{i+1} - x_i)$ where $x$ is an element of the time bins and $i = 0, 1, 2, \ldots$ are the indices of the time bins. The indices of the peaks flagged as RFI based on the properties is then used to mask in the unaltered data with its median value.

Next, the narrow-band and broad-band RFI excision algorithms are implemented in order. The spectra from the previous process is passed to the algorithms specific to removing these RFI sequentially. The channel grouping parameter for narrow-band removal is set to 5 channels and the threshold for the SNR mask is set to 10; the kernel size for the Sav-Gol filter is set to 10 seconds. The spectra, now cleaned of both periodic and narrow-band RFI are provided to the algorithm to clean broad-band RFI. The window size for the median filtering is set to 10 seconds and the threshold for the peak prominence is set to 10.

The raw dynamic spectrum has been processed using the above algorithms. Data points that were masked with zeros during processing have now been replaced with the median of the raw data.

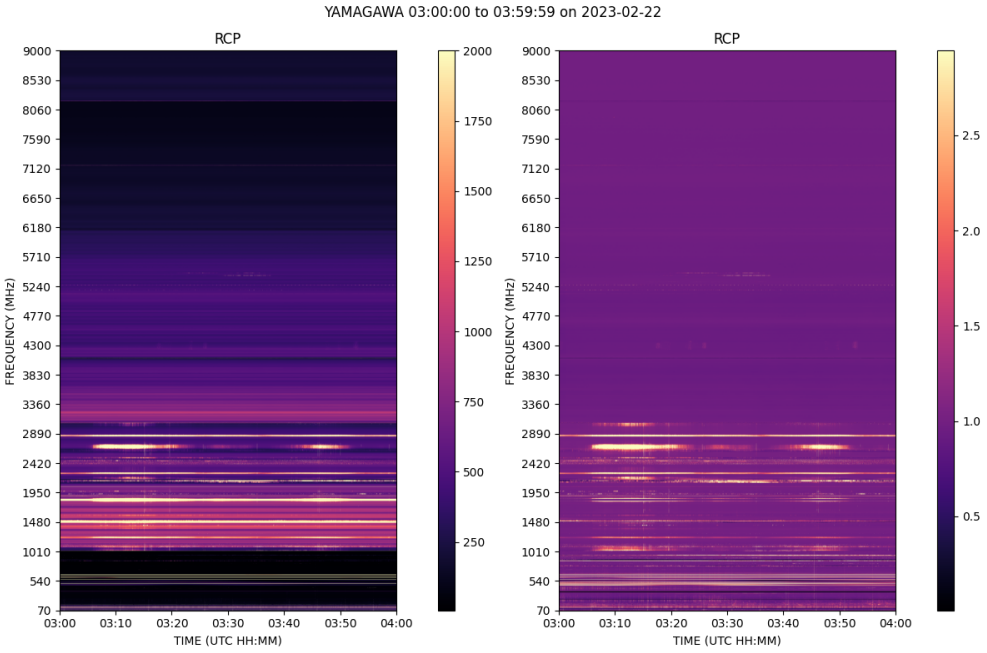



Figure 2.9: Left: RAW data saturated at 2000 units, saturated at 3 dB, which is three times the median value. Right: Dynamic Spectra pre-processed using the above algorithms for RFI mitigation; here the median de-noising algorithms has not been called due to its computational cost.

## 2.4 Feature Detection

### 2.4.1 Edge detection

The first method for the detection of edges is done by computing the first derivative or the gradient of the image. The gradient of the image here describes the change in Intensity at each pixel, this might help us capture the features in the image like edges and boundaries. Computing the gradient will provide us the information the change in Intensity and the direction of the steepest ascent. Finding the gradient map will help us find regions with substantial intensity changes. The solar activity signal presents itself as deviations from the quiet time background, computing the gradients is a promising step in detecting the events and extracting features.

Algorithm for extracting features through edge detection applies a Sobel operator to detect edges. The filter is applied along both the axes. Before applying the Sobel operator, there is also need to smooth the spectrum to reduce the flagging of edges due to the background fluctuations. First the gradients ($G_x$ and $G_y$) are computed along both the axes capturing the edges in their respective directions. Then the magnitude of the gradients is found by computing the Euclidean norm.

$$G_x = \frac{\partial I}{\partial x} \qquad G_y = \frac{\partial I}{\partial y} \qquad G = \sqrt{G_x + G_y} \tag{2.4}$$

The Sobel operator to compute this gradients is given by $3 \times 3$ kernels, $S_x$ and $S_y$ are applied to detect edges in the $x$ and $y$ direction respectively.

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \qquad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{2.5}$$

The kernels convolved with image to approximate the gradient and the final magnitude is found by

$$G = \sqrt{(I * S_x)^2 + (I * S_y)^2} \tag{2.6}$$

where the $*$ represents the convolution operator. Here, the terms image and dynamic spectrum can be used interchangeably, as the axes of the spectra are not relevant in the above computations.

Before, implementing the Sobel operator on the image, the image is first smoothed to remove fluctuations from the quiet sun. The random fluctuations can also appear as prominent edges after the convolution, as the gradient between the two pixels might be comparable to the gradient of a true edge. A gaussian smoothing function is used to smooth out the background emission.

$$g = n_\sigma * f \tag{2.7}$$

In the above equation, $g$ is the the gaussian smoothed function, smoothed with a gaussian ($n_\sigma$) of variance $\sigma$ and function $f$ represents the intensities at a frequency channel $\nu$ through time-axis.

$$\nabla g = \nabla(n_\sigma) * f \tag{2.8}$$

From eq. 2.8, we see that $g$ can be found by convolving the gradient of the gaussian function with the image. The gaussian kernel for smoothing uses a standard deviation of $\sigma^2$; this value is unique for a frequency channel based on the trends of the signal. The value of $\sigma$ can be calculated from the largest Full Width at Half Maxima (FWHM) in the time-series data. The relation between the standard deviation and FWHM [25] is given by

$$FWHM = 2\sqrt{2\ln 2}\sigma \approx 2.355\sigma_{FWHM} \tag{2.9}$$

here, the $\sigma_{FWHM}$ is standard deviation of the gaussian for which the FWHM is calculated. For smoothing the image, standard deviation of the gaussian kernel is chosen such that it is less than half the value of the standard deviation obtained from FWHM. This ensures that the signal is not completely smoothed, and that information in the signal is preserved. The kernel size of the gaussian is calculated based on $\sigma$ and can be given approximately as the $2(\lceil 2\sigma \rceil) + 1$.

The image is finally, convolved with the Sobel operator after being smoothed. This presents with the edges in the dynamic spectrum. Apart from Sobel operator, a more advanced algorithm for edge detection - canny edge detector was also used to detect edges; canny algorithm returns a 2-D array with 1 in locations corresponding to edges and 0 elsewhere. The ones in the array are then grouped based on well-known solar signal to create regions of the burst. The above methods were not used in the pipeline due to the difficulties arising from improper thresholding, high false alarm rates and the difficulties in extracting features.

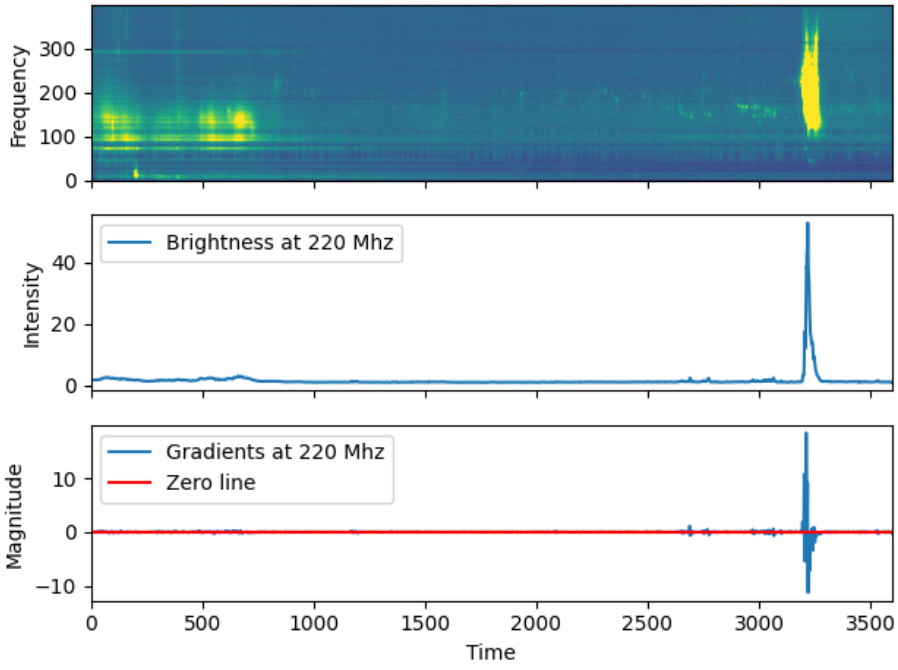



Figure 2.10: Plots show the smoothing operator convolved with one-dimensional array corresponding to observations from various smoothed frequency channels.

### 2.4.2 Blob detection

Blobs are defined as regions with properties which differ largely from the rest of the image. Two methods were used for detecting blobs - Laplacian of Gaussian and Difference of Gaussian. Both the methods were called to identify blobs after thresholding the image at $\tilde{x} + 3\sigma_{\tilde{x}}$, but were unsuccessful in identifying the blobs that are required to extract features from the image. These blobs, mentioned above, were also presented as circles, which differs from the nature of the solar signal leading to incorrect masking.

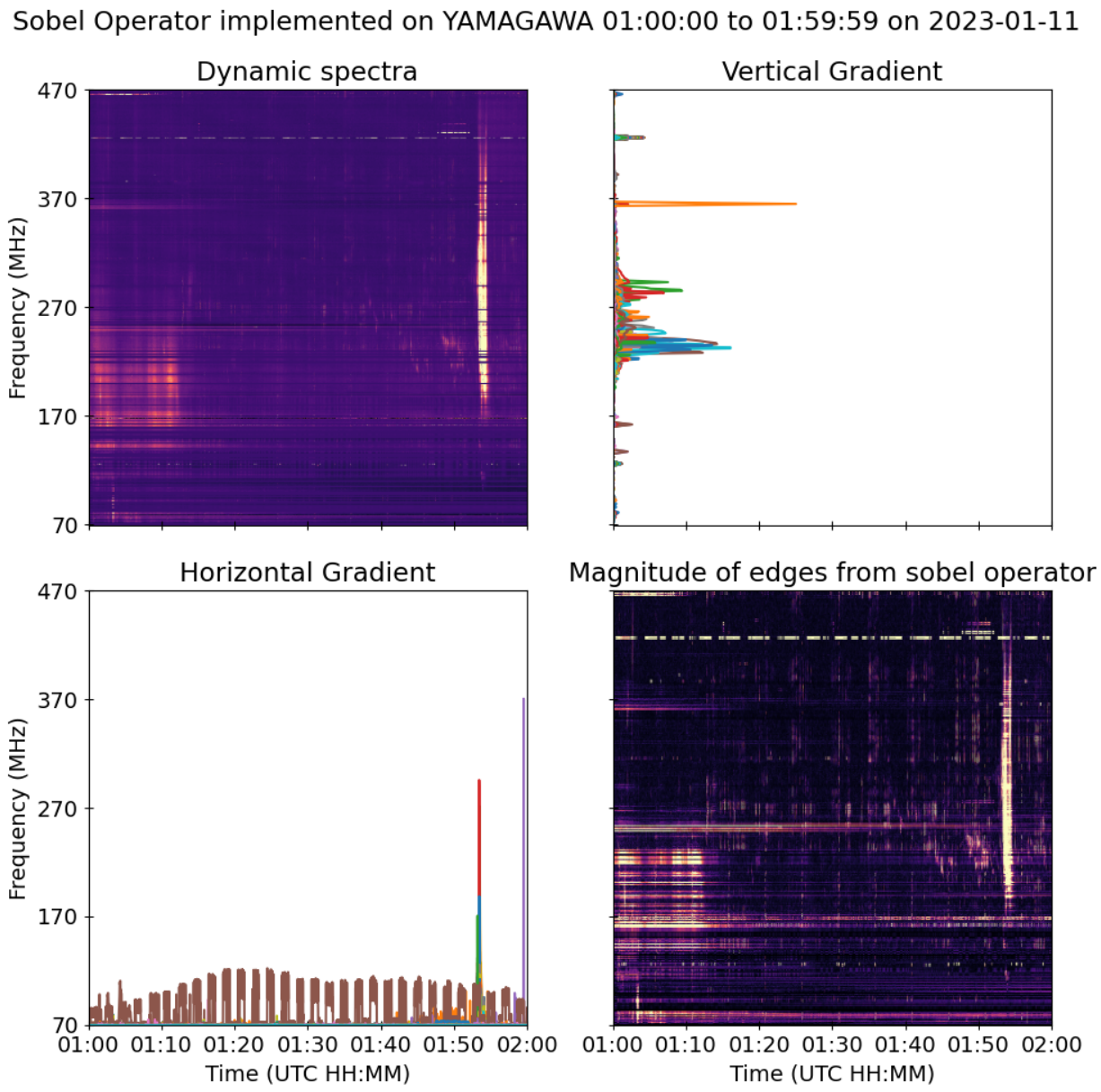



Figure 2.11: Top right and Bottom left panels show the gradient along the frequency and time axis respectively. The Bottom right panel displays gradients sum in quadrature, eq. 2.6, magnitude of the detected edges.

## 2.5 Contour Detection

### 2.5.1 Masking

Masking is an important step in pre-processing for contour detection, as it simplifies the image into a binary format where the background emission and fluctuations are represented by zeros and the foreground signal is represented by ones. As the background emission is completely removed from the spectra, the increase in the contrast helps identify meaningful structures in the dynamic spectrum. Without proper masking, the contours might have excessive and fragmented edges; the contours would also be incorrectly detected due to the small variations in the intensity. Thresholding ensures that only the significant features are outlined, improving the accuracy of the contouring algorithm and other feature extraction processes.

A threshold mask is implemented on the dynamic spectrum which returns the masked array. The threshold for the mask was set at $I_{min} = \tilde{x} + 10\sigma$, where $\sigma$ is the median absolute deviation. In other words, we assert that a pixel on the dynamic spectrum exceeding the threshold is significant deviation from the background and can be considered for feature extraction. Both RFI and solar signals lie well above this limit and get flagged while tracing contours.

Other methods such as the Otsu method have been used for thresholding the spectrum. Though the Otsu method is typically used for bimodal distributions, it can be applied to a unimodal or an exponential distribution of intensities, as observed in this scenario. Otsu method operates by maximizing the variance between the classes. All possible values of the thresholds are iteratively explored dividing the histogram into two parts - above and below the threshold. Next, the inter-class variance for the two groups is calculated, and the threshold value where the variance is maximum is selected.

The median absolute deviation method is selected to generate the threshold and mask the image over the Otsu method, as it better tends to the data in our scenario.

### 2.5.2 Contours

Contours play an important role in demarcating the boundaries between the background and the regions of interest in the spectrum; contours reduce the complexity of relationships between indi-

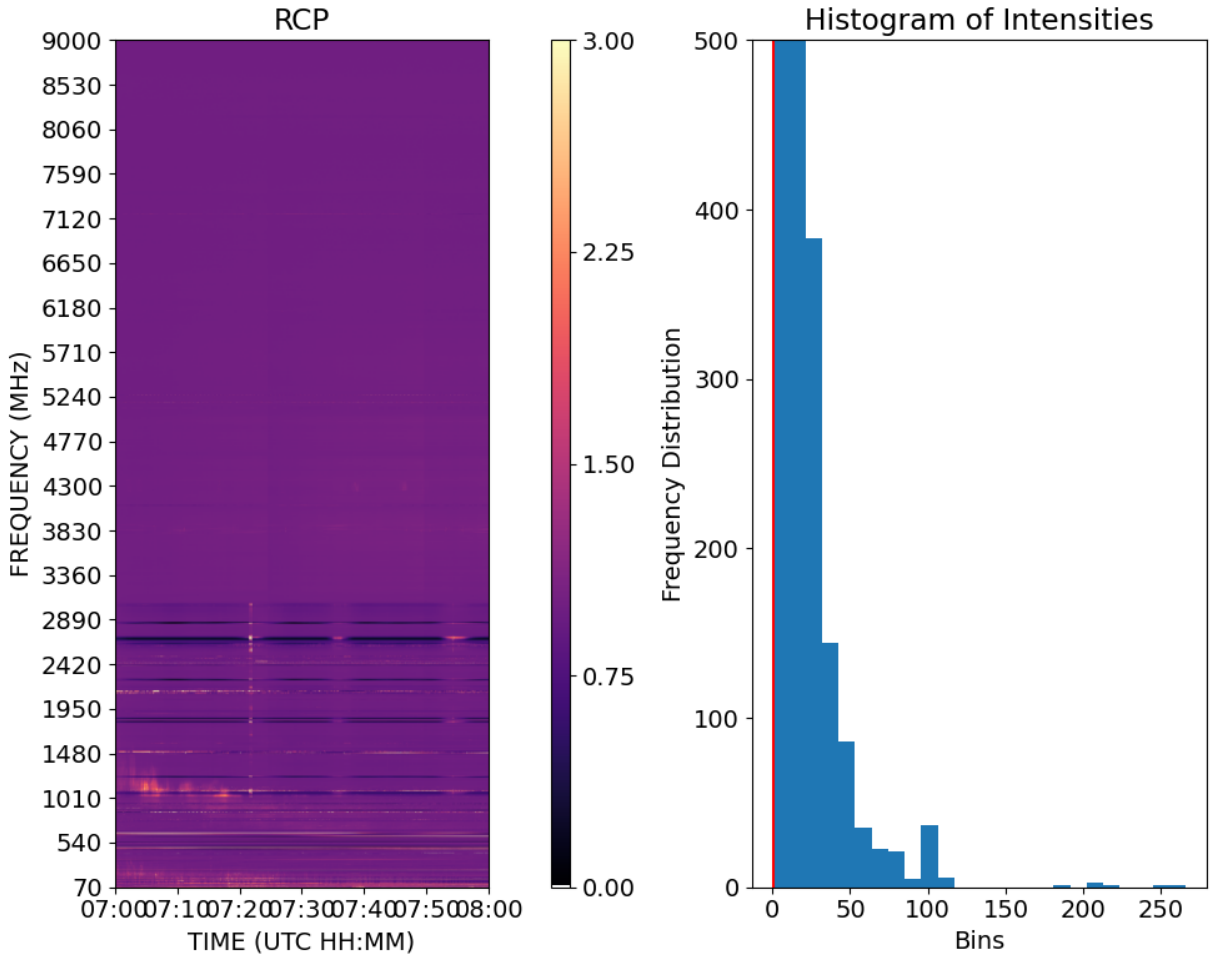



Figure 2.12: Otsu method for thresholding based on the histogram, the histogram is saturated at x. The histogram shows the uni-modal distribution of intensities in the dynamic spectrum. Threshold placed by Otsu method is $\approx 0.52$, which is less than unity, from the statistical analysis of the dynamic spectra the median is found to be unity. Hence, this method is not preferred for thresholding in case of exponential distributions.

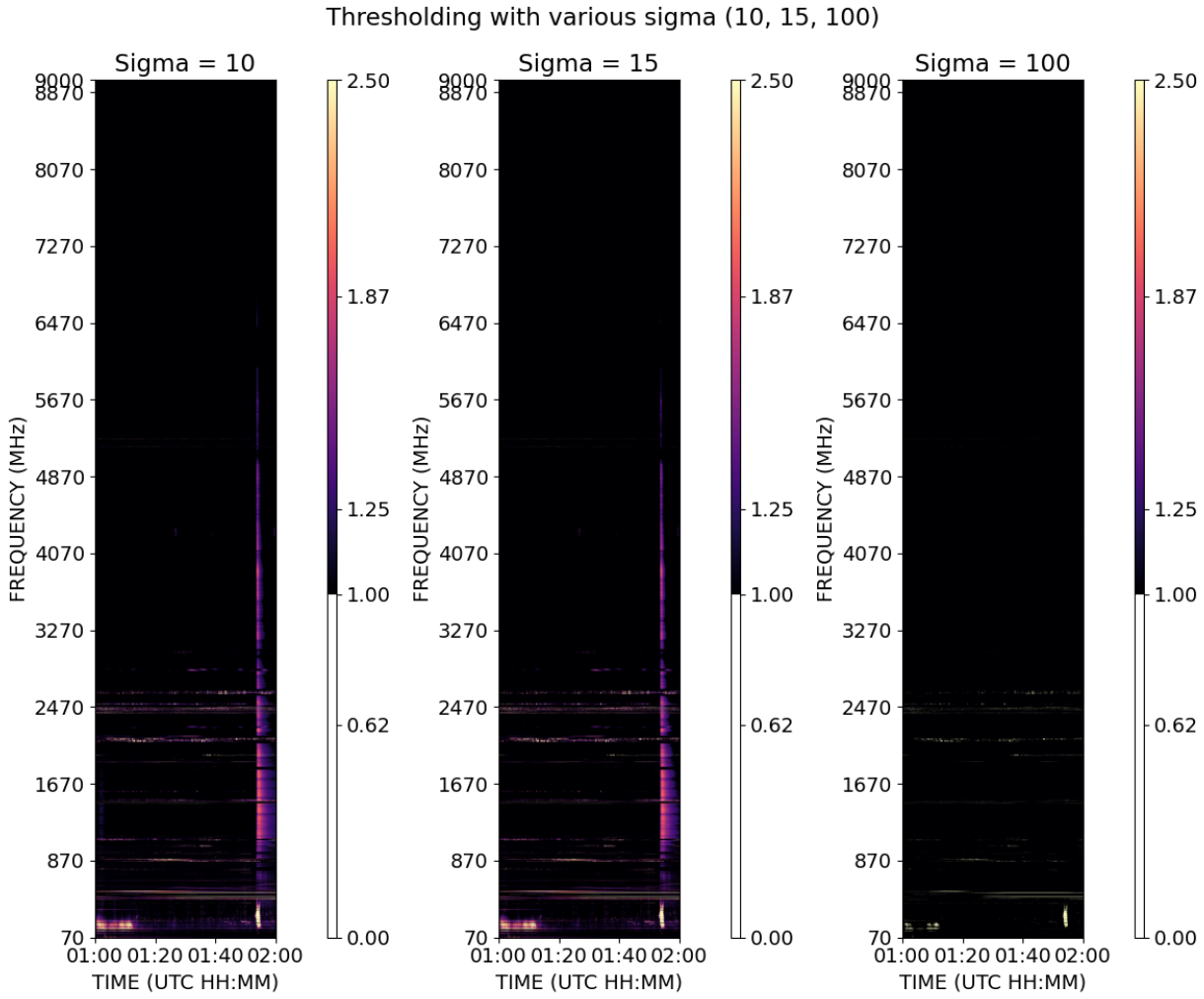



Figure 2.13: The plot show the various median absolute deviation ($\sigma$) used for thresholding the image. The sigma values chosen for the plots are $\sigma = 10, 15, 100$; the abnormally high value of 100 was chosen to demonstrate the strength of the signal above the median.

vidual pixels by converting them into continuous curves. The high-resolution data provided by the spectrograph can now be converted to geometric information, such as the perimeter of the contour, the area enclosed, the shape, etc., making it essential for feature extraction. Information about the distribution of pixels inside the contour can help distinguish between RFI and solar signal; contours help in the identification by reducing the total number of pixels in the spectrum that have to be analyzed. Drawing contours around the boundaries of the binary map after thresholding through Median Absolute Deviation is implemented through `measure.find_contours` method in `skimage` Python module that uses marching squares algorithms (a special case of marching cubes algorithm) [18] to identify contours in a 2D scalar field or *grayscale* images. The dynamic spectra received from the spectrograph can be mathematically represented as a 2D scalar field as every pixel holds one value for intensity, unlike a conventional RGBA model where the image has four values of intensities for a pixel – three corresponding to different channels and one value corresponding to the transparency. Hence, the implementation of the contouring algorithm is justified.

The marching squares algorithm used for the detection works by examining the $2 \times 2$ cell formed by the neighboring pixels. Ones are assigned to points when the intensity is above or below a threshold, often called the iso-value, here the since the algorithm is applied on the masked spectra, the values of ones and zeros are already assigned. This approach can be used to draw the internal sub-contours inside the base contour to map the equipotential surfaces. The algorithm then defines a look-up table determining the line segments in the cell. A linear interpolation is used to find the contour intersections and to captures the intensity transitions; it is given by

$$p_I = p_1 + \frac{iso - f_1}{f_2 - f_1}(p_2 - p_1) \tag{2.10}$$

where $p_I$ is the intersection point, $p_1, p_2$ are the endpoints of the cell with values $f_1, f_2$ and $iso$ is the threshold value, $iso = 1$ in this scenario. Finally, this method returns the list of points $(x, y)$, corresponding to time and frequency axis respectively, lying on the contour. An additional parameter can be set which mentions the height above the base (background at 0) at which the contour should be drawn; this is particularly useful when finding contours for at different thresholds (e.g. $10\sigma$, $20\sigma$, ...).

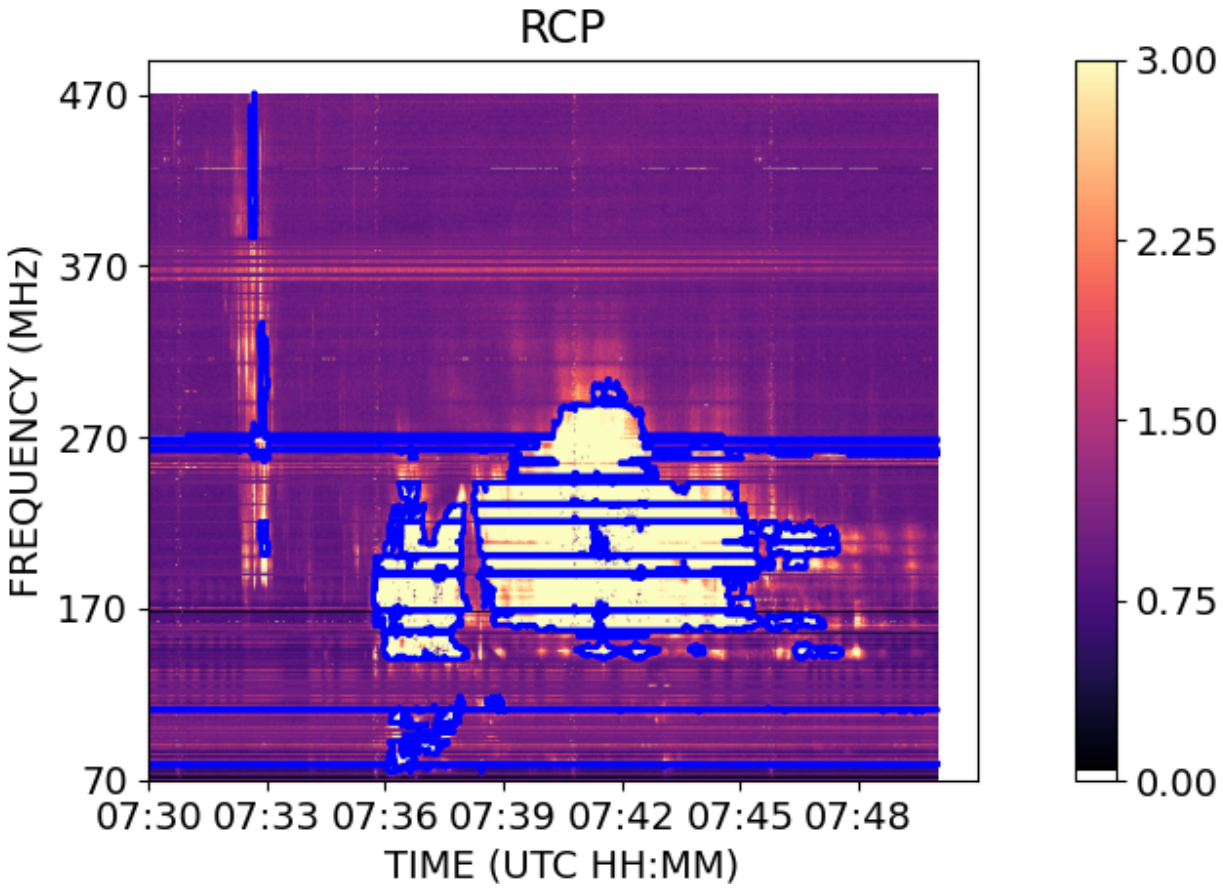



Figure 2.14: The blue lines here denote the contours as detected by the algorithm. The plot only shows the 50 largest contours in the dynamic spectra.

## 2.6 Machine Learning models and approaches

### 2.6.1 Feature vector

The machine learning model is trained on the features of the image for clustering and identification of various morphologies of the bursts. The model can be configured to accept a slice of the image or use the present features. The first approach provides a cropped image that encapsulates the signal/burst to train the model [29]. While training the model, existence of dense layers or fully connected layers would require the input layer to have a fixed number of dimensions. The fixed nature of the input limits our ability to fully exploit the dataset's temporal and spectral ranges without significantly increasing computational time and memory requirements for training; the challenges of the above approach leads us to generate a feature vector describing the the signal. A brief description of the features are given below.

This section lists and briefly describes the components of the feature vector.

1. **Contour :** The contour is drawn around the base of the signal to demarcate the event from the background. A binary map is created by thresholding the image at $3\sigma$ from the median, and $\sigma$ denotes the median absolute deviation. The base contour is presented as a list of tuples and is drawn on the signal rather than the background, i.e., the point on the contour is a part of the signal.As a part of the `skimage.draw.find_contours` implementation, the contours at the zeroth index aren't complete, and the image needs to be padded with zeros at the boundaries to ensure closed contours. The contour is returned as a nested list containing the tuples of contour points. The list is sorted by the length in descending order.

2. **Sub-contours :** In the above discussion, the contours only represented the lines drawn at the base of the burst to demarcate the signal from the background. The sub-contours are obtained by slicing the signal at specified levels. These levels can be expressed in terms of the fraction of the peak value inside the contour (e.g. 90%, 50%, etc. of peak value), and the deviation from the median value (contours at $10\sigma$, $20\sigma$, $30\sigma$, etc.). Sub-contours are drawn with a tolerance to the variation in the values equal to $\sim 10\%$ to the median value, i.e., 0.1 dB.

3. **Area and Perimeter :** The number of pixels inside the regions enclosed by the contour and the contour's length constitute the contour's area and perimeter, respectively. The calculation

of the area includes the points that lie on the contour, and the perimeter is calculated as the number of tuples present in the set of points given as a contour. The ratio between the perimeter and area is the same for all the closed curves of the same shapes. This ratio is also called the iso-perimetric inequality and states that the circle has the smallest value for the ratio. A measure of this ratio informs about the deviation from the shape of the circle. The area parameter also helps reduce the number of contours by filtering out the smaller contours.

4. **Centroid and Center of Mass (Intensity):** The centroid and the center of mass are given by

$$\bar{x} = \frac{\sum x_i}{N}, \quad \bar{y} = \frac{\sum y_i}{N} \tag{2.11}$$

$$x_{\text{COM}} = \frac{\sum I_i x_i}{N}, \quad y_{\text{COM}} = \frac{\sum I_i y_i}{N} \tag{2.12}$$

respectively, using intensity and mass interchangeably. After computing these quantities, the contours, area and perimeter quantities are transformed to the Center of Mass coordinate system. The transformation would ensure that the model does not learn the time stamps and invalidate the translational symmetry.

5. **Shape of the blob:** The blob's shape can be described by the spectral and the temporal extent of the blob from the centroid. The blob is rotated, and the process is repeated to obtain the the .

6. **Euler Characteristics:** For polyhedra, it is defined as the $\chi = V - E + F$, where the $V, E, F$ denote the number of vertices, edges and faces, respectively.

### 2.6.2 YOLO models

Object detection is a computer vision task that involves identifying and locating objects within an image. Unlike Image classification, which assigns single label to an entire image. Object detection provides both class labels and bounding boxes for multiple objects in a scene. There are various pretrained object detection models are available online. YOLO object detection models are one of

the popular object detection models known for there speed and accuracy.
The characteristics of the model are described below:

1. YOLO (You Only Look Once) is a family of deep learning models designed for real-time object detection.

2. Unlike traditional methods that use region proposals and multiple passes through an image, YOLO performs detection in a single forward pass of the neural network, making it efficient and fast.

3. In this experiment, we are using YOLOv-10-small model for our task. This model improved over earlier versions. Optimized backbone (DCNN) for better feature extraction and a refined detection head (final stage of an object detection model) for improved precision

4. Adaptive advanced training techniques and achieves a better latency-accuracy trade-off.

5. The YOLO-v10 small variant was selected due to its balance between efficiency and accuracy. Smaller model size makes it feasible for real time without requiring high end hardware.

6. This version of the model is not as powerful as other variants yet does not compromise on the performance, making it good choice for this task with limited computational resources.

The nuances of the model used are:

1. Dataset Details: The dataset consists of images generated from FITS files, which represent frequency-time graphs where amplitude is visualized using a color bar. These images illustrate different types of solar radio bursts that are significant in space weather research and astrophysics. Each image spans a time duration of two hours and covers frequencies up to 1000 Hz. The dataset includes four distinct classes of solar radio bursts - type I noise storms were excluded due to longer emission duration and difficulty in flagging. All four classes were included in the training process. However, class V was more challenging to detect due to its characteristics and potentially lower representation in the dataset. Data Pre-processing and Augmentation Techniques Applied -

    - Annotation: A total of 440 images were manually annotated using the online Make Sense tool. Annotations were saved in YOLO format for multi-label object detection.

- Image Properties: The images used for training had dimensions of 1200x1000 pixels - 1200 seconds and 1000 MHz
- Pre-processing: No explicit pre-processing techniques were applied before training.
- Augmentation: YOLOv10s built-in augmentation techniques were utilized, including:
  - Colour space transformations (HSV modifications)
  - Scaling and translation
  - Random erasing and mosaic augmentation

2. Model Training set-up:

   - Hardware Used - GPU: NVIDIA RTX 3050 (4GB VRAM), CPU: Intel i5 (12th Gen), RAM: 16GB, Operating System: Windows
   - Training Parameters - Model: YOLOv10 small (yolov10s.pt), Epochs: 50, Batch Size: 8, Image Size: 640, Optimizer: Not manually set (default auto-selected), Loss Function: Default YOLO loss components: (i)Box Loss: 7.5 (ii)Class Loss: 0.5 (iii)DFL Loss: 1.5
   - Learning Rate: (i)Initial (lr0): 0.01 (ii)Final (lrf): 0.01
   - Momentum: 0.937
   - Weight Decay: 0.0005
   - Warm up Strategy: (i)Epochs: 3.0 (ii)Momentum: 0.8 (iii)Bias Learning Rate: 0.1
   - Validation: Enabled (val: true)
   - IoU Threshold: 0.7

3. Challenges

   - Difficulties in Detecting Certain Burst Patterns
     - Class V was particularly difficult to detect. This may be due to its visual characteristics, lower representation, or similarity with background noise.
     - The overall model performance for the other three classes was stable.
   - Class Imbalance Handling
     - No explicit class balancing techniques were applied, as most classes had sufficient representation.
     - Class V had relatively fewer instances, which may have impacted its detection performance.

### 2.6.3 Classifiers RFI detection

Even after the pre-processing and thresholding of the images is not devoid of RFI. The contours are also drawn around the RFI in most cases, and a need arises to differentiate between RFI and the solar signal. To simplify this task, a machine-learning approach is considered to classify RFI and to flag the solar signal from the features extracted by drawing contours. The features used for the extraction of contours are similar to those that are provided in section 2.6.1. A Random Forest Classifier (RFC) is used as a preliminary model for classification. It operates by training each tree with a randomly sampled subset of the data, and each node at the tree consists of a randomly sampled subset from the feature list, ensuring that the entire range of feature vectors is spanned. The prediction is made by casting votes; each tree casts a vote for a class, and the majority vote determines the final class. The majority votes ensure that the overfitting is reduced compared to individual trees.

The efficiency of RFC for handling large datasets makes it a preferred method for classification compared to methods like utilizing Convolution Neural Networks or Multi-Layer Perceptrons; apart from the efficient handling of large data, the classifier is robust to missing data points and noise and provides a good generalization of the classes.

The classifier is trained on data divided into two classes - RFI and not-RFI represented by arbitrary numbers 0 and 2 respectively. Contours were drawn around the binary map made after thresholding dynamic spectrum (interchangeable with image) observed on $5^{th}$ May, 2024. The data for the entire observation time was considered, and $\sim 10^6$ contours were detected; processing large number of contours is computationally expensive and futile. The number of contours were reduced by analyzing the frequency distribution of the area enclosed by the contour, all the contours below 5 pixels in area were removed as they were suspected to be RFI; RFI occupied the majority of the contours drawn. Bounding the area enclosed by the contours brought down the number of valid contours to 28699, and 500 contours were randomly sampled for labeling. The process has been repeated with 1000 and 1500 samples for better results during training and prediction.

The features used for training are

- Number of points: The area or the number of points enclosed by the contour
- Length of contour: The number of points on the perimeter of the contour and the area of the contour together provide an estimate for the *roundness* or the deviation from a circle

- Local maximum value: Value of the maxima inside the contour
- Mean maxima value: Maxima inside the contour is calculated with a tolerance, and hence this quantity is used to capture the mean of the local maxima that lie in the specified tolerance.
- Range in frequency axis: Captures the span on the frequency axis which is important for flagging different types of bursts which often have a well defined frequency limits
- Zero and undefined slope: After computing the two-point gradient using the `numpy.gradient` method, counting the instances of zero slope can help identify contours that might be RFI
- Extents of the contour: Calculating the distance between the centroid of the contour and furthest distance to a point on the contour. This is calculated for all the cardinal direction, i.e., along the positive and negative of x and y axis, assuming centroid as the origin and the intermediate direction, where the contour has been rotated with the rotation matrix

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.13}$$

  where $\theta = 45^o$ is the angle of rotation and the distance along the axes are captured, hence giving the distances in the intermediate direction that lie in between the cardinal directions.

The labeled data along with the feature vector is split into two parts, one for training the model and one to test the predictions. Around 80% of the data is reserved for training, i.e., 400 samples are used training and the remaining is used for validation. After the training the algorithm returns the accuracy, and other statistical scores. A particular example data taken on $5^{th}$ May, 2024, the labeled data consisted of 333 cases of RFI and 67 cases of true solar signal. Results, scores and the statical quantities are saved to a file in `h5py` format for efficient storage.

| | | | | |
|---|---|---|---|---|
| Total Labelled Samples | 500 | | | |
| Training Samples | 400 | | | |
| Validation Samples | 100 | | | |
| Class Distribution of Training set | | | | |
| RFI | 0 | 333 | | |
| Not-RFI | 1 | 67 | | |
| | | | | |
| Accuracy | 0.74 | | | |
| Classification Report | | | | |
| Class | Precision | Recall | F1 - score | Support |
| RFI (0) | 0.77 | 0.93 | 0.84 | 74 |
| Not-RFI(1) | 0.5 | 0.19 | 0.28 | 26 |
| Accuracy | | | 0.74 | 100 |
| Marco Average | 0.63 | 0.56 | 0.56 | 100 |
| Weighted Average | 0.70 | 0.74 | 0.69 | 100 |

Table 2.1: Random Forest classifier results for 500 samples. The recall and the precision for RFI is very high where as, the value is low for Not-RFI, due to less number of instances of Not-RFI in the training data.

# Chapter 3

# Results and Analysis

## 3.1 Hot pixels and other non-conformity in the data

Hot pixels are bright spots in an image where the intensity value is the maximum, often occurring due to defects in the instrument and prolonged exposure times. Unlike random noise, which does not stay confined to the same pixel, hot pixels appear at fixed positions in the spectra. These anomalies in the data can be identified by their abnormally large values compared to their neighboring pixels. Hot pixels can be corrected by using methods like median filtering, but these methods are not preferred. Median filtering methods which convolve with a window of size $(n \times n)$ and replaces the value of the center pixel with the median value of the kernel; though the process is effective in removing hot pixels, it could also lead to flagging of small transient bursts like type III bursts that have time durations as short as a few seconds.

The spectral data also contain other eccentricities; the values in the dynamic spectra correspond to arbitrary flux units, the instrument is set to saturate at 65535 units, which is abnormally high; these values are can also be used to describe the points in time and frequency where data might not be collected. The Radio Frequency Interference typically saturates the instruments response in a particular channel, and usually presents itself with very large values. These large values could correspond to either hot pixels or Radio Frequency Interference, but solar signals cannot achieve such intensities; moreover, solar signals also tend have a drift associated with them, where the frequency of emission decreases with time. Features exhibited by RFI are contradictory to solar signals; usually limited to band and time, and having sharp well defined, straight boundaries

without frequency drift.

## 3.2 Median Subtraction vs Division

As seen in section **??**, the normalization was performed by dividing by the median instead of subtraction, though both the subtraction and division can be used as valid methods for normalization. In the subtraction method, the data get shifted by the value of the median, centering the data around 0; leading to the spectra containing negative values. During the downstream processes like median denoising and tracing contours, the negative values in the dataset could cause low-power channels to gain significance synthetically. Subtracting the median does not affect the fluctuations in the data and will preserve the absolute variations in the spectrum.

$$M'_{i,j} = M_{i,j} - \widetilde{M}_i \tag{3.1}$$

On the other hand, median division ensures that all the values in the normalized spectrum are positive, and values are scaled relative to the median. The median division also scales all the variations in the data proportionally.

$$M'_{i,j} = M_{i,j} / \widetilde{M}_i \tag{3.2}$$

Unlike the median subtraction, median division process does not amplify the weaker signals synthetically, where the contour detection algorithm would flag the contour in such cases; as all the variations and values are scaled relatively, is helpful when there is a need to preserve the relative difference in intensity between the pixels.

Elaborating on the synthetic increase of low-intensity channels; in case of channels with low-intensity emissions, subtracting the median from that channel would bring all these values close to zero. At the same time, the channels with stronger signal, where the subtraction would take away a larger value and also bring these value closer zero. The values for both high-intensity and low-intensity are close to zero and would be given similar preference by the masking or contouring algorithm. Above situation can be avoided by considering the normalization through division, instead of subtraction. Dividing the values treats the higher and lower intensity signals proportionally, making the making and the contouring more efficient.

## 3.3 Stokes I dataset

The Stokes I is the first component of the Stokes vector and it represents the total light intensity or the brightness. This quantity is given by the quadrature sum of the Right and Left Circular Polarizations $I = \sqrt{R^2 + L^2}$. The Stokes I parameter is given after pre-processing from the spectrograph. It is mentioned that the dynamic spectra is normalized by subtracting 33% quantile ($Q_{33}$) from the sum of two circular components.

$$M'_{i,j} = M_{i,j} - Q_{33} \tag{3.3}$$

Since the dataset is already normalized and the median is set to zero, further normalization is unnecessary. A verification can be done to check whether the normalized data and the method to obtain the normalized data give the same result. A quick check to see if the median of the both, given and synthesized, spectra are identical is performed by computing the median of the datasets. The median of the given Stokes I spectra varies largely, ranging from values between $0.00 - 1$. Whereas, synthesized median from the quadrature sum of the polarizations gives a median of $\sim 41$. The above check was performed considering 20 dynamic spectra obtained from various days in 2023.

## 3.4 Constant bandshape assumption

It is already mentioned that the bandshape is the equivalent to the gains of the respective spectral channels. While calculating the bandshape, an assumption that the bandshape was constant with respect to time was considered. The assumption that the bandshape is constant can be relaxed to verify the claim about the constant instrumental response. The median bandshape is calculated for consecutive observing dates, and the absolute difference between a reference value is plotted against the observing days; the difference between consecutive entries is plotted to check the daily variation in the bandshape.

The median values differ around the point *mean* and with a standard deviation of $\sigma$.

## 3.5 Performance of RFI filtering algorithms

The RFI filtering algorithms have limitations when identifying RFI in the spectrum. Complete flagging of data might not occur when using the RFI algorithm, due to reasons like improper thresholds for the detection of peaks, determining the number of channels to consider for flagging, etc.

### 3.5.1 Sav-Gol Filter

In case of the persistent narrow-band RFI, the excision is described in section 2.3; the most important processed used in this algorithm is the smoothing of data with the Savitsky-Golay or Sav-Gol filter. The Sav-Gol filter is a good choice here for the smoothing operation and is better than the other moving average counterparts.

The Savitzky-Golay filter is a smoothing operator extensively used in digital signal processing to enhance the quality of the data which preserving the signal trends and characteristics. Unlike the other moving average counterparts where the signal is blurred and fine details in the spectrum are distorted. To smooth the spectra, the Sav-Gol filter fits a low-degree polynomial over moving window using the linear least squares method. The values in the center of the window are replaced according to the result of the least square fitting. As the Sav-Gol filter can preserve signal well, it is ideal for use in spectroscopy and smoothing of spectrographs.

Sav-Gol filter also preserves higher moments of the signal, like the peak width and the curvature; hence preserving the underlying trends and fine details about the spectra. Other smoothing techniques reduces noise at the cost of distorting sharp signals in the spectra. The polynomial function used for the fitting assuming a locally continuous signal, and any strong or abrupt changes in the signal would cause the filter to perform sub-optimally. Moreover, the window size and the degree of the polynomial are significant quantities that decide the quality of smoothing.

On the other hand, the Gaussian filter smooths data by convolving it with the Gaussian function. The gaussian function calculated a weighted average by assigning higher weights to the points that are closer to the center, increasing their contribution and smaller weights to the ones that are away, hence decreasing the contribution from points that are further from the center. The primary advantage of the Gaussian filter is the suppression of high-frequency noise, (i.e., noise that is

Figure 3.1

rapidly fluctuating), which occurs at the cost of blurring and distorting sharp signals and changes, leading to edge degradation.

The Sav-Gol filter is not a true low-pass filter and it does not remove noise with high frequency components (rapidly varying) in the Fourier plane; depending the type of noise and the situation can enhance the noise present in the spectra. Gaussian Filters are low-pass filters and remove the high frequency components systematically. The scenario under consideration - the dynamic spectra of the Sun, requires us to preserve the edges and the features of the signal for further analysis of the bursts, making the use of Sav-Gol filter necessary in the RFI mitigation algorithm.

### 3.5.2 SNR

The median based methods are preferred over the averaging methods as the former approach provides a handle on the outliers better than the latter method. The calculation of SNR from the standard deviation is straightforward with respect to the definition of SNR, and is given by

$$SNR = \frac{S}{\sigma_\mu} \tag{3.4}$$

where $S$ is the signal and $\sigma_\mu$ is the background fluctuations. The calculations of SNR from the Median Absolute deviation (MAD) is given by eq. 2.3. Let's assume a gaussian distribution with $\mathcal{N}(0,1)$ for $X$ such that

$$MAD = median(|X - median(X)|) \tag{3.5}$$

$$MAD = median(|X|) \tag{3.6}$$

the above equation follows from a symmetry argument where the left and right halves from of the mean of distribution have the same area, thus the mean is equal to the median.

The probability distribution for $X$ is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{3.7}$$

and the median $m$ is given by, and from the above argument, we know that the median $m = 0$.

$$P(x \leq m) = 0.5 \tag{3.8}$$

The probability that $|X|$ is less than some value $q$

$$P(|X| \leq q) = 2P(X \leq q) - 1 \tag{3.9}$$

To find the median of $|X|$, setting the above equation to 0.5 and simplifying,

$$P(X \leq q) = 0.75 \tag{3.10}$$

After computing the values, $q = 0.6745$, and for a standard normal $X$, we have $MAD = 0.6745$. Scaling the above equation for a arbitrary $\sigma$, we get

$$MAD = 0.6745\sigma \tag{3.11}$$

This is the same as eq. 2.3.

### 3.5.3 RFI excision

The RFI excision algorithm performs fairly well in identifying and mitigating RFI in the dynamic spectrum. The result of the plots after the implementation of the algorithm are given below.

From fig. 2.9, it is evident that a fairly large portion of the RFI that are long-lived and in a narrow-band were removed. Removing this RFI can also remove bursts like type I bursts and noise storms which are long duration bursts that last for a time period of a few hours to a few days. The window size for Sav-Gol filter and the threshold for flagging were set based on the values observed routinely in the spectra. The values for the window size and threshold are 10 and $10\sigma$. The polynomial order for the Sav-Gol filter is chosen to be a quadratic, i.e., order 2 polynomial, to reduce the computation cost and time.

The plots show the performance of the RFI excision algorithm with a window size of 5 and a threshold of 5.

The removal of periodic RFI is often precise and complete in the lower frequency spectral

channels; the periodicity of the RFI is well defined in these channels, and occurs at a time interval of 300 seconds. There are other instances of periodic RFI where the periodicity is not well defined. Periodic RFI at higher spectral channels are also sporadic and do not occur at a fixed frequency channels or times of the day.

### 3.5.4 Presence of RFI post excision

## 3.6 Contour Detection

After the excision of RFI, the contour detection, as shown in fig. 2.14 is the next step which is responsible for the drawing boundaries demarcating the signal and the background in the masked dynamic spectra.

## 3.7 Machine Learning Models - reports accuracies and statistics

### 3.7.1 Challenges and Considerations

#### Difficulties in Detecting Certain Burst Patterns

- Class V was particularly difficult to detect due to its visual characteristics, lower representation, or similarity with background noise.
- The overall model performance for the other three classes was stable.

#### Class Imbalance Handling

- No explicit class balancing techniques were applied, as most classes had sufficient representation.
- Class V had relatively fewer instances, which may have impacted its detection performance.

### 3.7.2 Results

#### Detection Performance Metrics

The performance of the YOLOv10 small model was evaluated using key detection metrics:

- Precision (B): 0.3088
- Recall (B): 0.3865
- mAP@50 (B): 0.3914
- mAP@50-95 (B): 0.2263

These values indicate moderate detection performance, with relatively low precision and recall, suggesting that the model may struggle with false positives and false negatives. The mAP values further highlight the difficulty of accurately detecting solar radio burst patterns across different IoU thresholds.

#### Loss Analysis

**Training Losses:**

- Box Loss: 2.7619
- Classification Loss: 2.5575
- Distribution Focal Loss (DFL): 2.7260

**Validation Losses:**

- Box Loss: 3.8014
- Classification Loss: 4.4687
- DFL: 3.5587

The higher validation losses compared to training losses suggest potential overfitting.

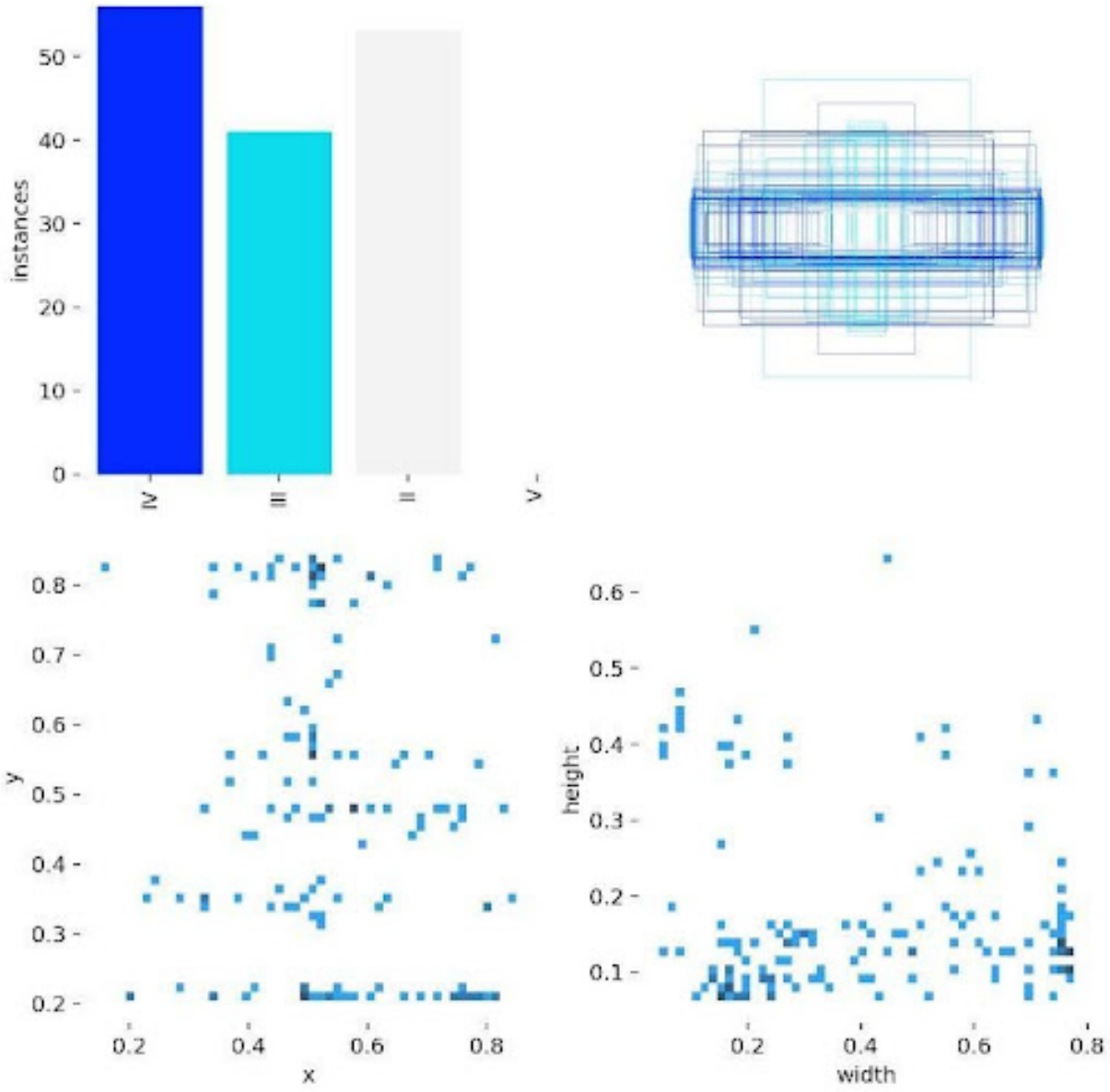



Figure 3.2: Top Left: Bar graph showing the distribution of classes. Top Right: Bounding box distribution in the image. Bottom Left: Spatial distribution of objects (x-y scatter). Bottom Right: Size distribution of bounding box (height-width scatter).

# Chapter 4

# Conclusions

## 4.1 RFI excision

Mitigation and the excision of RFI has been executed sufficiently well after observing and consulting various properties of RFI. The parameters, thresholds and other quantities provided to the algorithm ensure that RFI over flagging is prevented; over flagging increases the risk of removing transient broad-band events like type III bursts. The RFI mitigation algorithm has been optimized for fast execution and to deliver high performance even with limited computational resources. The omission of the denoising algorithm for RFI mitigation has also reduced the compute time and resources required. For the study only the RCP and LCP components were utilized, due to the nature of its normalization and pre-processing. A synthetic I data was created from the RCP and LCP components, following the instruction in the header files, but later was found that the instructions were incomplete and the dataset is heavily pre-processed before packaging, hence distorting the statistical measures used in other parts of the pipeline.

## 4.2 Feature detection and extraction

### 4.2.1 Contours

Contour detection has been carried out by robust algorithms - marching squares algorithm, and the implementation achieves the desired results by identifying the contours after the masking from the background. Apart from the contours at the base level, the sub-contours drawn at higher deviation values from the median also provided information about the distribution of intensities in the signal region, this is a part of the feature vector for the training of the Machine Learning models. Other features (section 2.6.1) can also be extracted from the contours; these properties also help determine the relationship between the pixels or the intensities inside the contour.

### 4.2.2 Machine Learning explorations

Machine Learning approaches were used for two purposes in this explorative study:

1. RFI classification:
   Detected contours in the dynamic spectra were rife with RFI, and other artifacts. The Random Forest Classifier was trained on 500, 1000, and 1500 feature samples from observations made on different days, including the quiet periods of the solar cycle. The prediction and the training dataset were split and the model was about $\approx 75\%$ accurate. Further statistical reports have been mentioned in section 3.7. Hence, promising a moderately accurate description of RFI.

2. YOLO for burst detection:
   The YOLO model was trained on the image files rather than features. The model performs poorly with current set of samples, implying the need for a better robust dataset and additional tuning. The statistical reports on recall and precision of the model is provided in section 3.7.

## 4.3 Future Work

The study was set out with an objective to study the spectrographs provided by YAMAGAWA observatory; their clean, contiguous spectra made it possible to explore the features of the sun at radio frequencies. Coming to the end of thesis, a proposal for the plausible investigations that can be carried out with the YAMAGAWA dataset and the algorithms and methods are listed below:

1. Exploration of stoke I parameter, as the stokes I data holds a information missing in either one of the frequency channels; it has been seen earlier that the degree of polarization in type III bursts are often $< 0.5$ [8].

2. The measurements carried out by observatories are referred to as Sun-as-Star measurements - where there is no localization of the events. Combining the data, especially at higher frequencies, with the X-ray emissions, would reveal interesting events and further exploration on the mechanism of emission is possible.

3. A complete survey of the dataset spanning $\approx 28$ years, or about two and half solar cycles, can be done to search for solar bursts occurring at higher spectral channels, deviating from the traditional study of solar bursts below 500 MHz.

4. Robust Machine Learning models need to be developed to identify and flag bursts in such a voluminous data.

# Bibliography


[1] A. O. Benz and G. L. Tarnstrom. "Synchrotron or plasma process emission in narrow-band type IV$_{dm}$ bursts?" In: 204 (Mar. 1976), pp. 597–603. DOI: 10.1086/154208.

[2] Arnold O. Benz. "Decimeter Burst Emission and Particle Acceleration". In: *Solar and Space Weather Radiophysics: Current Status and Future Developments*. Ed. by Dale E. Gary and Christoph U. Keller. Dordrecht: Springer Netherlands, 2005, pp. 203–221. ISBN: 978-1-4020-2814-4. DOI: 10.1007/1-4020-2814-8_10. URL: https://doi.org/10.1007/1-4020-2814-8_10.

[3] Benz, A. O. et al. "A broadband FFT spectrometer for radio and millimeter astronomy". In: *AA* 442.2 (2005), pp. 767–773. DOI: 10.1051/0004-6361:20053568. URL: https://doi.org/10.1051/0004-6361:20053568.

[4] A. Boischot and B. Clavelier. "Conditions of Acceleration of Solar Electrons, and Determination of the Magnetic Field in the High Corona from the Characteristics of a Type-Iv Burst". In: *Structure and Development of Solar Active Regions*. Ed. by Karl Otto Kiepenheuer. Vol. 35. IAU Symposium. Jan. 1968, p. 565.

[5] H. V. Cane and D. V. Reames. "Soft X-Ray Emissions, Meter-Wavelength Radio Bursts, and Particle Acceleration in Solar Flares". In: 325 (Feb. 1988), p. 895. DOI: 10.1086/166060.

[6] H. V. Cane and D. V. Reames. "Some Statistics of Solar Radio Bursts of Spectral Types II and IV". In: 325 (Feb. 1988), p. 901. DOI: 10.1086/166061.

[7] James J. Condon and Scott M. Ransom. *Essential Radio Astronomy*. 2016.

[8] G. A. Dulk, S. Suzuki, and K. V. Sheridan. "Solar noise storms - The polarization of storm Type III and related bursts". In: 130.1 (Jan. 1984), pp. 39–45.

[9] D. Gary and Christoph Keller. "Solar and Space Weather Radiophysics - Current Status and Future Developments". In: 314 (Aug. 2004).

[10] O. Hachenberg. “Radio Frequency Emission of the Sun in the Centimeter-Wavelength Range: Microwave Bursts”. In: *Solar System Radio Astronomy: Lectures presented at the NATO Advanced Study Institute of the National Observatory of Athens: Cape Sounion August 2–15, 1964*. Ed. by Jules Aarons. Boston, MA: Springer US, 1965, pp. 241–254. ISBN: 978-1-4615-8603-6. DOI: 10.1007/978-1-4615-8603-6_12. URL: https://doi.org/10.1007/978-1-4615-8603-6_12.

[11] https://www.spaceacademy.net.au/spacelink/solrfi/solrfi.htm. *SOLAR RADIO INTERFERENCE TO SATELLITE DOWNLINKS*. URL: https://www.spaceacademy.net.au/spacelink/solrfi/solrfi.htm.

[12] ESA Hubble. *Spectrograph and Spectroscopy*. URL: https://esahubble.org/wordbank/spectrograph-spectroscopy/.

[13] National Institute of Information and Japan Communications Technology. *YAMAGAWA Observatory Image*. URL: https://solarobs.nict.go.jp/.

[14] Kazumasa Iwai et al. “OCTAD-S: digital fast Fourier transform spectrometers by FPGA”. In: *Earth, Planets and Space* 69.1 (July 2017), p. 95. ISSN: 1880-5981. DOI: 10.1186/s40623-017-0681-8. URL: https://doi.org/10.1186/s40623-017-0681-8.

[15] WJ Karzas and Richard Latter. “Electron Radiative Transitions in a Coulomb Field.” In: *Astrophysical Journal Supplement, vol. 6, p. 167* 6 (1961), p. 167.

[16] Tetsuro Kondo et al. “The New Solar Radio Observation System At Hiraiso”. In: *Communications Research Laboratory Review* 40 (Mar. 1994), p. 85.

[17] Mukul R. Kundu. *Solar radio astronomy*. 1965.

[18] William E. Lorensen and Harvey E. Cline. “Marching cubes: A high resolution 3D surface construction algorithm.” In: *SIGGRAPH*. Ed. by Maureen C. Stone. ACM, 1987, pp. 163–169. ISBN: 0-89791-227-6. URL: http://dblp.uni-trier.de/db/conf/siggraph/siggraph1987.html#LorensenC87.

[19] D. B. Melrose. “The emission mechanisms for solar radio bursts”. In: *Space Science Reviews* 26.1 (May 1980), pp. 3–38. ISSN: 1572-9672. DOI: 10.1007/BF00212597. URL: https://doi.org/10.1007/BF00212597.

[20] Morosan, D. E. et al. “Variable emission mechanism of a Type IV radio burst”. In: *AA* 623 (2019), A63. DOI: 10.1051/0004-6361/201834510. URL: https://doi.org/10.1051/0004-6361/201834510.

[21] Alexander Nindos. “Incoherent Solar Radio Emission”. In: *Frontiers in Astronomy and Space Sciences* 7 (2020). ISSN: 2296-987X. DOI: 10.3389/fspas.2020.00057. URL: https://www.frontiersin.org/journals/astronomy-and-space-sciences/articles/10.3389/fspas.2020.00057.

[22] George B. Rybicki and Alan P. Lightman. *Radiative processes in astrophysics*. 1979.

[23] Shirsh Soni, Edwin Ebenezer, and Manohar Lal Yadav. “Multi-wavelength analysis of CME-driven shock and Type II solar radio burst band-splitting”. In: *Astrophysics and Space Science* 366 (Mar. 2021). DOI: 10.1007/s10509-021-03933-7.

[24] A. R. Thompson et al. “The Very Large Array.” In: 44 (Oct. 1980), pp. 151–167. DOI: 10.1086/190688.

[25] Erric W. Weisstein. *Gaussian Function*. URL: https://mathworld.wolfram.com/GaussianFunction.html.

[26] D. C. Wells, E. W. Greisen, and R. H. Harten. “FITS - a Flexible Image Transport System”. In: 44 (June 1981), p. 363.

[27] Stephen M. White. “Solar Radio Bursts and Space Weather”. In: *arXiv e-prints*, arXiv:2405.00959 (May 2024), arXiv:2405.00959. DOI: 10.48550/arXiv.2405.00959. arXiv: 2405.00959 [astro-ph.SR].

[28] J. P. Wild, S. F. Smerd, and A. A. Weiss. “Solar Bursts”. In: *Annual Review of Astronomy and Astrophysics* 1.Volume 1, 1963 (1963), pp. 291–366. ISSN: 1545-4282. DOI: https://doi.org/10.1146/annurev.aa.01.090163.001451. URL: https://www.annualreviews.org/content/journals/10.1146/annurev.aa.01.090163.001451.

[29] Zhang, Weidan et al. “Identification and extraction of type II and III radio bursts based on YOLOv7”. In: *AA* 683 (2024), A90. DOI: 10.1051/0004-6361/202348026. URL: https://doi.org/10.1051/0004-6361/202348026.

[30] V. V. Zheleznyakov and V. V. Zaitsev. “The Origin of Type-V Solar Radio Bursts.” In: 12 (Aug. 1968), p. 14.