

Identification of RF Interference in Astronomical Observations Using Weakly Supervised Machine Learning Classifiers

Arush S. Sharma
The University of Arizona
Tucson, USA

George Reiland
The University of Arizona
Tucson, USA

Marwan Krunz
The University of Arizona
Tucson, USA

Daniel P. Marrone
The University of Arizona
Tucson, USA

ABSTRACT

Radio frequency interference (RFI) is a major concern for passive radioastronomical observations. There is a great interest within the wireless and radioastronomy communities to identify in real time man-made RFI in the vicinity of a telescope. This paper proposes the use of weakly supervised machine learning (ML) techniques to detect the presence of RFI in captured astronomical scans. Weakly supervised training is particularly appropriate when only a small subset of captured data is labeled, as is the case with many radioastronomical datasets. Our study is based on scans obtained from the Arizona Radio Observatory (ARO) at Kitt Peak, Arizona. We rely on the experience of astronomical engineers to label ten 20 MHz channels of a small fraction of the captured scans as "clean" or "dirty". The remaining channels of 4 GHz of the observed spectrum are unlabeled. We first use human-labeled data as ground truth and train two ML classifiers in a supervised manner: a Convolutional Neural Networks - Bidirectional Long Short Term Memory (CNN-BiLSTM) classifier and a Deep CNN classifier. For the unlabeled channels, a semi-supervised technique is adopted, whereby the unlabeled data is first fed to the trained supervised classifier and the outputs with high confidence are assigned pseudo labels. These pseudo-labeled data are further used to train a semi-supervised classifier. To test the performance of the semi-supervised technique, the two classifiers are considered again. We observe test accuracies of 94.55% and 93.69% respectively under weakly supervised training.

CCS CONCEPTS

• **Computing methodologies** → **Semi-supervised learning settings.**

KEYWORDS

Radio Frequency Interference, Arizona Radio Observatory, Machine Learning, CNN-BiLSTMs, Deep CNN

ACM Reference Format:

Arush S. Sharma, Marwan Krunz, George Reiland, and Daniel P. Marrone. 2023. Identification of RF Interference in Astronomical Observations Using Weakly Supervised Machine Learning Classifiers. In *Proceedings of the Int'l*

ACM Conference on Modeling Analysis and Simulation of Wireless and Mobile Systems (MSWiM '23), October 30–November 3, 2023, Montreal, QC, Canada. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3616388.3617545>

1 INTRODUCTION

The ever-growing demand for wireless capacity is pushing the telecom industry to explore the possibilities of utilizing the spectrum above 100 GHz. Next-generation wireless communications, such as 6G and beyond, require tens of GHz of channel bandwidth, which can only be harvested in the sub-THz region. This requires more research on spectrum sharing with "passive" devices as parts of it are used by radioastronomy and weather satellites. The Federal Communications Commission (FCC) has opened a part of this spectrum for experimentation [5]. Despite higher propagation loss, technologies such as improved directional antennas make wireless operation viable above 100 GHz.

Radioastronomy telescopes observe Earth's environment, the solar system, and more across all atmospheric windows from 2 MHz to 1000 GHz and beyond. The ITU-R RA.314-10 [11] provides the list of preferred frequency bands for astronomical measurements (mainly molecules). Unwanted emissions in radioastronomy bands due to terrestrial transmissions result in Radio Frequency Interference (RFI) which disrupts observations, leading to redundancy and costly re-measurements. This comes at a high cost in terms of telescope time and manpower. Radio telescopes, being highly sensitive, are located far from urban areas to detect faint signals, unlike typical communication receivers. With advances in radioastronomy [3], there has been an enormous improvement in sensitivity to detect very faint cosmic signals. Hence it is essential to minimize RFI effects as much as possible.

A key challenge is real-time RFI identification to enable coexistence between wireless systems and passive devices. After gathering observations, a telescope engineer often visually inspects for RFI, which is time-consuming. Machine learning (ML) based techniques can be used to perform inspection, real-time identification of RFI. Several neural network based classifiers have already been proposed in the literature for this purpose. In this paper, we study two classifiers: Convolutional Neural Networks - Bidirectional Long Short Term Memory (CNN-BiLSTMs) [4] and Deep CNN networks [13]. The main contributions of this paper are as follows:

- **Radioastronomical Measurements:** We obtain 515 scans taken by the 12-meter ARO telescope at Kitt Peak, Arizona. Each scan represents temperature vs frequency over 4000 MHz spectrum at sky frequencies between 80 GHz and 104 GHz.



This work is licensed under a Creative Commons Attribution International 4.0 License.

The scans contain both 'clean' and 'dirty' channels (due to RFI) which can be identified in 200 MHz of the observed spectrum. This RFI is due to the modulated waveforms matching licensed transmissions from the FCC database [6]. As a result, only 200 MHz subpart of the scan, divided into 10 channels, is considered for labeling by a telescope engineer. The remaining 3800 MHz data of each scans are unlabeled.

- **ML classifiers:** Thereafter, we consider two neural network classifiers (CNN-BiLSTMs and Deep CNN) and train them in a supervised manner using ground truth data. The testing accuracy of CNN-BiLSTMs and Deep CNN are found to be 95.94% and 94.12%, respectively.
- **Weakly supervised training:** After training, the two classifiers are fed with the unlabeled dataset. Samples with prediction confidence above a threshold provide pseudo-labels for retraining the classifiers, forming a semi-supervised approach. Performance is evaluated using the confusion matrix, Area under Receiver Operating Characteristic (AUROC) curve, and Area under Precision-Recall curve. We observe that the CNN-BiLSTM-based semi-supervised classifier achieves a test accuracy of 94.55% and AUROC value of 0.951. On the other hand, Deep CNN based semi-supervised classifier achieves a test accuracy of 93.69% and AUROC value of 0.927.

2 RELATED WORK

Several previous works presented RFI detection approaches involving signal processing or machine learning. Ford et al. [7] summarized RFI mitigation techniques in radio astronomy. These techniques involve time domain processing using Mean Absolute Deviation (MAD) estimation, spectral kurtosis for identifying gaussian or nongaussian components, adaptive beamforming for spatial excision, and cancellation, where RFI is subtracted from corrupted data after separate detection and estimation.

Taking advantage of deep learning techniques, Czech et al. [4] proposed a combination of CNN along with LSTM to classify transient RFI along with its source. The dataset consists of 63,130 experimentally recorded RFI signals from 8 different sources by switching a number of common potential RFI sources on or off. The classifier achieved a test accuracy of 96.36%. It is worth noting that the work in [4] did not deal with any astronomical data, although the author's goal was to eventually build an RFI monitoring station near the vicinity of the Square Kilometer Array (SKA) telescope site in South Africa.

RFI mitigation using deep CNNs was proposed by Akeret et al. [1]. Their approach uses U-Net to classify clean signals and RFI signatures in time-domain data obtained from a radio telescope. U-Net first found its application in biomedical image segmentation [10]. The authors initially used open source HIDE & SEEK package for generating simulated data with perfect ground truth information. The U-Net model achieved an AUROC value of 0.959 on this simulated dataset. They later applied the model to real astronomical data obtained from Bleien Observatory with imperfect ground truth information, resulting in an AUROC value of 0.88.

cGAN model-based RFI mitigation for radio astronomy data was proposed by Islam et al. [8]. A generative adversarial network (GAN) is composed of a generator and discriminator. The generator produces synthetic RFI-free astronomical data from noisy input

while the discriminator improves the generator's performance by classifying the generated data as real or fake. In addition to the synthetic data, the author proposed using real data from the Green Bank Telescope.

Chakraborty et al. [2] proposed collaborating with cellular networks to cancel RFI, utilizing signal characterization and eigenspaces. This approach improved data quality and throughput for radio astronomy. The authors tested it with real-world astronomical data and simulated LTE signals, achieving an 89.04% reduction in cellular RFI.

Sobjerg et al. [12] introduced spectral kurtosis for RFI detection, which excels at detecting RFI with duty cycles above 15%. However, it has reduced sensitivity to RFI with short duty cycles. The authors suggested that combining spectral kurtosis with normal kurtosis can enhance RFI detection.

3 ASTRONOMICAL DATASET

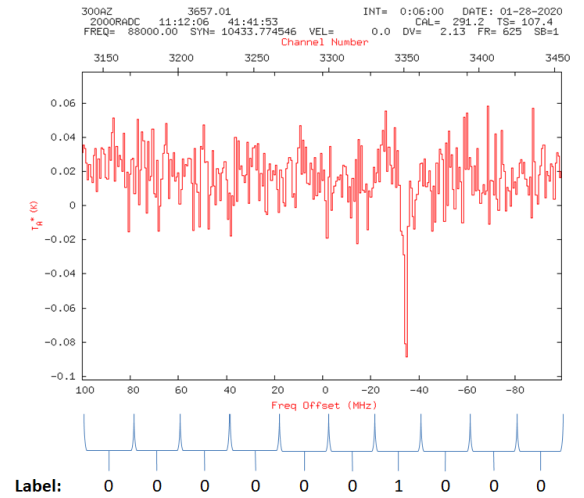


Figure 1: Ten 20 MHz labeled channels in a given scan (0 signifies a clean channel and 1 signifies a dirty channel).

3.1 ARO Telescope at Kitt Peak

The Arizona Radioastronomy Observatory (ARO) operates several telescopes, including a 12-meter Alma-like telescope at Kitt Peak. This telescope has operating frequencies from 68 GHz to 180 GHz [9]. It supports both spectral lines and continuum observations. We gathered astronomical observations for our analysis using the position-switched mode of the telescope. In this mode, the telescope switches between ON and OFF positions according to a given azimuthal offset. The OFF position is free of any emissions. While in the ON position, the telescope points at the astronomical source and records the temperature emitted by this source (if any). The same is done when it points toward the OFF position. The difference in the temperatures between ON and OFF is normalized, i.e., (ON-OFF)/OFF, which signifies the source's antenna/brightness temperature. This position-switched mode consists of an OFF-ON-ON-OFF pattern, where the ON and OFF periods last for 30 seconds

each. Each ON-OFF pair is called a repeat. In a typical scan, the antenna temperature is averaged over 6 such repeats.

3.2 Measurement Setup

The astronomical scans are in Single Dish Data (SDD) format, which can be processed using a proprietary Linuxpops CLASS package on a Unix machine. The dataset consists of 515 scans with various sky frequencies in the ranges 78-82 GHz, 86-90 GHz, and 102.7-106.7 GHz. Each scan is 4000 MHz wide. These scans provide normalized temperature measurements in units of Kelvin for a particular region of the sky. This region is determined by the telescope pointing at specific azimuth and elevation angles. The observations for a given sky frequency are taken by considering both vertical and horizontal polarization. By default, the duration of a scan is six minutes. The interference is clearly visible in our scans when the azimuth angle is greater than 200 degrees. As shown in the example in Fig. 1, interference can be easily eyeballed by a human. A channel of 20 MHz wide can capture such interference, which is why we segment the scans into nonoverlapping 20 MHz channels. Each channel is assigned a binary label, where 0 signifies RFI-free channel and 1 signifies a channel with an unacceptable RFI. For each scan, the frequency resolution is set to 0.625 MHz, which implies that a 20 MHz wide channel contains 32 samples.

4 RFI ANALYSIS

For most of the scans plagued by interference, we could confidently identify the RFI over a small portion of the spectrum (± 100 MHz) from the center sky frequency of the observation. This is because, from a qualitative perspective, the distortion effects in the brightness temperature of an astronomical source due to RFI are visible. Each scan is divided into nonoverlapping 20 MHz wide channels, as these channels could capture these interference effects. Furthermore, from the quantitative point of view, the standard deviation of the temperature of such channels is computed where we found out that the difference in standard deviation between the RFI channel and adjacent clean channel is at least 0.2.

Subsequently, we assign binary labels (0 or 1) for channels that lie within ± 100 MHz of the center frequency of the observation. The remaining 190 channels are unlabeled.

Algorithm 1 : Pseudocode of Semi-supervised Learning

- 1: $X \leftarrow \text{threshold}; \quad ch \leftarrow \text{channel}; \quad L(ch) \leftarrow \text{pseudo-label}$
 - 2: Consider two datasets: Labeled ground truth data and unlabeled data
 - 3: Train the ML model with a labeled dataset (supervised classifier)
 - 4: Iterate over the unlabeled dataset by considering one channel at a time
 - 5: Test the supervised classifier on an unlabeled channel
 - 6: **if** prediction accuracy $\geq X\%$, assign $L(ch) := 1$
 - 7: **else if** prediction accuracy $\leq (100 - X)\%$, assign $L(ch) := 0$
 - 8: **else** skip the channel
 - 9: Append the pseudo-labeled channel to the new dataset
 - 10: Train the new classifier with the same architecture and hyper-parameters on semi-supervised training dataset
-

In total, 4,673 channels are labeled and 98,327 channels are unlabeled. Due to the large number of unlabeled channels, a semi-supervised approach is considered. Let $\mathcal{L}$ be the labeled dataset; $|\mathcal{L}| = 4673 * 33$ (recall that each channel is characterized by 32 temperature values and a binary label). In our semi-supervised approach, we first train an ML classifier in a supervised manner using a subset of $\mathcal{L}$ and test it over the unlabeled dataset. As described in Section 5, the output layer of the model uses a sigmoid activation function signifying the probability of the positive class (RFI-based channels). We choose two thresholds for assigning a binary pseudo label. For the model's outcome greater than 90%, a pseudo label of 1 is assigned to the channel. Similarly, when the model's outcome is less than 10%, a pseudo label of 0 is assigned signifying a clean channel. Furthermore, the channels which don't satisfy the above criteria are discarded. We observe that out of 98,327 unlabeled channels, 94,971 channels are pseudo-labeled when CNN-BiLSTM classifier is used. Similarly, for the Deep CNN classifier, the number of pseudo-labeled channels are 92,158. Likewise, this procedure is repeated, and a separate database of such highly confident samples is created which can be thought of as semi-supervised training data.

Finally, new classifier(s) with the same architecture and hyper-parameters are trained using semi-supervised training data as described in the next section. The pseudo code of the semi-supervised learning technique is shown in Algorithm 1.

5 RESULTS

We consider two approaches in the automatic identification of RFI in the scans. These approaches are based on machine learning-based mechanism where the details of supervised as well as semi-supervised techniques are presented.

5.1 Machine Learning Classifiers

We use open source TensorFlow library [14] for our analysis. Several deep learning based classifiers in the existing literature have been used in identifying RFI from astronomical data. We use two types of classifiers with their design similar to their recent works and compare their performance. The input to the classifiers is a channel of 20 MHz wide consisting of 32 temperature values.

5.2 CNN-BiLSTM Classifier

CNN-BiLSTM based classifier takes advantage of both CNN and LSTM in learning the hidden patterns of the data. CNN layer identifies the salient features from the input data through kernel and pooling. In bidirectional LSTMs the input flows in both directions (past to future and future to past) which helps in a better understanding of the patterns. Fig. 2(a) illustrates the CNN-BiLSTM classifier used in our analysis.

Following the input, 1-dimensional convolutional layers with 64 units, stride of 1, and a kernel size of 4×1 is used. The output of CNN is fed to bidirectional LSTM with 384 hidden units. Finally, a single node with a sigmoid activation function serves as an output layer. All other layers prior to output use ReLU activation function. Adam optimizer is used in this classifier. A batch size of 64 is considered while training. In comparison to [4], our model has a different kernel size and stride which is better suited to our data.

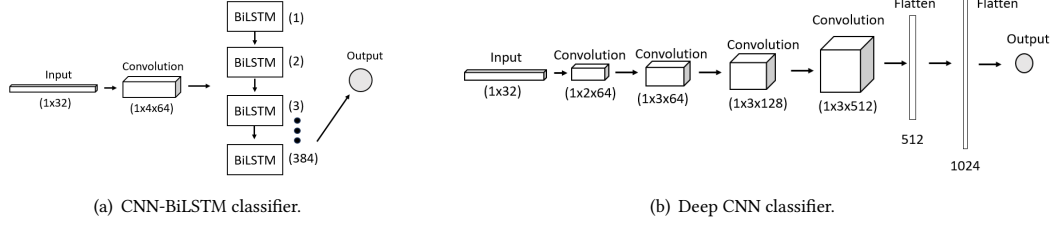


Figure 2: Neural network based classifiers for our analysis.

5.3 Deep CNN Classifier

The authors [13] proposed a robust CNN model to identify the interference based on a simulated data of the SKA telescope. The model has four convolutional layers each followed by max pooling. The layer prior to the output layer is flattened and is passed through the dense layer. All other layers prior to output use ReLU activation function. The output layer uses a sigmoid activation function. Since our data is 1 dimensional, the original 2-dimensional kernel size are replaced with one dimension. Batch normalization is used as a regularization. Adam optimizer is used in this model. A batch size of 32 is considered while training. Fig. 2(b) illustrates the CNN based classifier used in our analysis.

5.4 Supervised Learning Analysis

After the preparation of the ground truth data, the total number of clean and dirty channels are found to be 3898 and 775 respectively. This dataset is split into 80/20 for training and testing and is stratified to preserve the proportions of samples in each class. To account for such class imbalance, different class weights are assigned to each class. The minority class is assigned a higher class weight resulting in imposing a higher cost on the classifier when it misclassifies a sample during the training process. The total number of epochs considered for training are 200. Early stopping criteria with a patience value of 30 is applied to prevent the model from overfitting.

Figures 3(a) and 3(b) show the confusion matrices for CNN-BiLSTM and Deep CNN classifiers respectively. The test accuracy of CNN-BiLSTM and Deep CNN classifiers are found to be 95.94% and 94.12% respectively. The CNN-BiLSTM classifier comparatively has a higher true positive rate of 0.86.

Figure 4(a) indicates the ROC curve which signifies how well the classifier can differentiate between the binary classes. The higher the area under the curve, the better the performance of the classifier. Similarly, the Precision-Recall curve as indicated in Figure 4(b), shows how well the classifier is able to identify the true positives (in our case, dirty channels) while minimizing the false positives. CNN-BiLSTM classifier performs better than Deep CNN classifiers as it is able to distinguish the binary classes better.

Table 1 summarizes the comparison of CNN-BiLSTM and Deep CNN supervised classifiers.

5.5 Semi-supervised Learning Analysis

In this section, the new classifiers with similar architecture and hyperparameters but different initialized weights are trained with

Table 1: Comparison of supervised classifiers

Performance Metrics	CNN-BiLSTM	Deep CNN
True positive rate	0.86	0.78
False negative rate	0.14	0.22
True negative rate	0.98	0.97
False positive rate	0.021	0.027
Area under ROC curve	0.961	0.967
Area under Precision-Recall curve	0.925	0.916
Test Accuracy	95.94%	94.12%

pseudolabeled data and are tested on the same truly labeled testing data (20%) similar to supervised analysis. The testing accuracy for CNN-BiLSTM and Deep CNN classifiers are found to be 94.55% and 93.69% respectively. Figures 3(c) and 3(d) show the confusion matrices for CNN-BiLSTM and Deep CNN classifiers respectively.

CNN-BiLSTM classifier comparatively has the highest number of true positives which is what we desire in our system as we care more about identifying dirty channels as dirty. Figures 4(a) and 4(b) show the ROC and precision-recall curves for the semi-supervised classifiers where the CNN-BiLSTM classifier in comparison is better able to identify the RFI channels.

Table 2 summarizes the comparison of CNN-BiLSTM and Deep CNN semi-supervised classifiers.

Table 2: Comparison of semi-supervised classifiers

Performance Metrics	CNN-BiLSTM	Deep CNN
True positive rate	0.8	0.71
False negative rate	0.2	0.29
True negative rate	0.97	0.98
False positive rate	0.026	0.018
Area under ROC curve	0.951	0.927
Area under Precision-Recall curve	0.901	0.865
Test Accuracy	94.55%	93.69%

6 CONCLUSIONS AND FUTURE WORK

In this work, we analyzed the performance of semi-supervised technique in identifying RFI from astronomical data. At first, we prepared the ground truth database for a supervised classifier by analyzing the RFI within 100 MHz from the center sky frequency of observation. We ended up with fewer samples of labeled data

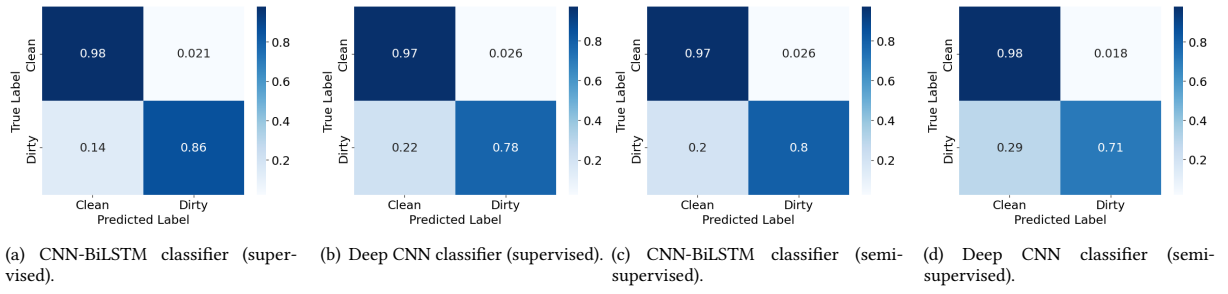


Figure 3: Confusion matrices for supervised and semi-supervised classifiers.

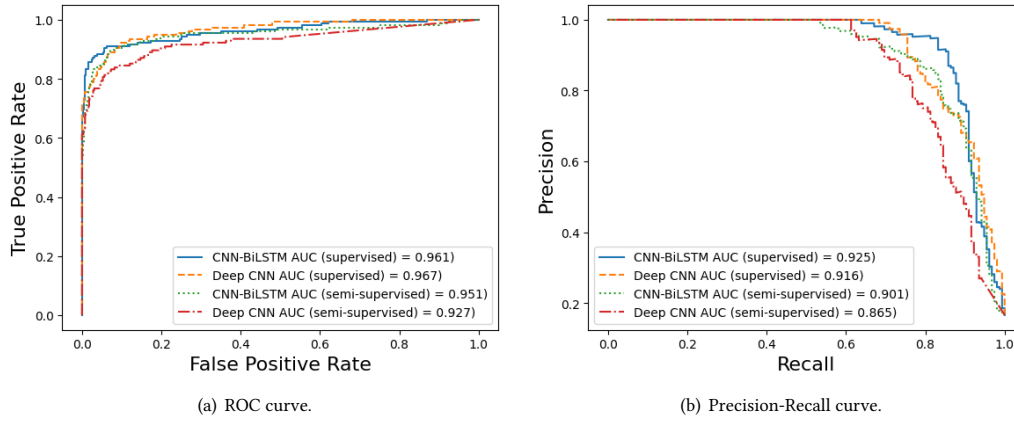


Figure 4: ROC and precision-recall curves for supervised and semi-supervised classifiers.

and relatively many samples of unlabeled data. This motivated us to consider the semi-supervised technique where new classifiers (CNN-BiLSTM and Deep CNN) were trained with highly confident samples and their performance matches closely with the corresponding supervised models. This indicates that semi-supervised models have learned useful patterns from the unlabeled data. In the future, we will perform statistical analysis of weakly supervised classifiers by considering different proportions of the ground truth data. Also, we plan to analyze scans using spectrograms in the time-frequency domain to visualize various RFI signatures, aiding in the localization of RFI.

ACKNOWLEDGMENTS

This research was supported in part by NSF (grants # 2229386 and 1822071) and by the Broadband Wireless Access & Applications Center (BWAC). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of NSF.

REFERENCES

- [1] Joel Akeret, Chihway Chang, Aurelien Lucchi, and Alexandre Refregier. 2017. Radio frequency interference mitigation using deep convolutional neural networks. *Astronomy and Computing* 18 (2017), 35–39.
- [2] Shuvam Chakraborty, Gregory Hellbourg, Maqsood Careem, Dola Saha, and Aavek Dutta. 2023. Collaboration with Cellular Networks for RFI Cancellation at Radio Telescope. *IEEE Transactions on Cognitive Communications and Networking* (2023).
- [3] National Research Council et al. 2010. *Spectrum Management For Science in the 21st century*. National Academies Press.
- [4] Daniel Czech, Amit Mishra, and Michael Inggs. 2018. A CNN and LSTM-based approach to classifying transient radio frequency interference. *Astronomy and Computing* 25 (2018), 52–57.
- [5] FCC. 2023. <https://docs.fcc.gov/public/attachments/FCC-19-19A1.pdf>.
- [6] FCC Universal Licensing System. 2023. <https://wireless2.fcc.gov/UlsApp/UlsSearch/licensePathsSum.jsp?licKey=1052169>.
- [7] John M Ford and Kaushal D Buch. 2014. RFI mitigation techniques in radio astronomy. In *Proceedings of the IEEE Geoscience and Remote Sensing Symposium*. 231–234.
- [8] Islam Helmy and Wooyeol Choi. 2023. cGAN Model-Based Radio Frequency Interference Mitigation for Radio Astronomy Data. In *Proceedings of the International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*. IEEE, 748–750.
- [9] JG Mangum. 2000. *User's Manual for the NRAO 12 Meter Millimeter-Wave Telescope Kitt Peak, Arizona*. NRAO Tucson, AZ.
- [10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention – MICCAI*. Springer, 234–241.
- [11] RA Series. 2010. *Preferred frequency bands for radio astronomical measurements*. Technical Report. Technical Report 10, International Telecommunication Union (ITU).
- [12] Sten Schmidl Søbjaerg, Jan Svoboda, Jan E Balling, and Niels Skou. 2012. Detection of radio-frequency interference in microwave radiometers using spectral kurtosis. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*. 7141–7144.
- [13] Haomin Sun, Hui Deng, Feng Wang, Ying Mei, Tingting Xu, Oleg Smirnov, Linhua Deng, and Shoulin Wei. 2022. A robust RFI identification for radio interferometry based on a convolutional neural network. *Monthly Notices of the Royal Astronomical Society* 512, 2 (2022), 2025–2033.
- [14] TensorFlow. 2023. <https://www.tensorflow.org/>.