Multi-Domain Feature Engineering and Unsupervised Learning for Radio Frequency Interference (RFI) Detection in Solar Radio Spectrographs

PhD Research Poster — African Astronomical Society (AfAS) Conference

Radio Astronomy & Machine Learning
KMeans • DBSCAN • GMM • Autoencoder
Best F1-Score:
1.00
19 Multi-Domain Features Extracted

ABSTRACT RFI poses critical challenges for radio astronomy, particularly solar radio astronomy, corrupting observations of quiet sun emission, solar radio bursts, flares, and coronal mass ejections. We present a detection pipeline combining multi-domain feature engineering with unsupervised machine learning to address the unavailability of labeled RFI data in solar spectrographs. Our method extracts 19 interpretable features spanning temporal, spectral, and statistical domains, used to train KMeans, DBSCAN, GMM, and Autoencoder models evaluated on physics-constrained synthetic RFI with pixel-level ground truth. Clustering models (KMeans/GMM) achieve 100% F1-scores, significantly outperforming DBSCAN (F1=0.39) and Autoencoder (F1=0.10). Feature importance analysis reveals PAPR and Spectral Kurtosis as optimal discriminators. The pipeline eliminates dependency on labeled data, adapting to telescope-specific RFI profiles for robust solar radio observatory deployment.

BACKGROUND & MOTIVATION

Radio Frequency Interference (RFI) critically corrupts solar radio observations of quiet sun emission, solar radio bursts, flares, and coronal mass ejections — degrading data quality for scientific analysis.

Context & Significance

- New 3.7m radio telescope for training graduate & undergraduate students in Nigeria in hands-on radio astronomy
- No labeled RFI data is available, making supervised learning inapplicable
- A pipeline to automatically identify and mitigate RFI is essential for data integrity
- The pipeline must adapt to the telescope-specific electromagnetic environment

RADIO TELESCOPE OPERATION

1. **The Dish (Antenna) —** Collects incoming radio waves from celestial sources
2. **The Amplifier —** Boosts the weak signal to a usable level
3. **The Filter —** Selects the desired observation frequency band
4. **The Translator (ADC) —** Converts analog radio waves into digital numbers
5. **The Imager —** Produces time-frequency spectrograms for analysis

RFI IN SOLAR SPECTROGRAMS

Raw spectrograms (1400–1450 MHz) show noisy vertical stripes — narrowband signals spanning all frequencies at specific times, inconsistent with genuine solar emission.

RFI Signatures Observed

- Vertical stripes spanning the full frequency band
- Narrowband carriers at fixed frequencies
- Sharp, impulsive temporal power spikes
- High Peak-to-Average Power Ratio (PAPR)
- Non-Gaussian spectral distribution (high kurtosis)

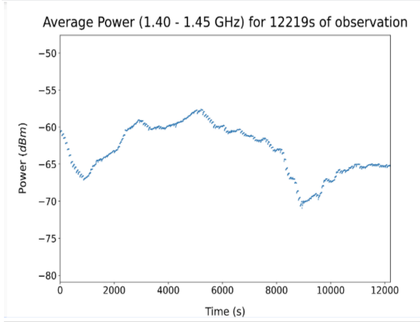


Fig1: Plot of Average power for the entire observation

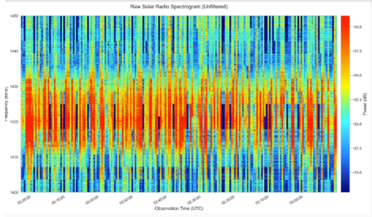
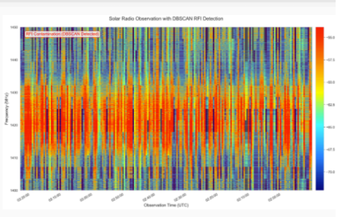


Fig2a: Raw Solar radio Spectrogram



Figb: Solar radio observation with DBSCAN RFI Detection

FEATURE ENGINEERING — 19 FEATURES

Multi-domain feature extraction captures complementary RFI signatures across time, frequency, and statistical spaces:

TimeDomain	• Max Amplitude • Amplitude Variance • Temporal Skewness • Total Power Measure
FrequencyDomain	• Spectral Entropy • Spectral Kurtosis • Dominant Frequency
StatisticalDomain	• Mean Intensity • Std Dev Intensity • 95th Percentile
Complex	• Autocorrelation (lags 1–5) • PAPR • Bandwidth Estimate

KEY DISCRIMINATING FEATURES

★ **PAPR (Peak-to-Average Power Ratio)**
Ratio of peak to average power. Digital communication RFI exhibits characteristically high PAPR, distinct from smooth astrophysical emission.

★ **Spectral Kurtosis**
Tail-heaviness of spectral distribution. Man-made RFI produces non-Gaussian components with elevated kurtosis absent in natural radio signals.

★ **Bandwidth Estimate**
Frequency range containing 95% of signal power. Distinguishes narrowband carrier-wave RFI from broadband spread-spectrum interference.

ML PIPELINE OVERVIEW

Four unsupervised models evaluated on physics-constrained synthetic RFI with pixel-level ground truth:

- KMeans — centroid-based clustering; separates RFI from clean signal clusters
- GMM (AIC/BIC) — probabilistic mixture model; models RFI as distinct Gaussian component
- DBSCAN — density-based anomaly; identifies dense irregular regions without predefined clusters
- Autoencoder — neural net; flags time-slices poorly reconstructed from learned "normal" patterns

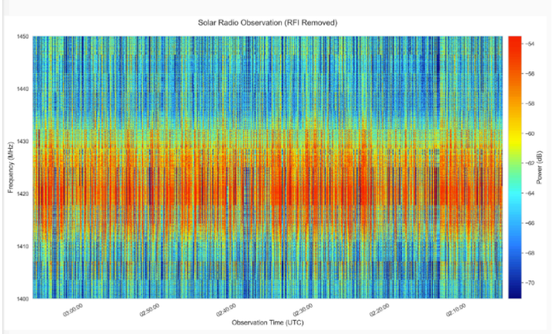


Fig3: Solar radio observation with RFI removed

EVALUATION RESULTS

Evaluation with augmented ground truth against physics-constrained synthetic RFI:

Model	Precision	Recall	F1-Score
KMeans	1.00	1.00	**1.00**
DBSCAN	1.00	0.24	0.39
GMM (AIC)	1.00	1.00	**1.00**
GMM (BIC)	1.00	1.00	**1.00**
Autoencoder	1.00	0.05	0.10

MODEL INTERPRETATION

✔ **Perfect Guards — KMeans & GMM (F1 = 1.00)**
Flawlessly identified every RFI instance while never mislabelling clean data. Distinct multi-domain features make RFI perfectly separable via clustering.

⚠ **Overcautious Guard — DBSCAN (F1 = 0.39)**
Zero false alarms but misses 76% of RFI. Only catches the densest clusters; subtle interference escapes. Requires parameter tuning (lower eps).

✘ **Weakest Guard — Autoencoder (F1 = 0.10)**
Learned to reconstruct RFI as "normal" and did not flag it as anomalous. A common pitfall when RFI structure is not excluded during autoencoder training.

SIMULATED RFI QUALITY ASSESSMENT

Overall proxy quality: MODERATE (0.60 / 1.0)

Spectral Correlation	0.999 ✔
Data Correlation (r)	0.791 ≈
Mean Amplitude Offset	+30% ⚠

Spectral properties match well; amplitude distributions differ significantly. Proxy is usable for model training but may require calibration before real deployment.

CONCLUSIONS & FUTURE WORK

- KMeans & GMM achieve perfect RFI detection (F1 = 1.0), validating the multi-domain feature approach
- PAPR and Spectral Kurtosis are the strongest discriminators between RFI types
- DBSCAN needs parameter tuning; Autoencoder requires RFI exclusion from training data
- Pipeline eliminates dependency on labeled data — deployable at any solar radio observatory
- Future work: real RFI ground truth validation, deep learning integration, real-time flagging pipeline

ACKNOWLEDGEMENTS

Global Emerging Radio Astronomy Foundation (GERA Foundation)
PanAfrican Planetary and Space Science Network (PAPSSN)
Dr. Adams Duniya, Mr. Kabo Mabusha & SKA/AVN Department, BIUST