

Automated Interference Detection in Radio Spectrum under Data Scarcity: A Comparative Analysis of Machine Learning and Deterministic Rule-Based Approaches

Abstract. Efficient radio spectrum management is a critical challenge for regulatory agencies. In Argentina, the National Communications Agency (ENACOM) currently relies on manual processes for interference detection, which limits its operational capacity.

This work presents the development and evaluation of an automated system for detecting interference in radio frequency (RF) signals under a data-scarce setting (approximately 800 labeled samples). Three approaches are compared: Machine Learning models, deterministic rules based on spectral properties, and a hybrid architecture that combines both.

Experiments were conducted on a dataset of 1,624 spectral samples validated by ENACOM specialists in the FM band (88–108 MHz). The results show that the hybrid approach achieves a recall of 94.10% and an accuracy of 93.90%, outperforming the Machine Learning models when used in isolation.

These findings suggest that, in low-data scenarios, incorporating domain knowledge through physically motivated rules is key to achieving robust performance in real operational environments.

Keywords: interference detection, radio spectrum monitoring, machine learning, deterministic methods, spectral analysis, ENACOM.

Detección Automatizada de Interferencias en Espectro Radioeléctrico bajo Escasez de Datos: Un Análisis Comparativo de Enfoques de Machine Learning y Reglas Determinísticas

Resumen. La gestión eficiente del espectro radioeléctrico representa un desafío crítico para los organismos reguladores. En Argentina, el Ente Nacional de Comunicaciones (ENACOM) depende actualmente de procesos manuales para la detección de interferencias, lo que limita su capacidad operativa.

Este trabajo presenta el desarrollo y evaluación de un sistema automatizado de detección de interferencias en señales de radiofrecuencia (RF), en un contexto de disponibilidad limitada de datos (aproximadamente 800 muestras etiquetadas). Se comparan tres enfoques: modelos de Machine Learning, reglas determinísticas basadas en propiedades espectrales y una arquitectura híbrida que combina ambos.

Los experimentos se realizaron sobre un conjunto de 1.624 muestras validadas por especialistas del ENACOM en la banda FM (88–108 MHz). Los resultados muestran que el enfoque híbrido alcanza una sensibilidad del 94,10% y una exactitud del 93,90%, superando a los modelos de Machine Learning evaluados de forma aislada.

Estos resultados sugieren que la incorporación de conocimiento del dominio mediante reglas físicas resulta clave para lograr un desempeño robusto en contextos operativos reales.

Palabras clave: detección de interferencias, espectro radioeléctrico, Machine Learning, reglas determinísticas, ENACOM, procesamiento espectral.

1 Introducción

La gestión del espectro radioeléctrico representa uno de los principales desafíos operativos de los organismos reguladores. En Argentina, el ENACOM es responsable de detectar y controlar interferencias —perturbaciones que degradan la calidad de transmisiones legítimas en servicios esenciales como comunicaciones aeronáuticas y radiodifusión FM—. Actualmente, este proceso depende de la inspección manual por técnicos especializados, resultando lento, reactivo y susceptible al error humano.

La literatura demuestra que técnicas de Deep Learning pueden clasificar señales RF con alta precisión con grandes datasets etiquetados (O'Shea et al., 2016; West & O'Shea, 2017; Zhang et al., 2020); sin embargo, el contexto operativo del ENACOM impone una restricción crítica: apenas 800 muestras únicas etiquetadas, un volumen incompatible con estos enfoques. Este escenario de datos escasos es común en dominios regulatorios donde la captura y etiquetado de anomalías resulta costosa y dependiente de expertos.

La contribución de este trabajo es triple: (i) comparación empírica de ML puro, reglas físicas determinísticas y arquitecturas híbridas bajo escasez de datos; (ii) diseño de reglas espectrales adaptadas al dominio FM del ENACOM; y (iii) análisis del comportamiento y limitaciones de cada enfoque en condiciones operativas reales.

Además de sus aspectos técnicos, el proyecto posee una dimensión institucional relevante. La solución fue desarrollada utilizando herramientas abiertas y configuraciones adaptadas al contexto operativo del ENACOM, permitiendo que el organismo conserve control sobre la evolución del sistema, los criterios de detección y los datos utilizados para su entrenamiento. Esto reduce la dependencia de soluciones propietarias externas y fortalece las capacidades tecnológicas internas del organismo regulador (Hughes, 1987).

2 Metodología

2.1 Configuración experimental y datos

Las señales fueron capturadas mediante un sensor Signal Hound BB60C conectado a una antena en instalaciones del ENACOM, con barridos de 250 kHz de ancho de banda. Aunque el sistema fue concebido para múltiples bandas, este trabajo se centra en la banda FM (88–108 MHz), por la mayor disponibilidad de datos etiquetados en este rango.¹

El dataset de evaluación contiene 1.624 muestras (1.483 interferencias, 141 normales), clasificadas por técnicos del ENACOM, con desbalance representativo del entorno real. Para XGBoost —único modelo supervisado— se conformó un conjunto balanceado con las 800 muestras normales disponibles y anomalías sintéticas hasta igualar ese número. Los modelos no supervisados (Isolation Forest y PCA+Mahalanobis) se

¹ Signal Hound BB60C real-time spectrum analyzer.
<https://signalhound.com/products/bb60c/>

entrenan exclusivamente sobre señales normales. En la fase de evaluación, para mitigar el sesgo estadístico derivado del desbalance, todas las métricas se calculan con promedio balanceado (Macro Average). Se reportan Precisión, Sensibilidad (Recall), F1-Score y Exactitud (Accuracy).

La selección inicial de algoritmos se realizó mediante técnicas estándar de validación, utilizadas para descartar enfoques preliminares y definir los modelos incorporados al pipeline final.

2.2 Representación de señales y modelos evaluados

Todos los modelos reciben como entrada representaciones derivadas del espectro de 1.640 puntos de frecuencia (barrido de 250 kHz, normalizado individualmente con Min-Max para desacoplar la morfología de la amplitud absoluta). Se evaluaron dos representaciones: (a) `clf_raw`, vector crudo normalizado de 1.640 puntos, que preserva la forma completa de la señal; y (b) `clf_feat`, vector de 11 características numéricas que resumen la forma del espectro.

Se entrenaron tres modelos: Isolation Forest (ISO) (Liu et al., 2008) sobre `clf_raw`; XGBoost supervisado (XGB) (Chen & Guestrin, 2016) con muestras reales y sintéticas; y PCA + Mahalanobis (MH) (Mahalanobis, 1936) sobre `clf_feat`.

2.3 Enfoques determinísticos (reglas físicas)

Se desarrollaron cuatro reglas complementarias basadas en las propiedades físicas esperadas de una transmisión FM normal: (a) Peak Rule (PK): Detecta picos espurios mediante dos medias móviles (ventana corta y larga); clasifica como interferencia cuando la diferencia supera un umbral de 8 dB. (b) Gaussian Rule (GS): Ajusta una función gaussiana a la envolvente normalizada de la señal FM. El Error Cuadrático Medio (RMSE) del ajuste actúa como indicador de calidad morfológica: valores altos revelan irregularidades incompatibles con una transmisión normal. (c) Ripple Rule (RPL): Aplica un filtro Savitzky-Golay (Savitzky & Golay, 1964) para obtener el residuo de alta frecuencia; detecta oscilaciones periódicas o picos secundarios en el residuo mediante análisis de prominencia. (d) Asymmetry Rule (AS): Explora la simetría inherente de una transmisión FM respecto a su frecuencia central. Compara, ventana a ventana, la diferencia media absoluta entre las mitades izquierda y derecha del residuo (espejado).

2.4 Sistema integrado

El sistema completo se implementó como una app web que integra: captura en tiempo real vía Signal Hound BB60C (script Python/Windows), comunicación mediante MQTT (broker HiveMQ), backend Python/FastAPI desplegado en Render, frontend React en Vercel, y base de datos Supabase para registrar las predicciones y habilitar la

retroalimentación de operadores del ENACOM. Este sistema permitió obtener las muestras clasificadas que se utilizaron para la evaluación final de los modelos.²³⁴⁵⁶ La plataforma incorpora además mecanismos de validación humana mediante la intervención de operadores especializados del ENACOM. Las clasificaciones realizadas por los técnicos permiten corregir resultados, validar eventos detectados y generar nuevos ejemplos etiquetados. De esta forma, el sistema no reemplaza el conocimiento experto, sino que lo integra al ciclo de mejora continua del proceso de detección.

3 Resultados y Discusión

Se evaluaron ocho configuraciones de pipeline de forma incremental, construyendo cada una sobre la anterior. La Tabla 1 resume las métricas obtenidas.

Tabla 1. Métricas de rendimiento de los pipelines evaluados (Macro Average, n=1.624)

Pipeline	Descripción	Precisión	Sensibilidad	F1	Accuracy
ISO	Isolation Forest solo	64,55%	82,22%	72,32%	82,76%
PK-ISO	Peak Rule + ISO	66,00%	83,26%	73,63%	84,67%
PK-XGB-ISO	Peak Rule + XGBoost + ISO	70,66%	85,69%	77,45%	89,10%
PK-GS	Peak + Gaussian	56,49%	68,14%	61,77%	41,81%
PK-GS-RPL	Peak + Gaussian + Ripple	64,45%	88,08%	74,43%	78,81%
PK-GS-RPL-AS	Anterior + Asimetría	75,67%	94,87%	84,19%	91,81%
PK-GS-RPL-AS-MH	Anterior + Mahalanobis	79,40%	94,10%	86,13%	93,90%
PK-GS-RPL-AS-MH-ISO	Anterior + ISO	80,56%	88,17%	84,19%	94,21%

Los modelos de ML puros (P1–3) establecieron una línea base razonable: Isolation Forest alcanzó 82,22% de Sensibilidad, con Precisión limitada por falsos positivos. XGBoost mejoró la detección de espurias, pero fallaba ante anomalías locales no representadas en el entrenamiento sintético.

Las reglas determinísticas evidenciaron que la Gaussian Rule en aislamiento es insuficiente (Sensibilidad: 68,14%). La adición secuencial de Ripple Rule y, especialmente, la Asymmetry Rule marcó el punto de inflexión, elevando la Sensibilidad al 94,87%.

El Pipeline 7 (PK-GS-RPL-AS-MH), solución definitiva, combina reglas físicas con Mahalanobis como refinamiento estadístico: 94,10% de Sensibilidad, 79,40% de Precisión y 93,90% de Accuracy. El P8 maximiza Precisión (80,56%) pero penaliza el Recall de forma inaceptable para un sistema donde omitir interferencias es crítico.

² HiveMQ MQTT broker. <https://www.hivemq.com/>

³ FastAPI web framework. <https://fastapi.tiangolo.com/>

⁴ Render cloud platform. <https://render.com/>

⁵ Vercel deployment platform. <https://vercel.com/>

⁶ Supabase open-source database platform. <https://supabase.com/>

4 Conclusiones y Trabajo Futuro

Los resultados validan que, aun sin datasets masivos, es posible lograr automatización efectiva mediante una arquitectura híbrida que complementa inferencia estadística con conocimiento determinístico del dominio. Los modelos de ML clásicos exhiben limitaciones estructurales ante la variabilidad morfológica real. El Deep Learning — estado del arte (O'Shea et al., 2016; West & O'Shea, 2017; Zhang et al., 2020)— requiere volúmenes (>100.000 muestras) incompatibles con el entorno operativo del ENACOM.

La solución propuesta no reemplaza al especialista, sino que constituye una herramienta de apoyo a la decisión que combina automatización y experiencia humana. La participación de los operadores permite preservar y reutilizar conocimiento operativo acumulado, fortaleciendo las capacidades técnicas del organismo y favoreciendo su autonomía tecnológica.

Como trabajo futuro, se propone la extensión del pipeline a otras bandas del espectro, incluyendo el Servicio Móvil Aeronáutico (108–136 MHz), donde la disponibilidad de datos etiquetados presenta desafíos adicionales.

Referencias

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Hughes, T. P. (1987). The evolution of large technological systems. En W. E. Bijker, T. P. Hughes, & T. J. Pinch (Eds.), *The Social Construction of Technological Systems* (pp. 51–82). MIT Press.
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. En 2008 Eighth IEEE International Conference on Data Mining (pp. 413–422). IEEE. <https://doi.org/10.1109/ICDM.2008.17>
- Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1), 49–55.
- O'Shea, T. J., Corgan, J., & Clancy, T. C. (2016). Convolutional radio modulation recognition networks. En *Engineering Applications of Neural Networks* (pp. 213–226). Springer. https://doi.org/10.1007/978-3-319-44188-7_16
- Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), 1627–1639. <https://doi.org/10.1021/ac60214a047>
- West, N. E., & O'Shea, T. J. (2017). Deep architectures for modulation recognition. En *2017 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)* (pp. 1–6). IEEE. <https://doi.org/10.1109/DySPAN.2017.7920754>
- Zhang, Y., Li, J., Zakharov, Y., Li, X., & Li, J. (2020). Deep learning based spectrum sensing for cognitive radio networks: State-of-the-art, challenges and future directions. *IEEE Transactions on Cognitive Communications and Networking*, 6(2), 588–599. <https://doi.org/10.1109/TCCN.2020.2982270>