UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

GRADO EN FÍSICA

TRABAJO FIN DE GRADO

DETECCIÓN DE RADIO GALAXIAS EN MAPAS DE LA RADIACIÓN DEL FONDO CÓSMICO DE MICROONDAS

MIGUEL ÁNGEL REMEDIOS SANJURJO

Dirigido por

Dr. MARCOS LÓPEZ CANIEGO ALCARRIA

CURSO 2024-2025

TÍTULO: DETECCIÓN DE RADIO GALAXIAS EN MAPAS DE LA RADIACIÓN DEL FONDO CÓSMICO DE MICROONDAS.

AUTOR: MIGUEL ÁNGEL REMEDIOS SANJURJO

TITULACIÓN: GRADO EN FÍSICA

DIRECTOR/ES DEL PROYECTO: Dr. MARCOS LÓPEZ CANIEGO ALCARRIA

FECHA: JUNIO de 2025

RESUMEN

El presente trabajo tiene como objetivo detectar fuentes espurias en los catálogos del satélite Planck, haciendo uso de redes neuronales convolucionales (CNN). Para ello, se han elaborado de manera realista diferentes simulaciones de mapas del cielo a diferentes frecuencias con las bibliotecas de Python *Pysm3* y *Healpy*. Para la detección de fuentes, se crearon cuatro modelos predictivos entrenados con parches centrados en las coordenadas de las fuentes del catálogo PCCS2 (fuentes compactas) a 30 GHz de Planck. Tras ello, los modelos se aplicaron al catálogo multifrecuencial PCNT (fuentes no térmicas) para encontrar las posibles fuentes espurias.

Los modelos que se han desarrollado han predicho la presencia de fuentes verdaderas y de posibles fuentes espurias. Posteriormente, se evaluó la fiabilidad de cada modelo con distintas métricas. Se observó que los modelos A1 y B1 tenían un sesgo al predecir la detección de una fuente verdadera debido a que, el propio entrenamiento se realizó solamente con fuentes reales. Mientras, el modelo A2 ha realizado predicciones más realistas, otorgando al 13,55% de las fuentes del catálogo PCNTb (fuentes no térmicas brillantes) una probabilidad de que sean espurias superior al 50%. Los resultados de tres de los cuatro modelos se analizaron mediante métodos estadísticos (hipótesis y cuartiles), identificando posibles "outliers" que podrían ser fuentes espurias.

El presente trabajo pretende demostrar que las redes convolucionales pueden ser útiles para aumentar la fiabilidad de los catálogos astronómicos, mejorando así la calidad de los parámetros cosmológicos.

Palabras clave: Fondo Cósmico de Microondas, redes neuronales convolucionales, fuentes espurias, Catálogo Planck (PCNT, PCCS2).

ABSTRACT

This work aims to detect spurious sources in the Planck satellite catalogs using convolutional neural networks (CNNs). To this end, realistic simulations of sky maps at different frequencies were developed using the Python libraries Pysm3 and Healpy. For source detection, four predictive models were trained with patches centered on the source coordinates of the Planck satellite PCCS2 catalog (compact sources) at 30 GHz. The models were then applied to the multi-frequency PCNT catalog (non-thermal sources) to identify potential spurious sources.

The developed models predicted the presence of both true sources and potential spurious sources. The reliability of each model was subsequently evaluated using different metrics. It was noted that models A1 and B1 were biased in predicting the detection of a true source because they were trained only with real sources. Meanwhile, the A2 model made more realistic predictions, giving 13.55% of the sources in the PCNTb catalog (bright non-thermal sources) a probability greater than 50% of being spurious. The results of three of the four models were analyzed using statistical methods (hypotheses and quartiles), identifying potential outliers that could be spurious sources.

This work try demonstrate that convolutional networks can be useful for checking the reliability of astronomical catalogs, thereby improving the quality of cosmological parameter.

Keywords: Cosmic Microwave Background, convolutional neuronal networks, spurious sources, Planck Catalogue (PCNT, PCCS2).

AGRADECIMIENTOS

En primer lugar, me gustaría mostrar mi agradecimiento a Marcos López Caniego, tutor de este TFG, por estar siempre disponible para atender mis dudas y preocupaciones, por muy tontas que fueran o por muy ocupado que él estuviera.

A mi familia, en especial a mis padres, muchas gracias por permitirme estudiar esta hermosa carrera. Por alentarme y entenderme en todo momento. Muchas gracias a mi abuela, por todo el cariño que me ha dado y por ponerme siempre un plato de comida cuando llegaba de la universidad a las cuatro de la tarde.

También quiero agradecer a todas las personas que han formado parte de mi vida universitaria. Muchas gracias a todos los profesores del grado en Física; Raquel, Aijon, Rodrigo, Darío, María, Gabriel, Miguel y muchos otros que no menciono, por impartir con tanto entusiasmo asignaturas que sin vosotros no serían tan bonitas. No me olvido de mis compañeros de clase, que a medida que han pasado los años, se han convertido en verdaderos amigos y han hecho más amenas las horas de clase. Gracias a Silvia, por ser el mejor apoyo que he podido tener, tanto dentro de la universidad como fuera.

Por último, quiero agradecer a los amigos que me han acompañado siempre, en especial al grupo de *Mentalidad do Tiburao.* Por estar siempre conmigo, dispuestos a tomarnos unas cañas mientras vemos un partido o una batalla de gallos, y permitirme despreocuparme de la universidad.

Sin todas estas personas no hubiera podido finalizar el grado en física, por lo que este mérito es vuestro casi tanto como mío.

Apostasteis por el niño, y ahora por el hombre que descubro cada día en el espejo.
Melendi.

TABLA RESUMEN

	DATOS
Nombre y apellidos:	Miguel Ángel Remedios Sanjurjo
Título del proyecto:	Detección de radio galaxias en mapas de la radiación del fondo cósmico de microondas
Directores del proyecto:	Dr. Marcos López Caniego Alcarria
El proyecto ha consistido en el desarrollo de una investigación o innovación:	SI
Objetivo general del proyecto:	Desarrollo de técnicas de detección de objetos en mapas de la radiación del fondo cósmico de microondas mediante la aplicación de técnicas de aprendizaje profundo y redes neuronales convolucionales. Además de la detección, el proyecto busca estudiar la fiabilidad de estas detecciones, para lo cual se van a usar catálogos de objetos de la misión Planck de la Agencia Espacial Europea.

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1. INTRODUCCIÓN

Este trabajo busca aumentar la fiabilidad de los métodos de detección de fuentes en mapas de radiación del fondo cósmico de microondas usando técnicas de aprendizaje profundo y redes neuronales convolucionales (CNN). Se han entrenado cuatro modelos con dos arquitecturas diferentes y dos conjuntos de datos distintos; posteriormente, se evaluó el rendimiento de los modelos a la hora de predecir si la fuente de la imagen es verdadera con un set de imágenes de las fuentes del catálogo multifrecuencial de fuentes no térmicas (PCNT) de Planck. Finalmente se han analizado los "outliers" que hay en las distribuciones de probabilidad que otorgan los modelos para encontrar fuentes que se han denominado como sospechosas y posibles de ser espurias.

Se han simulado mapas del cielo con diferentes componentes y a diversas frecuencias para familiarizarse con las bibliotecas de Python: *Pysm3* y *Healpy,* fundamentales para conformar los sets de entrenamiento y análisis de las redes neuronales.

1.1 ESTADO DEL ARTE

Penzias y Wilson detectaron en 1960 por primera vez una señal isotrópica que provenía de todas las direcciones del espacio, con un valor aproximado de 3,5 Kº, la denominaron "exceso de temperatura de la antena". En paralelo, un grupo de astrofísicos de Princeton formado por Dicke, Peebles y Wilkinson creían en la teoría del Big Bang, argumentando que esta explosión aparte de dispersar materia también debería haber liberado radiación en longitud de microondas. Ambos grupos tras ponerse en contacto interpretaron que la radiación capturada por Penzias y Wilson era la radiación que confirmaba el Big Bang.

Unos años más tarde, algunos cosmólogos como Zel'dovich, Harrison y Peebles indicaban en sus teorías que para la formación del universo observable tenía que haber pequeñas inhomogeneidades en el "exceso de temperatura". Sin embargo, no fue hasta la década de 1980 que se descubrió que esta señal tiene fluctuaciones del orden de $\Delta T/T \leq 10^{-4}$ (Durrer, 2015).

Actualmente, a esta señal se le denomina radiación del Fondo Cósmico de Microondas (por sus siglas en inglés, CMB), una radiación electromagnética que se encuentra por todo el universo. Su espectro es el cuerpo negro medido más preciso en la naturaleza, con una temperatura de 2,725 Kº y una frecuencia asociada al rango de microondas (Durrer, 2015). Está radiación es una de las pruebas más contundentes con las que se sustenta el modelo cosmológico del Big Bang, y por tanto, objeto de una gran cantidad de estudios ya que permite estudiar el universo primitivo (Dodelson, 2003; Weinberg, 2008). Para la eliminación de fuentes que contaminan la señal del CMB se han usado una gran variedad de técnicas estadísticas como el Matched Filtering para la separación de fuentes compactas o el Independent Component Analysis (ICA) y el Needlet ILC (NILC), usadas en las misiones WMAP y Planck para separar la señal primordial del ruido instrumental y de las emisiones difusas como son la emisión térmica del polvo o la emisión de sincrotrón (Bennett et al., 2003; Planck Collaboration, 2020a)

En especial, la información de la misión Planck ha sido crucial para poder crear catálogos de fuentes compactas, como son: PCCS, PCCS2 o PCNT(Planck Collaboration, 2014b, 2016d,

2018). Estos catálogos contienen información de las fuentes, que permiten crear máscaras que separan la contaminación que producen de la señal primordial del CMB.

Los experimentos terrestres más recientes que han estudian el fondo cósmico de microondas con alta resolución han sido: Atacama Cosmology Telescope (Chile) , dedicado a determinar las condiciones del universo primitivo y a observar las galaxias distantes para poder cómo ha evolucionado el universo después de las condiciones iniciales (Swetz et al., 2011); y el South Pole Telescope (Polo sur), capaz de detectar anisotropías en longitudes de onda milimétricas y submilimétricas, y así poder analizar la señal del CMB (Carlstrom et al., 2011).

En los últimos años nuevos enfoques basados en Deep Learning han surgido para el análisis de mapas cosmológicos. Por un lado, CosmoPower (Mancini et al. 2022) se basa en redes neuronales que emulan el espectro de potencia, reduciendo la carga computacional al acelerar la inferencia de parámetros cosmológicos. Por otro lado, DeepSphere (Perraudin et al., 2018) usa redes convolucionales para identificar patrones espaciales en mapas del cielo, mostrando ser prometedor para definir de forma precisa la señal primordial y separarla de emisiones contaminantes en simulaciones.

Respecto a las misiones espaciales, después del telescopio Planck han surgido novedosos y ambiciosos proyectos como el Observatorio Simons que se encuentra en el desierto de Atacama (Chile) y que busca mejorar nuestro conocimiento sobre la inflación cósmica, la masa de neutrinos y la reionización del universo o el LITEBIRD, una misión de alta sensibilidad, liderada por la agencia espacial japonesa (JAXA), con importantes contribuciones de Europa, EEUU y Canadá, que analizará la polarización del fondo cósmico de microondas para detectar ondas gravitacionales (Ade et al., 2019; Ghigna et al., 2024).

Las actuales y futuras misiones espaciales generan una cantidad ingente de datos. La necesidad de recompilarlos, procesarlos y analizarlos para obtener información útil hace necesario el uso de la minería de datos y de modelos predictivos que permitan identificar patrones y conexiones que podrían pasar desapercibidos para las personas (Aggarwal, 2015).

1.2 CONTEXTO Y JUSTIFICACIÓN

El estudio del CMB ha sido y sigue siendo fundamental en la cosmología moderna, proporcionando información sobre el origen, composición y evolución de nuestro universo. La caracterización de sus anisotropías ha permitido confirmar con gran robustez que el universo está dominado por energía y materia oscura, y una pequeña fracción de materia bariónica, supuestos en el modelo cosmológico estándar ΛCDM. Algunas de las misiones espaciales que más información han aportado han sido: COBE, detectando por primera vez las anisotropías primordiales del fondo cósmico de microondas; WMAP, que incrementó la precisión de está detección aumentando la resolución angular de los mapas; y recientemente Planck de la Agencia Espacial Europea (ESA), la primera misión en mapear todo el cielo con una sensibilidad inferior a miliJansky (densidad de flujo espectral) y una resolución angular de 10 minutos de arco. Esta misión generó mapas en nueve bandas de frecuencia, desde los 30 GHz hasta los 337 GHz (European Space Agency, 2013; NASA Goddard Space Flight Cente, 2010; NASA Goddard Space Flight Center, 2022; Planck Collaboration, 2020a).

No obstante, la calidad de estos mapas se ve comprometida en cierta manera por la presencia de fuentes contaminantes que dificultan la obtención de la señal cosmológica primordial del CMB.

Por ello, es fundamental la correcta identificación, modelado y eliminación de estas fuentes contaminantes para la cosmología observacional; conocer las emisiones "foreground", es decir, emisiones que se encuentran entre nosotros y la señal del CMB como son: la emisión de sincrotrón, la libre-libre o la emisión de polvo térmica. Identificar su espectro y localización permite crear máscaras galácticas que se ajustan de forma más precisa, permitiendo eliminar la contaminación de las señales astrofísicas cosmológicas. Por otra parte, las fuentes extragalácticas puntuales introducen contaminación a pequeñas escalas angulares que pueden confundirse con las anisotropías de la emisión del CMB. Identificar estas fuentes puntuales permite aplicar técnicas de separación que eliminan sesgos en las señales primordiales.

A medida que se preparan nuevas generaciones de observatorios, también surgen en paralelo nuevos métodos para separar tanto la emision difusa como la compacta de la radiación del fondo cósmico de microondas. En este contexto, herramientas basadas en inteligencia artificial se entrenan con datos reales para estudiar su capacidad de detectar las emisiones contaminantes frente a los algoritmos clásicos, teniendo una proyección más eficiente y veloz al ser capaz de procesar grandes cantidades de datos. En este trabajo se muestra como las CNNs, con un entrenamiento adecuado, pueden ser una buena alternativa y/o complemento de otros algoritmos de detección clásicos, como el matched filter o la mexican hat wavelet entre otros, para la detección y clasificación de fuentes puntuales espurias en catálogos que contienen grandes volúmenes de datos. Estas redes aprenden e identifican patrones complejos en los mapas del cielo, incluso ante una posible presencia de ruido, demostrando ser efectivas en la tarea de detección automática de fuentes y su posterior clasificación como contaminantes.

El presente trabajo se justifica en el marco de la transición de métodos tradicionales al uso de aprendizaje profundo en la metodología de detección. En concreto, se analiza el rendimiento al detectar fuentes compactas mediante el uso de diferentes modelos basados en "Deep Learning". Esta clasificación no solo es útil para depurar los catálogos y comprobar la fiabilidad de estos, sino que también facilita la limpieza de los mapas mediante un enmascaramiento preciso. Esto permite un proceso robusto de extracción de información cosmológica libre de contaminaciones, fundamental para el análisis estadístico de emisiones primordiales y la estimación precisa de los parámetros cosmológicos.

1.3 PLANTEAMIENTO DEL PROBLEMA

Los catálogos de fuentes compactas de Planck en frecuencias de radio y microondas incluyen galaxias que tienen emisión de sincrotrón y térmica. Sin embargo, también pueden contener fuentes espurias correspondientes a mezclas de otros tipos de fuentes compactas que a su vez pueden estar superpuestas con emisión difusa, y qué de forma casual, cumplen con los criterios de inclusión en los catálogos. Este trabajo de fin de grado se centra en la creación e implementación de redes neuronales convolucionales para comprobar la fiabilidad de los catálogos, con el fin de detectar posibles fuentes espurias, mejorando la calidad de los catálogos y el análisis cosmológico posterior.

1.4 PLANIFICACIÓN DEL TFG

Inicialmente se trabajó en profundizar en el campo de la astrofísica, en particular en el fondo cósmico de microondas y sus anisotropías. Se llevó a cabo una documentación previa sobre la motivación y el mecanismo de separación de la emisión primordial del CMB y las señales

contaminantes presentes en los datos observacionales, y se realizó la introducción a la teoría de redes neuronales, enfatizando el papel que tienen las distintas capas y cómo éstas influyen en las predicciones.

Durante las siguientes semanas, se indagó en las bibliotecas de Python: *Pysm3* y *Healpy*, efectuando simulaciones del cielo. Se testearon mapas a diferentes frecuencias, con diferentes componentes y escalas de colores, en busca de una visualización óptima. Se experimentó con diferentes métodos de recorte de parches (regiones más pequeñas extraídas del mapa completo) con el fin de evaluar cuál de ellos se adaptaba de manera más adecuada para convertir los datos esféricos del cielo a una en geometría cartesiana en el parche bidimensional. Se compararon algunos parches con el software Aladin Sky Atlas (Bonnarel et al., 2000) con el objeto de comprobar que el recorte era correcto.
A posteriori, se construyeron las arquitecturas de los distintos modelos de redes neuronales convolucionales, entrenados para clasificar las fuentes del catálogo como reales o posibles espurias utilizando las bibliotecas: *Numpy* y *Tensorflow*; se localizaron las fuentes puntuales del catálogo de fuentes compactas de Planck (PCCS2) a 30 GHz sobre el mapa de Planck a 30 GHz (European Space Agency, 2015b), se extrajeron los parches y se entrenaron en los modelos que previamente se construyeron.
Finalmente, se evaluaron los modelos realizando ajustes en los parámetros y umbrales de detección positiva para obtener unos resultados satisfactorios. Para el análisis de los datos y su interpretación se elaboraron gráficas mediante el uso de las bibliotecas *Matplotlib* y *Seaborn.*

Para concluir, con los modelos entrenados y analizados, se validaron los catálogos de las fuentes más brillantes del catálogo multifrecuencial de fuentes de emisión no-térmica de Planck (PCNT) sobre el mismo mapa de Planck a 30 GHz. Este análisis otorgó a cada fuente una probabilidad de que sea verdadera y pertenezca al catálogo.

1.5 RECURSOS REQUERIDOS

Este trabajo ha necesitado la plataforma online Google Colab para hacer el código de la simulación, construcción de modelos y análisis de estos mediante las bibliotecas ya mencionadas: *Pysm3*, *Healpy*, *Tensorflow* y *Matplotlib* entre otros. También se ha hecho uso del software Aladin Sky Atlas (Bonnarel et al., 2000)

Asimismo, ha sido fundamental la información generada por la misión Planck de la Agencia Espacial Europea (ESA), n particular, los artículos científicos, los mapas de observación y los catálogos de datos proporcionados por dicha misión han constituido una fuente esencial para la elaboración de este proyecto (European Space Agency, 2013, 2015a, 2015b).

Las herramientas de Inteligencia Artificial usadas en este trabajo han sido: ChatGPT, como apoyo en la redacción académica, búsqueda de fuentes académicas y optimizador de código; y Mendeley para la gestión de referencias.

Capítulo 2. OBJETIVOS

2.1 OBJETIVOS GENERALES

El objetivo del presente trabajo es evaluar la fiabilidad del catálogo multifrecuencial PCNT de Planck mediante el uso de modelos de redes neuronales convolucionales entrenados a partir de "cutouts" de las fuentes del catálogo de fuentes compactas PCCS2 a 30 GHz, para poder identificar posibles fuentes espurias.

2.2 OBJETIVOS ESPECÍFICOS

Se presenta un marco teórico que abarca los temas principales del TFG incluyendo: una breve introducción al estudio de las anisotropías del fondo cósmico de microondas, los modelos teóricos que rigen las distintas emisiones de los componentes del cielo como son el fondo cósmico de microondas, la emisión térmica del polvo o la emisión de sincrotrón, y un breve inciso en modelos predictivos de redes neuronales y estadística.

Simulación de mapas del cielo mediante las bibliotecas *Pysm3* y *Healpy*. Se muestran los mapas a diferentes frecuencias incluyendo componentes contaminantes y ruido instrumental para hacer realistas los mapas del fondo cósmico de microondas.

La construcción de diferentes modelos predictivos de Deep Learning entrenados con un conjunto de parches centrados en las coordenadas de las fuentes puntuales del catálogo de fuentes compactas de Planck en la banda de 30 GHz (PCCS2) (European Space Agency, 2015a).

Como último objetivo específico, se buscan fuentes espurias con los modelos previamente entrenados en el catálogo multifrecuencial de fuentes no termales (PCNT), fuentes detectadas en las 9 bandas del Planck (30-857 GHz) (Planck Collaboration, 2018) Se analizan los resultados y se discuten los métodos estadísticos que muestran las fuentes más sospechosas.

2.3 BENEFICIOS DEL PROYECTO

El principal beneficio de este trabajo es construir un método alternativo que permita dar un paso más en la detección de fuentes compactas en mapas de la radiación del fondo cósmico de microondas, maximizando la fiabilidad de las fuentes detectadas.

Las fuentes espurias dificultan el proceso de separación del CMB del resto de emisiones difusas. Un exceso de estas fuentes en los catálogos reduce la cantidad efectiva de cielo donde separar la señal primordial del CMB, y por tanto, sin poder construir el espectro de potencias ni obtener de forma precisa los parámetros cosmológicos.

La identificación y clasificación de las fuentes de emisión no térmicas extragalácticas y su distinción de las emisiones térmicas galácticas es fundamental para mejorar el análisis mencionado. En particular, la emisión difusa procedente de nuestra propia galaxia como: la emisión sincrotrón, la emisión "free-free" o el polvo térmico; constituye un fondo contaminante que puede ocultar señales de interés cosmológico. Además, existen fuentes compactas como las radio galaxias y/o galaxias infrarrojas lejanas, cuya emisión no térmica puede interferir con la detección de otras señales al estar en la misma longitud de onda que la señal primordial del fondo cósmico de microondas. Esto permite generar máscaras más eficientes que mitigan

contaminaciones específicas, obteniendo así mapas del fondo cósmico de microondas más limpios y fiables.

Con la arquitectura y entrenamiento adecuados los modelos pueden distinguir entre una fuente puntual y el ruido instrumental o la emisión difusa, que distorsionan el espectro de emisión de la fuente principal. Además, los modelos acaban siendo útiles para analizar catálogos rápidamente sin necesidad de rediseñar el modelo en función del catálogo a filtrar.

Capítulo 3. MARCO TEÓRICO

3.1 FONDO CÓSMICO DE MICROONDAS

El modelo cosmológico estándar (ΛCDM) es el más aceptado en cosmología, sustentándose en tres pilares observacionales: la expansión del universo observada en el corrimiento al rojo de los espectros de emisión de las galaxias, las abundancias de elementos ligeros que concuerdan con la teoría de la nucleosíntesis y la radiación de cuerpo negro también conocido como fondo cósmico de microondas o CMB (Dodelson, 2003).

En el universo primitivo (después del Big Bang), la temperatura era tan alta que los electrones no permanecían ligados a los núcleos atómicos. La energía térmica que poseían los fotones, dada por $k_B T$, era muy superior a la energía de enlace del hidrógeno (13,6 eV), provocando que cualquier átomo que se formara fuera ionizado de manera instantánea. Como consecuencia, existía un equilibrio térmico entre la radiación y la materia, permaneciendo en un estado altamente ionizado y denso. A medida que el universo se expandía también se enfriaba debido a la transformación de la energía térmica en energía asociada al trabajo de la propia expansión.

Con el tiempo, esta expansión conllevó a un momento en el que los fotones ya no poseían la energía suficiente para ionizar a los átomos de hidrógeno, permitiendo que los electrones y los protones formaran átomos neutros. A este periodo se le conoce como la época de la recombinación. Desde ese momento, el universo se volvió transparente a la radiación, que comenzó a propagarse libremente, originando lo que hoy se observa como el fondo cósmico de microondas (Weinberg, 2008).

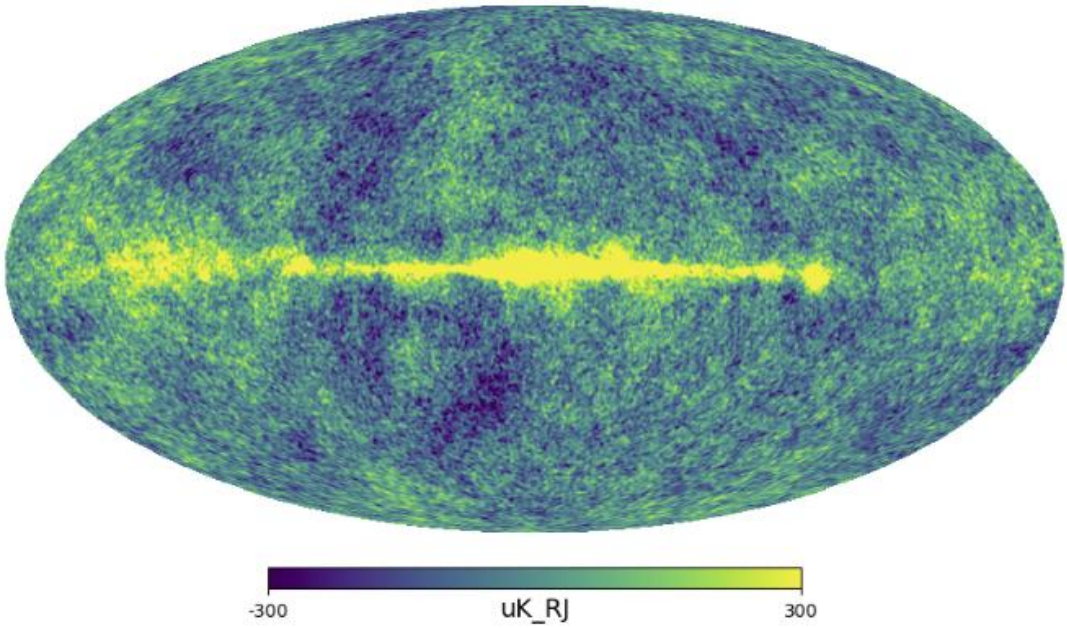

Figura 1: Mapa simulado del cielo a 30 GHz. Nside = 1024 y fwhm = 32.26 arcmin. Componentes de la simulación: CMB, polvo y emisión del sincrotrón. Se ha removido el efecto del monopolo y del dipolo. Rango de temperaturas de [-300, 300] microKelvin.

3.2 Anisotropías del fondo cósmico

Para analizar la radiación térmica del CMB y sus fluctuaciones se usan métodos matemáticos que permiten descomponer los mapas del cielo y extraer la información clave. En este contexto, las funciones matemáticas de armónicos esféricos forman distribuciones en un espacio esférico como suma de funciones que forman una base ortogonal, permitiendo expresar la temperatura T medida en una dirección $\hat{n}$ expresada en los ángulos (θ, ϕ) como:

$$\Delta T(\theta,\phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\theta,\phi) \qquad [1]$$

Donde Y_{lm} es el armónico esférico y a_{lm} el coeficiente de expansión que indica la contribución de cada modo (l, m) (Dodelson, 2003). Para calcular estos coeficientes de expansión se usa el espectro angular de potencias que mide la intensidad de las fluctuaciones del CMB, definiéndose como:

$$C_l = \frac{1}{2l+1} \sum_{m=-l}^{l} |a_{lm}|^2, \qquad l \approx \frac{180}{\theta} \qquad [2]$$

Donde l es el número multipolo, relacionado con la escala angular de la que se calcula la temperatura. Algunos multipolos relevantes son:

- Si l = 0, se trata del componente monopolar y es la temperatura promedio del CMB.
- Si l = 1 (θ = 180º), este valor indica cómo la temperatura varía a lo largo de dos puntos opuestos. En los mapas de CMB, la mayor parte de la señal dipolar se le atribuye al movimiento de la Tierra y del Sistema Solar. Este movimiento provoca un desplazamiento Doppler de las ondas produciéndose un patrón dipolar 'artificial' (Dodelson, 2003). El último valor dado fue por la misión COBE, obteniendo $(\Delta T/T) = (1{,}231 \pm 0{,}003) \cdot 10^{-3}$ (Planck Collaboration, 2020a)
- Si l es grande, se trata de estructuras pequeñas del cielo.

Los multipolos l, también tienen una relación indirecta con el momento en el que se generaron las diferentes anisotropías, a través del tamaño de las perturbaciones que aparecen en el cielo con sus escalas angulares. Aunque los fotones del CMB se emitieron todos a la vez en el "last scattering" ($z \approx 1100$), su energía ha podido alterarse por diferentes procesos. Cada multipolo l, representa un ángulo sustentado por una onda con cierta longitud. Las perturbaciones de gran escala (multipolos bajos) tienen longitudes de onda más grandes, por lo que estaban fuera del horizonte cosmológico en el momento del desacople, haciendo que empezaran a evolucionar en épocas más tardías del universo, y por tanto, siendo sensibles a los efectos más recientes. En el caso contrario, las perturbaciones de pequeña escala (multipolos grandes) entraron en el horizonte antes o durante el desacople de energía-materia, siendo influidas por los fenómenos que transcurrieron en el universo temprano, como son las oscilaciones

acústicas producidas por la presión de las variaciones de densidad de radiación-materia (Weinberg, 2008).

Las anisotropías del fondo cósmico de microondas se dividen en dos grupos según su origen: en perturbaciones primordiales, las cuales se generaron en el plasma primitivo o en la superficie de último "scattering" y que durante la posterior inflación se ha manifestado en la distribución tanto de radiación como de materia, y por otro lado, las anisotropías secundarias, que se han dado por interacciones de los fotones del CMB con potenciales gravitaciones u otras interacciones después del "last scattering" (Dodelson, 2003).

3.2.1 Anisotropías primordiales

Como se ha mencionado, la consecuencia directa de las anisotropías primordiales es la fluctuación de materia en el universo actual. Es decir, las fluctuaciones superdensas se corresponden actualmente con las estructuras de materia como son las galaxias o cúmulos de galaxias. Dentro de este grupo de anisotropías primordiales, se distinguen dos tipos (Lanz, 2016):

- Las fluctuaciones adiabáticas, se corresponden a la nula variación de la entropía de cada componente en cada punto:

$$\delta\left(\frac{n_b}{n_\gamma}\right) = \delta\left(\frac{n_{DM}}{n_\gamma}\right) = 0 \qquad [3]$$

  donde n_b, n_{DM} y n_γ son las densidades de materia bariónica, materia oscura y radiación respectivamente.

- Las fluctuaciones de isocurvatura se caracterizan por tener una energía neta nula en cada punto, describiéndose como:

$$\delta(\rho_\gamma + \rho_b + \rho_{DM}) = 0 \qquad [4]$$

Desde el punto de vista del origen de estas fluctuaciones, hay teorías que exponen diversos orígenes, lo más aceptados en la comunidad científica son:

- Fluctuaciones intrínsecas de la temperatura en el plasma primigenio con un corrimiento al rojo de $z \approx 1090$. Estas fluctuaciones tanto en la densidad como en la temperatura generaron ondas de presión denominadas oscilaciones acústicas, que se han mantenido desde última dispersión, dejando un patrón característico en el espectro de potencia angular (picos acústicos) en las escalas grandes y medias ($l \lesssim 200$) (Weinberg, 2008).
- El efecto Doppler debido a las fluctuaciones de velocidad de los electrones en el plasma en el momento de la dispersión. Este movimiento fue coherente para escalas intermedias, predominando en los picos secundarios del espectro ($l \approx 200 - 600$) correspondiendo a los detalles finos (1º – 0,1º) en los mapas del fondo cósmico de microondas (Hu & Sugiyama, 1995).
- El efecto Sachs-Wolfe, la alteración de la energía de los fotones del plasma primigenio por estar imbuidos en un potencial gravitatorio de los bariones, provocando que en el momento del desacoplamiento hubiera fotones que saliesen de regiones de menor

densidad, ganando energía o de regiones de mayor densidad, perdiendo energía (Sachs & Wolfe, 1967). Este efecto predomina en los multipolos $l \approx 30$, que corresponde a estructuras similares o mayores a 6º en los mapas del cielo (Hu & White, 1997).

Finalmente, en los modelos inflacionarios que se describen en Hawking (1982) o Bardeen et al. (1983) se argumenta que el universo experimentó una transición de fase de primer orden a una fase *de Sitter* de expansión exponencial, donde las perturbaciones ya existentes se vieron amplificadas.

3.2.2 Anisotropías secundarias

Después de la última dispersión o "last scattering", los fotones han estado viajando libremente interaccionando con otras partículas o campos hasta llegar a nosotros. En ese viaje las interacciones y alteraciones que sufren los fotones se denominan anisotropías secundarias (Lanz, 2016).

El efecto Sachs-Wolfe Integrado (ISW) es la anisotropía secundaria más relevante. Se produce cuando el potencial gravitatorio en el que se encuentran los fotones cambia en el tiempo después del desacople materia-energía. El efecto ISW temprano sucede en la transición del dominio de radiación a materia en el universo, es decir, un corrimiento al rojo $z \approx 1000 - 100$. Durante esta época los potenciales gravitatorios decaen ligeramente, porque la presión de la radiación impide el crecimiento eficiente de estructuras, debilitando los pozos gravitacionales. Este efecto modifica escalas angulares pequeñas ($l > 10$), ya que estas fluctuaciones entraron en el horizonte durante esta transición siendo afectadas por la variación de los potenciales. Por otro lado, el efecto ISW tardío afecta a los fotones en épocas recientes, donde la expansión acelerada del universo está dominada por la energía oscura ($z < 1$), lo que suprime el crecimiento de estructuras y hace que los pozos gravitacionales se debiliten o incluso desaparezcan con el tiempo. Si un fotón entra a un potencial gravitatorio y al salir el pozo es más débil, entonces el fotón ha ganado energía neta. Este efecto manipula escalas angulares grandes ($l < 10$), porque solo las fluctuaciones de gran tamaño han tenido tiempo de cruzar estos potenciales durante la era de energía oscura (Weinberg, 2008).

3.3 MAPAS DEL FONDO CÓSMICO DE MICROONDAS

Las misiones e instrumentos que miden la temperatura del CMB y sus anisotropías lidian con diversas fuentes de contaminación. Observan el cielo en frecuencias de radio y microondas para obtener los datos en bruto, para posteriormente aplicar técnicas de separación que permiten aislar la señal del CMB de la del resto de componentes. La tarea de separación se hace a través de diferentes herramientas y algoritmos. Para el caso del Planck: el algoritmo Commander, fue usado para la separación de componentes difusos. Específicamente, este algoritmo opera en el dominio de los píxeles mediante un enfoque bayesiano, ajustando modelos físicos a los datos obtenidos mediante técnicas de muestreo de Monte Carlo. Este enfoque ha permitido tener en cuenta tanto las señales astrofísicas como los parámetros instrumentales, las correcciones de banda y las calibraciones (Planck Collaboration, 2020b; Planck Collaboration, 2015).

Además de Commander, la misión Planck empleó otros tres métodos principales para la separación de la emisión difusa del cielo: NILC (Needlet Internal Linear Combination), SMICA (Spectral Matching Independent Component Analysis) y SEVEM (Spectral Estimation Via

Expectation Maximization). Las técnicas NILC y SMICA combinan mapas de distintas frecuencias y diferentes pesos para extraer la señal del CMB, mientras que SEVEM utiliza plantillas construidas a partir de las diferencias entre canales de frecuencia para identificar y eliminar componentes no cosmológicos (Planck Collaboration, 2015, 2020b). Por otro lado, la técnica de Matched Filtering fue utilizada por Planck para la detección de fuentes compactas, como cúmulos de galaxias o galaxias individuales, ya que permite optimizar la detección de señales con perfiles conocidos sobre un fondo ruidoso (Planck Collaboration, 2014b)

3.3.1 Componentes de la contaminación

La señal del CMB es máxima en el rango de frecuencias de microondas y submilimétricas (10 – 100 GHz), sin embargo, no todo este rango es óptimo para la captura de la señal primordial. En este rango de frecuencias se encuentran diversas fuentes de contaminación, como la atmósfera, la Vía Láctea y fuentes extragalácticas que también emiten fotones en este rango de frecuencias alterando la señal del CMB. Los tipos de contaminación más abundantes son:

- La emisión de sincrotrón es generada por las partículas cargadas que se mueven en el seno de un campo magnético galáctico. Para partículas clásicas, la frecuencia con la que gira un electrón por estar dentro de un campo magnético se obtiene al igualar la fuerza de Lorentz y la fuerza centrípeta, obteniendo:

$$\omega_B = \frac{qB}{\gamma mc} \qquad [5]$$

  Donde q y m son la carga y masa de la partícula, B la intensidad del campo magnético, c la velocidad de la luz en el vacío y γ el factor relativista.
  Por el contrario, si las partículas son relativistas, la emisión de sincrotrón se emite en un cono de radio $\Delta\theta = 2/\gamma$, con un espectro más complejo y un pulso más corto. Como se explica en Lanz (2016), la luminosidad expresada en términos de temperatura para el espectro de los electrones relativistas es:

$$T(\nu) \propto \nu^{(\rho+3)/2} = \nu^{-\beta} \qquad [6]$$

  Donde ρ es una función del espectro de energía del electrón, y por tanto β, es el índice espectral, con valores típicos comprendidos entre -2,7 y -3,3. Esto implica que la emisión de sincrotrón es intensa a frecuencias bajas (por debajo de 70 GHz), mientras que para las frecuencias altas es casi imperceptible.(Fuskeland et al., 2021).

- Otra fuente contaminante es la emisión de polvo térmico, detectable a medida que aumenta la frecuencia. Los granos de polvo se encuentran en el medio interestelar absorbiendo la luz visible y ultravioleta, calentándose y remitiendo como radiación térmica, siguiendo una distribución de cuerpo negro modificado (Planck Collaboration, 2014a). Esta alteración viene dada por diferentes factores como: los tamaños microscópicos de los granos de polvo en comparación a las longitudes de onda que absorben y emiten; composición del polvo, formado por silicatos, carbonáceos y hielo que no absorben toda la radiación como si hace un cuerpo negro; la forma de los granos de polvo, la gran mayoría de los granos polvo son amorfos por lo que no emiten la radiación de manera uniforme.

La intensidad de la emisión de radiación de los granos de polvo se suele modelar como:

$$I_\nu = \kappa_\nu \cdot B_\nu(T) \qquad [7]$$

Donde el primer término es la emisividad del grano de polvo, dependiente de la frecuencia, y el segundo término es la intensidad espectral, conocida como la ley de Planck del cuerpo negro, que describe la energía emitida por unidad de tiempo, área, frecuencia y ángulo sólido:

$$B_\nu(T) = \frac{2h\nu^3}{c^2} \cdot \frac{1}{e^{\frac{hc}{kT}} - 1} \qquad [8]$$

Las constantes h, k son la de Planck y la de Boltzmann respectivamente. Finalmente, la temperatura a la que se encuentra el grano de polvo se puede definir, si está en equilibrio térmico con la radiación que lo rodea, como:

$$T(r) = \left(\frac{4\pi}{\sigma_{SB}} J(r)\right)^{1/4} \qquad [9]$$

Siendo σ_{SB} la constante de Stefan-Boltzmann y $J(r)$ la intensidad media de la radiación donde se encuentra el grano (Gail & Sedlmayr, s. f.).

Hay otras fuentes contaminantes que no son tan relevantes, pero que siguen influyendo en medidas tan precisas como las que realizan misiones espaciales como Planck o COBE. Una de estas fuentes contaminantes es: "Bremsstrahlung", también conocida como "free-free emision". Está radiación se produce cuando las partículas cargadas se desaceleran en el campo eléctrico de otra partícula cargada, dominando en frecuencias bajas a 50 GHz aproximadamente (Rybicki &

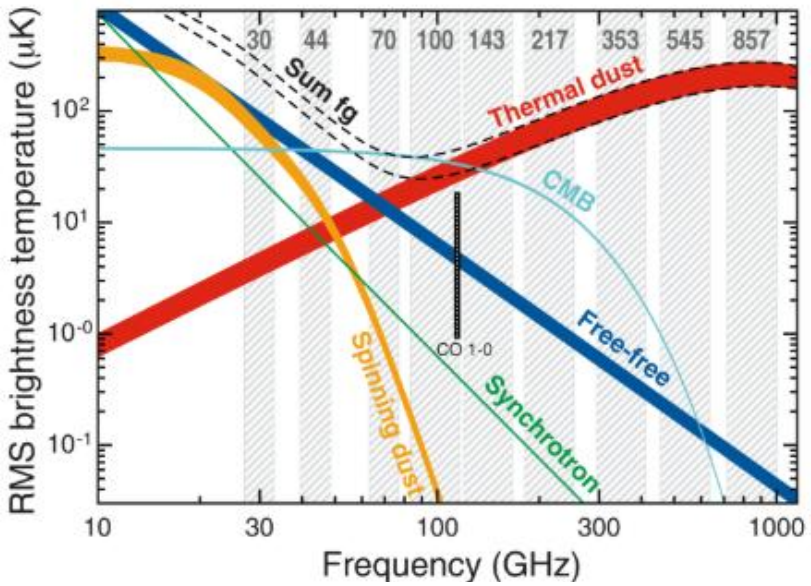

Figura 2: Espectro de brillo (temperatura) en función de la frecuencia para los diferentes componentes del cielo (Planck Collaboration, 2016c).

Lightman, 2004). Otra fuente de contaminación a frecuencias bajas también es la denominada emisión de "polvo giratorio", teniendo el mismo rango de emisión que la "free-free", pero con mayor intensidad y un momento dipolar eléctrico rotacional (J. A. Génova-Santos, 2015).

En cualquier mapa del cielo, independientemente de la frecuencia, se encuentran el monopolo y el dipolo, que deben ser tratados cuidadosamente ya que introducen contaminaciones sistemáticas. El monopolo representa el valor promedio de la temperatura del cielo (2,725 Kº). Aunque a priori este valor es importante, la información más valiosa del fondo cósmico de microondas viene dado por las anisotropías, por lo que es fundamental eliminar el monopolo de los mapas para no ocultar estas pequeñas fluctuaciones. Si no se elimina puede complicar la separación de los componentes o afectar en el calibrado de los instrumentos.

Del mismo modo, el dipolo representa la anisotropía inducida por el movimiento de la Tierra respecto al marco de referencia del CMB, y en menor el movimiento del baricentro del sistema solar. Este movimiento provoca un efecto Doppler de las ondas produciéndose un patrón artificial (Dodelson, 2003). El último valor dado fue por la misión COBE y WMAP, obteniendo $(\Delta T/T) = (1{,}231 \pm 0{,}003) \cdot 10^{-3}$ (Durrer, 2015)

3.3.2 Fuentes puntuales

Otra fuente de contaminación para los mapas del CMB son las fuentes puntuales, ya que introducen señales adicionales tanto en la temperatura como la polarización. Los principales problemas son las alteraciones en el espectro angular de potencia, alterando la homogeneidad del CMB. Estas fuentes puntuales introducen errores al estimar parámetros cosmológicos, modificando los estudios de modo B de polarización del CMB que son útiles para detectar la inflación del Universo (Planck Collaboration, 2016b, 2016a). Si estas fuentes puntuales se detectan y se caracterizan, tanto espacial como espectralmente, se pueden crear máscaras ajustadas que eliminan la radiación que emiten.

En la frecuencia de radio, las fuentes puntuales están dominadas por la emisión de sincrotrón. Las fuentes extragalácticas más comunes son los núcleos de galaxias activos, que emiten grandes cantidades de energía en forma de radiación en todo el espectro electromagnético. Según la orientación de los "jets" relativistas (chorro de partículas) pueden ser, blazares si el jet apunta hacia la tierra o cuásares si la orientación es otra. Por otro lado, entre las fuentes puntuales galácticas destacan los pulsares y estrellas de neutrones, que emiten en radio (Bianchi et al., 2022).

3.4 CATÁLOGOS

Los catálogos del satélite Planck recopilan conjuntos de datos astronómicos sobre la radiación del fondo cósmico de microondas y otras fuentes astrofísicas. Los principales catálogos que se han manejado para este trabajo han sido el catálogo de fuentes compactas (PCCS2) y el catálogo multifrecuencial de fuentes no térmicas (PCNT). La información típica que se muestra de una fuente es: nombre, coordenadas, el flujo y su error asociado, entre otros.

El catálogo PCCS2 contiene solamente fuentes compactas. Para este trabajo se ha manejado la información del catálogo de 30 GHz, perteneciente al modo de baja frecuencia del Planck LFI. EL catálogo cubre la mayor parte del cielo y posee un valor superior al 84% de fiabilidad para sus 2.039 fuentes detectadas en las tres bandas de baja frecuencia (30, 44, 70 GHz). La evaluación

de esta fiabilidad se realizó comparando las fuentes detectadas por Planck con catálogos independientes ya existentes como NED, CRATES, NEWPS y AT20G. Si la fuente aparece en uno de estos catálogos, se la considera confirmada externamente y fiable. Específicamente, las fuentes que se encuentran en este catálogo para la banda de 30 GHz tienen una relación señal/ruido mayor que 4, conteniendo así 1.560 fuentes. Las fuentes extragalácticas que predominan son en su mayoría los núcleos activos de galaxias o cuásares, mientras que las fuentes galácticas pueden ser estrellas envueltas en polvo o regiones de gas ionizado compactas (Planck Collaboration, 2016d)

El catálogo PCNT recoge las fuentes no térmicas detectadas en los mapas de temperatura del Planck en todas las bandas de Planck (30-857 GHz). Las 29.400 fuentes candidatas cumplen que su relación señal/ruido es mayor que 3 para uno de los canales de 30 o 143 GHz, sin embargo, no se pueden considerar fuentes confirmadas ya que una parte de ellas pueden corresponder a emisiones difusas, fuentes térmicas mal clasificadas o ruido (Planck Collaboration, 2018). El espectro de las fuentes térmicas sigue la ley de cuerpo negro, sin embargo, las fuentes no térmicas tienen una distribución de radiación modificada, conocida como ley de potencia. Una gran parte de las fuentes candidatas poseen emisión de sincrotrón, los ejemplos más comunes son los núcleos activos (AGN) (Rybicki & Lightman, 2004). Del catálogo se pueden extraer dos subcatálogos, el catálogo de fuentes no térmicas brillantes (PCNTb) y el de alta significancia (PCNThs). El PCNTb está compuesto por las fuentes más brillantes que cumplen la condición de que la relación señal/ruido es mayor a 4 para los canales 30 y 143 GHz simultáneamente. En total 1.424 fuentes cumplen está condición, siendo la gran mayoría fuentes de radio extragalácticas que se han identificado en otros catálogos como núcleos activos de galaxias que emiten grandes cantidades de energía en todo el espectro. En concreto, 913 de las fuentes son posibles blazares que han sido catalogados previamente en BZCAT5 a una diferencia menor de 7’ en las coordenadas de detección (Massaro et al., 2015; Planck Collaboration, 2018). Por otro lado, el catálogo PCNThs está compuesto por 151 fuentes que cumplen la relación señal/ruido mayor que cuatro para las 9 bandas de Planck, siendo este subconjunto de datos el más fiable (Planck Collaboration, 2018).

3.5 REDES NEURONALES

Los modelos de redes neuronales procesan datos simulando la estructura del sistema nervioso humano. La unidad de computación de este modelo es la “neurona”, la cual recibe información de otra neurona, realiza cálculos en función del peso de la conexión que hay entre ambas, y posteriormente, vuelve a enviar la información procesada a otra nueva neurona (Aggarwal, 2015).

Una red neuronal está compuesta por capas que se combinan para formar un modelo. Cada modelo tiene una “función de pérdida”, que mide cuánto se equivoca el modelo al realizar predicciones frente a los valores reales. Por otro lado, el “optimizador” rige la actualización de los pesos de las diferentes capas en cada repetición con el objetivo de minimizar la función de pérdida.

Como se menciona en Chollet (2018), "elegir la arquitectura de red correcta es más un arte que una ciencia", esto hace referencia a que no hay unas reglas determinadas para construir ni apilar las capas del modelo. Lo más importante es la calidad de los datos que son entrenados en el modelo (*Input X)*, una correcta elección de la función de pérdida y del optimizador en función de los objetivos.

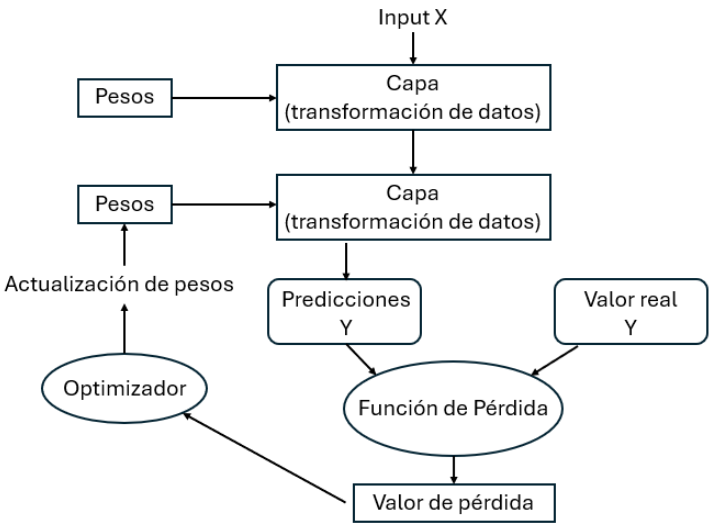

Figura 3: Esquema del proceso simplificado que realiza un modelo típico de red neuronal (Chollet, s. f.).

Para las predicciones binarias, como es el caso del objetivo del trabajo, al tratar de detectar si el catálogo contiene fuentes falsas, los datos que provienen de las imágenes se almacenan en tensores 4D (número de imágenes simultáneas, ancho en pixeles, largo en pixeles, canal) para ser tratadas mediante capas de convolución 2D. Este tipo de capas trabajan con los mapas bidimensionales, aplicando diferentes filtros o "kernels" para detectar patrones (Hinton et al., 2012). Mientras que, los datos vectoriales simples se almacenan en tensores 2D (muestras, características) que se procesan en capas de neuronas densamente conectadas (Chollet, s. f.).

3.5.1 Capas de Convolución 2D

Con la biblioteca de *Tensorflow* de Python, se construyen capas de convolución 2D con la siguiente estructura:

$$\rightarrow Conv2D(32\,,(3,3), activation = 'relu', padding = 'same')$$

esta capa aplica 32 filtros de tamaño 3x3 pixeles a cada imagen de entrada, siendo capaces de detectar patrones espaciales como pueden ser bordes, texturas o formas. Normalmente se desea conservar las dimensiones espaciales de la imagen de entrada como se ve representado en: $padding = 'same'$, aparte de poder introducir no linealidades para que el modelo aprenda relaciones más complejas en: $activation = 'relu'$.

Otro tipo de capa muy usada es:

$$\rightarrow MaxPooling2D((2,2))$$

que tiene como función reducir la resolución espacial de la imagen a la mitad. Para ello, toma cuadrículas de 2x2 pixeles y las sustituye por el valor máximo, pasando de una entrada (32,32, canal) a (16,16, canal), por ejemplo.

3.5.2 Capas densamente conectadas

Cuando se tiene un conjunto de datos dispuestos en un vector (1D), es habitual encontrarse con las siguientes estructuras:

$$\rightarrow Dense\,(32, activation =' relu')$$

$$\rightarrow Dropout\,(0.3)$$

esta capa contiene 32 neuronas totalmente conectadas que analizan las diferentes relaciones que pueden tener los datos, mientras que, la función $Dropout$ apaga un porcentaje aleatorio de las neuronas de cada subconjunto reduciendo el sobreajuste y la posible incapacidad para predecir datos fuera del conjunto de entrenamiento.

3.5.3 Compilación del modelo

Cuando se compila el modelo completo se establecen las funciones de pérdida y optimizadores que van a hacer que los datos del conjunto de entrenamiento predigan mejor a los del conjunto de prueba. Por norma general, para problemas de clasificación binaria se usa la entropía cruzada binaria como función de pérdida, mientras que para los problemas de regresión es más habitual usar el error cuadrático medio (mse) o el error absoluto medio (mae). Una arquitectura recurrente para problema de clasificación es:

$$\rightarrow model.compile(optimizer = 'adam', loss = 'mse', metrics = ['mae'])$$

donde se usa *adam,* un optimizador que adapta cada peso individualmente, actualizándolos con medias y varianza del gradiente, permitiendo una buena estabilidad para un conjunto de grandes datos o parámetros con mucho ruido (Kingma & Ba, 2014). La métrica evalúa el rendimiento del modelo, en este caso usa "mae", que calcula cuánto se equivoca el modelo respecto al promedio sin penalizar los *"outliers"*. Esta decisión hace al modelo más robusto frente a ruido extremo, mientras que, la función de pérdida es "mse" basada en la entropía cruzada, no hace diferencias entre errores pequeños y grandes dentro de la clasificación.

3.6 ESTADÍSTICA

Los modelos de red neuronal creados con el objetivo de detectar fuentes espurias en los catálogos del Planck arrojan probabilidades de que una imagen contenga una fuente verdadera. Las probabilidades no son binarias representando un 0% o 100% (análogo a 0 o 1), si no que pueden comprender cualquier calor entre estas dos probabilidades. Los distintos modelos evalúan los resultados de diferentes maneras. Los "outliers" aunque tengan probabilidades altas de contener una fuente verdadera, pueden representar posibles fuentes espurias, que a priori, el modelo detecta como verdaderas pero que estadísticamente poseen una probabilidad significativamente alejada de la media de la probabilidad del catálogo.

Una forma sensata de abordar las interpretaciones de los modelos es aplicar pruebas de hipótesis con la distribución normal estándar. El estadístico *z* permite ver cuánto se aleja cada probabilidad de la media en función de la desviación estándar, permitiendo así la detección de valores atípicos.

Aunque las predicciones de los distintos modelos no sigan exactamente una distribución normal, es una buena práctica asumiendo que el número de las fuentes es suficientemente grande como para que se pueda aplicar el Teorema Central del Límite. Este teorema establece que si se toman muchas muestras aleatorias de una población, la distribución de los promedios se debe asemejar a una distribución normal aunque está población no siga una distribución normal.

El estadístico *z* se calcula como (Walpole, 2012):

$$z_i = \frac{x_i - \mu}{\sigma} \qquad [10]$$

Si el valor del estadístico *z* de una fuente es inferior a una *z* crítica, se puede determinar que la probabilidad que el modelo a otorgado a esa fuente en concreto es un valor "outlier" del conjunto de probabilidades del catálogo. Esto permite poder clasificar a esa fuente como una posible fuente espuria.

Otra forma de encontrar "outliers" si la distribución de las probabilidades no sigue una distribución normal es hacer una detección de fuentes falsas en función de los cuartiles. Una fuente se puede interpretar como "outlier" si su probabilidad 'x' para un modelo cumple (Montgomery, 2020):

$$x < Q_1 - 1{,}5 \cdot (Q_3 - Q_1) \qquad [11]$$

Al igual que en el caso anterior, si una fuente tiene un valor de la probabilidad 'x' que cumple la ecuación superior, se puede catalogar como un "outlier", y entonces, como posible fuente espuria.

Capítulo 4. METODOLOGÍA

La simulación de mapas del fondo cósmico de microondas ha sido realizada en Python. El módulo PySM3 se ha usado para la simulación de mapas del cielo a diferentes frecuencias incluyendo diferentes componentes contaminantes (Pan-Experiment Galactic Science Group, 2025; Thorne et al., 2017; Zonca et al., 2021). La simulación de mapas se compagina con el módulo Healpy de HEALPix que pixela el mapa en función de la resolución que se requiera. Además de poder realizar otras acciones como visualizar, almacenar y manipular los mapas en formato 'fits' (K.M. Górski, 2005; Zonca et al., 2019).

4.1 SIMULACIÓN DE MAPAS

La biblioteca de PySM3 permite generar simulaciones de los mapas del cielo a diferentes frecuencias mediante una apilación de plantillas o "templates". Estas plantillas se construyen a partir de los datos de observaciones reales (principalmente datos de Planck), para poder simular los diferentes componentes de emisión que hay en el cielo, pudiendo así extrapolar la emisión difusa observada a otras frecuencias.

Los mapas del cielo que se han simulado en este trabajo incluyen los componentes de: fondo cósmico de microondas, emisión térmica de polvo y emisión de sincrotrón; para poder recrear de forma realista el cielo a diferentes frecuencias. Al generar las simulaciones se manejan dos parámetros clave:

- El parámetro *Nside* se relaciona con el número de pixeles que contiene el mapa, siendo fundamental para la resolución angular de cada píxel, tal que:

$$N_{pixeles} = 10 \cdot N_{side}^2 \qquad [12]$$

  Para este trabajo se han simulado diferentes mapas con diferente cantidad de número de pixeles:

Nside	Resolución angular (arcmin)
256	13,7
512	6,87
1024	3,44

Tabla 1: Resolución angular en arcominutos, por píxel en función del *Nside* del mapa simulado.

- El *Full Width at Half Maximum* o *fwhm* sirve para ajustar de forma más realista la resolución angular de cada píxel. Se aplica un suavizado gaussiano que difumina los detalles más nítidos. Este suavizado refleja la calidad de resolución angular máxima que poseen los instrumentos de las misiones espaciales.

En relación con los modelos de cada componente, el polvo térmico está basado en un cuerpo negro modificado, usando plantillas del análisis de Planck – 2015 procesados con Commander.

Por otro lado, la de emisión del sincrotrón se rige por un modelo de ley de potencia [6] que depende de la frecuencia y del índice espectral, que a su vez depende de la dirección. Finalmente, para el CMB, el modelo de simulación tiene en cuenta el efecto de lente gravitacional (Pan-Experiment Galactic Science Group, 2025; Thorne et al., 2017; Zonca et al., 2021).

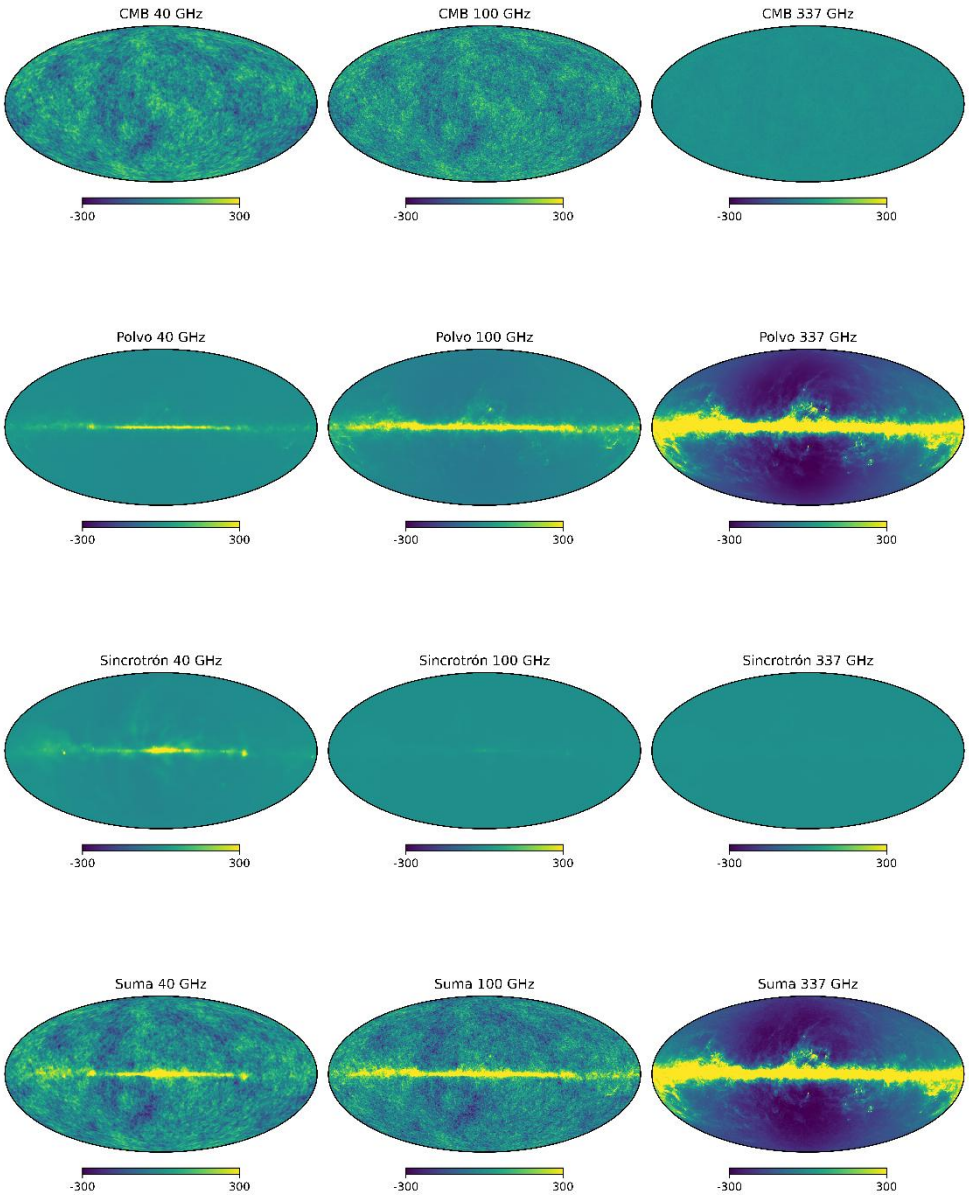

Figura 4: Simulaciones de 3 de las 15 bandas que tendrá el telescopio LiteBIRD. Mapas de Nside = 512. Por columnas, las diferentes frecuencias: 40 GHz (LFT), 100 GHz (MFT) y 337 GHz (HFT), con suavizado fwhm: 60, 29 y 20 arcominutos (Ghigna et al., 2024) Por fila, los componentes que se simulan: CMB, polvo térmico, emisión de sincrotrón y la suma de los tres componentes. Para una mejor visualización se ha filtrado el monopolo y el dipolo, aparte el rango de temperaturas para todos los mapas es [-300,300] microKelvin.

En la Figura 4, se observan las contribuciones de los diferentes componentes a los mapas del cielo. La señal del CMB y del sincrotrón es más potente a frecuencias bajas, siendo a frecuencias

alta como 337 GHz casi imperceptible. Mientras, la emisión térmica del polvo aumenta con la frecuencia como se puede ver en la segunda fila.

Los mapas en los que se visualiza el CMB de forma más eficiente con menos contaminación son a una frecuencia baja y media, tal y como indica la Figura 2. La contaminación a estas frecuencias es principalmente galáctica, por lo que aplicando una máscara como puede ser GAL070 (máscara que ocupa el 30% del cielo, específicamente en la zona galáctica) mejora la visualización limpia de la señal del CMB (Planck Collaboration, 2018)

4.2 MODELOS DE RED NEURONAL

Se han construido dos modelos diferentes para la evaluación y predicción. Ambos modelos se han entrenado con parches del catálogo de fuentes compactas a 30 GHz de Planck (PCCS2) (European Space Agency, 2015a). Estos parches (1.559 específicamente) se han hecho centrados en la fuente y recortados con una proyección gnomónica, permitiendo trabajar la geometría esférica de los mapas en un plano cartesiano, convirtiendo las trayectorias geodésicas en líneas rectas de la imagen. Este método resulta óptimo para que la red neuronal pueda analizarlas eficientemente (Kluger et al., 2017).

Todos los parches tienen un tamaño de 32x32 pixeles, siendo realizados sobre el mapa del cielo de frecuencia de 30 GHz NS1024, que posee un *Nside* = 1024 y un *fwhm* = 32,29 arcmin. (Collaboration, 2020a). Con este número de píxeles, cada parche captura 1,8º por lado del cielo.

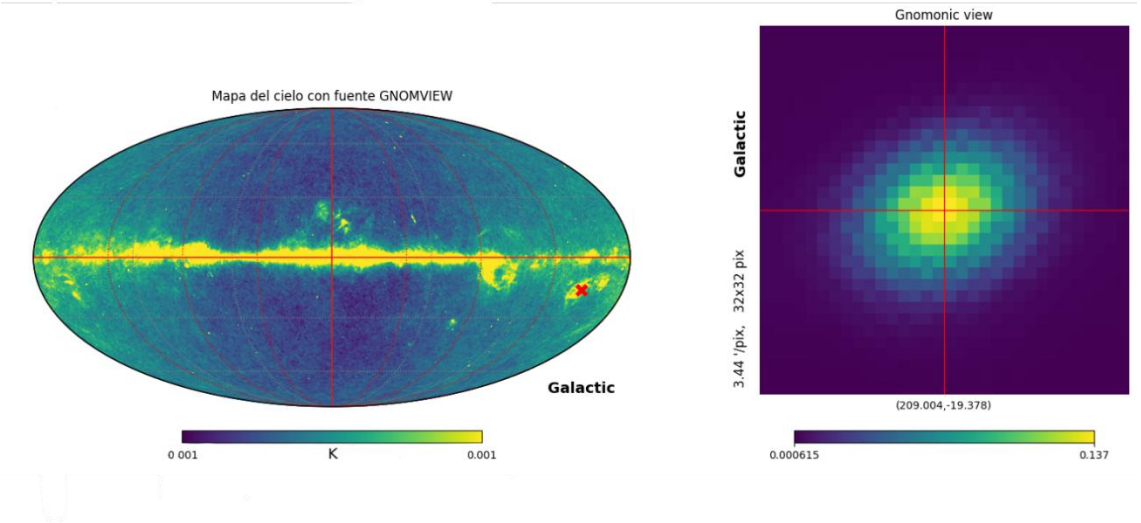

Figura 5: A la izquierda el mapa de Planck NS1024 a 30 GHz con un suavizado de 32,29 arcmin, con el monopolo y dipolo removidos. El rango de temperaturas es [-0,001 ;0,001] Kelvin. Con una X se marca el "cutout" hecho en el mapa que se puede observar en la parte de la derecha. El parche se realiza sobre una fuente del subcatálogo PCNThs, localizada en las coordenadas galácticas que se encuentran debajo del propio recorte.

4.2.1 Modelo A

El primer modelo se ha entrenado solamente con la información de las imágenes 2D. La red neuronal consiste en una apilación de capas convolucionales 2D que aplican filtros de tamaño 3x3 pixeles y capas que reducen las dimensiones espaciales para extraer las características más

dominantes. Posteriormente los mapas se convierten en vectores de una dimensión para aplicar un par de capas densamente conectadas. Finalmente, la capa de salida cuenta con 8 valores que son: Probabilidad de que la fuente sea positiva, centro de la imagen en dos dimensiones (X,Y), posición del píxel más brillante en dos dimensiones (X,Y), el flujo de la fuente y su ruido asociado; todos ellos expresados como valores comprendidos entre 0 y 1.

Posteriormente, el modelo se compila con el optimizador "adam", la función de pérdida "mse" y la métrica "mae", siguiendo el mismo ejemplo que se ha explicado en el apartado Compilación del modelo. El código del modelo A se puede visualizar en el Anexo 10.1.

En resumen, el modelo A toma una imagen, extrae toda su información con las capas convolucionales, las aplana, y luego la información convertida en vector unidimensional pasa por capas densas para generar una salida de 8 valores, útiles para la detección de la fuente y el posicionamiento de esta dentro del parche evaluando hasta un total de 281.544 parámetros. Cabe destacar, que aunque la capa de salida devuelve un vector que contiene 8 valores, solo se han tratado de predecir: la probabilidad de encontrar una fuente verdadera en la imagen y la posición del píxel más brillante.

4.2.2 Modelo B

El segundo modelo aparte de las imágenes 2D, también se ha entrenado con los valores del flujo y su error asociados a las fuentes. Esta decisión se toma ya que, al hacer un catálogo, una condición importante para incluir una fuente es que su proporción de señal/ruido sea más alta que un umbral (Planck Collaboration, 2018). La arquitectura de este modelo contiene las mismas capas convolucionales 2D que el anterior, pero se le añade una capa densa que adjunta los datos correspondientes al flujo y su error. Después, se aplana la salida de la imagen para combinarla con la salida de los datos tabulares, permitiendo que la red use tanto información visual como contextual. Con el conjunto de datos comprimidos en un solo vector 1D, se pasa por otra capa densa.

Finalmente, hay dos salidas simultáneas, la primera devuelve la probabilidad de encontrar una fuente verdadera en la imagen, mientras que la segunda salida predice la posición de la fuente en las coordenadas cartesianas (X,Y).

El modelo B se compila con el optimizador "adam". Para la salida de detección de fuente se usa como función de pérdida "binary_crossentropy", usada en clasificación binaria, mientras que la métrica usada fue "accuracy" que proporciona el número de predicciones correctas. Por el contrario, para la segunda salida se usa la función de pérdida "mse" y la métrica "mae", como en el modelo A. El código del modelo B se puede visualizar en Anexo 10.1.

También se compiló el modelo A con la función de pérdida "binary_crossentropy", sin embargo no se encontraron cambios significativos en las predicciones si se realizaba con la función de pérdida "mse".

En resumen, el modelo B detecta si la fuente de la imagen es espuria y estima su posición (coordenadas X, Y), utilizando tanto la imagen como los datos numéricos complementarios siendo analizados través de 167.603 parámetros.

4.3 ENTRENAMIENTO DE LOS MODELOS

El objetivo de los modelos ha sido detectar las fuentes falsas en el catálogo multifrecuencial de fuentes no térmicas del Planck (PCNT) (European Space Agency, 2015a; Planck Collaboration, 2018)

Para ello, cada uno de los modelos se ha entrenado con dos conjuntos de entrenamiento. Esto permitirá comparar la eficacia en función no solo de la arquitectura del modelo, sino también del tipo de datos de entrenamiento.

Para entrenar los modelos, todos los parches se han realizado sobre el mapa del cielo de frecuencia de 30 GHz NS1024 (European Space Agency, 2015b).

4.3.1 Modelo A1

Este modelo A se ha entenado solamente con el conjunto de parches del catálogo PCCS2 a 30 GHz (European Space Agency, 2015a). Al tomarse todas las fuentes de este catálogo como verdaderas es posible que este modelo esté totalmente sesgado hacia una detección positiva de las fuentes. El modelo se entrena con un "subset" de 1.247 imágenes en un solo canal (escala de grises).

4.3.2 Modelo A2

Este modelo A se entrena utilizando un conjunto de datos compuesto por dos tipos de parches: la primera mitad corresponde a los "cutouts" del cielo que contienen las fuentes del catálogo PCCS2 a 30 GHz (European Space Agency, 2015a), como el modelo anterior; mientras que la otra mitad está compuesta por parches seleccionados con coordenadas aleatorias. Estos últimos parches se han realizado centrados en coordenadas aleatorias, con una mínima de un grado a las coordenadas de cualquier fuente del catálogo PCCS2. Este valor de separación mínima asegura tener un conjunto de datos negativos confiables, y por lo tanto, eliminar un posible sesgo en el entrenamiento del modelo. La cantidad de píxeles y la resolución de estos parches 'falsos' son la misma que tienen los parches centrados en fuentes del catálogo PCCS2. Este modelo se entrena con un "subset" de 2.495 imágenes en un solo canal.

4.3.3 Modelo B1

Este modelo B se ha entrenado con el mismo conjunto de entrenamiento descrito para el modelo A1, es decir, fuentes centradas en las coordenadas del catálogo PCCS2 a 30 GHz (European Space Agency, 2015a). Este modelo se diferencia de A1 en que las imágenes se evalúan también con los valores del flujo y error asociados. El modelo se entrena con un "subset" de 1.247 imágenes en un solo canal.

4.3.4 Modelo B2

Este modelo B se entrena utilizando el mismo "subset" de entrenamiento que el modelo A2, pero añadiendo el valor del flujo y error asociados a cada imagen. Sin embargo, como este conjunto de entrenamiento contiene parches con regiones que no contienen fuentes, el valor del flujo y su error se ha determinado como '-1', ya que es un valor imposible para este tipo de característica. Este modelo se entrena con un "subset" de 2.495 imágenes en un solo canal .

Capítulo 5. RESULTADOS

5.1 Análisis de los modelos

Se analizan los modelos de predicción en base a sus resultados tras ser entrenados con el conjunto de datos de prueba del catálogo PCCS2. La evolución del error cuadrático medio entre los conjuntos de datos de entrenamiento y prueba durante el entrenamiento del modelo sirve para comprobar que el modelo no está sobreajustado. Estos modelos presentan un rendimiento muy alto en el set de entrenamiento, mientras que frente a grupos de datos nuevos el porcentaje de predicción realizada correctamente desciende, siendo poco robusto para los conjuntos de datos de prueba. Para mejorar esto, todos los modelos se han entrenado realizando 20 repeticiones en su entrenamiento. El modelo que se guarda finalmente es el que posee un menor valor de pérdida en el conjunto de prueba, demostrando una validación óptima sin sobreajuste sobre los datos de entrenamiento. Las gráficas se pueden consultar en: Figura 23 y Figura 24, situadas en el Anexo 10.3.

5.1.1 Modelo A1

Este modelo ha sido entrenado solamente con parches que contienen fuentes. Aunque el modelo esta sesgado, el objetivo es detectar fuentes espurias de un subset de un catálogo que contiene fuentes con una relación S/R > 4, es decir, que son las fuentes más brillantes del catálogo de fuentes compactas no térmicas en todas las frecuencias. Por lo que entrenar un modelo con este conjunto de datos sesgados podría predecir mejor los outlayers, que serán fuentes que tienen más probabilidades de ser fuentes falsas.

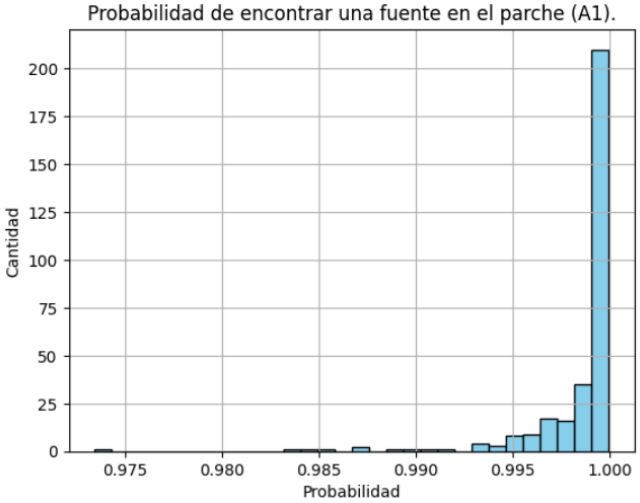

Figura 6: Probabilidad de encontrar una fuente no térmica en cada parche en el set de entrenamiento con el modelo A1.

En la gráfica superior se observan altas probabilidades generalizadas, el modelo A1 da una probabilidad mínima de encontrar la fuente en el parche por encima del 97%. Aunque a priori parece que el modelo está prediciendo correctamente el conjunto de datos de prueba, se encuentran unos pocos “outliers” que poseen una probabilidad desplazada de la media.

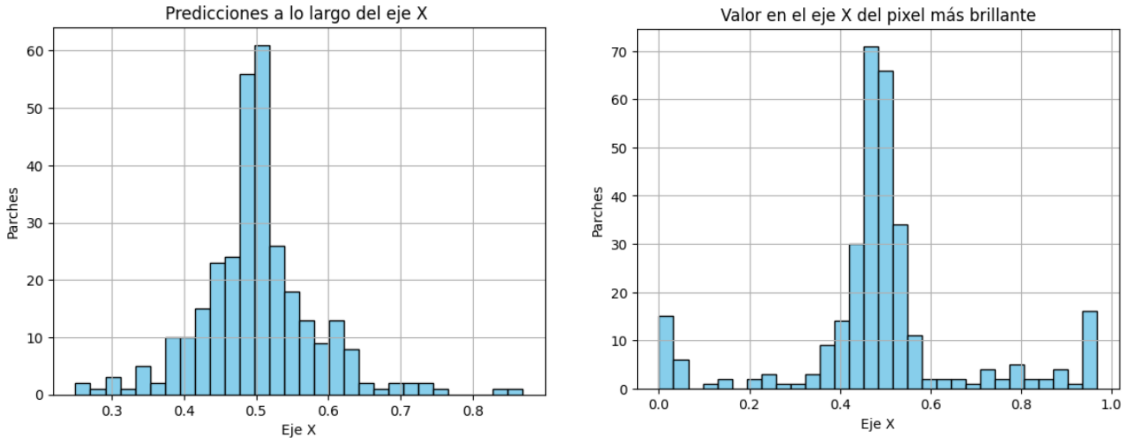

Figura 7: A la izquierda, la predicción de A1 sobre el conjunto de prueba de la posición del píxel más brillante a lo largo del eje X. A la derecha, el valor real sobre el eje X del píxel más brillante del conjunto de prueba.

Para comprobar que A1 realmente está encontrando las fuentes centradas en el parche se grafica la predicción de encontrar el píxel más brillante (la fuente) a lo largo de los ejes X e Y.

Como se observa en la gráfica que muestra la predicción de encontrar al píxel más brillante a lo largo del eje X y la que enseña el valor real, hay una concentración de parches con la fuente centrada. Este es un buen síntoma de que realmente las fuentes están centradas y de que la red neuronal las está identificando correctamente. Esto también pasa en el eje Y, como se puede observar:

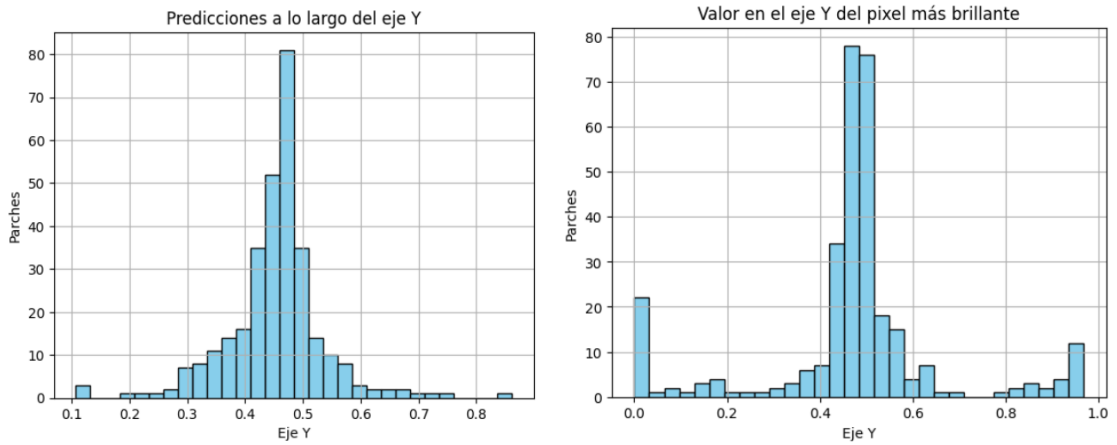

Figura 8: A la izquierda, la predicción de A1 sobre el conjunto de prueba de la posición del píxel más brillante a lo largo del eje Y. A la derecha, el valor real sobre el eje Y del píxel más brillante del conjunto de prueba.

En las gráficas que muestran los valores reales del píxel más brillante, se observa una concentración de parches en ambos extremos de los ejes, es decir 0 y 1. Esto puede ser debido a que en algunos parches existe la posibilidad de que haya una segunda fuente de emisión aún más brillante que la fuente principal del catálogo, representa así un segundo píxel. Es probable que esa segunda fuente no provenga del catálogo PCCS2, sino de otras fuentes de emisión, como el CMB o el sincrotrón. En cambio, en las gráficas que predicen el valor, esa acumulación en los extremos

desaparece. Este suceso puede estar debido al sesgo del entrenamiento, ya que la mayoría de los parches contienen una fuente en el centro. Esto provoca que la red prediga posiciones cercanas al centro, ya que esa es la posición que minimiza el error.

5.1.2 Modelo A2

Este modelo se entrena con un set compuesto por imágenes que poseen fuentes y por imágenes aleatorias que no la contienen. Esto permite al modelo reconocer una fuente y separar las imágenes en dos grupos. Las imágenes que son cercanas al 0% de probabilidad deben ser todas, o al menos en su gran mayoría, fuentes aleatorias; mientras que las imágenes cercanas al 100% de probabilidad deben ser las fuentes del catálogo.

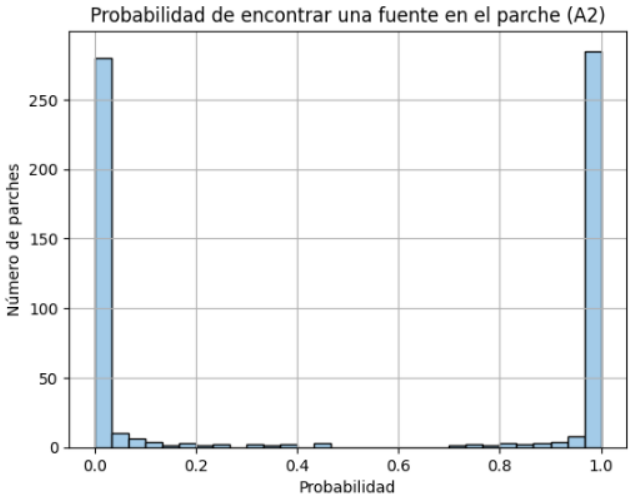

Figura 9: Probabilidad de encontrar una fuente no térmica en cada parche en el set de entrenamiento con el modelo A2.

Se puede observar que, aunque existen fuentes con probabilidades intermedias, la red neuronal ha logrado agrupar eficazmente las imágenes en dos categorías diferenciadas: aquellas que contienen una fuente y aquellas que no. Esta separación sugiere que el modelo ha aprendido a distinguir entre ambas clases. Para verificar que esta clasificación no es resultado de una separación aleatoria, se evalúa el rendimiento del modelo utilizando los parámetros derivados de la matriz de confusión. En este análisis, todas las imágenes a las que el modelo asigna una probabilidad inferior a 0,5 se consideran como pertenecientes a la clase '0' (sin fuente), mientras que aquellas con una probabilidad superior se clasifican como clase '1' (con fuente). La matriz de confusión correspondiente puede visualizarse en la Figura 22 localizada en el Anexo 10.2.

Las métricas de rendimiento obtenidas han sido:

Modelo A2	Umbral al 50%	Umbral al 80%
Accuracy	0,9824	0,9856
Precision	0,9773	0,9869
Recall	0,9869	0,9837
F1-score	0,9821	0,9853

Tabla 2: Métricas de la matriz de confusión del modelo A2.

Dado que el objetivo principal es identificar las fuentes falsas dentro del catálogo PCNT, resulta fundamental maximizar el valor de la "Precision". Esta métrica indica qué proporción de las predicciones positivas corresponde realmente a fuentes verdaderas. Si el umbral de detección se sitúa en el 50%, el modelo presenta un error promedio del 2,3 %, lo que significa que por cada 100 fuentes clasificadas como verdaderas, aproximadamente 2,3 son en falsos positivos. En un catálogo que contiene miles de fuentes, esto podría traducirse en varias decenas o incluso centenas de entradas incorrectas que afectarían a los parámetros cosmológicos.

No obstante, también es importante considerar que es preferible clasificar erróneamente una fuente falsa como verdadera antes que omitir una fuente verdadera. Por esta razón, es necesario mantener un buen nivel de "Recall", que mide la capacidad del modelo para identificar correctamente las fuentes reales como es el caso de este modelo obteniendo 0,9869. En consecuencia, debe buscarse un equilibrio entre estas métricas, lo que se refleja en el valor de "Accuracy", siendo 0,9824.

Si para este modelo se escoge como umbral el 80% de probabilidad para separar entre las imágenes con y sin fuente, el modelo maximiza la métrica de acierto global, "Accuracy" al 0,9856 y la métrica del equilibrio armónico "F1-score" al 0,9853.

5.1.3 Modelo B1

Este modelo otorga a la gran mayoría de las imágenes una probabilidad mayor al 99,95 % de contener una fuente no térmica. Al igual que pasa con el modelo A1, esto se debe a que el modelo está totalmente sesgado hacia una detección positiva de las fuentes. Se puede destacar que el modelo B1 está más 'confiado' con sus predicciones que el modelo A1, ya que aunque coincidan en la forma de la distribución, las probabilidades generales que da este modelo son más cercanas a 1. Esta confianza proviene de la información numérica con la que se ha entrenado el modelo, la relación señal/ruido que poseen todas las fuentes del catálogo PCCS2. Como todas las fuentes tienen un valor SNR > 3, la red neuronal encuentra está relación y empieza a sesgar en función de la relación. La distribución de las probabilidades del modelo B2 se puede visualizar en la gráfica de la siguiente página:

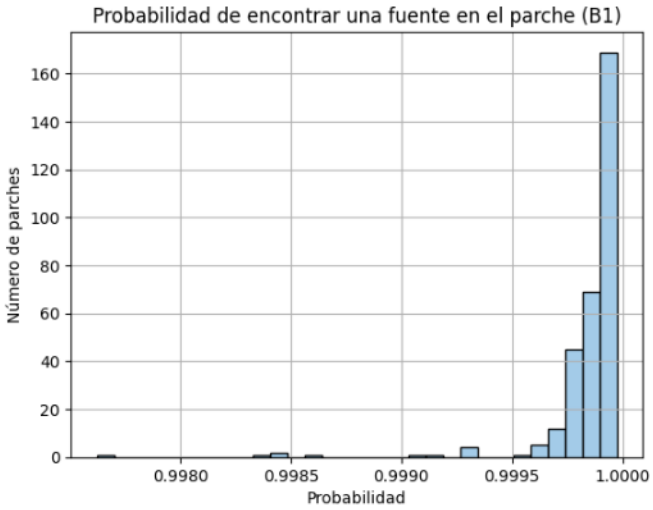

Figura 10: Probabilidad de encontrar una fuente no térmica en cada parche en el set de entrenamiento con el modelo B1.

5.2 Modelo B2

El modelo B2 se ha entrenado con la información de las imágenes y valores del flujo de las fuentes del catálogo PCCS2 a 30 GHz, y con imágenes aleatorias del mapa del cielo a 30 GHz. Sin embargo, como se ha mencionado antes, las imágenes aleatorias que no contienen fuente no se les ha podido adjuntado un valor de flujo ni error asociado que sean realistas. Se representa la distribución de las probabilidades del modelo B2 en la siguiente figura:

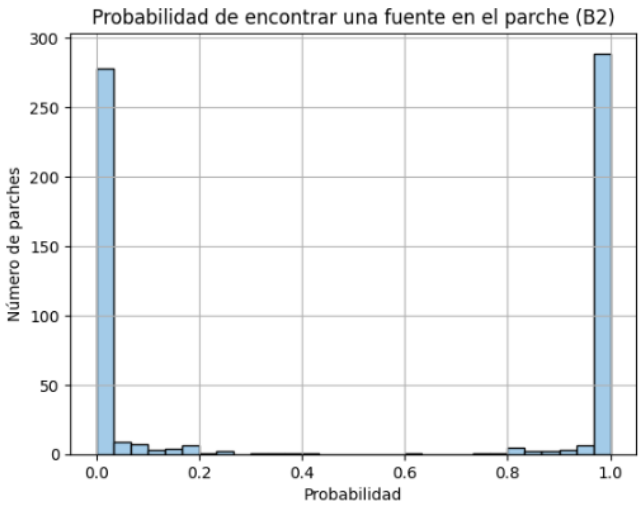

Figura 11: Probabilidad de encontrar una fuente no térmica en cada parche en el set de entrenamiento con el modelo B2.

A priori, la red neuronal ha identificado de forma eficiente que las imágenes que tienen asociada un flujo negativo son las fuentes falsas, mientras que cualquier imagen que tenga asociada un flujo positivo es una fuente verdadera.

Eventualmente, se muestra el rendimiento que ha tenido B2 al detectar las fuentes en las imágenes del set de entrenamiento compuesto por los "cutouts" del catálogo PCCS2 a 30 GHz e imágenes con fuentes falsas.

Como se observa en Figura 11, la distribución de probabilidades es muy parecida a la del modelo A2. Para comprobar cuál de los dos es más eficiente, lo más adecuado es comparar las métricas de rendimiento del modelo. Al igual que con el modelo A2, las imágenes se dividen en dos grupos, a partir del umbral del 50% de probabilidad de contener la fuente y el valor en el cual el modelo alcanza la mayor exactitud. La matriz de confusión asociada al modelo B2 se puede ver en la Figura 21, que se encuentra en el Anexo 10.2.

Modelo B2	Umbral al 50%	Umbral al 87%
Accuracy	0,9840	0,9872
Precision	0,9774	0,9967
Recall	0,9902	0,9771
F1-score	0,9838	0,9868

Tabla 3: Métricas de la matriz de confusión del modelo B2.

Este modelo es levemente más preciso y exacto que el modelo A2, al tratar de predecir y detectar las fuentes verdaderas y falsas. Aunque los dos presentan bastantes buenos resultados, los modelos deben estar preparados no solo para detectar imágenes que no contienen fuentes, sino en detectar fuentes que han sido catalogadas como fuentes verdaderas, ya sea por su forma o por su relación señal/ruido, y que en realidad sean fuentes espurias.

5.3 DETECCIÓN DE FUENTES FALSAS EN EL CATÁLOGO PCNTb

Para detectar las posibles fuentes espurias dentro del catálogo PCNTb se han procesado las imágenes en tres de los cuatro modelos entrenados. Las fuentes del PCNTb podrían haberse confundido con emisión térmica de polvo o con estructuras difusas que casualmente cumplían con las exigencias del catálogo.

5.3.1 Detección con A1

El modelo confirma la existencia de una fuente brillante en todos los parches con una confianza mayor al 95%. De las 1.424 fuentes solo 83 tienen una probabilidad inferior al 99,5%, de las cuales 33 también son inferiores al 99%.

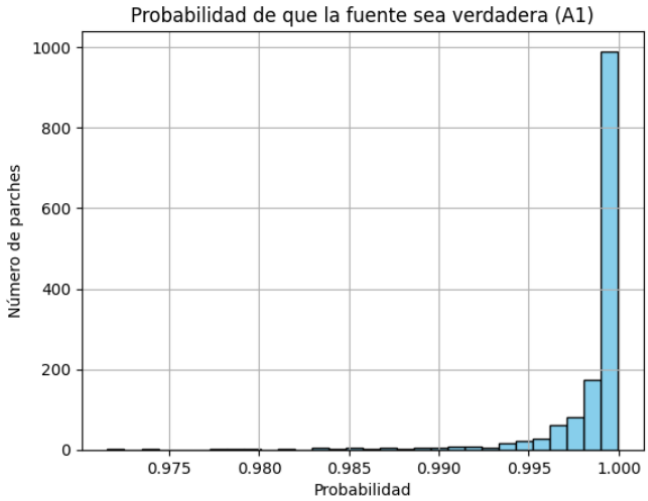

Figura 12: Probabilidad de encontrar una fuente no térmica en el subcatálogo PCNTb con el modelo A1.

Aunque el modelo no le atribuye a ningún parche una probabilidad baja de contener una fuente verdadera, esto no significa que no haya ninguna, ya que se debe recordar que este modelo está positivamente sesgado.

Se determina si una fuente tiene una probabilidad "outlier" haciendo dos pruebas estadísticas. La primera, es una prueba de hipótesis estadística, donde se supondrá que la distribución de probabilidades del modelo A1 es una distribución normal. La hipótesis nula H_o es que la imagen contiene una fuente verdadera perteneciente al catálogo PCNTb, mientras que la hipótesis alternativa H_1 afirma que la imagen tiene una probabilidad significativamente menor de pertenecer al catálogo PCNTb, que la probabilidad media del modelo A1 de las fuentes que conforman el catálogo. Por lo tanto, si la fuente cumple la hipótesis alternativa H_1 se puede considerar como "outlier" y posible fuente espuria.

La media del conjunto de probabilidades es del 99,86%, mientras que la desviación estándar es 0,27%. Calculando el estadístico z para todas las fuentes con una confianza del 95%, es decir $\alpha = 0{,}05$ y un valor crítico asociado $z < -1{,}96$, se encuentran un total de 55 fuentes con un valor del estadístico z menor que el crítico.

Por otro lado, se realiza una segunda prueba estadística basada en cuartiles. A diferencia de la prueba de hipótesis esta se puede calcular aún sin asumir que la distribución de probabilidades del modelo A1 sigue una distribución normal. Usando la ecuación [11], se obtiene una probabilidad umbral del 99,71%, y por tanto, 176 posibles fuentes espurias que tienen una probabilidad inferior a la probabilidad umbral.

Los resultados del modelo A1 al realizar las predicciones sobre el conjunto de fuentes del catálogo PCNTb se resumen en la siguiente tabla:

Fuentes espurias	Umbral al 99%	Umbral al 99,5%	Hipótesis	Cuartiles
Cantidad	33	83	55	176
Porcentaje	2,32%	5,83%	3,86%	12,36%

Tabla 4: Fuentes espurias que detecta el modelo A1 para el catálogo PCNTb en función del umbral, y los métodos estadísticos de hipótesis y cuartiles.

5.3.2 Detección con A2

Este modelo detecta un elevado número de fuentes espurias con nula posibilidad de ser verdaderas. La gran mayoría tienen una alta probabilidad de contener una fuente no térmica, mientras que, hay unas pocas fuentes con las que tiene incertidumbre, otorgándolas probabilidades intermedias.

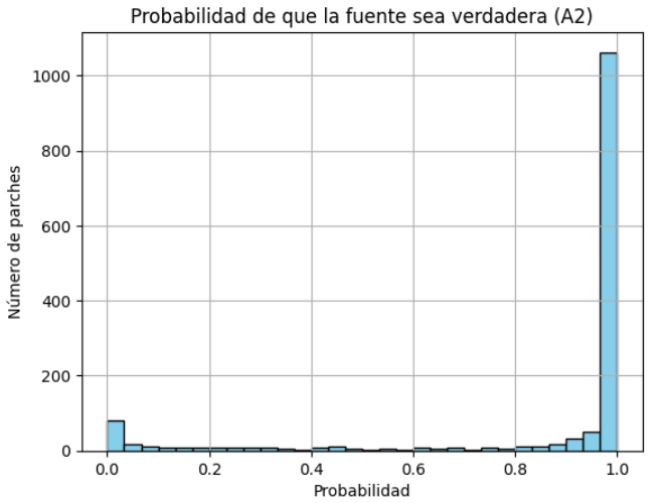

Figura 13: Probabilidad de encontrar una fuente no térmica en el subcatálogo PCNTb con el modelo A2.

Si el umbral de detección positiva se sitúa en el 50% de probabilidad, este modelo detecta 193 fuentes una probabilidad inferior al umbral, o sea posibles espurias. Si el umbral de confirmación se sitúa en el 80% de probabilidad, valor que maximizaba la exactitud del modelo, entonces A2 detecta un total de 240 fuentes falsas. Respecto a las 1.424 fuentes que conforman el catálogo PCNTb, el modelo como mínimo detecta que el 13,5% de las fuentes son espurias.

Comparando con el modelo anterior, A2 otorga una confianza menor al 99% a 982 fuentes de las 1.424 que hay en el catálogo PCNTb.

La clasificación del modelo A2, se resumen en la siguiente tabla:

Fuentes espurias	Umbral 50%	Umbral 80%	Inferior al 99%
Cantidad	193	240	982
Porcentaje	13,55%	16,85%	68,96%

Tabla 5: Fuentes espurias que detecta el modelo A2 para el catálogo PCNTb en función del umbral, y cantidad de fuentes con probabilidad superior al 99% de ser verdaderas.

Aunque este modelo no asigna probabilidades extremadamente altas como lo hacía el modelo anterior, mantiene una gran capacidad para identificar la presencia de una fuente en la imagen. En la siguiente figura se presentan las fuentes con las mayor y menor probabilidad (100% y 0,005 %) de contener una fuente verdadera según el modelo A2.

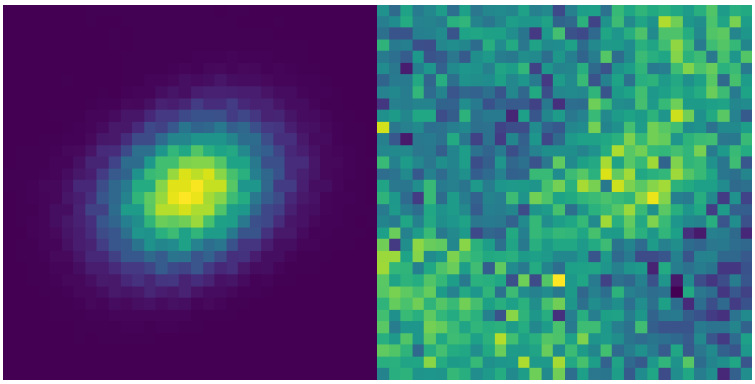

Figura 14 A la izquierda, la fuente a la que el modelo A2 asigna la mayor probabilidad de ser una fuente verdadera; a la derecha, la fuente con la probabilidad más baja según el mismo modelo.

Interpretar por qué el modelo clasifica una imagen con menor o mayor probabilidad es complejo debido a que una red neuronal es un modelo de tipo caja negra. No obstante, sí es posible analizar cómo evalúa los parámetros asociados a las fuentes, como la relación señal/ruido (SNR). Tal y como se observa en la Figura 25 situada en el Anexo 10.4, existe una clara correlación: a mayor SNR, mayor es la probabilidad que el modelo asigna a que la fuente sea verdadera. Este comportamiento es esperable, ya que una mayor SNR implica una fuente más nítida y, por lo tanto, una morfología mejor definida. Sin embargo, por debajo de un cierto umbral (aproximadamente un valor de 10, según la gráfica), la distribución de probabilidades se vuelve más uniforme, siendo posible que el modelo por debajo de este umbral de SNR, base su decisión en la imagen y no en los datos adjuntados.

5.3.3 Detección con B.1

La distribución de las predicciones del modelo B1 es muy parecida a la de A1. Esto se debe principalmente a que han sido entrenados con el mismo set de imágenes, sin embargo, se puede observar que este modelo está más "confiado" al hacer sus predicciones.

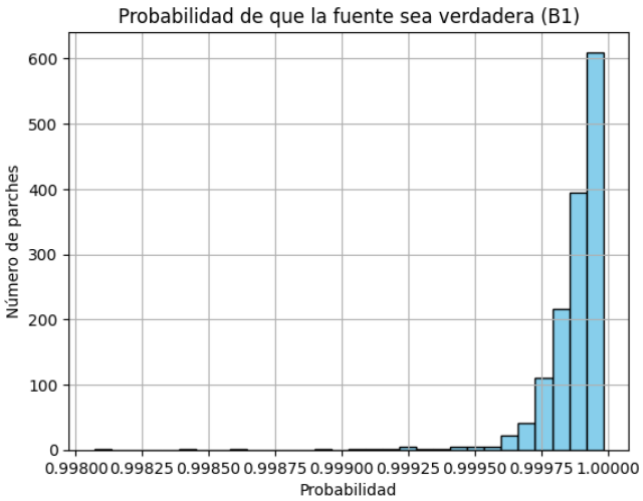

Figura 15: Probabilidad de encontrar una fuente no térmica en el subcatálogo PCNTb con el modelo B1.

A diferencia del modelo A1, este modelo otorga una probabilidad de contener una fuente verdadera menor al 99,9% a solo 4 parches. Esto resulta una gran diferencia respecto a los otros dos modelos. Esto se puede deber a que todos los parches contienen una relación señal/ruido mayor a 4.

Aunque el modelo no le atribuye a ninguna imagen una probabilidad baja de ser una fuente verdadera, los "outliers" pueden ser interpretados como fuentes falsas al estar alejadas de la distribución normalizada unilateral de las probabilidades.

Se realiza el mismo procedimiento que se ha llevado a cabo para el modelo A1. La media de la probabilidad de la población es del 99,99%, mientras que la desviación estándar es del 0,01%. Calculando el estadístico z para todas las fuentes con una confianza del 95%, y teniendo en cuenta que el valor crítico es $z = -1{,}96$, se obtienen 37 fuentes sospechosas que se pueden considerar "outliers". Mientras tanto, la detección de "outliers" por el método estadístico de cuartiles, muestra que hay 59 fuentes espurias, que no han superado el umbral del 99,97% de probabilidad. La detección de posibles fuentes espurias según el modelo B1 se puede resumir en la siguiente tabla:

Fuentes espurias	Umbral al 99,9%	Umbral al 99,95%	Hipótesis	Cuartiles
Cantidad	4	25	37	59
Porcentaje	0,28%	1,76%	2,60%	4,14%

Tabla 6: Fuentes espurias que detecta el modelo B1 para el catálogo PCNTb en función del umbral, y los métodos estadísticos de hipótesis y cuartiles.

5.4 DETECCIÓN DE FUENTES FALSAS EN EL CATÁLOGO PCNThs

Para verificar el rendimiento de los modelos al tratar de detectar fuentes espurias, se realiza el mismo procedimiento para el subcatálogo PCNThs. Este el subcatálogo es aún más robusto que el PCNTb ya que sus exigencias son más restrictivas, lo que a priori sugiere, que no debe tener ninguna fuente espuria. Es por esto que, los tres modelos deben predecir inequívocamente la detección de la fuente en todos los parches, o al menos, un porcentaje mucho menor de detección de fuentes espurias que el catálogo PCNTb.

5.4.1 Detección con A1

El modelo A1, igual que para el anterior catálogo, asigna probabilidades que siguen una distribución normal. Se detectan todas las fuentes con una probabilidad mínima cercana al 97,5%. Mientras que solo 22 de las 151 fuentes tienen una probabilidad inferior al 99%.

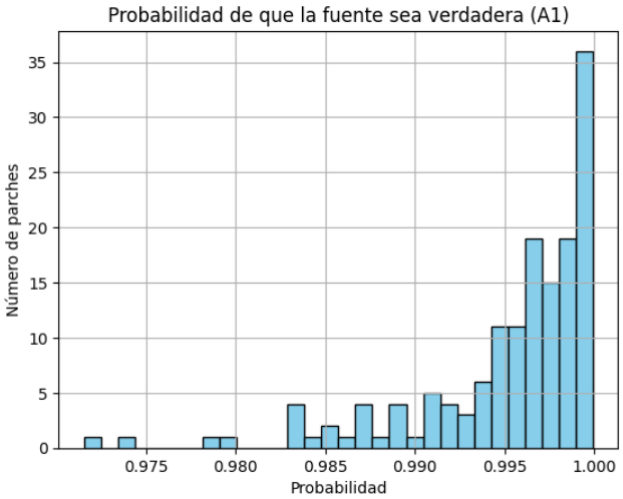

Figura 16: Probabilidad de encontrar una fuente no térmica en el subcatálogo PCNTb con el modelo A1.

Siguiendo con el mismo procedimiento que para el catálogo PCNTb, se calcula la cantidad de "cutouts" que son determinados "outliers" mediante los dos métodos estadísticos. Con la media de las predicciones situada en el 99,52% y una desviación estándar de 0,52%, por el método de hipótesis alternativa H_1 con una confianza del 95%, se han detectado 11 fuentes espurias.

Mediante el análisis de cuartiles, usando la ecuación [11], se obtiene una probabilidad umbral del 98,67%, y por tanto, 12 fuentes dudosas. Las predicciones del modelo A1 sobre el conjunto de fuentes del subcatálogo PCNThs se resumen:

Fuentes espurias	Umbral al 99%	Umbral al 99,5%	Hipótesis	Cuartiles
Cantidad	22	46	11	12
Porcentaje	14,57%	30,46%	7,28%	7,95%

Tabla 7: Fuentes espurias que detecta el modelo A1 para el catálogo PCNThs en función del umbral, y los métodos estadísticos de hipótesis y cuartiles.

5.4.2 Detección con A2

A diferencia de las predicciones sobre el catálogo PCNTb, para este conjunto de fuentes de alta significancia PCNThs el modelo A2 prácticamente no predice ninguna fuente con probabilidad nula de ser verdadera. Esta predicción es totalmente coherente con el catálogo.

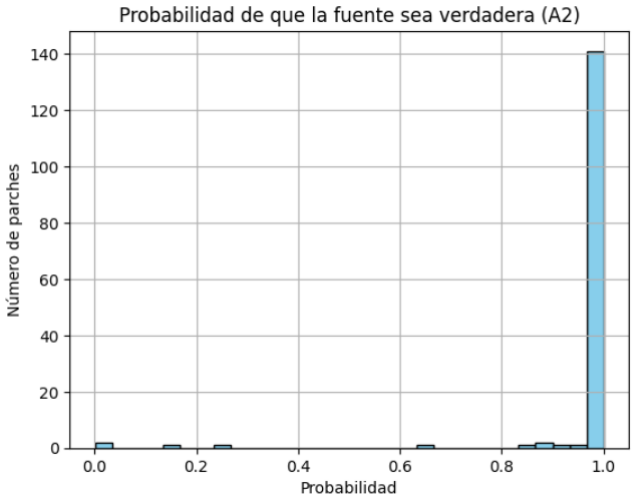

Figura 17: Probabilidad de encontrar una fuente no térmica en el subcatálogo PCNThs con el modelo A2.

Aunque hay algunos parches sobre los que el modelo sigue teniendo incertidumbre, la gran mayoría de parches se aglutinan en una probabilidad alta, siendo coherente con la calidad que contienen los datos del catálogo PCNThs. La distribución de probabilidades del modelo A2 es:

Fuentes espurias	Umbral al 50%	Umbral al 80%	Umbral al 99%
Cantidad	4	5	13
Porcentaje	2,65%	3,31%	8,61%

Tabla 8: Fuentes espurias que detecta el modelo A2 para el catálogo PCNThs en función del umbral.

5.4.3 Detección con B1

El modelo B1, al igual que ocurría con el catálogo PCNTb, predice con la misma estructura que el modelo A1 pero con mayor confianza en sí mismo. En este caso, 139 de las 141 fuentes las atribuye una probabilidad mayor al 99,95%.

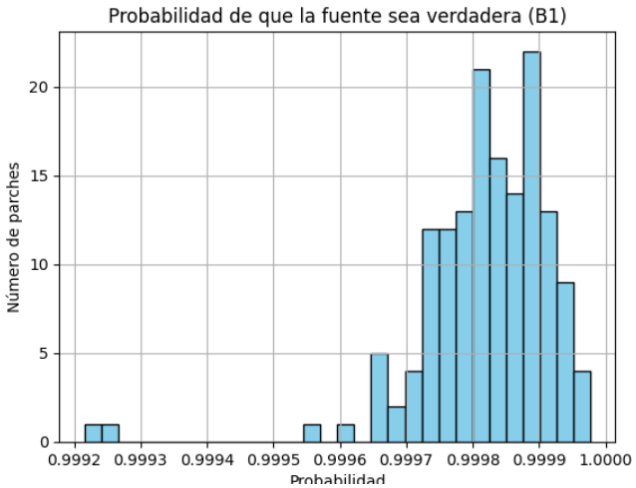

Figura 18: Probabilidad de encontrar una fuente no térmica en el subcatálogo PCNThs con el modelo A1.

La distribución de probabilidades se asemeja a una distribución normal, y además, proporciona probabilidades mucho más altas que las que daba este mismo modelo pero analizando el catálogo PCNTb. La media del conjunto de probabilidad es 99,98%, mientras que la desviación estándar es 0,01%. Del catálogo PCNThs, un total de 4 fuentes cumplen la hipótesis alternativa H_1 con una confianza del 95%. Mientras que, la detección de "outliers" por el método estadístico de cuartiles, muestra que hay 3 fuentes que no superan el umbral obtenido de la ecuación [11] siendo, $x = 99{,}96\%$.

Fuentes espurias	Umbral al 99,9%	Umbral al 99,95%	Hipótesis	Cuartiles
Cantidad	0	2	4	3
Porcentaje	0%	1,32%	2,65%	1,99%

Tabla 9: Fuentes espurias que detecta el modelo B1 para el catálogo PCNTb en función del umbral, y los métodos estadísticos de hipótesis y cuartiles.

Capítulo 6. DISCUSIÓN

Los modelos basados en redes neuronales convolucionales han demostrado tener un gran rendimiento. Se han obtenido los resultados esperados en función del tipo de modelo y del conjunto de datos.

El modelo A1 ha sido el más simple. Durante su entrenamiento, ha determinado que todas las fuentes del catálogo PCCS2 son verdaderas, posteriormente, al evaluar los subcatálogos PCNTb y PCNThs también ha clasificado a todas las fuentes como verdaderas, en ambos casos con una probabilidad mínima del 97%. Sin detectar rigurosamente fuentes espurias, se han encontrado fuentes con probabilidades que se pueden considerar "outliers" según dos métodos estadísticos diferentes, por lo que estas fuentes se pueden considerar como fuentes dudosas.

En principio, el subcatálogo de fuentes brillantes no compactas (PCNTb) contiene las fuentes más brillantes de las casi 30.000 que posee el catálogo principal PCNT. Este subcatálogo es bastante fiable al contener fuentes con una relación señal/ruido superior a 4. Las predicciones que realiza el modelo A1 es coherente con el contenido del catálogo, sin embargo, las altas probabilidades también se deben al sesgo introducido en el conjunto de entrenamiento, que solo contiene fuentes verdaderas.

Para comprobar que este modelo cataloga de forma correcta se probó sobre el catálogo PCNThs. Los resultados se pueden considerar satisfactorios al haber detectado todas las fuentes como verdaderas. Sin embargo, al comparar las predicciones que hace el modelo A1 sobre ambos catálogos (Tabla *4* y Tabla *7*), se observa que el primer conjunto de datos (PCNTb) tiene una menor proporción de fuentes por debajo de los umbrales, es decir, que sean posibles fuentes espurias. Esto es incongruente con la calidad respectiva de cada catálogo, ya que se esperaba que el modelo PCNTb tuviera una mayor proporción de fuentes por debajo de los umbrales que el catálogo PCNThs. Este catálogo, al recoger las fuentes más brillantes en todos los canales del Planck es muy improbable que contenga fuentes falsas, por lo que debería tener muy pocas fuentes clasificadas como espurias, o como mínimo, menos que el catálogo PCNTb. Esto es un indicio de que el modelo, posiblemente debido al sesgo, no clasifica correctamente las fuentes por lo que no es idóneo para la detección de fuentes espurias.

Respecto al modelo B1, los resultados son similares a los de A1. El modelo afirma que todos los parches del catálogo PCNTb contienen una fuente de emisión no térmica, con una probabilidad mínima superior al 99,80 %. La razón por la que el modelo se muestra más seguro que el anterior es por el entrenamiento de las imágenes junto a los datos tabulares. Se ha comprobado con la Figura 25 que a mayor SNR, mayores han sido las probabilidades otorgadas por el modelo, lo cual era esperable. Valorando las detecciones sobre el catálogo PCNThs, solo dos fuentes poseen una probabilidad inferior al 99,95 %, siendo superior a la del catálogo PCNTb. Por último, al comparar las proporciones de fuentes con probabilidades inferiores al umbral, el catálogo PCNThs presenta una menor proporción posibles fuentes espurias que el PCNTb, lo cual vuelve a concordar con el resultado esperado en base a la calidad de ambos catálogos.

El entrenamiento realizado por el modelo A2 incluía un conjunto de imágenes con una fuente del catálogo PCCS2 y otro conjunto de imágenes sin fuente, eliminando el sesgo presente en los modelos A1 y B1. Para el catálogo PCNTb, el modelo detecto 193 posibles fuentes espurias, al situar el umbral al 50% de probabilidad, representando el 13,5 % de las fuentes totales del

conjunto. Este valor puede ser, a priori, demasiado grande para el catálogo PCNTb, ya que son fuentes detectadas a varias frecuencias y es subconjunto razonablemente fiable.

Para comprobar que el modelo A2 está clasificando incorrectamente las fuentes, se realiza el análisis sobre el conjunto de datos del catálogo PCNThs. Los porcentajes de detección de posibles fuentes falsas del modelo para este catálogo (Tabla *8*) es significativamente superior al compararlo con los del catálogo PCNTb de este mismo modelo (Tabla *5*). El modelo solo detecta como fuentes falsas un 2,65 % (posicionando el umbral al 50 %) frente al 13,55 % del catálogo PCNTb. Estos resultados son compatibles con la calidad de ambos catálogos. Aunque pueden seguir siendo levemente altos para ambos casos, parece que el modelo A2 sí está detectando correctamente las fuentes espurias.

A diferencia del modelo B1, las proporciones de detección del modelo A2 se reducen notablemente entre los catálogos PCNTb y PCNThs, mientras que para B1, estás proporciones se reducen mínimamente. Que la proporción de fuentes espurias sea parecida para ambos, puede ser sospechoso, ya que como se ha mencionado anteriormente, los criterios del PCNThs son mucho más restrictivos que para el modelo PCNTb.

Finalmente, no se ha testeado el modelo B2, ya que el conjunto de entrenamiento contenía las fuentes falsas con un flujo y error de flujo de valor '-1'. Al tener todas las fuentes de los catálogos PCCS2 y PCNT un flujo y su error definidos con números positivos, el modelo ha detectado sin error alguno este patrón, haciendo poco útil su uso para el objetivo del trabajo.

En resumen, los modelos A1 y B1 presentan unas predicciones positivas muy altas, sin embargo, esto se puede deber a que son modelos sesgados positivamente. El modelo B1 se ajusta de manera eficiente a la calidad relativa de ambos catálogos, mientras que el modelo A1 presenta mejores resultados para el PCNTb que para PCNThs, lo cual es incongruente con el contenido de cada catálogo. Por otro lado, el modelo A2 no tiene sesgo pero detecta demasiadas fuentes espurias para la calidad que debe tener el catálogo PCNTb. Sin embargo, la diferencia de proporciones de fuentes falsas entre ambos catálogos si presenta un valor que se puede corresponder con las calidades respectivas.

En mi opinión, para determinar si una fuente sospechosa/posible fuente falsa es realmente una fuente espuria, lo más adecuado sería visualizar estas fuentes de forma manual y analizarlas mediante otros algoritmos tradicionales de detección compacta.

Capítulo 7. CONCLUSIONES

7.1 CONCLUSIONES DEL TRABAJO

Este Trabajo de Fin de Grado ha tenido como objetivo principal la detección de fuentes espurias en catálogos astronómicos, concretamente en el catálogo multifrecuencial de fuentes no térmicas (PCNT) de la misión Planck, mediante el uso de CNNs. A lo largo del trabajo, se han simulado mapas del cielo en distintas frecuencias, se han construido y entrenado modelos predictivos, y se ha realizado un análisis estadístico detallado de los resultados obtenidos.

Los cuatro modelos desarrollados han mostrado comportamientos claramente diferenciados en función de su arquitectura y del conjunto de datos de entrenamiento utilizados. En particular, el modelo A2, entrenado con un conjunto balanceado de imágenes, ha demostrado ser eficaz en la detección de posibles fuentes espurias. Este modelo identificó un total de 193 fuentes sospechosas dentro del catálogo PCNTb, una cifra notablemente elevada dado el carácter del catálogo. En cambio, al aplicarse sobre el PCNThs, aún más fiable, la tasa de detección descendió a un 2,4%, concordando con las expectativas. Por el contrario, los modelos A1 y B1, entrenados exclusivamente con fuentes reales, evidenciaron un sesgo hacia la clasificación positiva de las fuentes, reduciendo su capacidad de detección. Al confirmar todas las fuentes de ambos catálogos como positivas se aplicaron métodos estadísticos para encontrar las fuentes con probabilidades más alejadas de la media, detectando “outliers” y por tanto, fuentes dudosas o posibles fuentes espurias.

Los resultados obtenidos abren una puerta al uso de redes neuronales como una herramienta adicional a las tradicionales para comprobar la fiabilidad de los catálogos astronómicos.

7.2 CONCLUSIONES PERSONALES

Respecto al trabajo, aunque los resultados de los modelos han sido satisfactorios, la detección de posibles fuentes espurias en los catálogos debe complementarse con métodos tradicionales. Por su complejidad y su impacto en la interpretación cosmológica, creo que una revisión manual y visual de las fuentes clasificadas como dudosas podría ser también una buena alternativa además de los métodos clásicos.

Personalmente, este trabajo ha representado un gran desafío académico, una experiencia de aprendizaje muy enriquecedora mediante la que he podido integrar los conocimientos que he adquirido en la carrera en un proyecto real con aplicaciones directas. Trabajar con datos reales del satélite Planck y poder construir modelos funcionales desde cero ha sido especialmente motivador, cumpliendo con mis expectativas sobre mi punto de vista acerca de esta carrera científica.

Capítulo 8. FUTURAS LÍNEAS DE TRABAJO

A partir de los métodos aplicados y resultados obtenidos, se abren varias líneas de trabajo que podrían complementar y/o mejorar la investigación iniciada en este proyecto:

- Exploración y detección de fuentes con modelos más avanzados como redes residuales (ResNet), redes neuronales recurrentes (RNN) o transformers aplicados a imágenes. Estas redes podrían mejorar la capacidad de detección.
- Entrenar los modelos con una mayor cantidad de imágenes puede ser un beneficio para mejorar la precisión de detección. Se puede plantear realizar recortes de fuentes compactas detectadas a diferentes frecuencias sobre sus mapas correspondientes.
- Además de una clasificación binaria como la que se ha planteado en este trabajo, desarrollar modelos capaces de clasificar distintos tipos de fuentes (blazar, estrella del infrarrojo lejana, …).
- En un futuro, validar los modelos con datos del Observatorio Simons, LiteBIRD o futuras misiones. Esto permitirá comprobar la adaptabilidad de los modelos y si las predicciones se han realizado correctamente.

Estas posibles líneas de investigación se plantean como un desarrollo natural de este trabajo de fin de grado, con el fin de estudiar la fiabilidad de los catálogos astronómicos, haciendo uso de modelos basados en aprendizaje profundo.

Capítulo 9. REFERENCIAS

Ade, P., Aguirre, J., Ahmed, Z., Aiola, S., Ali, A., Alonso, D., Alvarez, M. A., Arnold, K., Ashton, P., Austermann, J., Awan, H., Baccigalupi, C., Baildon, T., Barron, D., Battaglia, N., Battye, R., Baxter, E., Bazarko, A., Beall, J. A., … Zhu, N. (2019). The Simons Observatory: science goals and forecasts. *Journal of Cosmology and Astroparticle Physics*, *2019*(02), 056-056. https://doi.org/10.1088/1475-7516/2019/02/056

Aggarwal, C. C. (2015). *Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14142-8

Bardeen, J. M., Steinhardt, P. J., & Turner, M. S. (1983). Spontaneous creation of almost scale-free density perturbations in an inflationary universe. *Phys. Rev. D*, *28*(4), 679-693. https://doi.org/10.1103/PhysRevD.28.679

Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Spergel, D. N., Tucker, G. S., Wollack, E., Wright, E. L., Barnes, C., Greason, M. R., Hill, R. S., Komatsu, E., Nolta, M. R., Odegard, N., Peiris, H. V., … Weiland, J. L. (2003). First-Year Wilkinson Microwave Anisotropy Probe (WMAP). Observations: Preliminary Maps and Basic Results. *The Astrophysical Journal Supplement Series*, *148*(1), 1-27. https://doi.org/10.1086/377253

Bianchi, S., Mainieri, V., & Padovani, P. (2022). *Active Galactic Nuclei and their demography through cosmic time*. https://doi.org/10.1007/978-981-16-4544-0_113-1

Bonnarel, F., Fernique, P., Bienaymé, O., Egret, D., Genova, F., Louys, M., Ochsenbein, F., Wenger, M., & Bartlett, J. G. (2000). The ALADIN interactive sky atlas. *Astronomy and Astrophysics Supplement Series*, *143*(1), 33-40. https://doi.org/10.1051/aas:2000331

Carlstrom, J. E., Ade, P. A. R., Aird, K. A., Benson, B. A., Bleem, L. E., Busetti, S., Chang, C. L., Chauvin, E., Cho, H.-M., Crawford, T. M., Crites, A. T., Dobbs, M. A., Halverson, N. W., Heimsath, S., Holzapfel, W. L., Hrubes, J. D., Joy, M., Keisler, R., Lanting, T. M., … Williamson, R. (2011). The 10 Meter South Pole Telescope. *Publications of the Astronomical Society of the Pacific*, *123*(903), 568-581. https://doi.org/10.1086/659879

Chollet, F. (s. f.). *M A N N I N G*.

Dodelson, S. (2003). *Modern Cosmology* (J. Hayhurst, Ed.). Academic Press.

Durrer, R. (2015). The Cosmic Microwave Background: The history of its experimental investigation and its significance for cosmology. *Classical and Quantum Gravity*, *32*(12). https://doi.org/10.1088/0264-9381/32/12/124007

European Space Agency. (2013). *Planck*. https://www.esa.int/Enabling_Support/Operations/Planck

European Space Agency. (2015a). *Planck Legacy Archive - Catalogues*. https://pla.esac.esa.int/pla/#catalogues

European Space Agency. (2015b). *Planck Legacy Archive - Maps*. https://pla.esac.esa.int/#maps

Fuskeland, U., Wehus, I. K., Eriksen, H. K., Krachmalnicoff, N., & Baccigalupi, C. (2021). Constraints on the spectral index of polarized synchrotron emission from WMAP and Faraday-corrected S-PASS data. *Astronomy & Astrophysics*, *645*. https://doi.org/10.1051/0004-6361/201937330

Gail, H., & Sedlmayr, E. (s. f.). *Physics and Chemistry of Circumstellar Dust Shell*.

Ghigna, T., Adler, A., Aizawa, K., Akamatsu, H., Akizawa, R., Allys, E., Anand, A., Aumont, J., Austermann, J., Azzoni, S., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Basyrov, A., Beckman, S., Bersanelli, M., … Collaboration, the L. (2024). *The LiteBIRD mission to explore cosmic inflation*.

Hawking, S. W. (1982). The development of irregularities in a single bubble inflationary universe. *Physics Letters B*, *115*(4), 295-297. https://doi.org/https://doi.org/10.1016/0370-2693(82)90373-2

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). *Improving neural networks by preventing co-adaptation of feature detectors*.

Hu, W., & Sugiyama, N. (1995). Toward Understanding CMB Anisotropies and Their Implications. *Physical Review D*, *51*(6). https://doi.org/10.1103/PhysRevD.51.2599

Hu, W., & White, M. (1997). The Damping Tail of Cosmic Microwave Background Anisotropies. *The Astrophysical Journal*, *479*, 568-579. https://doi.org/10.1086/303928

J. A. Génova-Santos, J. A. R.-M. et al. (2015). QUIJOTE scientific results - I. Measurements of the intensity and polarisation of the anomalous microwave emission in the Perseus molecular complex. *Monthly Notices of the Royal Astronomical Society*, *452*, 4169-4182.

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*.

Kluger, F., Ackermann, H., Yang, M. Y., & Rosenhahn, B. (2017). *Deep Learning for Vanishing Point Detection Using an Inverse Gnomonic Projection* (pp. 17-28). https://doi.org/10.1007/978-3-319-66709-6_2

K.M. Górski, E. H. A. J. B. B. D. W. F. K. H. M. R. M. B. (2005). *HEALPix: Hierarchical Equal Area isoLatitude Pixelization of the Sphere*. https://healpix.sourceforge.io/

Lanz, L. F. (2016). *Detección de fuentes extragalácticas en mapas de la radiación del fondo cósmico de microondas mediante técnicas multifrecuenciales*.

Mancini, A. S., Piras, D., Alsing, J., Joachimi, B., & Hobson, M. P. (2022). CosmoPower: Emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys. *Monthly Notices of the Royal Astronomical Society*, *511*(2), 1771-1788. https://doi.org/10.1093/mnras/stac064

Massaro, E., Maselli, A., Leto, C., Marchegiani, P., Perri, M., Giommi, P., & Piranomonte, S. (2015). The 5th edition of the Roma-BZCAT. A short presentation. *Astrophysics and Space Science*, *357*(1), 75. https://doi.org/10.1007/s10509-015-2254-2

Montgomery, D. C. and R. G. C. (2020). *Applied Statistics and Probability for Engineers* (7.ª ed.). John Wiley & Sons.

Pan-Experiment Galactic Science Group. (2025). *Full-sky Models of Galactic Microwave Emission and Polarization at Sub-arcminute Scales for the Python Sky Model*.

Perraudin, N., Defferrard, M., Kacprzak, T., & Sgier, R. (2018). *DeepSphere: Efficient spherical Convolutional Neural Network with HEALPix sampling for cosmological applications*. https://doi.org/10.1016/j.ascom.2019.03.004

Planck Collaboration. (2014a). Planck 2013 results. XI. All-sky model of thermal dust emission. *Astronomy & Astrophysics*, *571*. https://doi.org/10.1051/0004-6361/201323195

Planck Collaboration. (2014b). Planck 2013 results. XXVIII. The Planck Catalogue of Compact Sources. *Astronomy & Astrophysics*. https://doi.org/10.1051/0004-6361/201321524

Planck Collaboration. (2015). Planck 2015 results. IX. Diffuse component separation: CMB maps. *Astronomy & Astrophysics*. https://doi.org/10.1051/0004-6361/201525936

Planck Collaboration. (2016a). Planck 2015 results. XI. CMB power spectra, likelihoods, and robustness of parameters. *Astronomy & Astrophysics*, *594*, A11. https://doi.org/10.1051/0004-6361/201526926

Planck Collaboration. (2016b). Planck 2015 results. XVI. Isotropy and statistics of the CMB. *Astronomy & Astrophysics*, *594*, A16. https://doi.org/10.1051/0004-6361/201526681

Planck Collaboration. (2016c). Planck 2015 results. X. Diffuse component separation: Foreground maps. *Astronomy & Astrophysics*, *594*, A10. https://doi.org/10.1051/0004-6361/201525967

Planck Collaboration. (2016d). Planck 2015 results. XXVI. The Second Planck Catalogue of Compact Sources. *Astronomy & Astrophysics*, *594*, A26. https://doi.org/10.1051/0004-6361/201526914

Planck Collaboration. (2018). Planck intermediate results: LIV. the Planck multi-frequency catalogue of non-thermal sources. *Astronomy and Astrophysics*, *619*. https://doi.org/10.1051/0004-6361/201832888

Planck Collaboration. (2020a). Planck 2018 results. I. Overview and the cosmological legacy of Planck. *Astronomy & Astrophysics*, *641*, A1. https://doi.org/10.1051/0004-6361/201833880

Planck Collaboration. (2020b). Planck 2018 results. IV. Diffuse component separation. *Astronomy & Astrophysics*, *641*. https://doi.org/10.1051/0004-6361/201833881

Rybicki, G. B. ., & Lightman, A. P. . (2004). *Radiative processes in astrophysics*. Wiley-VCH.

Sachs, R. K., & Wolfe, A. M. (1967). Perturbations of a Cosmological Model and Angular Variations of the Microwave Background. *The Astrophysical Journal*, *147*, 73-90. https://doi.org/10.1086/148982

Swetz, D. S., Ade, P. A. R., Amiri, M., Appel, J. W., Battistelli, E. S., Burger, B., Chervenak, J., Devlin, M. J., Dicker, S. R., Doriese, W. B., Dünner, R., Essinger-Hileman, T., Fisher, R. P., Fowler, J. W., Halpern, M., Hasselfield, M., Hilton, G. C., Hincks, A. D., Irwin, K. D., … Zhao, Y. (2011). OVERVIEW OF THE ATACAMA COSMOLOGY TELESCOPE: RECEIVER, INSTRUMENTATION, AND TELESCOPE SYSTEMS. *The Astrophysical Journal Supplement Series*, *194*(2), 41. https://doi.org/10.1088/0067-0049/194/2/41

Thorne, B., Dunkley, J., Alonso, D., & Næss, S. (2017). The Python Sky Model: software for simulating the Galactic microwave sky. *Monthly Notices of the Royal Astronomical Society*, *469*(3), 2821-2833. https://doi.org/10.1093/mnras/stx949

Walpole, R. E. and M. R. H. and M. S. L. and Y. K. E. (2012). *Probabilidad y estadística para ingeniería y ciencias* (9.ª ed.). Pearson Educación.

Weinberg, S. (2008). *Cosmology*. Oxford University Press.

Zonca, A., Singer, L. P., Lenz, D., Reinecke, M., Rosset, C., Hivon, E., & Gorski, K. M. (2019). Healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in Python. *Journal of Open Source Software*, *4*(35). https://doi.org/10.21105/joss.01298

Zonca, A., Thorne, B., Krachmalnicoff, N., & Borrill, J. (2021). The Python Sky Model 3 software. *Journal of Open Source Software*, *6*(63), 3783. https://doi.org/10.21105/joss.03783

Capítulo 10. ANEXO

10.1 Código de los modelos

```python
#Capas convolucionales 2D
inp = Input(shape=input_shape)
x = Conv2D(32, (3, 3), activation='relu', padding='same')(inp)
x = MaxPooling2D((2, 2))(x)
x = Conv2D(64, (3, 3), activation='relu', padding='same')(x)
x = MaxPooling2D((2, 2))(x)

#Capas densas
x = Flatten()(x)
x = Dense(64, activation='relu')(x)
x = Dropout(0.3)(x)
out = Dense(8, activation='sigmoid')(x)

#Compilación del modelo
model = Model(inputs=inp, outputs=out)
model.compile(optimizer='adam', loss='mse', metrics=['mae'])
```

Figura 20: Código en Python de la red neuronal, modelo A.

```python
image_input = Input(shape=(32, 32, 1), name='image_input')
x = Conv2D(32, (3,3), activation='relu')(image_input)
x = MaxPooling2D((2,2))(x)
x = Conv2D(64, (3,3), activation='relu')(x)
x = MaxPooling2D((2,2))(x)
x = Flatten()(x)

# Entrada 2: Datos tabulares (flujo y error_flujo)
meta_input = Input(shape=(2,), name='meta_input')
m = Dense(16, activation='relu')(meta_input)

# Concatenar ambas salidas
combined = concatenate([x, m])

# Capas densas finales
z = Dense(64, activation='relu')(combined)

# Salida 1: Clasificación binaria
output_class = Dense(1, activation='sigmoid', name='output_class')(z)

# Salida 2: Coordenadas (regresión)
output_coords = Dense(2, activation='linear', name='output_coords')(z)


# Modelo completo
model = Model(inputs=[image_input, meta_input],
              outputs=[output_class, output_coords])
#Compilar modelo
model.compile(optimizer='adam',
        loss={'output_class': 'binary_crossentropy','output_coords': 'mse'},
        metrics={'output_class': 'accuracy','output_coords': 'mae'})
```

Figura 19: Código en Python de la red neuronal, modelo B.

El código de este trabajo se puede ver en el siguiente repositorio de GitHub: https://github.com/MiguelRemedios/Trabajo-de-Fin-de-Grado.git

10.2 Matrices de Confusión

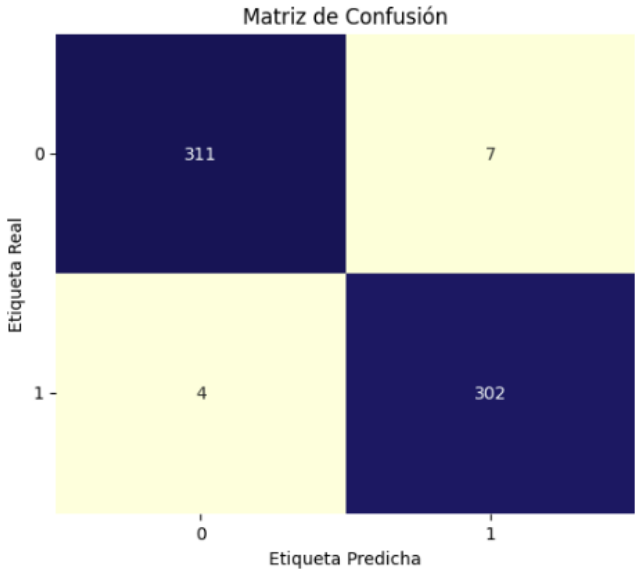

Figura 22: Matriz de confusión del Modelo A2 cuando el umbral se sitúa en el 50% de probabilidad.

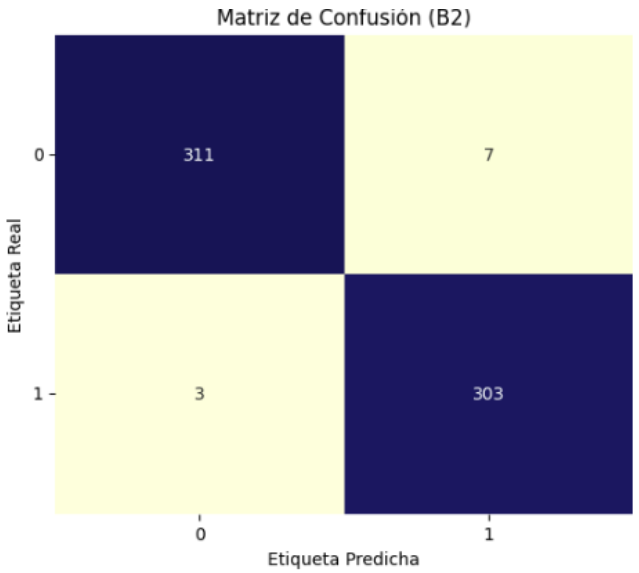

Figura 21: Matriz de confusión del Modelo B2 cuando el umbral se sitúa en el 50% de probabilidad.

10.3 Evolución de los modelos durante el entrenamiento

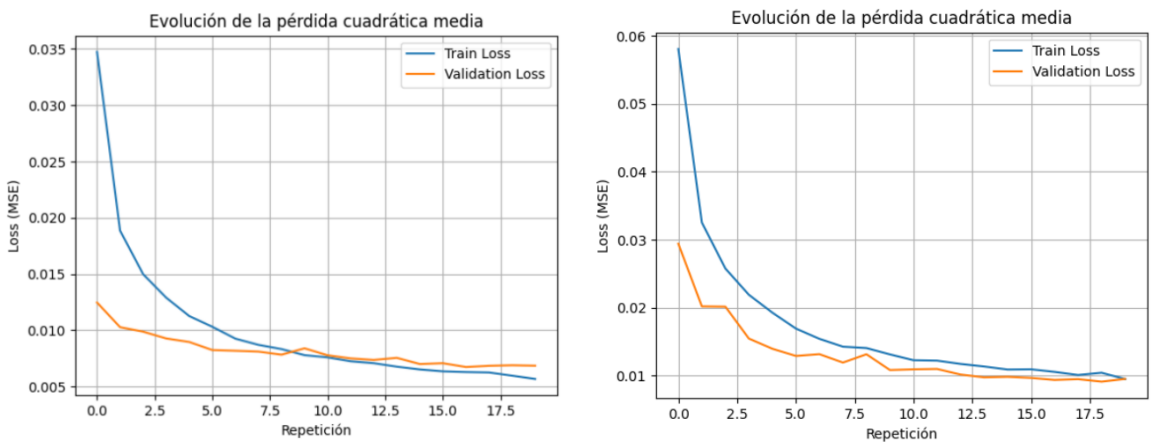

Figura 24: Se representa la pérdida del error medio para el conjunto de entrenamiento y para el de prueba en función de las repeticiones. A la izquierda el modelo A1, a la derecha el modelo A2.

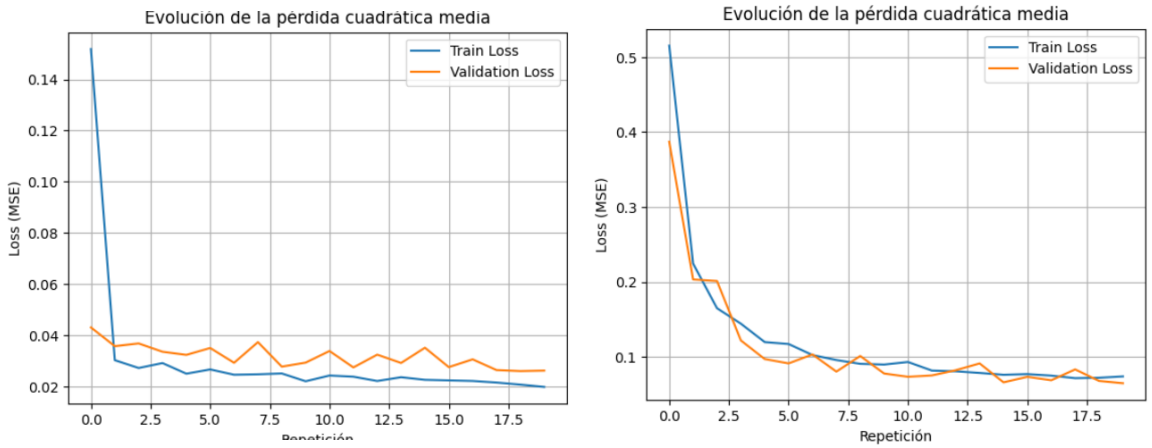

Figura 23: Se representa la pérdida del error medio para el conjunto de entrenamiento y para el de prueba en función de las repeticiones. A la izquierda el modelo B1, a la derecha el modelo B2.

10.4 Relación entre SNR y probabilidad. Modelo A2

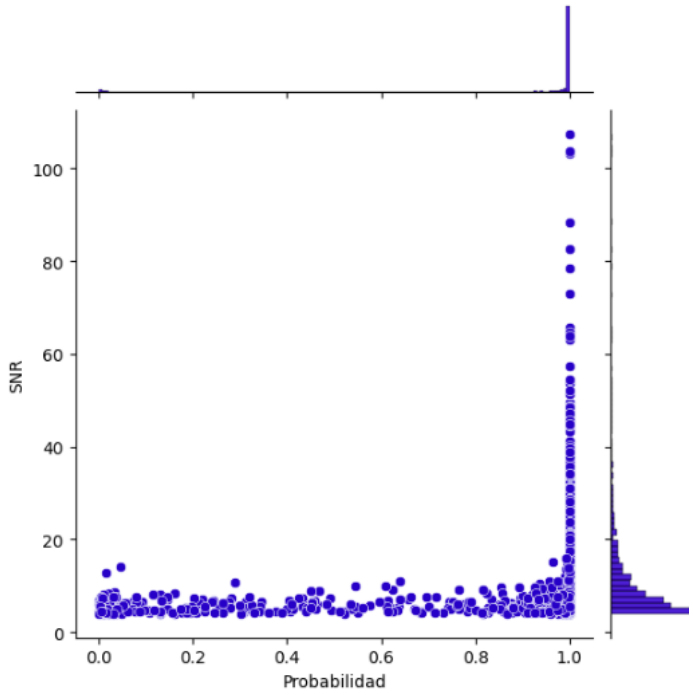

Figura 25: Relación entre la señal/ruido de las fuentes del catálogo PCNTb y la probabilidad asignada por el modelo A2 de que se trate de fuentes verdaderas.

10.5 Glosario de Abreviaturas

CNN – Red neuronal convolucional.

PCCS2 – Catálogo de fuentes compactas del Planck.

PCNT – Catálogo multifrecuencial de fuentes no termales.

GHz – Gigahercios.

Kº – Kelvin.

CMB – Fondo Cósmico de Microondas.

ICA – Independent Component Analysis.

NILC – Needlet Internal Linear Combination.

WMAP – Wilkinson Microwave Anisotropy Probe.

JAXA – Japan Aerospace Exploration Agency.

ΛCDM – Modelo Cosmológico Estándar Lambda Cold Dark Matter.

SMICA – Spectral Matching Independent Component Analysis.

SEVEM – Spectral Estimation Via Expectation Maximization.

ESA – Agencia Espacial Europea.

COBE – Cosmic Background Explorer.

NASA – National Aeronautics and Space Administration.

Fwhm – Full Width at Half Maximum.

ISW – Sachs-Wolfe Integrado.

LFI – Low Frecuency Instrument.

AGN – Núcleo activo de galaxia.

Mse – Error cuadrático medio.

Mae – Error absoluto medio.

SNR – Relación señal/ruido