# Evaluación de Riesgo de Incendios con FireRisk Reducido: Modelos Supervisados, Auto-supervisados y el Impacto de las Coordenadas

Rodrigo González Marcet
rodrigogm@correo.ugr.es

Raúl Martínez Alonso
e.raulmartinez03@go.ugr.es

Javier Trujillo Castro
javitrucas@correo.ugr.es

Angel Sánchez Guerrero
e.angelguesan@go.ugr.es

## Abstract

*En este trabajo replicamos los experimentos del artículo original de FireRisk [6], un dataset de teledetección para la evaluación del riesgo de incendio, utilizando una versión reducida al 20 % de las imágenes originales. Evaluamos los modelos utilizados en el paper original. Además, incorporamos mecanismos de atención para integrar coordenadas geográficas en los modelos.*

*Nuestros resultados muestran una disminución en la precisión debido al menor tamaño del dataset y el límite en épocas de entrenamiento, pero destacan la efectividad del uso de coordenadas y la ventaja de modelos preentrenados.*

## 1. Introduccion

Los incendios forestales representan un problema crítico con consecuencias devastadoras para ecosistemas y comunidades, incluyendo pérdida de biodiversidad, deterioro del suelo, contaminación del aire y altos costos económicos. Aunque los modelos tradicionales han avanzado usando datos geocientíficos, su dependencia de características específicas limita su aplicabilidad en distintas regiones.

Surge así la necesidad de enfoques más accesibles, basados exclusivamente en imágenes satelitales de alta resolución y técnicas modernas de aprendizaje automático. Estos permiten mapear el riesgo de incendio directamente desde datos visuales, sin integrar variables geocientíficas adicionales.

Este trabajo replica y expande el estudio del conjunto de datos FireRisk (91,872 imágenes etiquetadas en 7 clases de riesgo). Dicho estudio demostró, con modelos como ResNet, ViT, DINO y MAE preentrenado en ImageNet, una precisión del 65.29% al asociar imágenes satelitales con riesgo de incendio.

Nos enfocamos en dos preguntas clave:

¿Es posible mantener un buen desempeño con menor tiempo de entrenamiento? Replicamos los experimentos reduciendo el número de épocas para evaluar el impacto en el desempeño de modelos supervisados y auto-supervisados.

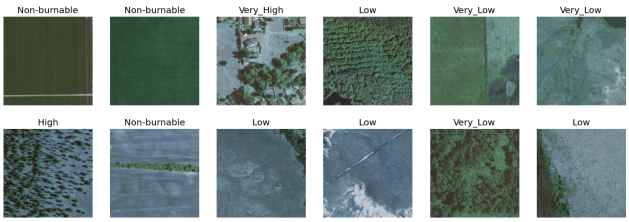


Figure 1. Muestra de imágenes con distintas etiquetas. Las dimensiones de las imágenes son 270 x 270 pixels.

¿Pueden las coordenadas geográficas mejorar la precisión? Exploramos mecanismos de atención para integrar coordenadas como características complementarias.

Nuestro trabajo valida y amplía las posibilidades prácticas del enfoque original, optimizando su uso en escenarios con datos limitados y variables adicionales.

## 2. Background

Para comprender los fundamentos de este trabajo, es necesario explorar los mecanismos de atención utilizados en aprendizaje profundo y su aplicación en la integración de coordenadas geográficas para la evaluación del riesgo de incendios forestales.

### 2.1. Mecanismos de atención en aprendizaje profundo

Los mecanismos de atención han revolucionado el campo del aprendizaje profundo al permitir que los modelos enfoquen su capacidad de procesamiento en partes relevantes de los datos de entrada. Originalmente introducidos en el procesamiento de lenguaje natural [2] (NLP), los mecanismos de atención se han extendido al análisis de imágenes, donde destacan por mejorar el desempeño en tareas como la clasificación, segmentación y detección de ob-

jetos.

Nuestro interés se ha centrado en las coordenadas geográficas, ya que son una fuente crucial de información en la evaluación del riesgo de incendios, pues encapsulan la ubicación espacial de las imágenes, reflejando factores contextuales como el clima regional, el tipo de vegetación y la cercanía a zonas urbanas. Sin embargo, integrar esta información con datos visuales no es trivial.

En este trabajo, las coordenadas geográficas se incorporan como un vector adicional en el modelo mediante mecanismos de atención, siguiendo un enfoque estructurado. Primero, las coordenadas se transforman en una representación aprendible que permite al modelo capturar su relación con los datos visuales. Posteriormente, estas representaciones codificadas se combinan con las características extraídas de las imágenes a través del mecanismo de atención, lo que facilita que el modelo conecte las características visuales con su contexto espacial. Efectivamente, asignando de manera adaptativa pesos a las características visuales, en función de la localización espacial, priorizando aquellas más relevantes para predecir con precisión la clase de riesgo de incendio. La inspiración para este acercamiento proviene del canal de 3Blue1Brown [1]

## 3. Trabajos relacionados

La evaluación del riesgo de incendios forestales utilizando imágenes de teledetección ha sido objeto de diversos estudios que exploran el potencial de las técnicas de aprendizaje automático y análisis espacial. Los trabajos relacionados se pueden dividir en tres áreas principales.

### 3.1. Modelos basados en datos geocientíficos satelitales para la prevención de incendios

Nuestro trabajo se plantea como una extensión del estudio realizado en FireRisk [6], que presentó el conjunto de datos homónimo para la clasificación de imágenes según el riesgo de incendio, junto con experimentos utilizando transformadores como ViT [4] y MAE [5]. Sin embargo, los resultados de estos enfoques fueron limitados, y los modelos propuestos requieren tanto un alto costo computacional como un considerable tamaño, lo que dificulta su implementación práctica en entornos con recursos restringidos, como es nuestro caso.

### 3.2. Transformers y técnicas de atención

La incorporación de modelos de aprendizaje profundo, como las redes convolucionales (CNN) y los transformadores, ha permitido avances significativos en la interpretación de imágenes satelitales. Estudios recientes han demostrado el potencial de enfoques supervisados y autosupervisados usando transformers, como ViT [4], y MAE [5], para asociar imágenes con el riesgo de incendios.

Adicionalmente, se ha explorado el uso de mecanismos de atención para integrar información espacial explícita, como coordenadas geográficas. Por ejemplo, la red SFANet [7] utilizó atención geográfica para mejorar la precisión en modelos de predicción climática, destacando el potencial de estas técnicas en problemas relacionados con datos espaciales.

## 4. Métodos

En esta sección describimos los modelos utilizados, las modificaciones realizadas y la metodología empleada para entrenar y evaluar los modelos en la tarea de clasificación de riesgo de incendios.

Para entrenar los modelos, utilizamos la biblioteca Fastai debido a su simplicidad y eficiencia en la implementación de técnicas de aprendizaje profundo.

### 4.1. Modelo ViT-B/16

ViT-B/16 es un modelo de visión por transformadores (Vision Transformer o ViT) [4]. Se basa en la arquitectura de transformadores, una tecnología inicialmente desarrollada para el procesamiento del lenguaje natural, que fue adaptada al dominio visual por Google Research. Es un modelo que aplica la arquitectura de transformadores directamente a imágenes dividiéndolas en pequeños parches, en lugar de usar convoluciones como en las redes neuronales convolucionales (CNN). En este caso, la imagen de entrada se divide en parches de 16x16 píxeles. Cada parche se trata como un ”token” (similar a una palabra en NLP) que se procesa por el transformador.

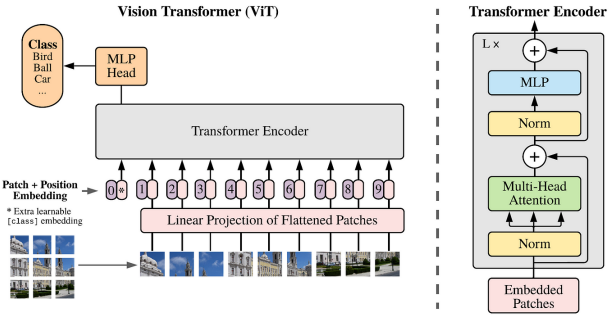


Figure 2. Arquitectura de la red ViT *Fuente: https://shtsang.medium.com/review-vision-transformer-vit-406568603de0*

### 4.2. Modelo MAE

MAE (Masked Autoencoder) es una arquitectura de autoencoder diseñada específicamente para tareas de aprendizaje auto-supervisado en tareas de visión por computador. Este enfoque introduce una estrategia de aprendizaje basada en enmascarar una gran proporción de las entradas de una imagen, el 75% en nuestro caso, y entrenar un modelo para

reconstruir los píxeles enmascarados utilizando únicamente los píxeles visibles. La clave de MAE radica en la eficiencia: al trabajar únicamente con los píxeles no enmascarados, reduce significativamente el costo computacional y permite el uso de modelos más grandes.

El modelo se divide en dos componentes principales:

- **Encoder:** Procesa los tokens visibles de la imagen, extrayendo representaciones compactas y útiles.
- **Decoder:** Reconstruye la imagen completa utilizando la información codificada y los tokens enmascarados.

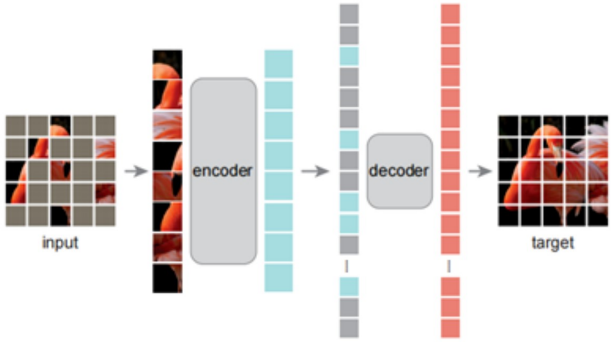


Figure 3. Arquitectura de la red MAE. Fuente: https://mmpretrain.readthedocs.io/en/dev/papers/mae.html

Como hemos visto en el apartado anterior, Vision Transformer (ViT) es un modelo basado en Transformers, adaptado para procesar imágenes.

Usar ViT como *backbone* significa que empleamos su encoder como extractor de características. En lugar de utilizar ViT para realizar directamente una tarea como clasificación o segmentación, aprovechamos su capacidad de aprender representaciones ricas de las imágenes para alimentar otra arquitectura, en nuestro caso, el encoder de MAE.

La principal diferencia entre ViT y MAE con ViT como backbone es el enfoque de entrenamiento. Mientras que ViT se entrena para tareas específicas como clasificación, MAE se entrena de forma no supervisada para predecir las partes enmascaradas, lo que permite una mayor eficiencia en el uso de datos no etiquetados.

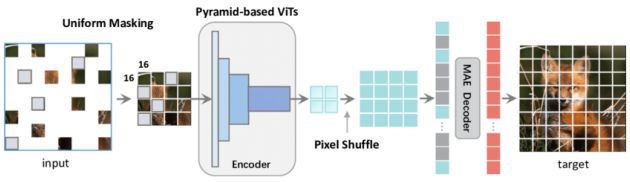


Figure 4. Arquitectura de la red MAE con ViT de Backbone

### 4.3. Modelo DINO

DINO (Distillation with no labels) es un método de aprendizaje auto-supervisado propuesto por Caron et al. [3] que emplea una destilación entre dos modelos, denominados *teacher* y *student*, sin requerir etiquetas. El *teacher* produce "pseudo-etiquetas" a partir de las imágenes, mientras que el *student* aprende a imitarlas. Para evitar soluciones triviales, el *teacher* se actualiza de forma asíncrona como una versión suavizada del *student*, y se incorpora un proceso de *centering* que estabiliza las activaciones. DINO también utiliza múltiples vistas aumentadas de cada imagen, lo que promueve la extracción de características invariantes.

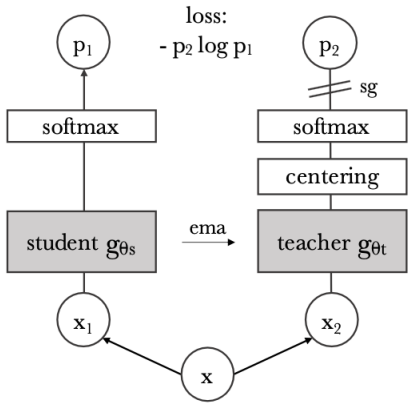


Figure 5. Esquema simplificado de la arquitectura DINO, mostrando el flujo de datos entre *teacher* y *student*.

DINO se basa en Vision Transformers y aprovecha su mecanismo de autoatención para modelar relaciones globales a través de parches en la imagen. A diferencia de MAE, que reconstruye píxeles enmascarados, DINO alinea las representaciones de *teacher* y *student* sin recurrir a una reconstrucción explícita. Esto le permite aprender rasgos discriminativos en contextos donde las anotaciones son limitadas, como en imágenes satelitales orientadas a la evaluación del riesgo de incendios. Una vez entrenado sin supervisión, DINO ofrece representaciones que pueden refinarse de forma supervisada con menos requerimientos de datos etiquetados.

### 4.4. Modelo ResNet

ResNet-50 es una arquitectura CNN de la familia ResNet, diseñada para abordar los desafíos del entrenamiento de redes profundas mediante bloques residuales. En lugar de aprender transformaciones completas, estos bloques aprenden las diferencias entre la entrada y la salida esperada, utilizando conexiones directas (skip connections) que facilitan el flujo de información y gradientes.

Con 50 capas de profundidad, ResNet-50 combina convoluciones, normalización por lotes (Batch Normalization), activaciones ReLU y bloques residuales organizados para mantener la fluidez de la información, lo que la convierte en un modelo eficiente y destacado desde su lanzamiento en 2015.

### 4.5. Aplicación de Atención para Contexto Geográfico

En el dataset FireRisk hemos identificado una limitación fundamental: existen imágenes visualmente indistinguibles que pertenecen a categorías de riesgo diametralmente opuestas. Esta ambigüedad complica significativamente la diferenciación entre las clases, reduciendo la eficacia de los modelos al basarse únicamente en las características visuales de las imágenes.

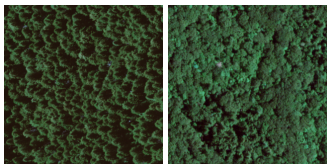

Figure 6. Muestra de imagenes de FireRisk, categorizadas con riesgo muy alto y muy bajo respectivamente.

Para abordar este desafío, proponemos incorporar las coordenadas geográficas como entradas adicionales para los modelos de predicción. Cada imagen en el dataset incluye información espacial en forma de coordenadas geográficas, las cuales pueden ser extraídas y utilizadas para complementar la información visual. Este enfoque permite enriquecer el contexto de la predicción, aprovechando la relación geográfica inherente entre las imágenes y sus categorías de riesgo, y mejora la capacidad del modelo para discriminar entre clases.

La incorporación de las coordenadas geográficas se realiza mediante la arquitectura descrita en el diagrama de la Figura 7. El enfoque combina un modelo de visión por computador clásico, como ResNet-18 o ResNet-50, que actúa como extractor de características visuales, con un extractor de características específico para las coordenadas geográficas, implementado como una red neuronal clásica. Este último transforma las coordenadas en un vector de características de mayor dimensionalidad, ajustado para integrarse con el modelo principal.

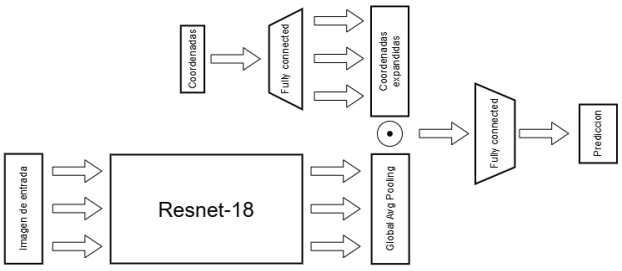


Figure 7. Diagrama de la arquitectura AResNet-18.

Posteriormente, las características derivadas de las coordenadas se utilizan como pesos para las características visuales, realizando una multiplicación elemento a elemento entre ambos vectores. El resultado es un nuevo vector de características que combina la información visual de la imagen con el contexto geográfico proporcionado por las coordenadas. Este vector combinado es finalmente procesado mediante un bloque neuronal clásico compuesto por capas densas para realizar la clasificación final. Esta arquitectura permite capturar la interacción entre el contenido visual y la información espacial. Estos modelos "híbridos" los denominamos AResNet-X (Atention ResNet-X).

## 5. Experimentos

En esta sección, presentamos los experimentos realizados para evaluar el desempeño de diversos modelos en la clasificación de riesgo de incendio utilizando el conjunto de datos FireRisk. Inicialmente, replicamos los experimentos del trabajo original, aplicando modelos supervisados y autosupervisados en una versión reducida del dataset. Posteriormente, incorporamos mecanismos de atención para integrar coordenadas geográficas al modelo, explorando cómo esta información adicional mejora la predicción. Los resultados obtenidos permiten analizar la eficacia de los enfoques propuestos y su sensibilidad al tamaño del conjunto de datos utilizado.

### 5.1. Dataset

El conjunto de datos FireRisk es una colección de imágenes de teledetección diseñadas para evaluar el riesgo de incendios forestales. Contiene un total de 91,872 imágenes etiquetadas en 7 clases de riesgo de incendio, basadas en la información proporcionada por el conjunto de datos Wildfire Hazard Potential (WHP). Las imágenes de alta resolución fueron recopiladas a través del programa National Agriculture Imagery Program (NAIP), que ofrece datos de teledetección de gran detalle espacial.

Cada imagen mide 320 × 320 píxeles, aunque nosotros reducimos su dimensionalidad a 224x224, y las etiquetas reflejan el nivel de riesgo de incendio, lo que permite analizar la relación entre las características geográficas observadas

y el peligro potencial de incendios. Además, dentro del dataset se incluyen las coordenadas de las imágenes, que usaremos mas adelante para entrenar nuestros modelos con mecanismos de atención.

### 5.2. Experimento ViT

Para el entrenamiento, cargamos la arquitectura preentrenada `vit_base_patch16_224` utilizando pesos iniciales entrenados en ImageNet. Se realizó un ajuste fino de todo el modelo, descongelando las capas preentrenadas y actualizando sus pesos. Hemos limitado a 5 y 10 épocas con un tamaño de batch de 64, utilizando la función de pérdida de entropía cruzada y el optimizador Adam con un learning rate inicial de $10^{-3}$. La métrica de evaluación ha sido la precisión global(accuracy).

| LR | Épocas | Test accuracy (%) |
|---|---|---|
| $1 \times 10^{-3}$ | 5 | 59.89 |
| $1 \times 10^{-3}$ | 10 | 55.53 |

Table 1. Resultados de accuracy en test usando el modelo ViT-B/16 preentrenado en ImageNet.

Podemos observar que extender el entrenamiento a 10 épocas no resultó en una mejora, sino en una ligera disminución del desempeño. Este comportamiento podría deberse a sobreajuste en el modelo.

### 5.3. Experimento ResNet-50

Durante el entrenamiento, se realizó un fine-tuning de 5 épocas, con los pesos preentrenados de ImageNet . Utilizamos el optimizador AdamW con un learning rate ajustado automáticamente por Fastai, combinando un enfoque de descongelamiento progresivo. Se emplearon la precisión global (accuracy) y el F1 Score ponderado para evaluar tanto el desempeño general como la capacidad del modelo de manejar clases desbalanceadas.

Se obtuvo una precisión global de 57.83% en test, indicando un aprendizaje moderado. Sin embargo, muestra señales de sobreajuste a partir de la época 2.

### 5.4. Experimentos MAE y su implementación en el problema de la clasificación

El Masked Autoencoder (MAE), aunque no diseñado para clasificación, puede ser útil para tareas de clasificación debido a las representaciones aprendidas durante la reconstrucción de imágenes enmascaradas [5]. En estos experimentos, buscamos entender el funcionamiento de MAE y su competencia con otros modelos.

Hemos realizado dos experimentos: El primero consistió en entrenar un MAE con un ViT preentrenado (utilizamos el modelo google/vit-base-patch16-224) [4]. En esta parte, es interesante observar cómo MAE reconstruye las imágenes y cómo el cálculo del error cuadrático medio (MSE) se realiza de forma personalizada, ya que solo se calcula en las áreas visibles, lo que guía al modelo a mejorar la reconstrucción.

El segundo experimento utilizó un modelo MAE preentrenado durante 80 épocas, ofrecido en el paper original [6], con el conjunto de datos UnlabelledNAIP. Este modelo preentrenado incluye un clasificador adicional con dos capas lineales y activación ReLU, lo que mejora la especificidad del modelo para tareas concretas. En comparación, el primer modelo es más simple y versátil, pero menos optimizado para tareas específicas.

| Name | Épocas | Val Acc (%) | Test Acc (%) |
|---|---|---|---|
| Original | 3 | 49.57 | 55.81 |
| Pretrained | 84 | 58.39 | 57.61 |

Table 2. Resultados de precisión en validación y test para modelos originales y preentrenados.

La comparación entre ambos enfoques muestra que, aunque el modelo preentrenado tiene un mejor rendimiento en entrenamiento y validación, su rendimiento en datos no vistos es limitado, con mejoras mínimas en los resultados de prueba. Esto sugiere que el modelo podría estar sobreajustandose durante el entrenamiento.

### 5.5. Experimentos con DINO

En esta sección, describimos dos experimentos realizados con modelos DINO para la clasificación del riesgo de incendio.

Primer experimento: DINO ViT-B/16 preentrenado en ImageNet: Se empleó un modelo ViT-B/16 entrenado con DINO sobre ImageNet, disponible en *facebookresearch/dino* a través de Torch Hub. Se llevaron a cabo cuatro configuraciones de ajuste donde se variaron el número de épocas y la tasa de aprendizaje (LR). En dos de ellas, se utilizó la herramienta `lr_find` de `fastai`, que recomendó un valor cercano a $1.2 \times 10^{-5}$.

Segundo experimento: Checkpoint DINO entrenado en UnlabelledNAIP: Para el segundo experimento, se partió de un *checkpoint* proporcionado por el trabajo original de FireRisk, entrenado durante 100 épocas en el conjunto de imágenes no etiquetadas (UnlabelledNAIP). Se realizaron tres configuraciones de ajuste, variando el número de épocas y la tasa de aprendizaje entre $1 \times 10^{-4}$ y $1 \times 10^{-3}$.

### 5.6. Experimentos con mecanismos de atención

Para evaluar el impacto de la incorporación de coordenadas geográficas mediante un mecanismo de atención, diseñamos una serie de experimentos comparativos.

Primero, utilizamos una ResNet-18 básica preentrenada en ImageNet como modelo de control, entrenada durante 15 épocas. Posteriormente, aplicamos el mecanismo

de atención geográfica en ResNet-18, creando el modelo AResNet-18, entrenado también durante 15 épocas, ajustando todos sus pesos desde el inicio. Los valores de learning rate de estos experimentos fueron obtenidos mediante la función $lr\_find()$ de FastAi.

En un tercer experimento, evaluamos AResNet-18 con ajuste fino: las capas convolucionales preentrenadas se congelaron durante las primeras 3 épocas, entrenando solo las capas adicionales del mecanismo de atención, seguido de 7 épocas de entrenamiento completo.

Finalmente, repetimos este esquema de ajuste fino con una AResNet-50, para analizar el impacto de una mayor profundidad en la integración de características visuales y geográficas.

## 6. Resultados

En nuestros experimentos iniciales, observamos que los modelos replicados presentan un rendimiento consistentemente inferior al reportado en el trabajo original. Esto puede atribuirse principalmente a la limitación en el número de épocas de entrenamiento, ya que los modelos entrenados para este estudio se limitaron a menos de 10 épocas, mientras que los modelos originales fueron entrenados durante 100 épocas. Este resultado destaca la importancia del tiempo de entrenamiento para alcanzar el máximo rendimiento de los modelos.

| Resultados | |
|---|---|
| Modelo | Acc (% ) |
| ResNet-50 | 61.37 |
| ResNet-50* | 57.82 |
| ViT-B/16 | 61.37 |
| ViT-B/16 | 59.89 |
| MAE | 62.51 |
| MAE* | 57.33 |
| DINO | 61.96 |
| DINO* | 57.92 |

Table 3. Resultados de los modelos entrenados en el paper original. * Resutados de nuestra mejor réplica de ese modelo.

Posteriormente, evaluamos el impacto de incorporar mecanismos de atención basados en las coordenadas geográficas. Los resultados muestran que la inclusión de las coordenadas como datos de entrada tiene un efecto significativo en el rendimiento de los modelos. Comparando los modelos básicos con las versiones mejoradas mediante atención, se observa un incremento de rendimiento sustancial (+3.52% como mejora mínima).

Además, el análisis demuestra que el ajuste fino (fine-tuning) desempeña un papel crucial en la mejora del rendimiento. En cuanto a la profundidad de las arquitecturas, se observa que su impacto es menos significativo en

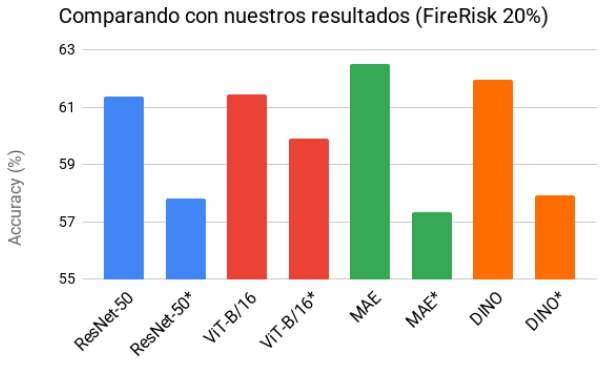


Figure 8. Resultados de los modelos entrenados en el paper original. * Resutados de nuestra mejor réplica de ese modelo.

comparación con el mecanismo de atención y el ajuste fino. AResNet-50, aunque ligeramente superior a AResNet-18, ofrece solo un incremento modesto del 0.48%. Esto sugiere que la inclusión de atención y un entrenamiento adecuado son factores más determinantes para el rendimiento que el aumento de la profundidad de la red.

| Resultados | |
|---|---|
| Modelo | Acc (% ) |
| ResNet-18ft | 57.25 |
| ResNet-50 | 57.82 |
| AResNet-18 M | 60.77 |
| AResNet-18 | 62.24 |
| AResNet-50 | 62.72 |

Table 4. Resultados de los modelos aplicando mecanismos de atención en comparación con ResNet-18 y ResNet-50 básicas.

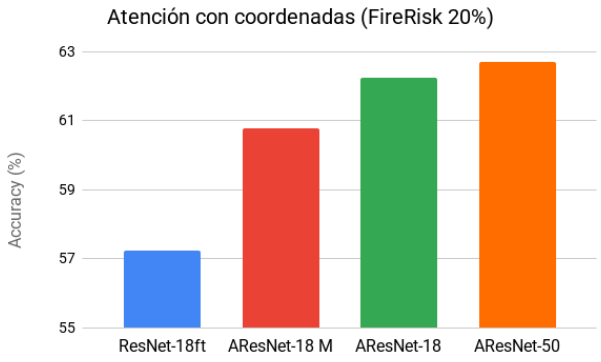


Figure 9. Comparacion visual entre los modelos básicos y los modificados con atención

Finalmente, al comparar los mejores modelos obtenidos en este estudio con aquellos del trabajo original, encontramos que tanto AResNet-18 como AResNet-50 alcanzan

un desempeño competitivo, posicionándose entre los tres mejores modelos.

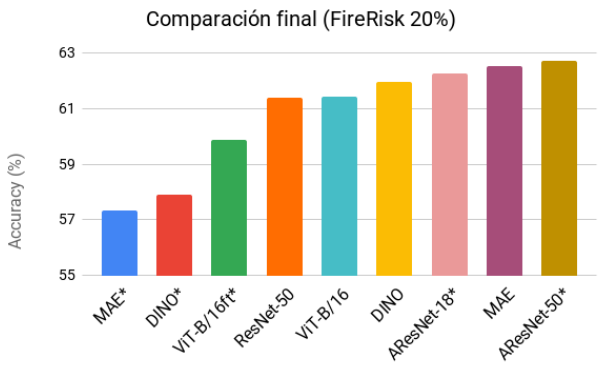


Figure 10. Resultados obtenidos de la mejor versión de cada modelo. * Réplicas realizadas por nosotros

Estas arquitecturas no solo igualan o superan los resultados de los modelos originales, sino que también logran estos resultados con un tiempo de entrenamiento significativamente menor (10 épocas frente a 100) y con un tamaño de modelo considerablemente más reducido en comparación con ViT-B/16, la base de las arquitecturas DINO y MAE.

## 7. Conclusiones

Este trabajo ha explorado la evaluación del riesgo de incendios mediante imágenes de teledetección, destacando la integración de coordenadas geográficas como un componente clave para mejorar la predicción. Los modelos mejorados con mecanismos de atención (AResNet) igualaron consistentemente a las arquitecturas estándar pese a su reducido tamaño y tiempo de entreno, demostrando que el contexto espacial es crucial en problemas de clasificación en teledetección.

El ajuste fino (fine-tuning) también se confirmó como esencial para adaptar redes preentrenadas a conjuntos de datos especializados, mientras que el aumento de la profundidad mostró un impacto menor en comparación con los beneficios del uso de atención.

Estos resultados no solo validan la efectividad del enfoque propuesto, sino que también sugieren nuevas oportunidades, como aplicar mecanismos de atención a arquitecturas avanzadas como ViT-B/16 o extender los experimentos a conjuntos de datos más diversificados. En conjunto, este estudio establece una base sólida para futuras investigaciones en la evaluación del riesgo de incendios y el uso de contexto geográfico en problemas de teledetección.

## References

[1] 3Blue1Brown. Transformers (how llms work) explained visually — dl5. 2

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. 2016. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2, 5

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 2, 5

[6] Shuchang Shen, Sachith Seneviratne, Xinye Wanyan, and Michael Kirley. Firerisk: A remote sensing dataset for fire risk assessment with benchmarks using supervised and self-supervised learning. 2023. 1, 2, 5

[7] Jiaze Wang, Hao Chen, Hongcan Xu, Jinpeng Li, Bowen Wang, Kun Shao, Furui Liu, Huaxi Chen, Guangyong Chen, and Pheng-Ann Heng. Sfanet: Spatial-frequency attention network for weather forecasting, 2024. 2