
Universidad Politécnica
de Madrid

**Escuela Técnica Superior de Ingenieros Informáticos**


Grado en Ingeniería Informática

Trabajo Fin de Grado

# Comparación del Desempeño de Transfer Learning en Transformers Visuales para la Clasificación de Imágenes de Satélite


Autor: Héctor García Barrado

Tutor(a): Consuelo Gonzalo Martín y Meryeme Boumahdi



Madrid, junio 2024

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

*Trabajo Fin de Grado*
*Grado en Ingeniería Informática*
*Título:* Comparación del Desempeño de Transfer Learning en Transformers Visuales para la Clasificación de Imágenes de Satélite
Junio 2024

*Autor:* Héctor García Barrado

*Tutor:*
Profesora Consuelo Gonzalo Martín, Ph.D.
Departamento de Arquitectura y Tecnología de Sistemas Informáticos
ETSI Informáticos
Universidad Politécnica de Madrid

*Cotutor:*
Meryeme Boumahdi
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid

# Resumen

El uso de imágenes satelitales es crucial para aplicaciones diversas como la monitorización ambiental, la planificación urbana y la agricultura de precisión. Sin embargo, la gran cantidad de datos generados y su complejidad presentan desafíos significativos en su análisis y clasificación. Este trabajo se justifica por la necesidad de mejorar la eficiencia y precisión en la clasificación de estas imágenes, utilizando técnicas avanzadas de aprendizaje automático con el uso de los modelos Transformer y técnicas de *Transfer Learning* (Aprendizaje por transferencia, TL).

El análisis de las imágenes satelitales es una tarea que requiere alta capacidad de procesamiento y grandes conjuntos de datos etiquetados, lo cual puede ser costoso y requerir una gran cantidad de trabajo. El trabajo se centra en ofrecer una solución eficiente para la clasificación de imágenes satelitales mediante el uso de modelos Transformers combinados con técnicas de *TL*.

En el trabajo se propone el uso de modelos Transformers para la clasificación de imágenes, específicamente los modelos Vision Transformer (ViT), Swin Transformer y Data-efficient Image Transformer (DeiT), debido a su capacidad para manejar datos secuenciales y capturar dependencias de largo alcance. Además, la técnica de *TL* se utiliza para optimizar el entrenamiento de estos modelos, aprovechando modelos preentrenados para reducir el tiempo y los recursos necesarios.

El desarrollo del trabajo fue dividido en diversas fases para organizar y facilitar su elaboración. En primer lugar, se estudió la arquitectura y capacidades de los modelos Transformer y las técnicas de *TL* durante la fase de investigación.

Después, se procedió a la preparación de los datos, donde se utilizó el dataset EuroSat el cual está compuesto por 27,000 imágenes satelitales clasificadas en 10 categorías, las cuales fueron divididas en los conjuntos de entrenamiento, validación y prueba.

Para la fase de desarrollo, en ella se entrenaron los modelos ViT, Swin y DeiT, tanto con y sin la utilización de la técnica de data augmentation (aumento de datos, DA). Además, en esta fase se definieron e implementaron la función de pérdida Cross Entropy Loss y el optimizador Adam, los cuales fueron usados para el entrenamiento de los modelos.

En la obtención de los resultados, después de analizar el desempeño de todos los modelos en todas las métricas evaluadas, el modelo DeiT sin la utilización de *DA* mostró el mejor desempeño en ellas, requiriendo además menos épocas en comparación con los modelos ViT y Swin. El análisis detallado reveló que, aunque DeiT sin el uso de *DA* fue el más eficiente, puede haber indicios de sobreajuste debido a la alta especifidad en las características del conjunto de datos de entrenamiento.

El trabajo muestra que los modelos Transformer son altamente efectivos para la clasificación de imágenes satelitales, especialmente el modelo DeiT. La técnica de *DA* mejora significativamente el rendimiento de los modelos ViT y Swin, lo cual sugiere su utilidad en escenarios con variabilidad de datos. Sin embargo, el posible sobreajuste del modelo DeiT sin el uso de *DA* indica la necesidad de una mayor investigación en técnicas de regularización. Las aplicaciones prácticas de estos hallazgos pueden optimizar operaciones en

sectores como la agricultura, minería y gestión de infraestructuras, contribuyendo además a la sostenibilidad ambiental y la innovación tecnológica.

# Abstract

The use of satellite images is crucial for various applications such as environmental monitoring, urban planning, and precision agriculture. However, the vast amount of data generated, and its complexity presents significant challenges in their analysis and classification. This work is justified by the need to improve the efficiency and accuracy in the classification of these images, using advanced machine learning techniques with the use of Transformer models and Transfer Learning (TL) Techniques.

Analyzing satellite requires high processing capacity and large labeled datasets, which can be costly and labor intensive. This work focuses on providing an efficient solution for the classification of satellite images by using Transformer models combined with TL techniques.

For the wok provided, the use of Transformer models for image classification is proposed, specifically Vision Transformer (ViT), Swin Transformer, and Data-efficient Image Transformer (DeiT) models, due to their ability to handle sequential data and capture long-range dependencies. Additionally, the TL technique is used to optimize the training of these models, leveraging pre-trained models to reduce the time and resources required.

The works development was divided into various phases to organize and facilitate its execution. First, the architecture and capabilities of the Transformer models and TL techniques were studied during the research phase.

Next, the data preparation phase involved using the EuroSat dataset, which consists of 27,000 satellite images classified into 10 categories. These images were divided into training, validation, and test sets.

During the development phase, the ViT, Swin, and DeiT models were trained, both with and without the use of data augmentation (DA) techniques. Additionally, the Cross Entropy Loss function and the Adam optimizer were defined and implemented for the training of the models.

For the results phase, after analyzing the performance of all models across all evaluated metrics, the DeiT model without the use of DA showed the best performance, also requiring fewer epochs compared to the ViT and Swin models. Detailed analysis revealed that although DeiT without DA was the most efficient, there may be signs of overfitting due to the high specificity in the characteristics of the training dataset.

This work shows that Transformer models are highly effective for satellite image classification, especially the DeiT model. The DA technique significantly improves the performance of the ViT and Swin models, suggesting its usefulness in scenarios with data variability. However, the potential overfitting of the DeiT model without DA indicates the need for further research in regularization techniques. The practical applications of these findings can optimize operations in sectors such as agriculture, mining, and infrastructure management, also contributing to environmental sustainability and technological innovation.

# Tabla de contenidos

# 1 Introducción

Las imágenes satelitales capturan información valiosa sobre la superficie de la Tierra desde una perspectiva única. Estas imágenes tienen una gran amplitud en cuanto a sus usos, como por ejemplo la monitorización ambiental [1], la planificación urbana [2] o la agrícultura de precisión [3]. Sin embargo, estas imágenes presentan un problema en cuanto a su análisis, debido a su gran volumen y complejidad. La cantidad masiva de datos generados por los satélites requiere grandes capacidades de almacenamiento y procesamiento. Además, los datos pueden incluir múltiples bandas espectrales y resoluciones espaciales, temporales y radiométricas, lo cual añade una capa adicional de complejidad al análisis [4].

La clasificación de imágenes satelitales es fundamental en la vigilancia de la Tierra, gracias a que permite identificar y monitorizar cambios en el uso del suelo, la expansión urbana, la cobertura vegetal, y otros fenómenos ambientales. Las imágenes satelitales también pueden ser utilizadas para detectar la deforestación, monitorear la salud de los cultivos, gestionar recursos hídricos y evaluar daños después de desastres naturales [5]. Debido a esto, la capacidad de analizar y clasificar grandes volúmenes de estas imágenes de forma precisa y rápida es esencial a la hora de tomar decisiones informadas en gestión ambiental y planificación territorial.

Para la clasificación y manejo de imágenes satelitales, el aprendizaje profundo ha demostrado ser extremadamente efectivo. Los modelos de las redes neuronales convolucionales (CNN) han mostrado un gran potencial en tareas de clasificación gracias a su capacidad de extraer características complejas y patrones de datos de imágenes [6]. Esto puede ser observado por ejemplo en el trabajo de dos Santos et al. (2019) [7], donde se emplearon técnicas de aprendizaje profundo para mejorar la precisión en la clasificación de imágenes de satélite en aplicaciones de monitoreo ambiental, y en un estudio de Kussul et al. (2017) [8] donde demostró cómo las CNN pueden ser utilizadas para la clasificación de tierras agrícolas con alta precisión.

Se puede decir que las arquitecturas de aprendizaje profundo como las CNN tienen algunas limitaciones en el manejo de tareas relacionadas con imágenes, en este contexto, los modelos Transformer muestran un buen desempeño en el procesamiento de datos secuenciales gracias a su capacidad para capturar relaciones de largo alcance, y a su flexibilidad para manejar diferentes tipos de datos [9]. Gracias a esto, estos modelos suponen una gran oportunidad para solucionar los problemas encontrados en el análisis de las imágenes satelitales como son las dificultades con las relaciones de largo alcance y la sensibilidad a la variación espacial.

A pesar de su utilidad, el entrenamiento desde cero de los modelos Transformer presenta varios desafíos. En primer lugar, debido a que son altamente parametrizados y pueden capturar relaciones complejas en los datos, estos necesitan de una gran cantidad de ejemplos para aprender representaciones efectivas, esto es especialmente desafiante en relación a la clasificación de imágenes satelitales, donde la recopilación de datos etiquetados a gran escala puede ser costosa y laboriosa. Además, la gran cantidad de parámetros presentes en estos modelos hacen necesario el requerimiento de una gran capacidad de computación para el procesamiento y la optimización eficientes durante el entrenamiento.

Otro factor que hay que considerar es el tiempo necesario para entrenar un modelo de alta calidad desde cero. El tiempo que puede llevar el entrenamiento

depende del tamaño del conjunto de datos y la complejidad del modelo, por lo que puede durar varios días, semanas o incluso meses.

La técnica de *TL* han surgido como una estrategia prometedora para lidiar con los desafíos planteados con el entrenamiento de modelos Transformers. Gracias a la capacidad de transferir conocimientos aprendidos en áreas relacionadas se puede mejorar significativamente el rendimiento de los modelos de aprendizaje profundo, especialmente en aquellos casos donde los conjuntos de datos de entrenamiento son limitados [10].

Este trabajo se centra en la evaluación comparativa del rendimiento de diferentes técnicas de *TL* sobre modelos Transformers en la clasificación de imágenes por satélite, obteniendo datos analizables para poder discernir las fortalezas y debilidades de los diferentes enfoques tomados durante la realización del trabajo.

El objetivo principal del trabajo será evaluar cómo se puede mejorar la eficacia de los modelos de clasificación de imágenes mediante la transferencia de conocimientos usando modelos Transformer preentrenados sobre otros conjuntos de datos.

# 2 Revisión de Literatura

En la actualidad, el campo de la inteligencia artificial ha sido revolucionado por los modelos de redes neuronales basados en la arquitectura Transformer, los cuales han demostrado una gran versatilidad y eficacia, especialmente en el procesamiento de secuencias de datos [11].

Desde el momento de su introducción en 2017, con la publicación "Attention is All You Need" de Vaswani et al. (2017) [12], los Transformers se han convertido en uno de los elementos de mayor importancia para una variedad de aplicaciones de aprendizaje profundo, superando incluso a las arquitecturas basadas en Redes Neuronales Recurrentes (RNN) y Memorias a Corto y Largo Plazo (LSTM) en tareas de procesamiento de lenguaje natural [13-14], generación de texto y comprensión del lenguaje natural [15].

Los modelos Transformer funcionan de forma eficaz a la hora de clasificar imágenes, esto es posible debido a su capacidad de capturar dependencias de largo alcance a través de la atención global, permitiendo un entendimiento más detallado del contenido visual [16-17]. Además, la técnica de *TL* ha emergido como un enfoque fundamental en el entrenamiento de modelos de aprendizaje profundo [18]. Gracias a que esta técnica aprovecha el conocimiento aprendido de una tarea para resolver otra relacionada, ha permitido realizar avances significativos en campos con datos limitados, o con obstáculos en la recolección de datos, como por ejemplo donde su alto coste hace que no sea factible la recolección [19-20].

La fusión de las técnicas de *TL* con los modelos *Transformer* ha abierto la oportunidad a numerosas e innovadoras aplicaciones en áreas que van más allá del procesamiento del lenguaje. En medicina, por ejemplo, los Transformers han sido implementados para su utilización a la hora de interpretar imágenes médicas y diagnosticar enfermedades con mayor precisión y velocidad que los métodos tradicionales [21-22]. En el campo de la robótica, esta tecnología permite a los sistemas aprender de simulaciones y aplicar los conocimientos obtenidos en el mundo real, logrando así adaptarse a situaciones y entornos cambiantes [23].

Asimismo, también se pueden ver beneficios en el ámbito ambiental, como la facilitación del monitoreo de cambios en las cubiertas terrestres, y la clasificación avanzada de imágenes satelitales, lo que da lugar a una mejor comprensión y gestión de los recursos naturales [24-25].

# 3 Materiales y Métodos

## 3.1 Definición de las Técnicas

### 3.1.1 Transfer Learning

El concepto de *TL* está basado en el concepto de reutilizar elementos de los modelos de aprendizaje automático (machine learning) preentrenados en nuevos modelos que se usaran para fines similares, es decir, al realizar una nueva tarea no se tiene que empezar desde cero, y se puede usar conocimiento de dominios ya existentes en el aprendizaje del nuevo dominio. Con esto se consigue optimizar los recursos y los datos necesarios para el entrenamiento.

La Figura 1 muestra las diferencias entre el aprendizaje automático tradicional y *TL*. Mientras que en el método tradicional se entrena un modelo separado para cada dominio, con la aplicación de *TL* se logra transferir el conocimiento aprendido de un dominio fuente a un dominio objetivo.

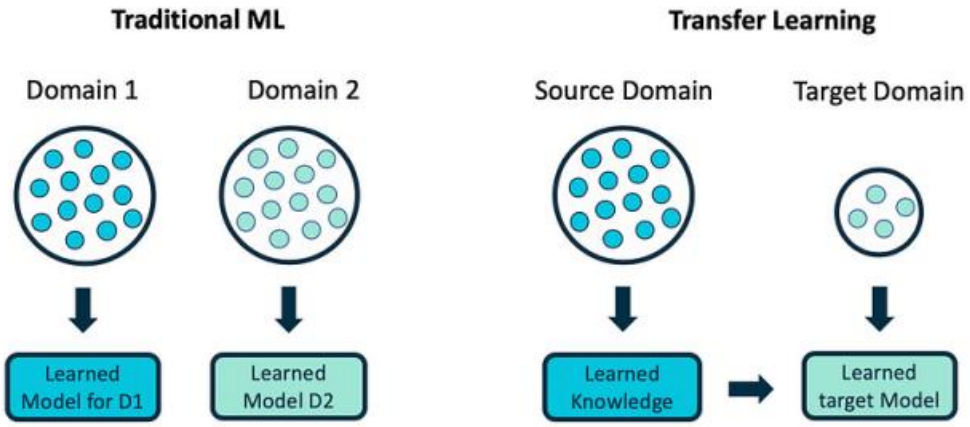


*Figura 1: Comparativa del aprendizaje automático tradicional a la izquierda, con el TL a la derecha. Se ve como en el aprendizaje tradicional cada dominio crea un modelo independiente, pero en TL se usa conocimiento del dominio de un modelo para generar un modelo nuevo (fuente: [10]).*

El uso de técnicas de *TL* en modelos basados en Transformers resulta especialmente útil, esto se debe a la gran cantidad de recursos computacionales que se necesitan para entrenar un modelo desde cero, así como de datos. Debido a esto, transferir datos de un modelo preentrenado permite reducir el tiempo y los recursos necesarios para el entrenamiento de un modelo de alta calidad.

La realización de la técnica se puede realizar de diversas maneras, entre las que se encuentran Fine-Tuning, Feature Extraction y Domain Adaptation. Fine-Tuning es la forma más común de *TL* y consiste en tomar un modelo preentrenado y ajustarlo (*fine-tune*) con datos específicos del nuevo dominio. De esta forma se entrena el modelo en el nuevo conjunto de datos partiendo de los pesos y parámetros del modelo preentrenado, logrando que el modelo aprenda las características del nuevo dominio sin olvidar el conocimiento general aprendido anteriormente [26].

Otra forma de lograrlo es mediante Feature Extraction, la cual utiliza el modelo preentrenado como un extractor de características. Además, se reemplazan las capas finales del modelo con capas nuevas adecuadas para la terea específica del nuevo dominio, logrando utilizar las características extraídas de las capas preentrenadas para entrenar las nuevas capas [27].

Por último, tenemos Domain Adaptation, la cual es una técnica enfocada en adaptar un modelo preentrenado para que funcione bien en un nuevo dominio.

Esto es logrado mediante el uso de técnicas como la regularización adversarial [28] o el aprendizaje contrastivo [29] para minimizar las diferencias entre el dominio fuente y el dominio objetivo.

### 3.1.2 Modelos Transformer

En el contexto del aprendizaje automático, se refiere como Transformers a una clase de modelos de lenguaje, los cuales están basados en redes neuronales desarrolladas para tareas de procesamiento de lenguaje natural. Estos modelos tienen la principal característica de no estar basados en recurrencia, y en su lugar, los modelos Transformer se basan en mecanismos de atención [30].

Los mecanismos de atención en los modelos Transformer son innovaciones clave, y gracias a estos se han logrado grandes avances en el procesamiento de datos secuenciales. Los datos secuenciales son un conjunto de datos que tienen un orden intrínseco en el cual cada elemento está relacionado con el siguiente. Entre los más comunes se encuentran el texto, donde los datos secuenciales suelen ser oraciones, párrafos o documentos completos donde cada palabra o token sigue un orden específico que es crucial para el significado del texto; y las series temporales, los cuales son datos donde las observaciones están ordenadas en el tiempo [31].

En el contexto de la clasificación de imágenes los datos secuenciales que se utilizan son las secuencias derivadas de la estructura espacial de las imágenes. Para ello, la imagen se divide en parches más pequeños y se tratan los parches como un token en una secuencia. Después se secuencian y alimentan los parches al modelo Transformer para su procesamiento, logrando de esta forma realizar la clasificación [32].

Estos mecanismos permiten al modelo ponderar de manera diferencial la importancia de diferentes partes de la entrada en cada paso de la predicción. Es decir, gracias a estos mecanismos, los Transformers pueden atender a todas las partes de la secuencia simultáneamente, y con eso asignar mayor o menor importancia a cada parte según sea necesario para la tarea dada.

Uno de los mayores beneficios que presentan estos mecanismos son su capacidad para manejar secuencias largas de datos de manera más efectiva que los modelos tradicionales. Además, estos modelos presentan un alto grado de paralelización, logrando que los Transformer puedan gestionar dependencias a largo plazo dentro de la secuencia con mayor facilidad, lo cual resulta útil para tareas como la clasificación de imágenes [33].

### 3.1.3 Cross Entropy Loss

*Cross Entropy Loss*, también conocido como *Log loss,* es una métrica utilizada para medir el rendimiento de un modelo de clasificación cuya salida es una probabilidad entre 0 y 1. Esta métrica es ampliamente utilizada en tareas de clasificación, en especial en aquellas con un enfoque en problemas de clasificación multiclase. El objetivo es minimizar esta pérdida durante el entrenamiento del modelo, lo que implica mejorar la precisión de las predicciones.

La entropía cruzada (Cross Entropy) proviene de la teoría de la información y mide la diferencia entre dos distribuciones de probabilidad: la distribución real y la distribución predicha por el modelo. Centrándonos en el contexto del aprendizaje automático, la entropía cruzada calcula la disimilitud entre la etiqueta verdadera (distribución real) y la probabilidad predicha [34].

Para el *TL* y el uso de Transformers visuales, Cross Entropy Loss es utilizada para optimizar el modelo en tareas de clasificación de imágenes. Durante su

entrenamiento, los modelos Transformer visuales calcularán la disimilitud entre las etiquetas verdaderas de las imágenes, y las probabilidades predichas por el modelo, ajustando consecuentemente los pesos del modelo para minimizar esta disimilitud [35].

La implementación de esta técnica presenta ventajas en su aplicación para la optimización del entrenamiento de modelos Transformers. Una de las ventajas significativas de la implementación del Cross Entropy Loss es su tendencia a converger más rápido que otras funciones de pérdida debido a que proporciona gradientes más grandes y significativos, especialmente cuando las predicciones están lejos de las etiquetas verdaderas. Gracias a esto, se facilita el ajuste eficiente de los parámetros del modelo durante el entrenamiento [36].

Otro beneficio que proporciona la implementación de esta función de pérdida es la penalización severa de las predicciones incorrectas, la cual se realiza de manera exponencial. En caso de que el modelo esté muy seguro de su predicción, pero equivocado, la pérdida será alta. Gracias a esto, el modelo se puede ajustar de manera efectiva mediante la reducción de la probabilidad de que se cometan errores graves [37].

Un último beneficio mencionable sería que la función de pérdida se alinea bien con la interpretación probabilística de las salidas del modelo. En el contexto de una clasificación multiclase, las salidas se pueden interpretar como probabilidades de pertenencia a cada clase, lo cual facilita una evolución clara del rendimiento del modelo [38].

### 3.1.4 Optimizador Adam

Adam (Adaptive Moment Estimation) es uno de los optimizadores más populares en el campo del aprendizaje profundo. Introducido por Diederik P. Kingma y Jimmy Ba en 2014 [39], en él se combinan las ventajas de dos métodos de optimización: AdaGrad yRMSProp. Adam tiene como objetivo mejorar tanto la velocidad de convergencia como la estabilidad del entrenamiento, lo cual logra adaptando los pasos de aprendizaje de cada parámetro del modelo de forma individual.

Este optimizador presenta grandes ventajas en el entrenamiento de modelos Transformer visuales para la categorización de imágenes, entre la que se destaca la adaptabilidad, la corrección de sesgo y la eficiencia computacional.

La adaptabilidad del optimizador Adam se basa en que este adapta la tasa de aprendizaje para cada parámetro, permitiendo de esta forma un ajuste fino y una convergencia rápida. Gracias a esta adaptabilidad, Adam es útil en problemas con datos dispersos o características poco frecuentes.

Las correcciones de sesgo que proporciona Adam mejoran la estabilidad del entrenamiento, especialmente en las primeras etapas. Con esta corrección, la implementación de este optimizador puede ser crucial para evitar inicios inestables.

En las aplicaciones a gran escala y en modelos grandes se requiere una gran cantidad de recursos computacionales, debido a esto, la implementación del optimizador Adam presenta una gran ventaja. Esta ventaja se debe a que Adam es computacionalmente eficiente, lo que hace que requiera poca memoria adicional.

### 3.1.5 Early Stopping

En el contexto del entrenamiento de los modelos de aprendizaje automático, Early Stopping es una técnica de regularización utilizada para prevenir el

sobreajuste durante el entrenamiento de los modelos. El uso de esta técnica implica detener el entrenamiento cuando el rendimiento del modelo en los datos de validación deja de mejorar después de un número determinado de iteraciones. De esta forma, se evita que el modelo se ajuste demasiado a los datos de entrenamiento.

El sobreajuste ocurre cuando un modelo aprende demasiado bien las características específicas del conjunto de entrenamiento, lo que incluye el ruido y las peculiaridades de los datos. Debido a esto, un modelo que ha sufrido sobreajuste tiene un peor rendimiento en aquellos datos que no ha visto (generalización).

Early Stopping afronta este problema mediante la monitorización del rendimiento del modelo en un conjunto de validación y deteniendo el entrenamiento cuando se observa que el rendimiento del modelo comienza a deteriorarse.

Para su correcto funcionamiento, es necesaria la división de los datos disponibles en tres conjuntos: entrenamiento, validación y prueba. Usando el conjunto de validación exclusivamente para evaluar el rendimiento del modelo durante el entrenamiento. Para ello, durante el entrenamiento del modelo del modelo se monitorea la función de pérdida en el conjunto de validación al final de cada época.

Además, es necesario determinar un criterio de parada el cual provoca que el entrenamiento se detenga en caso de no lograrse una mejora en la pérdida en el conjunto de validación después de un número determinado de épocas consecutivas, lo que es conocido como “paciencia”.

La implementación del Early Stopping tiene como ventaja una minimización del sobreajuste del modelo al detener el entrenamiento en el punto óptimo donde el modelo generaliza mejor en los datos de validación. Además, la implementación de esta técnica reduce el tiempo de entrenamiento y los recursos computacionales necesarios para el entrenamiento, debido a que evita épocas adicionales innecesarias una vez que el modelo ha alcanzado su rendimiento óptimo.

El Early Stopping es una técnica sencilla de implementar, la cual ha demostrado ser efectiva en una amplia variedad de aplicaciones de aprendizaje profundo, como la clasificación de imágenes y le procesamiento de lenguaje natural [40].

### 3.1.6 Data Augmentation

*DA* es una técnica utilizada para aumentar la cantidad y la diversidad de los datos de entrenamiento sin la necesidad de recopilar nuevos datos. Se logra mediante la aplicación de transformaciones aleatorias a los datos existentes, generando de esta forma muestras sintéticas que pueden ayudar a mejorar la robustez y la capacidad de generalización de los modelos de aprendizaje automático.

La utilización de *DA* tiene como objetivo prevenir el sobreajuste del modelo, y mejorar su generalización. Esto se logra al introducir variaciones en los datos de entrenamiento, lo que provoca que el modelo aprenda a ser más flexible y menos dependiente de características específicas de las muestras originales.

Para la implementación de *DA* en el trabajo, se utilizarán las técnicas de rotación y reflejo (flip), y se evitará el uso de técnicas como el recorte (crop). Las técnicas de rotación o reflejo crean una nueva imagen sin perder información de la imagen original. Sin embargo, técnicas como el recorte consisten en la

extracción de una subimagen de la imagen original, esto puede provocar conflictos en el entrenamiento del modelo a la hora de categorizar imágenes debido a que puede perderse información necesaria para la clasificación [41], [42].

## 3.2 Descripción del dataset EuroSat

El conjunto de datos del dataset ubicado en [43] presenta una recopilación de imágenes satelitales etiquetadas, las cuales sirven como recurso para la investigación y el análisis en el ámbito de la clasificación del uso del suelo y la cobertura terrestre.

En total, el dataset contiene 27,000 imágenes georreferenciadas y etiquetas en formato JPEG, cada imagen contiene 13 bandas espectrales cubriendo el espectro visible. Además, las imágenes se capturan en 3 canales (RGB) y Los parches tienen un tamaño de 64x64 píxeles. Estas imágenes corresponden a 10 clases de tipo de cubiertas y uso de suelo diferentes con una cantidad de entre 2,000 y 3,000 imágenes por clase. El criterio para la selección de estas 10 clases de superficie cubierta y superficie usada ha sido que sean visible a una resolución de 10 metros por píxel y estén lo suficientemente presentes en el European Urban Atlas [44] como para generar miles de parches de imagen.

Para diferenciar entre los distintos usos de la tierra agrícola, el dataset cubre las clases de cultivo anual, cultivo permanente y pasto. También diferencia entre las zonas construidas. Por lo tanto, cubre las clases de carretera, construcción residencial y construcción industrial. La clase residencial se crea usando las clases de tejido urbano descrito en el European Urban Atlas. Los diferentes cuerpos de agua aparecen en las clases río, mar y lago. Además, las zonas no desarrolladas se incluyen en las clases bosque y vegetación herbácea, una visualización de las clases se puede ver en la siguiente Figura 2 [45].

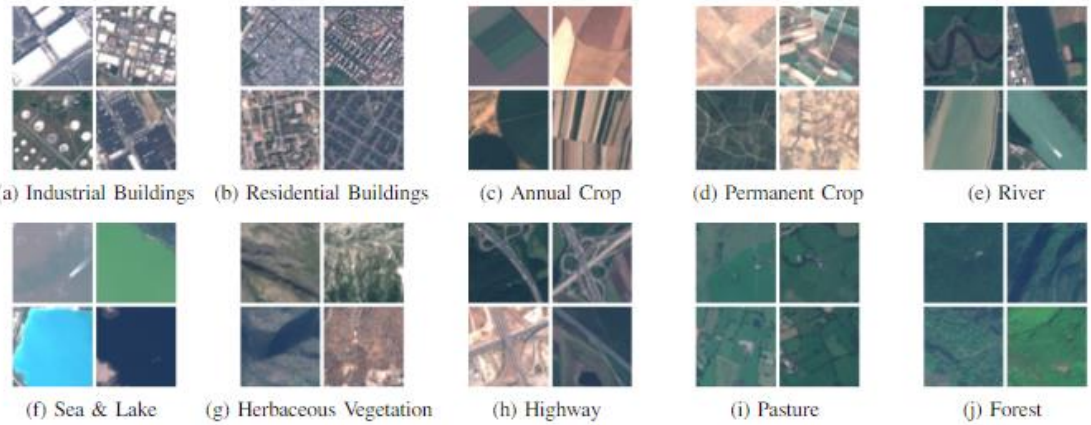


*Figura 2: Vision general de muestras de parches de imágenes de las 10 clases incluidas en el conjunto de datos EuroSat. Las imágenes tienen un tamaño de 64x62 píxeles. Cada clase contiene de 2,000 a 3,000 imágenes. En total, el conjunto de datos tiene 27,000 imágenes georreferenciadas.*

## 3.3 Selección y Configuración de Modelos

Se escogieron un total de 3 modelos entrenados en el contexto de la clasificación de imágenes: Vision Transformer [46] (Vit), Swin Transformer model [47] y Data-efficient Image Transformer [48] (Deit).

### 3.3.1 Modelo ViT

Este modelo usa una arquitectura de un Transformer diseñado originalmente para tareas relacionadas con texto, como se aprecia en la Figura 3 el modelo Vit representa la imagen entrante como una serie de partes de una imagen al igual que la serie de "embeddings" de palabras utilizados cuando se usan Tranformers para texto, y predice directamente las etiquetas de clase de la imagen.

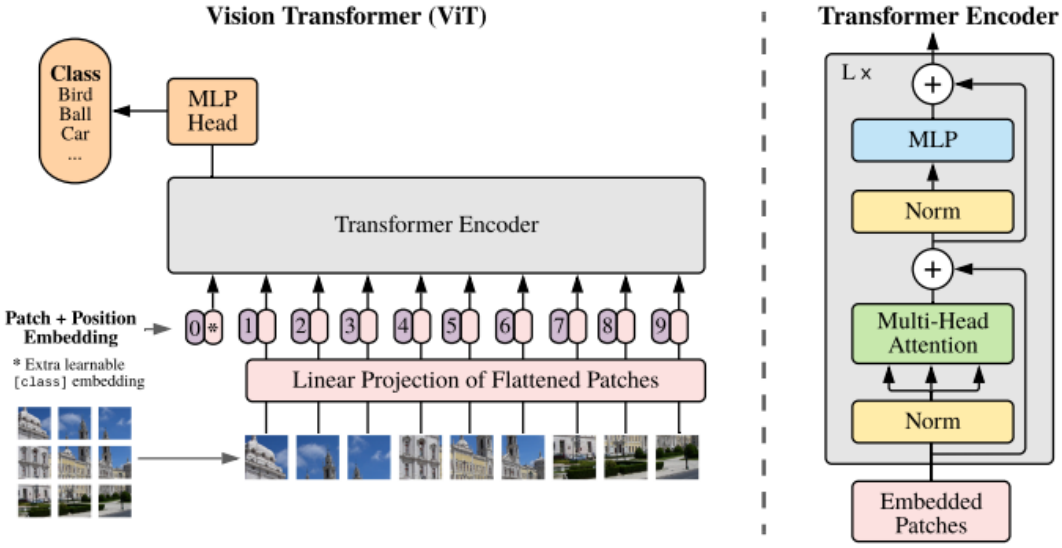


*Figura 3: Visión general de la arquitectura del modelo Vision Transformer (ViT). Se comienza dividiendo la imagen en parches de tamaño fijo. Cada parche es incrustado linealmente, complementado con incrustaciones posicionales y procesado a través de un codificador Transformer estándar. Para la clasificación, se agrega un "token de clasificación" aprendible a la secuencia. El diseño del codificador Transformer se inspira en Vaswani et al. (2017) [46].*

El modelo Vit presenta un enfoque innovador que difiere de las arquitecturas convencionales basadas en las redes convolucionales. Preentrenado sobre ImageNet-21k (14 millones de imágenes, 21,843 clases). Posteriormente, se realizó un fine-tuning en un subconjunto más restringido y altamente referenciado, ImageNet 2012, que consta de 1 millón de imágenes y 1,000 clases, ajustando así el modelo para optimizar su precisión y eficacia en tareas de clasificación de imágenes más específicas. a una resolución de 224x224. Este modelo ha demostrado una capacidad extraordinaria para desarrollar una comprensión profunda y generalizable del contenido visual a gran escala. El procedimiento de entrenamiento dual utilizado no solo mejora la precisión del modelo, sino que también lo hace más robusto frente a una variedad más amplia de tareas visuales [46].

### 3.3.2 Modelo Swin

El modelo Swin Transformer es una arquitectura de red neuronal que representa un gran avance para las tareas de visión por computadora, esto se debe gracias a su estructura jerárquica que permite una eficiente escalabilidad y versatilidad.

A diferencia de otros modelos, el modelo Swin trabaja realizando la construcción jerárquica de mapas conceptuales mediante la unión de partes de la imagen en capas más profundas, reduciendo la resolución de la imagen gradualmente y permitiendo el manejo eficiente de diferentes escalas y tamaños de imagen, como muestra en la Figura 4 [47].

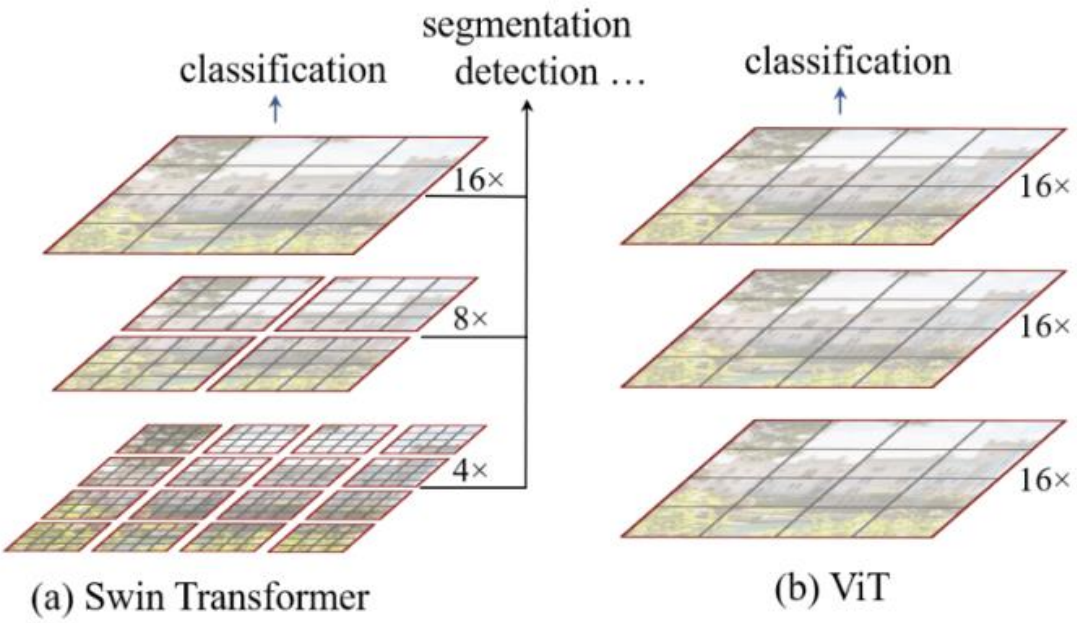


*Figura 4: Comparación de las arquitecturas Swin Transformer (a) y Vision Transformer (ViT) (b) para tareas de clasificación, segmentación y detección. Se observa que Swin Transformer procesa la imagen en diferentes niveles de resolución, mientras que ViT mantiene la misma resolución a lo largo de toda la arquitectura.*

Gracias a este enfoque, el modelo posibilita una atención localizada y eficiente desde ventanas de atención móviles que se desplazan sobre los parches de la imagen, logrando así reducir la complejidad computacional hasta un nivel linear con respecto al tamaño de la imagen entrante. Esto resulta en una reducción significativa en comparación con los métodos de atención global.

### 3.3.3 Modelo DeiT

Los modelos DeiT, abreviatura de Data-efficient Image Transformer, son una evolución de los modelos Transformer diseñados específicamente para la clasificación de imágenes. Estos modelos suponen una evolución clave en esta área aprovechando la eficiencia de datos en el entrenamiento de modelos Transformer. Esta eficiencia es lograda mediante la técnica de entrenamiento auto-supervisada conocida como “distillation”. Esta técnica consiste en el uso de un modelo preentrenado, denominado “teacher” [49] para guiar el aprendizaje del modelo Deit sin la necesidad de grandes cantidades de datos anotados. De esta forma, el modelo puede alcanzar un rendimiento similar a los

modelos supervisados usando una pequeña parte de los datos, reduciendo así significativamente los recursos y el tiempo necesario para el entrenamiento.

Los modelos DeiT También incorporan mecanismos de atención que permiten que el modelo se centre en las partes más relevantes de la imagen para la tarea de clasificación. A medida que la información avanza a través de las capas del modelo, se van construyendo los mapas conceptuales de forma más compleja y jerárquica, gracias a esto, el modelo permite representaciones cada vez más abstractas y poderosas de los datos visuales. Estos procesos pueden verse representados en la Figura 5 [50].

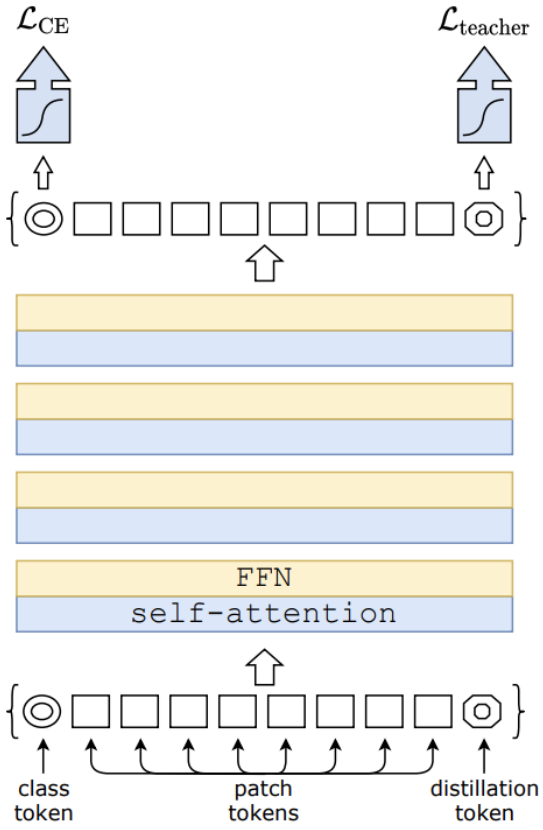


*Figura 5: Estructura de un modelo DeiT durante el proceso de entrenamiento. Se muestran los tokens de clase y de destilación al principio y al final de la secuencia de tokens de parche. La red utiliza tanto la pérdida de entropía cruzada LCE como la pérdida del profesor Lteacher para la destilación del conocimiento. El modelo aplica mecanismos de auto-atención y redes freedforward (FNN) en su arquitectura Transformer para procesar y clasificar imágenes.*

# 4 Desarrollo

## 4.1 Descripción de los experimentos

En esta sección se detallan las configuraciones, técnicas y bibliotecas usadas en el entrenamiento de los modelos Transformer. La información se organiza de manera cronológica y describiendo cada paso del proceso, desde la configuración del entorno hasta las fases de entrenamiento.

### 4.1.1 Versiones de las Bibliotecas

La implementación de los modelos se realizó en el lenguaje de programación Python debido a las librerías especializadas presentes en este ecosistema. Para las librerías utilizadas se incluyó Torch (versión 2.3.0+cu121), el núcleo de PyTorch, debido a que proporciona una plataforma flexible y potente para la computación de tensores y la diferenciación automática, presentando una interfaz intuitiva y una integración directa con aceleradores de hardware, lo cual es esencial para procesar eficientemente grandes volúmenes de datos.

Además, se utilizó la librería Torchvision (versión 0.18.0+cu121) que complementa a PyTorch, la cual ofrece un conjunto de herramientas para el procesamiento de imágenes. También se incluyó la librería Transformers (versión 4.41.1), desarrollada por Hugging Face, la cual permite el acceso a una variedad de modelos Transformer preentrenados.

### 4.1.2 Preparación de los Datos

Para la carga y preprocesamiento de las imágenes, las imágenes satelitales fueron cargadas y redimensionadas a un tamaño estándar de 224x224 píxeles mediante el uso del algoritmo de interpolación bilineal, el cual realiza una interpolación lineal en dos dimensiones, lo que permite redimensionar las imágenes obtenidas manteniendo una calidad visual adecuada.

Además, las imágenes fueron convertidas a tensores para que estén en un formato adecuado. Este proceso implica la normalización de los valores de los píxeles al rango entre 0 y 1, de esta forma se facilita el entrenamiento de los modelos al asegurar que los valores de entrada estén en un rango estándar y predecible.

En cuanto a la normalización de las imágenes, posterior a la carga y preprocesado, se aplicó una normalización con media 0.5 y desviación estándar 0.5. Este proceso ajusta los valores de los píxeles a un rango que facilita el entrenamiento gracias a que el optimizador funciona de manera más eficiente al reducir la varianza en las entradas.

Otros beneficios obtenidos de la normalización de las imágenes a una media de 0.5 y desviación estándar de 0.5 consiste en que los datos se centran en torno a cero y se distribuyen de manera uniforme en un rango de [-1,1]. Esto asegura que todas las características de entrada contribuyan de manera equilibrada al cálculo de los gradientes durante el entrenamiento. Además, este proceso asegura la consistencia de los datos para todos los lotes, lo que garantiza que estos tengan propiedades estadísticas similares. Esto evita que el modelo aprenda patrones basados en las diferencias en escala entre distintos lotes, mejorando la robustez y consistencia del entrenamiento.

### 4.1.3 Configuración del Modelo

Se seleccionaron modelos de Transformers visuales preentrenados disponibles en la biblioteca `Transformers`, en concreto se seleccionaron los modelos Vit (Vision Transformer), Deit (Data-efficient Image Transformer) y Swin Transformer.

Para el aprendizaje, los modelos preentrenados se ajustaron utilizando la biblioteca `torch`. Para todos los modelos se utilizó el optimizador Adam con una tasa de aprendizaje inicial de $5.10^{-5}$.

### 4.1.4 Entrenamiento del Modelo

Para la primera fase, se entrenaron los modelos sin el uso de técnicas de aumento de datos (*DA*), se dividieron las imágenes en los conjuntos de entrenamiento, validación y prueba, con un 75%, 15% y 10% de las imágenes en cada conjunto respectivamente, y se tuvo en cuenta que fueran conjuntos disjuntos. Además, se incluyeron las funciones de pérdida `CrossEntropyLoss` y el optimizador Adam mencionado anteriormente.

En la segunda fase de entrenamiento se incluyeron las técnicas de aumento de datos conocidas como *DA*, las técnicas que fueron utilizadas para mejorar la robustez del modelo fueron la rotación y el reflejo de las imágenes. La elección de estas técnicas y no otras de *DA* se basa en la pérdida parcial de información en la imagen mediante el uso de otras técnicas como, por ejemplo, el recorte. Por tanto, solo se usaron técnicas que no provoquen pérdida de información en la imagen.

Para esta segunda fase se mantuvieron las mismas divisiones de las imágenes en sus conjuntos respectivos, funciones de pérdida y optimizadores que en la primera fase con el objetivo de obtener métricas válidas para la comparación entre el rendimiento del uso de la técnica en los modelos.

Durante el entrenamiento, se fueron monitorizando las métricas de precisión y perdida, tanto del conjunto de entrenamiento como del conjunto de validación para asegurar el rendimiento del modelo y se usó la técnica de *early stopping* para evitar el sobreajuste.

## 4.2 Generación de Modelos

Durante el proceso de generación de modelos es importante definir las funciones de pérdida y el optimizador, los cuales juegan un papel necesario en el proceso. La función de pérdida tiene como función cuantificar la discrepancia entre las etiquetas predichas por el modelo y las etiquetas reales de los datos. Debido a esto, la función de pérdida va a ir indicando el rendimiento del modelo durante el entrenamiento, por lo que se va a tener como objetivo minimizarla a través del ajuste de los pesos de la red.

La función de perdida usada para la generación de los modelos, la CorssEntropyLoss, presenta grandes ventajas a la hora de la clasificación con múltiples clases, ajustándose así a la tarea de clasificación del dataset. Estas ventajas se deben a la asignación de un coste más alto cuando la probabilidad predicha para la clase verdadera es baja, mejorando de esta forma tanto la efectividad para aprender clasificaciones correctas, como la confianza del modelo en sus predicciones.

En cuanto al optimizador, se ha optado por la utilización del optimizador Adam (Adaptive Moment Estimation) debido a las ventajas que presenta para la clasificación de imágenes. Adam combina las mejores propiedades de AdaGrad y RSMProp para manejar tasas de aprendizaje adaptativas, dando lugar a una de sus principales características, el cálculo de tasas de aprendizaje individual para diferentes parámetros. Además, con el objetivo de mejorar la eficiencia del entrenamiento y prevenir el sobreajuste, se ha usado la técnica de *early stopping*.

Por último, el modelo es guardado en diferentes puntos del entrenamiento, permitiendo así la recuperación del modelo en un punto específico y poder seleccionar la mejor versión del modelo tras la conclusión del entrenamiento. Lo cual resulta de gran importancia para poder ser evaluado con el conjunto de prueba.

# 5 Resultados y conclusiones

En este apartado se va a realizar una comparativa del desempeño de los modelos, así como una explicación detallada sobre el rendimiento de estos, recalcando puntos fuertes e identificando las áreas donde los modelos han tenido los mayores problemas a la hora de la categorización de las imágenes satelitales. Primero se comparará el desempeño de cada modelo con y sin la utilización de *DA*, y finalmente se hará una comparativa del desempeño entre los modelos.

## 5.1 Modelo ViT

Para el entrenamiento del modelo y su posterior utilización para la realización de las predicciones, se han monitoreado y recopilado la pérdida y la precisión de los conjuntos de train y validación como se muestra en la Figura 6. El modelo ha presentado una rápida convergencia a lo largo de las épocas y se puede apreciar como la implementación de la técnica de *DA* ha mejorado la capacidad del modelo para generalizar desde el principio gracias al aumento de datos. Sin embargo, a pesar de un mayor rendimiento, el modelo en el que se ha implementado la técnica ha necesitado de 40 écpocas, un mayor número de épocas comparado con el modelo en el que no se implementó, el cual necesitó 30 épocas.

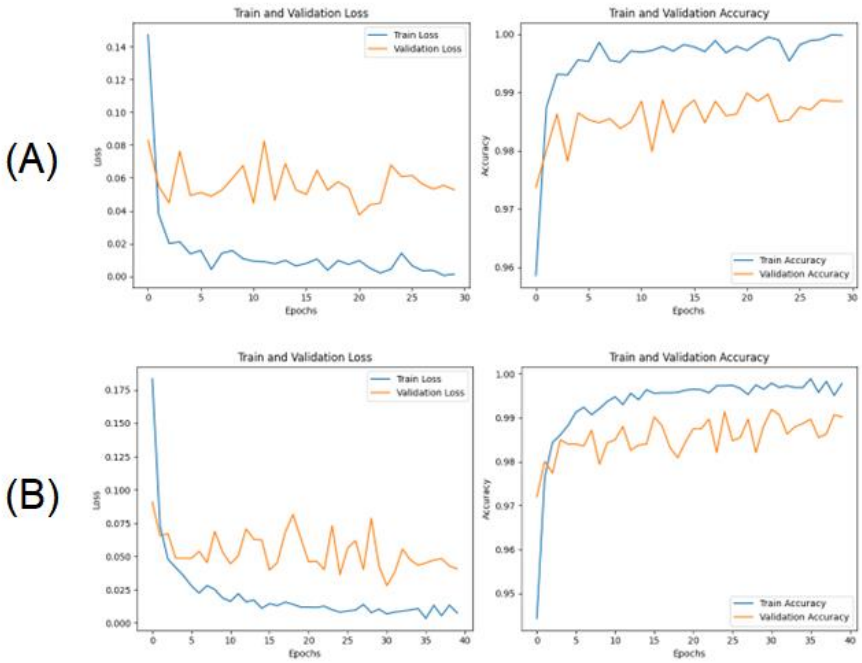


*Figura 6 Gráficas de pérdida y precisión durante el entrenamiento y la validación del modelo Vision Transformer (Vit). La Figura (A) corresponde al modelo sin la utilización de DA, y la Figura (B) pertenece al modelo en el que se utilizó. Las gráficas de la izquierda muestran la pérdida (loss) del conjunto de entrenamiento y validación a lo largo de las épocas. Las gráficas de la derecha muestran la precisión (accuracy) en los mismos conjuntos y durante las mismas épocas.*

El modelo ViT tras su entrenamiento sobre el dataset de las imágenes ha demostrado una precisión significativa y un rendimiento muy alto en las categorías predichas. Sin embargo, el modelo ha tenido dificultades en la predicción de ciertas clases en diferentes proporciones según la clase real, a continuación, vamos a hacer un análisis de los resultados obtenidos y a resaltar las causas principales por las que el modelo ha cometido los mayores errores demostrando una mayor confusión.

En cuanto al modelo ViT entrenado sin el uso de *DA*, como se puede apreciar en la Figura 7, ha demostrado los mejores resultados en la predicción de la clase Residential, solo una de las muestras ha sido mal clasificada, también ha demostrado gran eficiencia en la predicción de las clases Annual Crop, y Sea Lake, donde ha demostrado tasas de error muy bajas.

Mientras tanto, el modelo ha mostrado los mayores errores en las clases de Permanent Crop, la cual ha confundido 7 veces con Herbaceous Vegetation; y River, la cual ha sido confundida 6 veces con Highway.

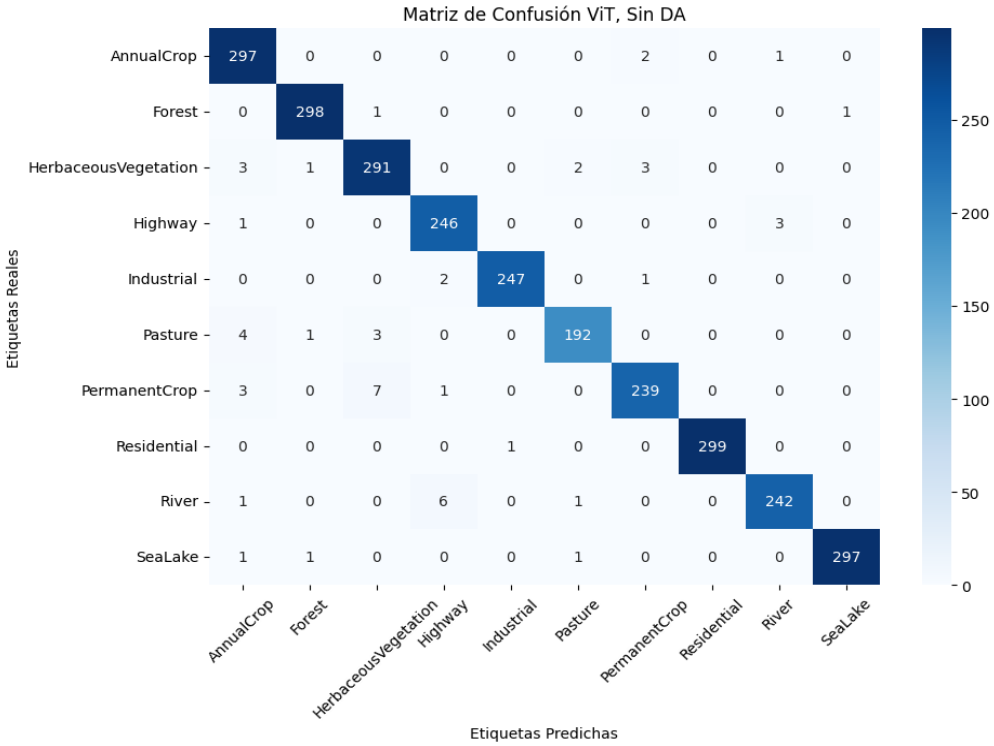


*Figura 7 Matriz de confusión del modelo ViT Transformer sin el uso de la técnica de DA. La matriz muestra el número de predicciones correctas y las confusiones entre las distintas clases. En el eje vertical se pueden observar las etiquetas reales, y en el eje horizontal las etiquetas predichas.*

Para el mayor error producido en el modelo, como se muestra en la Figura 8, podemos atribuir este error a la similitud en el color y la textura de las imágenes, así como a la baja resolución de estas. Una estrategia que podría resultar efectiva para combatir la confusión del modelo entre ambas clases sería la inclusión de imágenes con diferentes resoluciones. Así como aumentar el número de ejemplos de ambas clases en el conjunto de entrenamiento, asegurando que se incluyan imágenes donde se aprecian diversas condiciones de crecimiento y estaciones.

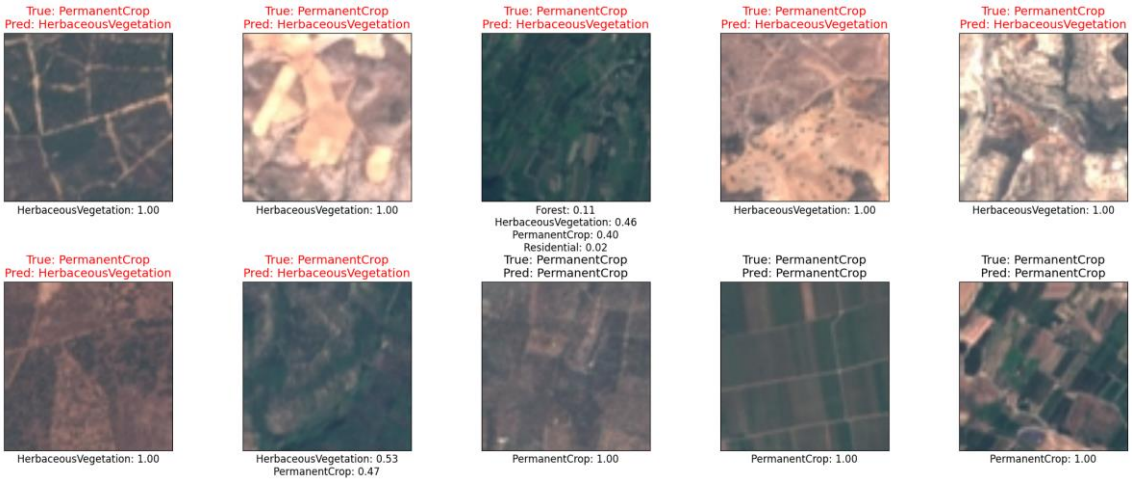


*Figura 8 Imágenes de la clase Permanent Crop con la predicción hecha por el modelo ViT sin el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase Permanent Crop con la clase Herbaceous Vegetation.*

En cuanto a la clase River, como podemos apreciar en la Figura 9, el mayor error ocurrido en el modelo tuvo lugar con la clase Highway. Uno de los factores principales para el error se debe a la similitud visual entre los ríos y las autopistas y a la resolución de las imágenes, tanto los ríos como las autopistas pueden aparecer como estructuras lineales y la baja resolución puede dificultar la diferenciación de texturas y detalles específicos. Como estrategia para remediar esta confusión, se podría incluir más ejemplos de ambas clases en diferentes condiciones.

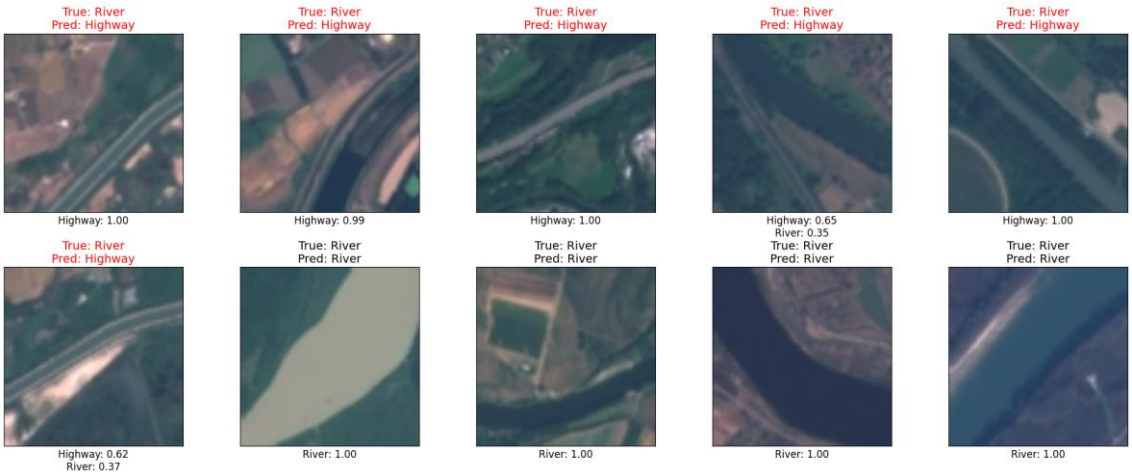


*Figura 9 Imágenes de la clase River con la predicción hecha por el modelo ViT sin el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase River con la clase Highway.*

Tras la implementación de las técnicas de *DA* obtenemos la matriz de confusión de la Figura 10, en ella podemos apreciar una mejora con respecto a su contraparte sin la utilización de esta técnica en la realización de las predicciones. Podemos ver como la clase Permanent crop ha tenido una reducción de su confusión con la clase Herbaceous Vegetation. Sin embargo, en la clase Herbaceous Vegetation ha tenido una mayor confusión con la clase Permanent Crop, esto se puede deber a que el aumento de datos introdujo variaciones que a pesar de mejorar el reconocimiento de Permanent Crop, no lograron

diferenciar adecuadamente las características específicas de la clase Herbaceous Vegetation.

Las imágenes confundidas entre las clases mencionadas y apreciables en la Figura 11 tienen una gran similitud en sus características visuales y en sus texturas, lo cual puede haber conducido al modelo a la confusión.

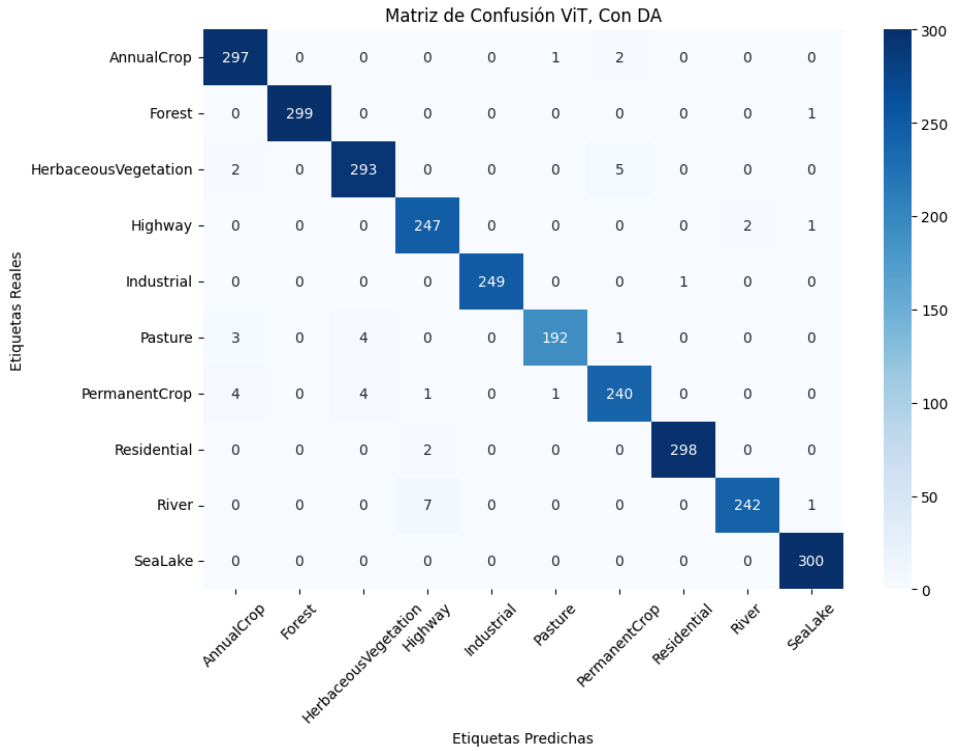


*Figura 10 Matriz de confusión del modelo ViT Transformer con el uso de la técnica de DA. La matriz muestra el número de predicciones correctas y las confusiones entre las distintas clases. En el eje vertical se pueden observar las etiquetas reales, y en el eje horizontal las etiquetas predichas.*

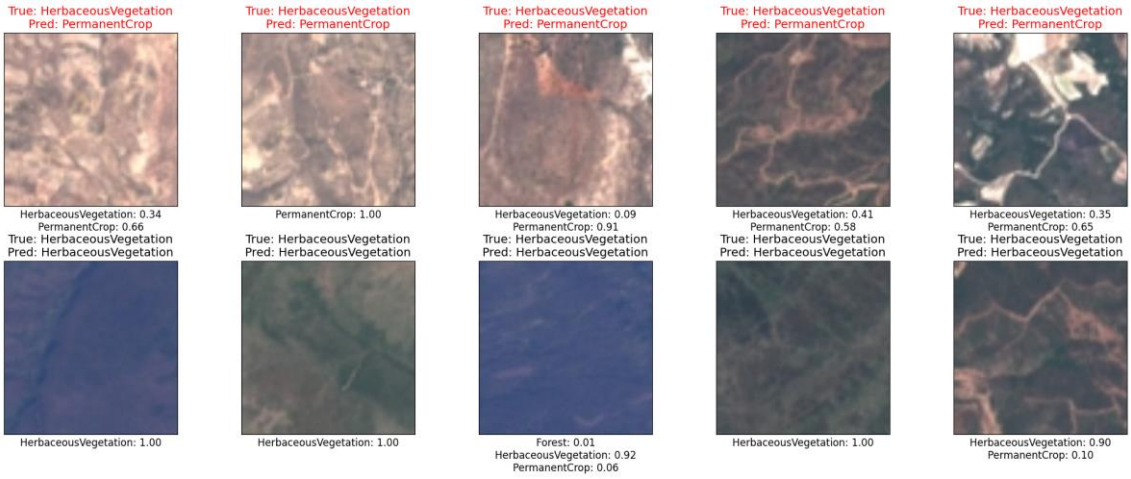


*Figura 11 Imágenes de la clase Herbaceous Vegetation con la predicción hecha por el modelo ViT con el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase Herbacecous Vegetation con la clase Permanent Crop.*

En cuanto a la clase River y su confusión con la clase Highway, como se puede ver en la Figura 12, ha mantenido la mayoría de los errores cometidos por el modelo sin la implementación de *DA*. Esto indica que el modelo no ha podido generalizar mejor mediante la implementación de la técnica en contextos donde la similitud visual y las texturas son muy similares.


*Figura 12 Imágenes de la clase River con la predicción hecha por el modelo ViT con el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase River con la clase Highway.*

## 5.2 Modelo Swin

La monitorizacion y recopilado la pérdida y la precisión de los conjuntos de train y validación como se muestra en la Figura 13. El modelo ha presentado una rápida convergencia, y se puede apreciar cómo la implementación de la técnica de *DA* ha tenido un impacto significativo en cuanto a la precisión y la pérdida del conjunto de validación, lo cual sugiere que el modelo se ha beneficiado de esto y ha logrado una mayor capacidad de generalización incluso desde las primeras épocas.

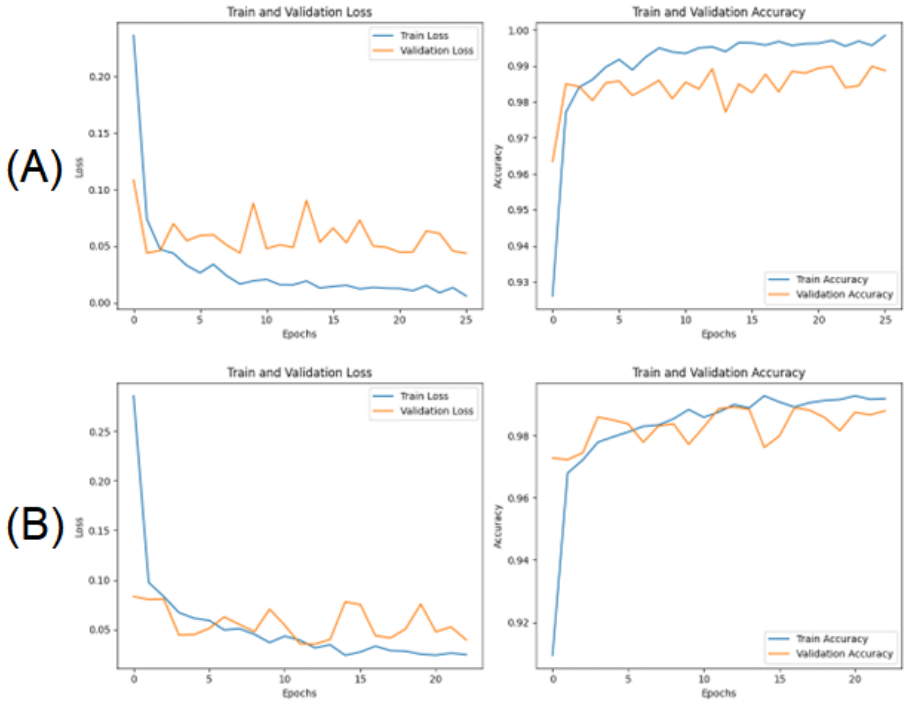


*Figura 13 Gráficas de pérdida y precisión durante el entrenamiento y la validación del modelo Swin Transformer. La Figura (A) corresponde al modelo sin DA, y la Figura (B) pertenece al modelo en el que se utilizó. Las gráficas de la izquierda muestran la pérdida (loss) del conjunto de entrenamiento y validación a lo largo de las épocas. Las gráficas de la derecha muestran la precisión (accuracy) en los mismos conjuntos y durante las mismas épocas.*

En cuanto a la confusión demostrada por el modelo tras finalizar su entrenamiento, para el modelo en el que no se ha implementado la técnica de *DA* podemos apreciar en la Figura 14 como el modelo demuestra una gran precisión a la hora de realizar las predicciones. Sin embargo, la precisión del modelo difiere con respecto a las clases, demostrando la mayor precisión en la clase Industrial y Herbaceous Vegetation, donde el modelo ha errado en la predicción un total de 0 y 1 veces respectivamente.

También podemos ver que el modelo ha tenido mayores dificultades, y por tanto mayor confusión en las clases Permanent Crop, Pasture y Annual Crop. Dando lugar a las mayores confusiones entre clases, las cuales son: Permanent Crop y Herbaceous Vegetation, donde el modelo ha confundido la primera con la segunda un total de 9 veces; y Pasture con Herbaceous Vegetation, donde el modelo ha confundido la primera con la segunda 8 veces.

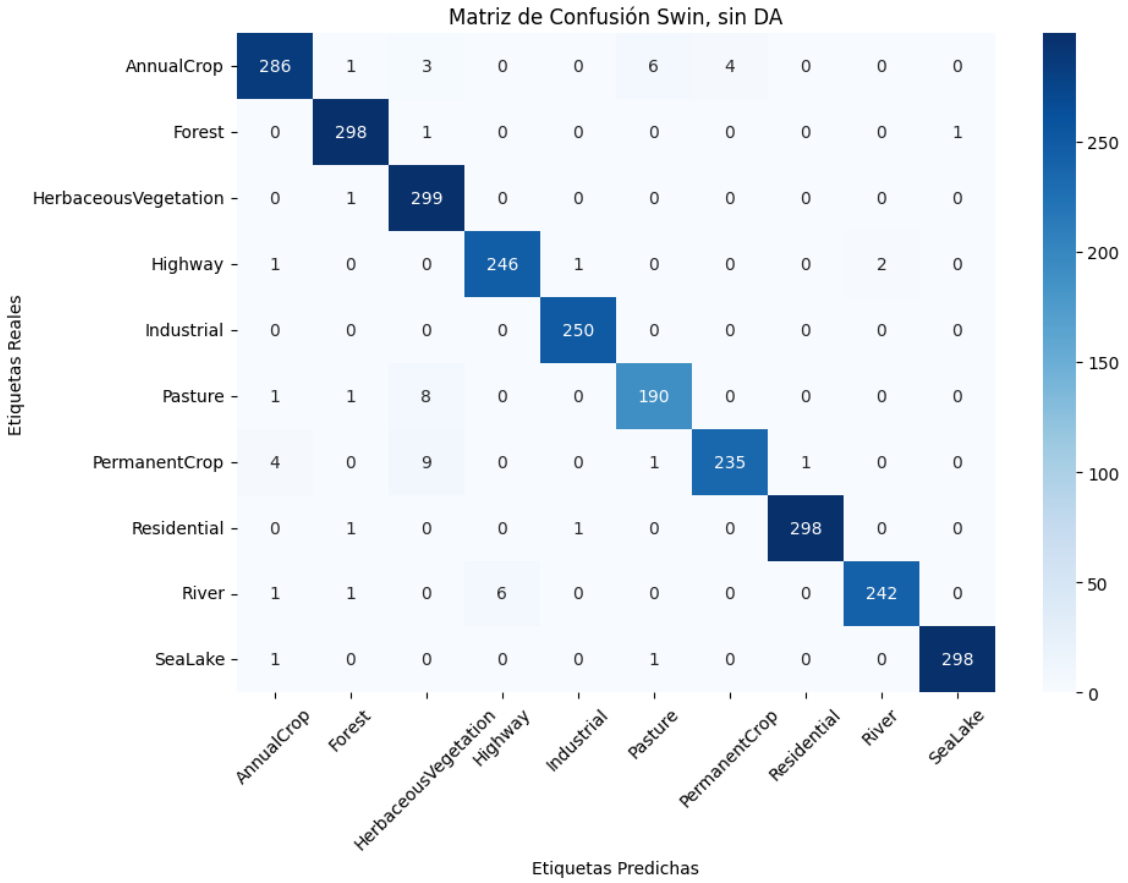


*Figura 14 Matriz de confusión del modelo Swin Transformer sin el uso de la técnica de DA. La matriz muestra el número de predicciones correctas y las confusiones entre las distintas clases. En el eje vertical se pueden observar las etiquetas reales, y en el eje horizontal las etiquetas predichas.*

En cuanto a la confusión del modelo de la clase Permanent Crop con Herbaceous Vegetation apreciable en la Figura 15, la confusión del modelo puede haberse dado lugar debido a la similitud en los colores y texturas entre ambas clases, así como a la falta de ejemplos diferenciadores específicos entre ambos. El modelo ha demostrado gran probabilidad para la predicción errónea, lo que indica que el modelo no ha podido diferenciar las características extremas de las clases. Por lo tanto, la inclusión de más ejemplos diferenciadores y la mejora de la resolución podrían reducir la confusión del modelo.

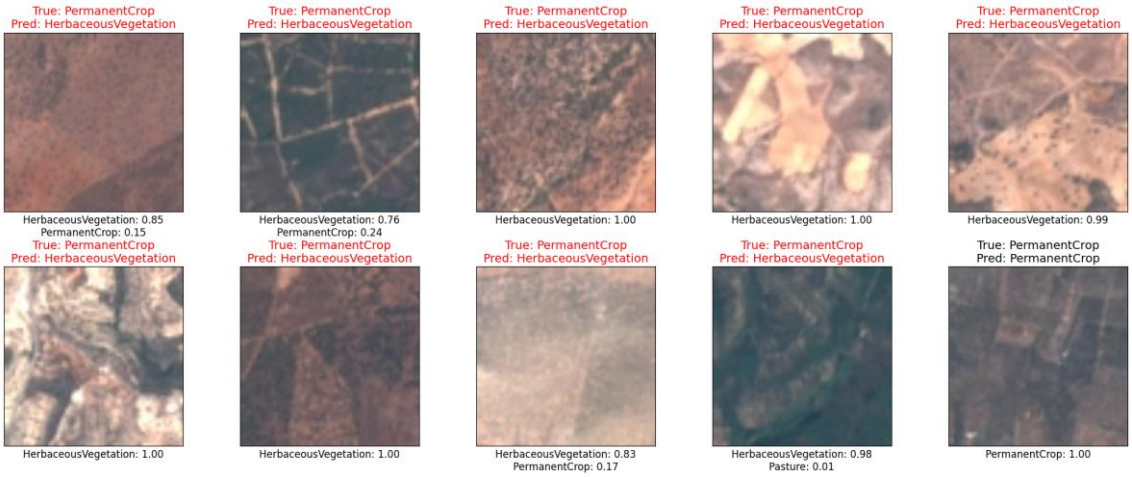


*Figura 15 Imágenes de la clase Permanent Crop con la predicción hecha por el modelo Swin sin el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase Permanent Crop con la clase Herbaceous Vegetation.*

En cuanto a la confusión de la clase Pasture con Herbaceous Vegetation que se pueden ver en la Figura 16, las imágenes confundidas por el modelo presentan un color y una textura similar a la clase Herbaceous Vegetation, lo cual puede haber sido la causa de la confusión. El modelo ha demostrado una mayor confianza en la predicción errónea en aquellas imágenes con una textura y color uniforme, lo cual también se da en la clase Herbaceous Vegetation. La inclusión de más ejemplos diferenciadores entre ambas clases podría reducir la confusión entre ambas.

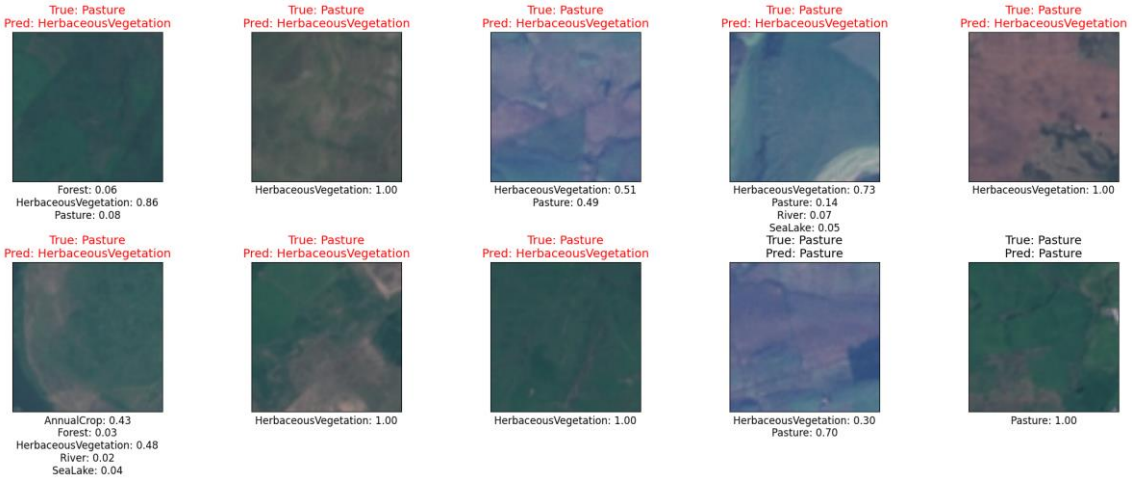


*Figura 16 Imágenes de la clase Pasture con la predicción hecha por el modelo Swin sin el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase Pasture con la clase Herbaceous Vegetation.*

Con la implementación de la técnica de *DA*, podemos apreciar en la Figura 17 la matriz de confusión del modelo tras su entrenamiento. En ella podemos ver una reducción significativa de las mayores confusiones producidas en el modelo sin la implementación de *DA*, así como una mejora en la precisión total.

Las mayores confusiones del modelo sin la implementación de la técnica se dieron entre las clases Permanent Crop y Herbaceous Vegetation, y Pasture y Herbaceous Vegetation. Para esas confusiones, podemos ver como el modelo ha reducido esas confusiones de 9 y 8 para las clases de Permanent Crop y Herbaceous Vegetation respectivamente, a 3 confusiones para cada una de ellas

Sin embargo, aunque el modelo ha presentado una mejora significativa en las clases más problemáticas sin la implementación de la técnica, éste también ha presentado un aumento significativo en la confusión de la clase Herbaceous Vegetation con Permanent Crop, y también ha mantenido la misma confusión de la clase River con Highway.

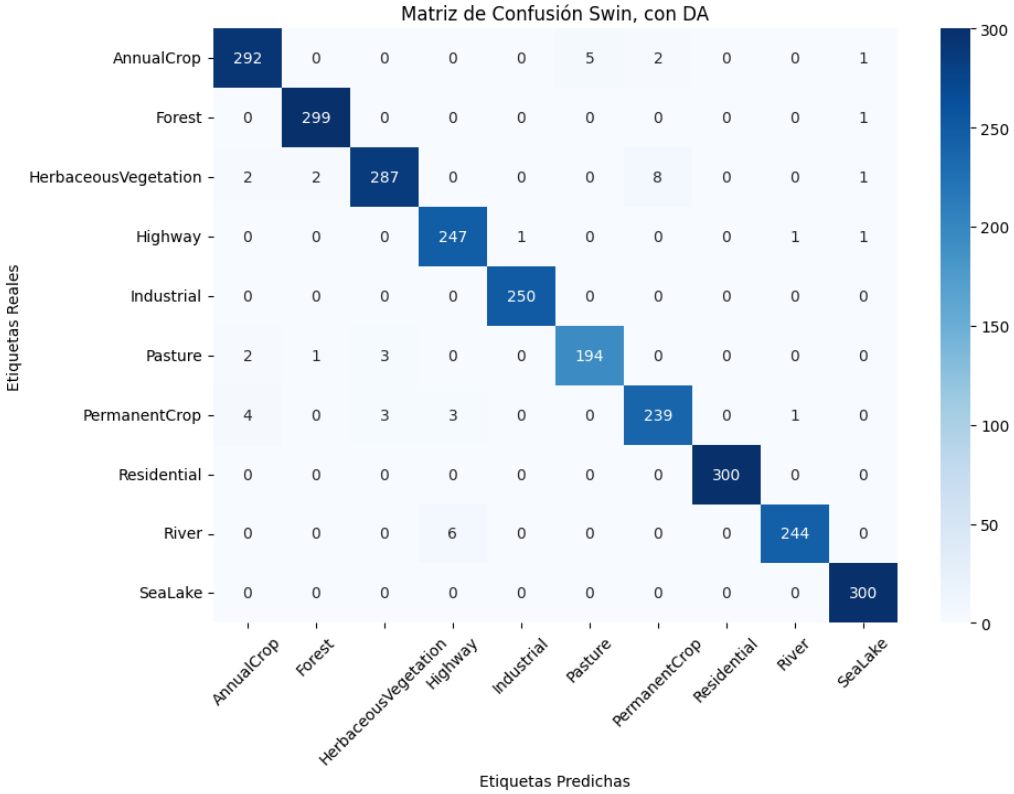


*Figura 17 Matriz de confusión del modelo Swin Transformer con el uso de la técnica de DA. La matriz muestra el número de predicciones correctas y las confusiones entre las distintas clases. En el eje vertical se pueden observar las etiquetas reales, y en el eje horizontal las etiquetas predichas.*

Las imágenes confundidas por el modelo en la Figura 18 presentan colores y texturas muy similares a la clase Permanent Crop. El aumento de confusión entre estas clases producidas en el modelo con la implementación de *DA* frente a su contraparte donde no se usó la técnica sugiere que el modelo ha sufrido una sobrecompensación. Esto puede haber ocurrido debido a que el modelo ha sufrido un sobreajuste a las Transformaciones, haciendo que el modelo aprenda patrones específicos de las transformaciones en lugar de las características intrínsecas de las clases.

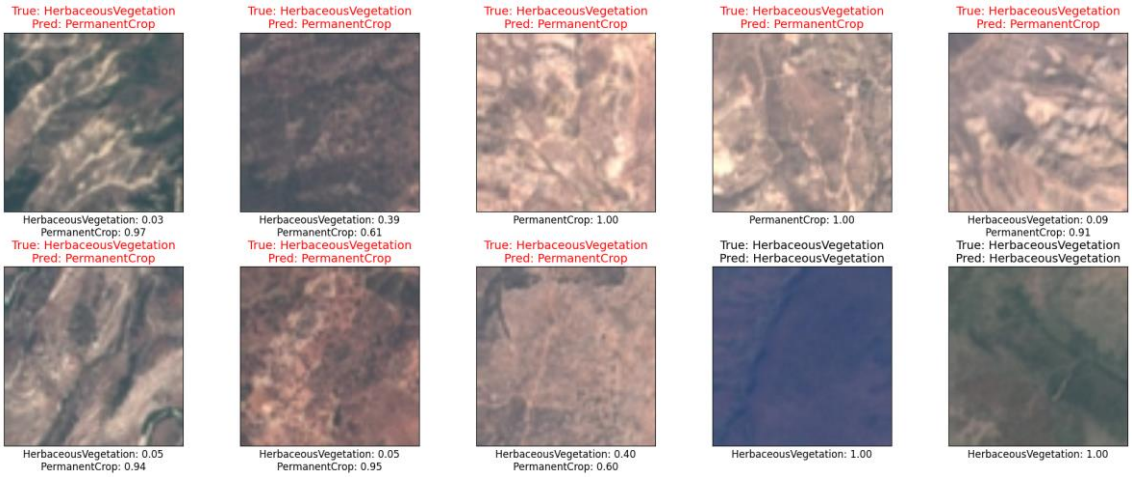


*Figura 18 Imágenes de la clase Herbaceous Vegetation con la predicción hecha por el modelo Swin con el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase Herbaceous Vegetation con la clase Permanent Crop.*

Para la confusión de la clase River con la clase Highway, podemos ver en la Figura 19 las imágenes confundidas por el modelo donde muestran una gran similitud en su visual y texturas. Debido a que el modelo ha mantenido el mismo nivel de confusión antes y después de la implementación de *DA*, podemos concluir que el modelo no ha podido generalizar mejor entre ambas clases con la utilización de la técnica.

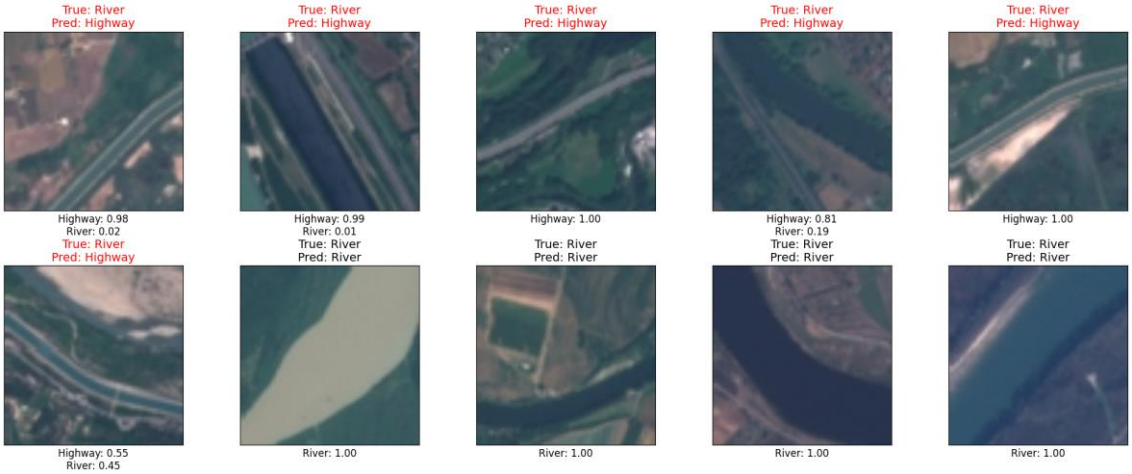


*Figura 19 Imágenes de la clase River con la predicción hecha por el modelo Swin con el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase River con la clase Highway.*

## 5.3 Modelo DeiT

El modelo DeiT ha demostrado la convergencia más rápida de todos los modelos implementados, así como el menor número de épocas usadas para su entrenamiento como muestra la Figura 20. Además, aunque el uso de la técnica de *DA* ha demostrado una mejora en la pérdida y precisión del conjunto de validación en las primeras épocas, el modelo al finalizar su entrenamiento ha obtenido peores métricas cuando se ha implementado la técnica en comparación a cuando no se implementó. Esto puede indicar un sobreajuste del modelo, lo cual sugiere que el modelo podría ser beneficiado de la aplicación de la técnica de early stopping de forma más severa.

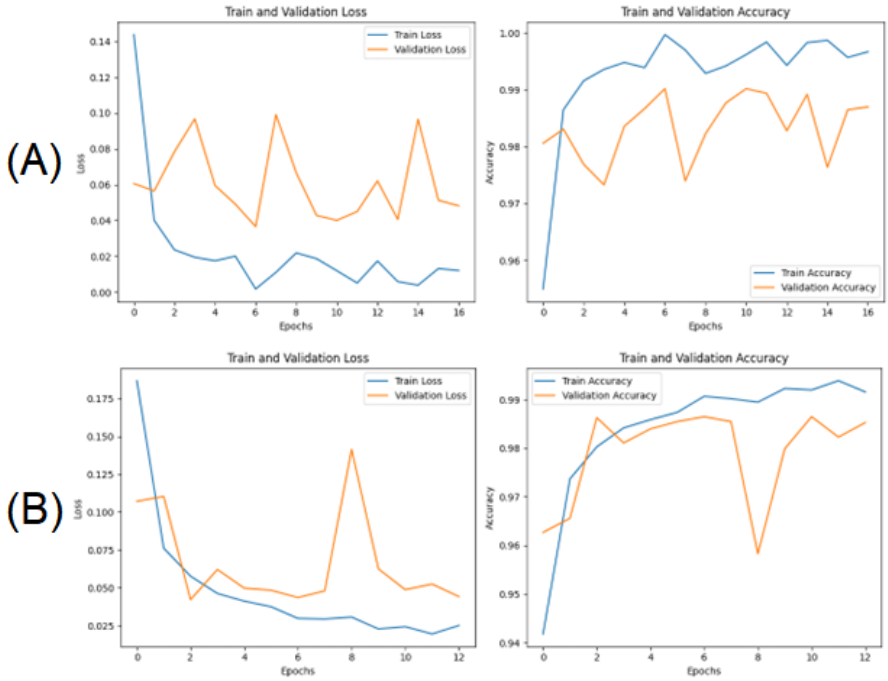


*Figura 20 Gráficas de pérdida y precisión durante el entrenamiento y la validación del modelo DeiT. La Figura (A) corresponde al modelo sin la utilización de la técnica de DA, y la Figura (B) pertenece al modelo en el que se utilizó. Las gráficas de la izquierda muestran la pérdida (loss) del conjunto de entrenamiento y validación a lo largo de las épocas. Las gráficas de la derecha muestran la precisión (accuracy) en los mismos conjuntos y durante las mismas épocas.*

Para el entrenamiento del modelo DeiT sin el uso de *DA*, podemos ver en la Figura 21 como el modelo ha mostrado una gran precisión y acierto a la hora de la realización de las predicciones, y por tanto una baja confusión. En el modelo, las clases en las que se produjo la menor confusión fueron SeaLake y Industrial, donde no se produjo ninguna confusión en ninguna de sus predicciones. Por el contrario, las clases donde más confusiones realizó el modelo fueron Herbaceous Vegetation y Permanent Crop. La mayor confusión entre clases se produjo en la clase Herbaceous Vegetation con Permanent Crop, donde el modelo se confundió en la predicción 6 veces.

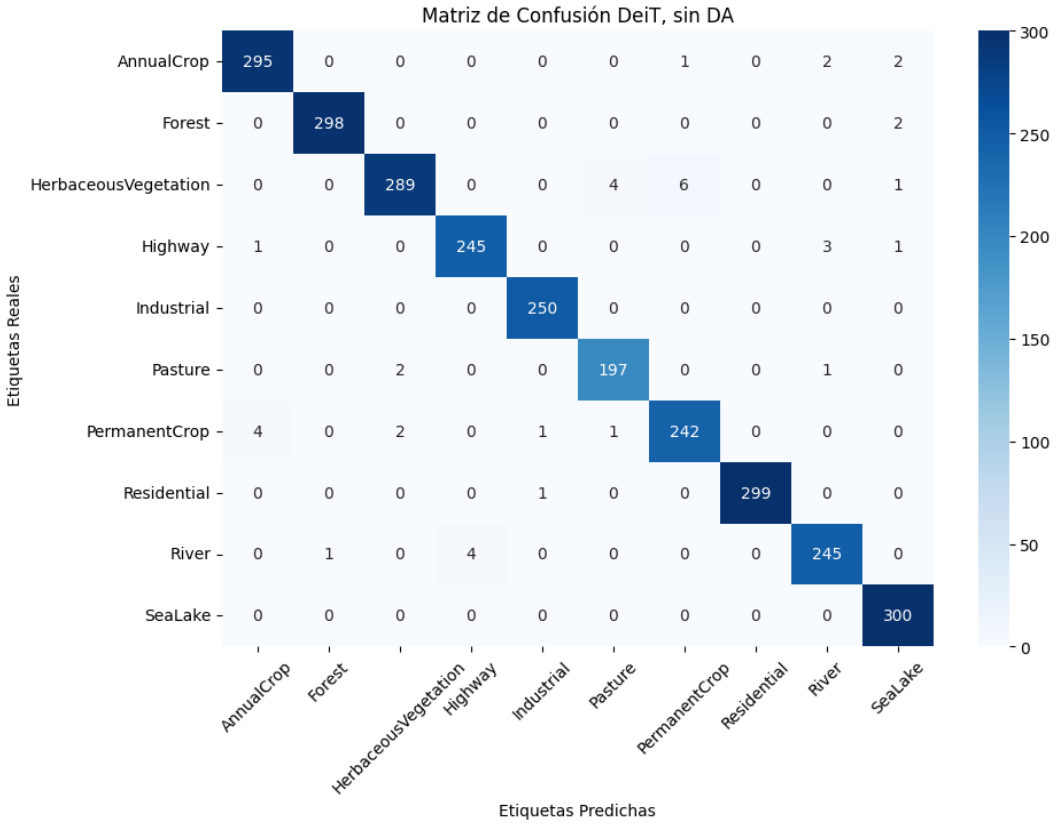


*Figura 21 Matriz de confusión del modelo DeiT Transformer sin el uso de la técnica de DA. La matriz muestra el número de predicciones correctas y las confusiones entre las distintas clases. En el eje vertical se pueden observar las etiquetas reales, y en el eje horizontal las etiquetas predichas.*

En la Figura 22 podemos observar las imágenes que provocaron la mayor confusión entre clases, las imágenes destacan por tener texturas y colores similares lo cual puede haber llevado el modelo a la confusión. Por tanto, el modelo podría beneficiarse en sus predicciones con la inclusión de más ejemplos distintivos entre ambas clases.

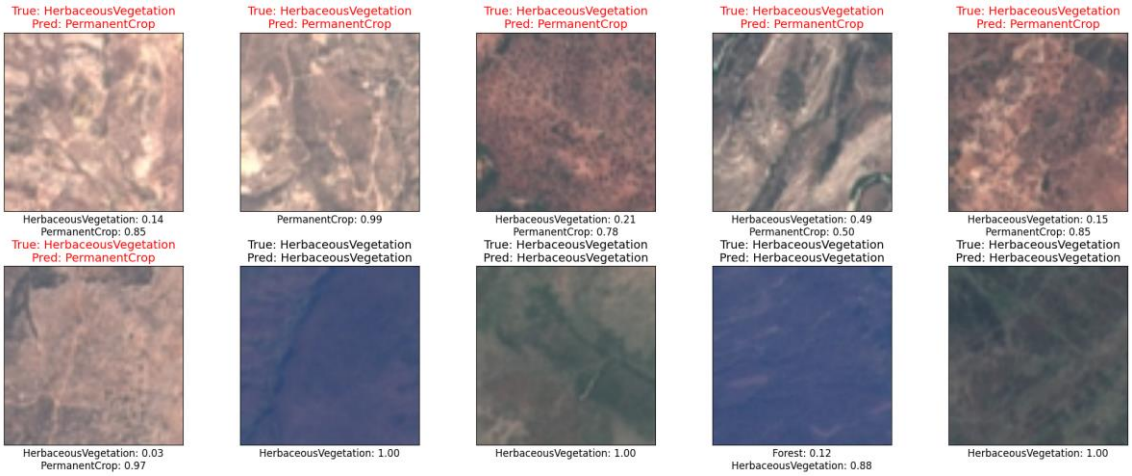


*Figura 22 Imágenes de la clase Herbaceous Vegetation con la predicción hecha por el modelo DeiT sin el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase Herbaceous Vegetation con la clase Permanent Crop.*

En cuanto al modelo DeiT con la implementación de *DA*, podemos observar en la Figura 23 como este modelo ha demostrado peores métricas que su contraparte donde no se utilizó la técnica, demostrando una mayor confusión general y peores métricas. Sin embargo, el modelo mejoró en sus predicciones en clases específicas, la más destacable es la clase de Herbaceous Vegetation, donde el modelo no realizó ninguna predicción errónea. Esto significa una gran mejora significativa en comparación al modelo donde no fue implementada la técnica de *DA*, en el cual la clase Herbaceous Vegetation fue la clase que resultó en la mayor confusión.

Para las clases entre las que se encontró la mayor confusión destacan Highway con River y Permanent Crop con Herbaceous Vegetation, las cuales resultaron en 8 y 7 confusiones respectivamente. Lo cual aumentó de 3 y 2 confusiones respectivamente del modelo sin la implementación de *DA*.

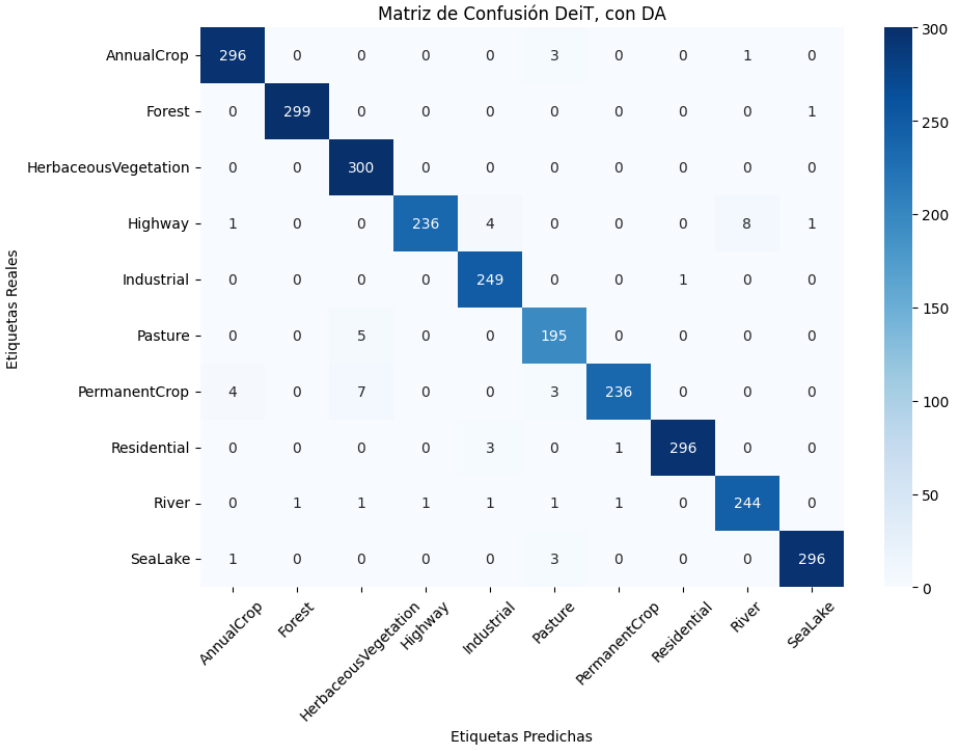


*Figura 23 Matriz de confusión del modelo DeiT Transformer con el uso de la técnica de DA. La matriz muestra el número de predicciones correctas y las confusiones entre las distintas clases. En el eje vertical se pueden observar las etiquetas reales, y en el eje horizontal las etiquetas predichas.*

Para las clases que resultaron con mayor confusión del modelo, observables en la Figura 24, ambas clases presentan tanto formas visuales como texturas similares lo cual ha llevado el modelo a la confusión. Para solucionarlo sería necesario la inclusión de más imágenes distintivas entre ambas clases, sobre todo en los casos donde las figuras visuales y las texturas son similares.

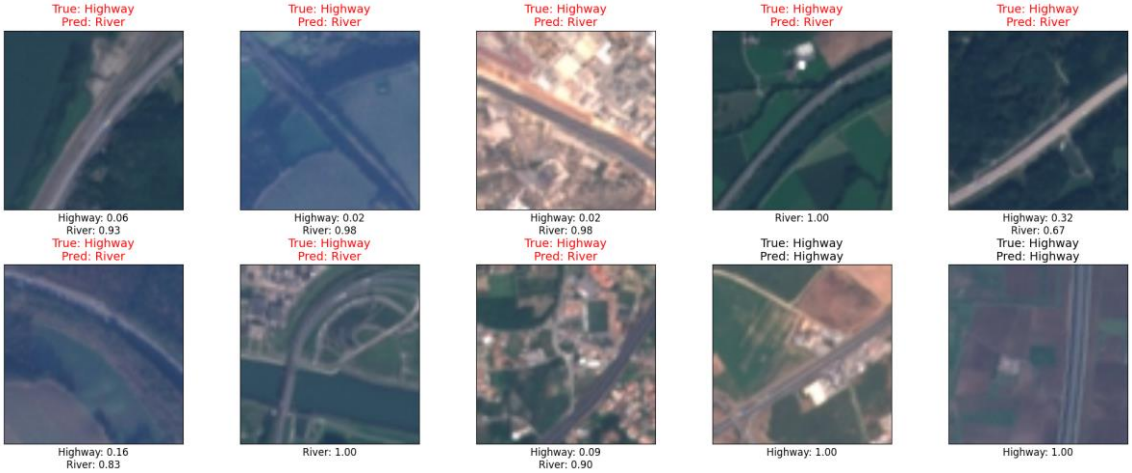


*Figura 24 Imágenes de la clase Highway con la predicción hecha por el modelo DeiT con el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase Highway con la clase River.*

En cuanto a la confusión de la clase Permanent Crop con Herbaceous Vegetation, podemos observar una similitud en el color y las texturas entre ambas clases en las imágenes de la Figura 25. Además, es necesario mencionar que el modelo predijo correctamente toda la clase Herbaceous Vegetation, lo cual puede indicar que el modelo otorga más peso a la clase Herbaceous Vegetation para estas texturas y colores. Debido a esto, para solventarlo sería necesario la inclusión de más imágenes que distingan entre ambas clases en el que se dé una textura y un color similar.

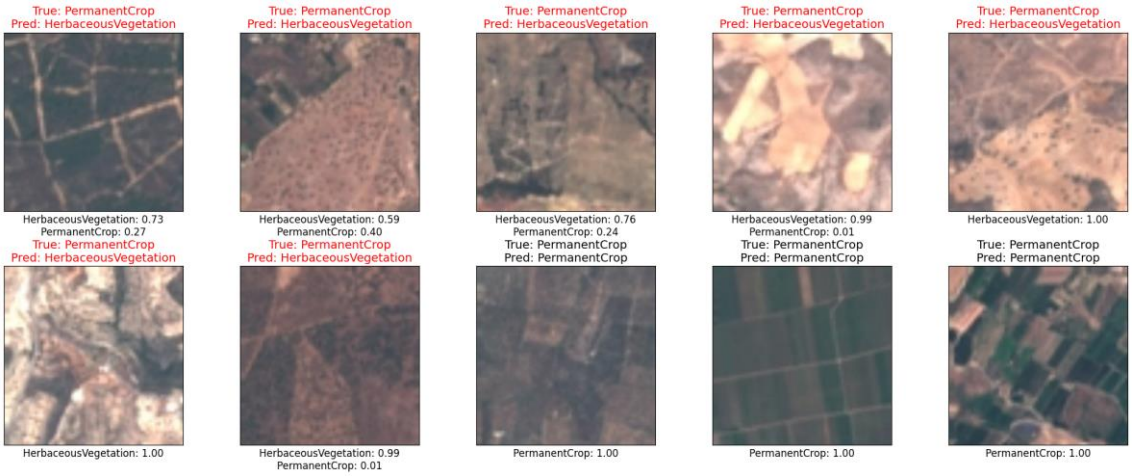


*Figura 25 Imágenes de la clase Permanent Crop con la predicción hecha por el modelo DeiT con el uso de DA, se destaca en rojo las predicciones incorrectas donde se ha confundido la clase Permanent Crop con la clase Herbaceous Vegetation.*

## 5.4 Comparación

Una vez analizados los resultados obtenidos por los diferentes modelos con y sin la implementación de la técnica de *DA*, se va a realizar una comparativa entre todos los experimentos, incluida en la Tabla 1. Las métricas Accuracy, Kappa de Cohen, Precision, Recall y F1-Score usadas en la comparación sirven para cuantificar la eficiencia del modelo, así como forma objetiva y válida de comparación entre ellos.

Podemos observar en la Tabla 1 cómo los resultados de los modelos mejoraron de forma general con la implementación de la técnica de *DA*, a excepción del modelo DeiT, el cual obtuvo peores resultados en las métricas en comparación a su contraparte donde no se usó la técnica. Además, el modelo DeiT sin la implementación de DA fue el modelo que obtuvo el mejor rendimiento en las métricas de Accuracy, Kappa de Cohen, Precision, Recall y F1-Score.

También podemos destacar como el número de épocas para el entrenamiento difiere según el modelo y según la utilización o no de *DA*. El modelo ViT vio un aumento considerable en el número de épocas necesitadas para su entrenamiento, mientras que el modelo Swin no aumentó en gran medida, y el modelo DeiT vio reducido el número de épocas necesarias.

*Tabla 1 Comparación de las métricas de desempeño de los tres modelos Transformers visuales (ViT, Swin y DeiT) en la clasificación de imágenes de satélite, tanto sin data augmentation (sin DA) como con data augmentation (con DA). Los valores resaltados indican el mejor desempeño para cada métrica en su respectiva categoría.*

| ***Métrica*** | ***sin Data Augmentation*** | | | ***con Data Augmentation*** | | |
|---|---|---|---|---|---|---|
| | **ViT** | **Swin** | **DeiT** | **ViT** | **Swin** | **DeiT** |
| ***Accuracy*** | 0.9807 | 0.9785 | **0.9852** | 0.9841 | 0.9822 | 0.9804 |
| ***Kappa de Cohen*** | 0.9786 | 0.9761 | **0.9835** | 0.9823 | 0.9802 | 0.9782 |
| ***Precision*** | 0.9809 | 0.9789 | **0.9852** | 0.9842 | 0.9822 | 0.9807 |
| ***Recall*** | 0.9807 | 0.9785 | **0.9852** | 0.9841 | 0.9822 | 0.9804 |
| ***F1-Score*** | 0.9807 | 0.9785 | **0.9852** | 0.9841 | 0.9822 | 0.9803 |
| ***Tiempo/Época Entrenamiento*** | 10min | 10min | 10min | 10min | 10min | 10min |
| ***Tiempo/Época Validación*** | 1min | 1min | 1min | 1min | 1min | 1min |
| ***Recursos Usados (Gb)*** | **5Gb** | 10Gb | 9Gb | **5Gb** | 10Gb | 9Gb |
| ***Número de Épocas*** | 30 | 26 | 17 | 40 | 23 | **13** |

La Tabla 1 muestra además la diferencia en recursos usados por los distintos modelos durante su entrenamiento producida por la diferencia entre las arquitecturas de los modelos, destacando en negrita el mejor resultado obtenidos en cada métrica. El modelo ViT utiliza 5Gb de Ram de GPU durante su entrenamiento, una cifra significativamente menor al resto de modelos. Esto se debe a que tiene una alta eficiencia en la utilización de datos y memoria gracias a su enfoque basado en parches de imagen y autoatención, lo cual logra

que generalmente requiera menos recursos que otros modelos más complejos en configuraciones equivalentes.

En cuanto al modelo Swin, éste utiliza un enfoque de ventanas desplazadas que mejora la escalabilidad y la eficiencia, pero debido a eso también introduce una complejidad adicional en la gestión de ventanas y conexiones cruzadas entre ventanas. Debido a esta mayor complejidad del modelo, se produce un mayor consumo de memoria durante el entrenamiento.

El modelo DeiT está optimizado para ser más eficiente en el uso de datos mediante técnicas de destilación de conocimiento. Aunque el modelo Deit es más eficiente que ViT en términos de uso de datos, la implementación de técnicas adicionales para mejorar la eficiencia de los datos y el rendimiento ha resultado en un consumo de memoria más alto durante el entrenamiento.

## 5.5 Discusión de los Resultados

En cuanto a los modelos Transformer entrenados sin *DA*, el modelo DeiT mostró el mejor desempeño en todas las métricas medidas, mientras que el modelo Swin obtuvo los peores resultados. Además, el modelo DeiT utilizo una menor cantidad de épocas para su entrenamiento completo que el resto de los modelos, mostrando una mayor eficiencia y desempeño en su aprendizaje que el resto de los modelos.

Los modelos ViT y Swin tuvieron un mejor rendimiento en las métricas utilizando la técnica de data agumentation en comparación a su rendimiento sin la utilización de la técnica, y el modelo ViT fue el que mejores calificaciones obtuvo en las métricas medidas para los modelos en los que se implementó *DA*, y el modelo DeiT las peores.

Es necesario destacar que el modelo DeiT entrenado con la técnica de *DA* tuvo un peor rendimiento que su contraparte que no la usó, el modelo que usó DA usó menos épocas para completar su entrenamiento, pero obtuvo peores resultados en las métricas usadas para su evaluación. Es decir, la capacidad de generalización del modelo DeiT empeoró con el uso de DA.

Además, el modelo ViT tuvo un aumento considerable en la cantidad de épocas requeridas para completar su entrenamiento, consumiendo más tiempo para lograrlo, mientras que el resto de los modelos no vio aumentado de forma apreciable su número de épocas necesarias para su entrenamiento.

Para la comparación de todos los modelos implementando y no implementando la técnica de *DA*, el modelo DeiT sin la implementación de la técnica logró los mejores resultados en todos los parámetros medidos, y además realizó una cantidad considerable menos de épocas que el resto de modelos para lograr su entrenamiento, excepto el modelo DeiT con *DA* el cual realizó una cantidad similar. Esto puede ser indicativo de la eficiencia del modelo Deit en la tarea de la clasificación de imágenes, aunque la técnica de *DA* generalmente mejora la capacidad de generalización del modelo, puede que esto no sea el caso del modelo DeiT, debido a que su diseñó ya incorpora métodos de eficiencia de datos.

Sin embargo, también es posible que los resultados obtenidos del modelo DeiT se deban a un sobreajuste. El modelo podría estar aprendiendo demasiado bien las características específicas del conjunto de entrenamiento, como el ruido del conjunto de entrenamiento. Esto tiene como consecuencia que la capacidad de generalizar a nuevos datos del modelo se ve impactada de forma negativa.

## 5.6 Conclusión

En cuanto a la aplicación de la técnica de Data Agumentation, podemos concluir que mejoró de forma notable el desempeño del modelo ViT en todas las métricas, lo cual es indicativo de que este modelo se benefició de la variabilidad introducida por el aumento de datos. En contraste, el desempeño del modelo Deit disminuyó ligeramente con la aplicación de *DA*, esto puede sugerir que el modelo ya estaba optimizado para el conjunto de datos original sin necesidad de aumento de datos.

Para aquellas aplicaciones en la que se dispone de datos adicionales y se pueda usar *DA*, la opción recomendada sería el modelo ViT debido a su capacidad de mejora en su desempeño con datos variados. En contextos donde se dispone de una cantidad limitada de datos, y no se puede aplicar *DA*, el modelo DeiT sería la opción más adecuada por su sólido desempeño sin necesidad de técnicas adicionales.

Como consideraciones futuras, es importante destacar la necesidad de realizar más experimentos con diferentes técnicas de *DA*, evaluando su impacto en los distintos modelos para entender mejor sus comportamientos. Adicionalmente, podría explorarse el ajuste fino de hiperparámetros y arquitecturas específicas para cada modelo, maximizando así su desempeño en la clasificación de imágenes satelitales.

# 6 Análisis de Impacto

En este capítulo, se va a analizar el impacto potencial de los resultados obtenidos durante la realización del Trabajo de Fin de Grado (TFG) en diversos contextos: personal, empresarial, social, económico, medioambiental y cultural. Se destacarán también los beneficios esperados, así como los posibles efectos perjudiciales. Además, se hará una referencia a los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030 que sean relevantes para la solución propuesta. También serán mencionadas las decisiones tomadas a lo largo del trabajo que tienen como base la consideración del impacto.

### Impacto Personal

El desarrollo de este proyecto ha permitido un significativo desarrollo de habilidades técnicas en técnicas avanzadas de *TL* y procesamiento de imágenes, fortaleciendo de esta forma tanto el perfil profesional como el académico. Además, la realización de experimentos y el análisis de datos han mejorado de forma notable la capacidad analítica, fomentando a su vez el pensamiento crítico y la resolución de problemas.

### Impacto Empresarial

La implementación de las técnicas desarrolladas durante este proyecto tienen la capacidad de mejorar de forma significativa la clasificación de imágenes satelitales, lo cual puede resultar en un beneficio para las empresas que se dedican al análisis de estas imágenes mediante un aumento de la precisión y eficiencia en tareas como el monitoreo ambiental, la agricultura de precisión y la gestión de desastres.

Además, la adopción de Transformers visuales presenta un avance tecnológico notable, posicionando a las empresas que las implementen como líderes en innovación tecnológica. Sin embargo, hay que tener en cuenta el coste de implementación, ya que la adopción de nuevas tecnologías puede implicar elevados costes en términos de infraestructura y formación personal.

### Impacto Social

La mejora en la clasificación de imágenes satelitales puede proporcionar información precisa y valiosa para la toma de decisiones en áreas como la gestión de recursos naturales y la planificación urbana. Además, la aplicación de esta tecnología puede aumentar la conciencia ambiental mediante un monitoreo más eficiente del cambio climático y la deforestación, destacando de esta forma la importancia de la sostenibilidad. Sin embargo, la aplicación de esta tecnología corre el riesgo de crear desigualdad en el acceso a estas, beneficiando de esta forma a regiones o comunidades con mayores recursos.

### Impacto económico

Con la mejora en la clasificación de imágenes satelitales se pueden optimizar operaciones en sectores como la agricultura, la minería y la gestión de infraestructuras, reduciendo de esta forma los costes y aumentando la productividad. Además, con la implementación y el desarrollo de estas nuevas tecnologías, se pueden crear nuevas oportunidades laborales en sectores relacionados con la inteligencia artificial y el análisis de datos. No obstante, la automatización y el uso de estas nuevas tecnologías también puede desplazar ciertos empleos, especialmente aquellos basados en tareas manuales y rutinarias.

### Impacto Medioambiental

Las mejoras en la clasificación de imágenes pueden ayudar en la monitorización de exosistemas y la conservación de la biodiversidad, contribuyendo de esta forma a la sostenibilidad ambiental. Además, con la optimización de procesos industriales mediante el uso de las imágenes satelitales se puede reducir las emisiones de gases de efecto invernadero. Sin embargo, debido al alto consumo energético requerido para el entrenamiento de los modelos de inteligencia artificial, puede darse un impacto negativo en el medio ambiente si no es gestionado adecuadamente.

### Impacto Cultural

Para el impacto cultural, podemos destacar que la publicación y difusión de los resultados del proyecto pueden fomentar el interés y la investigación en áreas relacionadas con la inteligencia artificial y el análisis de imágenes. Por otra parte, las técnicas de clasificación de imágenes también pueden ser aplicadas para la preservación y el monitoreo de sitios culturales y arqueológico.

Aunque el desarrollo de estas tecnologías pueda tener un gran impacto positivo en el contexto cultural, la rápida evolución de las tecnologías de inteligencia artificial puede ampliar la brecha digital entre diferentes grupos socio-económicos y culturales.

### Referencia a los Objetivos de Desarrollo Sostenible (ODS)

En el proyecto realizado se pueden observar las diferentes contribuciones y referencias a los Objetivos de Desarrollo Sostenible (ODS). Para el ODS 4: Educación de Calidad, podemos ver como el trabajo promueve el aprendizaje de tecnologías avanzadas y su aplicación en problemas reales; En cuanto al ODS 9: Industria, Innovación e Infraestructura, el trabajo fomenta la innovación tecnológica y la mejora de las infraestructuras mediante el uso de la inteligencia artificial; Para el ODS 13: Acción por el Clima, el desarrollo de las tecnologías y avance con el análisis de imágenes satelitales mediante el uso de inteligencia artificial contribuye al monitoreo y la mitigación del cambio climático; Por último tenemos el ODS 15: Vida de Ecosistemas Terrestres, para el cual el proyecto contribuye mediante la ayuda en la conservación de ecosistemas y biodiversidad mediante el monitoreo preciso de cambios ambientales.

### Decisiones Basadas en el Impacto

A lo largo del trabajo, se han tomado diversas decisiones considerando el impacto potencial, las cuales incluyen: La elección de modelos preentrenados, el uso de Google Colab, La división del conjunto de datos, la Normalización de las imágenes, la elección de las Métricas de evaluación, la aplicación de *DA* y la referencia a los Objetivos de Desarrollo Sostenible (ODS).

En cuanto a la elección de los modelos preentrenados, con ello se consigue reducir el tiempo y recursos computacionales necesarios para el entrenamiento desde cero, disminuyendo de esta forma el consumo energético y los costos asociados con la computación intensiva.

La utilización de los servicios de Google Colab permite el acceso a recursos computacionales avanzados, evitando de esa forma la necesidad de invertir en hardware costoso para la realización del trabajo.

Para la división del conjunto de datos, se eligió una división del 75% para el entrenamiento, 15% para validación y 10% para prueba. De esta forma se garantiza una evaluación justa y precisa del rendimiento del modelo, además,

esto facilita una estructura de trabajo organizada y metódica, mejorando de esta forma la gestión del proyecto.

La Normalización de las imágenes al rango de 0 a 1 y posteriormente a una media de 0.5 y una desviación estándar de 0.5 facilita el entrenamiento del modelo y mejora su convergencia.

En cuanto a la elección de las métricas de evaluación, las que se eligieron fueron Accuracy, Kappa de Cohen, Precision, Recalll y F1-Score. Estas métricas permiten evaluar el rendimiento del modelo de manera integral, asegurando de esta forma que sea confiable para aplicaciones críticas. Además, esta elección promueve la trasparencia y confiabilidad de los resultados, lo que es crucial para la aceptación y uso de la tecnología en diferentes contextos.

La aplicación de *DA* tuvo como objetivo buscar aumentar la robustez del modelo frente a variaciones en los datos, mejorando de esta forma su aplicación en diversos escenarios empresariales. Además, la capacidad del modelo para generalizar a nuevos datos reduce el riesgo de sobreajuste.

Por último, la alineación del trabajo con los ODS de la Agenda 2030 garantiza que el proyecto contribuya de forma positiva a los desafíos globales, y fomenta una conciencia global sobre la sostenibilidad y la responsabilidad social en el uso de tecnologías avanzadas.

# 7 Bibliografía

[1] "Application of Remote Sensing In Environmental Monitoring- Enhancing Sustainability," Spatial Post. [Online]. Available: https://www.spatialpost.com

[2] S. N. M. Saad y M. Mohan, "UAV Implementations in Urban Planning and Related Sectors of Rapidly Developing Nations: A Review and Future Perspectives for Malaysia," Remote Sensing, vol. 15, no. 11, p. 2845, May 2023. [Online]. Available: https://doi.org/10.3390/rs15112845

[3] R. P. Sishodia, R. L. Ray, y S. K. Singh, "Applications of Remote Sensing in Precision Agriculture: A Review," Remote Sensing, vol. 12, no. 19, p. 3136, Sep. 2020. [Online]. Available: https://doi.org/10.3390/rs12193136

[4] D. Lu y Q. Weng, "A survey of image classification methods and techniques for improving classification performance," International Journal of Remote Sensing, vol. 28, no. 5, pp. 823-870, Mar. 2007.

[5] X. Zhu, D. Tuia, L. Mou, G. S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources," IEEE Geoscience and Remote Sensing Magazine, vol. 5, no. 4, pp. 8-36, Dec. 2017.

[6] J. D. B. Nelson, "The use of convolutional neural networks for remote sensing applications," Geoscience and Remote Sensing Letters, vol. 14, no. 12, pp. 2255-2259, Dec. 2017.

[7] F. dos Santos, P. dos Santos, and R. da Silva, "Improving satellite image classification accuracy using deep learning techniques," Remote Sensing, vol. 11, no. 3, pp. 1-16, Jan. 2019.

[8] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 5, pp. 778-782, May 2017.

[9] N. Raghu, T. Unterthiner, A. Dosovitskiy, M. Gelly, J. Houlsby, "Do Vision Transformers See Like Convolutional Neural Networks?" arXiv preprint arXiv:2108.08810, 2021. [Online]. Available: https://ar5iv.org/2108.08810

[10] "An Introduction to Transfer Learning," Georgian Partners. [Online]. Disponible: https://georgian.io/an-introduction-to-transfer-learning/. [Accedido: 20-4-2024].

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," Advances in Neural Information Processing Systems, vol. 30, pp. 5998-6008, 2017.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

[13] J. Liu, X. Chu, Y. Wang, y M. Wang, "Deep Text Retrieval Models based on DNN, CNN, RNN and Transformer: A review," 2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS), 2022. Disponible en: https://www.semanticscholar.org/paper/2bf4dd6db2b19f312bba51543ed6d3a51647ec5d

[14] A. L. F. Novaes, R. Araujo, J. Figueiredo, y L. Pavanelli, "A New State-of-the-Art Transformers-Based Load Forecaster on the Smart Grid Domain," ArXiv, vol. abs/2108.02628, 2021. Disponible en: https://www.semanticscholar.org/paper/bee2383e797d3b6c3df5b1edc8d8049e53c3131b

[15] J. Li et al., "Developing RNN-T Models Surpassing High-Performance Hybrid Models with Customization Capability," ArXiv, vol. abs/2007.15188, 2020. Disponible en: https://www.semanticscholar.org/paper/51db55a30552ab4e2666879d9a28993e6b2acdb5

[16] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). "A Simple Framework for Contrastive Learning of Visual Representations." In International Conference on Machine Learning (pp. 1597-1607). PMLR.

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In Proceedings of the 9th International Conference on Learning Representations.

[18] M. Iman, H. R. Arabnia, y K. Rasheed, "A Review of Deep Transfer Learning and Recent Advancements," Technologies, vol. 11, no. 2, p. 40, Mar. 2023. [Online]. Available: https://doi.org/10.3390/technologies11020040

[19] Pan, S. J., & Yang, Q. (2010). "A Survey on Transfer Learning." IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359.

[20] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). "A Survey of Transfer Learning." Journal of Big Data, vol. 3, no. 1, pp. 1-40.

[21] S. Hossain et al., "Vision Transformers, Ensemble Model, and Transfer Learning Leveraging Explainable AI for Brain Tumor Detection and Classification," IEEE Journal of Biomedical and Health Informatics, 2023. Disponible: https://www.semanticscholar.org/paper/7be836d4a9ac385bc0832dd12d665daed89b05ba

[22] M. Usman, T. Zia, and S. A. Tariq, "Analyzing Transfer Learning of Vision Transformers for Interpreting Chest Radiography," Journal of Digital Imaging, vol. 35, pp. 1445-1462, 2022. Disponible: https://www.semanticscholar.org/paper/e9b06dfbad49a1c573af09a8a6e430a008f973cc

[23] A. Brohan et al., "RT-1: Robotics Transformer for Real-World Control at Scale," ArXiv, 2022. Disponible: https://www.semanticscholar.org/paper/fd1cf28a2b8caf2fe29af5e7fa9191cecfedf84d

[24] Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). "Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 7, pp. 2217-2226.

[25] Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). "Deep Learning in Remote Sensing: A Comprehensive Review and List

of Resources." IEEE Geoscience and Remote Sensing Magazine, vol. 5, no. 4, pp. 8-36.

[26] M. Iman, H. R. Arabnia, y K. Rasheed, "A Review of Deep Transfer Learning and Recent Advancements," Technologies, vol. 11, no. 2, p. 40, Mar. 2023. [Online]. Available: https://doi.org/10.3390/technologies11020040

[27] J. Brownlee, "A Gentle Introduction to Transfer Learning for Deep Learning," Machine Learning Mastery. [Online]. Available: https://machinelearningmastery.com/a-gentle-introduction-to-transfer-learning-for-deep-learning/

[28] A. Ganin, V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015, pp. 1180-1189.

[29] X. Chen, H. Fan, R. Girshick, and K. He, "Improved Baselines with Momentum Contrastive Learning," arXiv preprint arXiv:2003.04297, 2020. [Online]. Available: https://arxiv.org/abs/2003.04297

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, y I. Polosukhin, "Attention is All You Need," Advances in Neural Information Processing Systems, vol. 30, pp. 5998-6008, 2017.

[31] J. Devlin, M.-W. Chang, K. Lee, y K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," en Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2019, pp. 4171-4186.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, y N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

[33] N. Raghu, T. Unterthiner, A. Dosovitskiy, M. Gelly, y J. Houlsby, "Do Vision Transformers See Like Convolutional Neural Networks?" arXiv preprint arXiv:2108.08810, 2021.

[34] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, y N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

[36] D. P. Kingma y J. Ba, "Adam: A Method for Stochastic Optimization," en Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015.

[37] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[38] S. J. Pan y Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[39] D. P. Kingma y J. Ba, "Adam: A Method for Stochastic Optimization," en Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015.

[40] J. Yao, T. Wang, y D. Wipf, "Early Stopping and Hypothesis Testing: Benefits of Over-Parameterization in Function Learning," *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020. [Online]. Available: https://openreview.net/forum?id=r1xGO0EKDH

[41] A. Krizhevsky, I. Sutskever, y G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, vol. 25, pp. 1097-1105, 2012.

[42] L. Perez y J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," arXiv preprint arXiv:1712.04621, 2017.

[43] P. Helber et al., "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," 2017. [Online]. Available: https://www.researchgate.net/publication/319463676_EuroSAT_A_Novel_Dataset_and_Deep_Learning_Benchmark_for_Land_Use_and_Land_Cover_Classification. [Accedido: 20-4-2024].

[44] European Commission, "Mapping guide for a European urban atlas," 2012. [Online]. Disponible: https://ec.europa.eu/regional_policy/sources/tender/pdf/2012066/annexe2.pdf. [Accedido: 20-4-2024].

[45] "Dataset de imágenes satelitales para la clasificación del uso del suelo," Zenodo, 2023. [Online]. Disponible: https://doi.org/10.5281/zenodo.7711096. [Accedido: 20-4-2024].

[46] "ViT Base Patch16 224," Hugging Face Models. [Online]. Disponible: https://huggingface.co/google/vit-base-patch16-224. [Accedido: 20-4-2024].

[47] "Base Swin Transformer model pre-trained on ImageNet-22k," Hugging Face. [Online]. Disponible: https://huggingface.co/microsoft/swin-base-patch4-window7-224-in22k. [Accedido: 20-4-2024].

[48] "DeiT - Base model with patch size 16x16 pre-trained on ImageNet," Hugging Face. [Online]. Disponible: https://huggingface.co/facebook/deit-base-patch16-224. [Accedido: 20-4-2024].

[49] X. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, y H. Jégou, "Training data-efficient image transformers & distillation through attention," arXiv preprint arXiv:2012.12877, 2020.

[50] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," arXiv preprint arXiv:2012.12877v2, 2020. [Online]. Disponible: https://arxiv.org/abs/2012.12877v2. [Accedido: 20-4-2024].

Este documento esta firmado por

| | | |
|---|---|---|
|  | **Firmante** | CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES |
| | **Fecha/Hora** | Mon Jun 03 21:25:24 CEST 2024 |
| | **Emisor del Certificado** | EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES |
| | **Numero de Serie** | 561 |
| | **Metodo** | urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature) |