UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Modelo de clasificación de imágenes basado en Transformers para la evaluación de daños estructurales post-sísmicos de acuerdo con la Escala Macrosísmica Europea: comparativa con técnicas de aprendizaje de máquina

Johana Belén Duchi Tipán

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniera en Ciencias de la Computación

Quito, 12 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Modelo de clasificación de imágenes basado en Transformers para la evaluación de daños estructurales post-sísmicos de acuerdo con la Escala Macrosísmica Europea: comparativa con técnicas de aprendizaje de máquina

Johana Belén Duchi Tipán

Nombre del profesor, Título académico **Edison Fernando Loza Aguirre, Ph.D.**

Quito, 12 de diciembre de 2024

© DERECHOS DE AUTOR

Nombres y apellidos: Johana Belén Duchi Tipán

Código: 00321980

Cédula de identidad: 1724466584

Lugar y fecha: Quito, 12 de diciembre de 2024

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

El presente trabajo aborda la clasificación de daños estructurales en edificaciones de mampostería y hormigón armado, utilizando modelos basados en Data-Efficient Image Transformers (DeiT). La motivación principal surge de la necesidad de desarrollar sistemas automatizados y robustos que faciliten la evaluación post-sísmica, especialmente en contextos con datasets pequeños y desbalanceados. Para ello, se implementaron distintas técnicas de data augmentation, incluyendo enfoques agresivos y moderados, con el objetivo de analizar su impacto en el desempeño de los modelos. Los resultados obtenidos demuestran que los modelos DeiT, en particular la configuración DeiT Base Distilled 384, alcanzan un desempeño superior en comparación con modelos tradicionales basados en redes neuronales convolucionales (CNN), tanto en datos originales como en conjuntos aumentados. La capacidad de los Vision Transformers para capturar relaciones locales y globales permitió mejorar significativamente la clasificación en escenarios con desbalance de clases. Como contribución principal, este trabajo ofrece un modelo validado que puede servir de base para futuras implementaciones prácticas, proponiendo líneas de investigación orientadas al análisis multimodal, la ampliación de datasets y el desarrollo de sistemas de evaluación en tiempo real.

Palabras clave: clasificación de imágenes, Vision Transformers, DeiT, data augmentation, evaluación estructural, daño sísmico, aprendizaje automático, transferencia de aprendizaje, modelos de redes neuronales.

ABSTRACT

This work addresses the classification of structural damage in masonry and reinforced concrete buildings using models based on Data-Efficient Image Transformers (DeiT). The primary motivation stems from the need to develop automated and robust systems that facilitate post-seismic assessment. In particular, it is focused on scenarios with small and imbalanced datasets for which various data augmentation techniques were implemented. Most of these approaches are aggressive and moderate, to analyze their impact on model performance. The results show that DeiT models, particularly the DeiT Base Distilled 384 configuration, demonstrate to be superior to traditional approaches such as Convolutional Neural Network (CNN) models using the original and the augmented dataset. Notoriously, the ability of Vision Transformers to capture local and global relationships improved classification performance in class-imbalanced scenarios. Then, the main contribution of this work is a validated model that serves as a foundation for future practical implementations. It proposes future research lines focused on multimodal analysis, dataset expansion, and the development of real-time evaluation systems.

Keywords: image classification, Vision Transformers, DeiT, data augmentation, structural evaluation, seismic damage, machine learning, transfer learning, neural network models.

AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento a mis tutores, Edison Loza y Ricardo Flores, por su guía durante el desarrollo de esta tesis.

A mi madre, Dorys Tipán, le agradezco su apoyo, paciencia y confianza a lo largo de este proceso.

Finalmente, quiero reconocer a Fausto Pasmay, un profesor cuya influencia marcó significativamente mi formación académica y profesional.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La presente investigación plantea el desarrollo de un modelo de clasificación de imágenes basado en Transformers para evaluar el grado de daño estructural post-sísmico, de acuerdo con la Escala Macrosísmica Europea (EMS-98). Este trabajo surge en respuesta a la necesidad de realizar evaluaciones rápidas y precisas de los daños estructurales tras un terremoto, un proceso fundamental para la seguridad pública y la gestión eficaz de desastres. La importancia de desarrollar modelos automatizados para esta tarea se vuelve aún más crítica en países con alta actividad sísmica, como Ecuador.

Antecedentes

El contexto sísmico en Ecuador es alarmante. Entre 2012 y los primeros siete meses de 2022, se registraron aproximadamente 70.000 sismos en el país [1]. Esta alta actividad sísmica se debe a la ubicación de Ecuador en el Cinturón de Fuego del Pacífico, una de las regiones tectónicas más activas del mundo. Ecuador se encuentra en la zona de subducción donde interactúan las placas de Nazca y Sudamericana, una interacción que ha provocado numerosos terremotos de gran magnitud a lo largo de la historia [2]. En su historia el país ha sufrido los estragos de eventos como el devastador terremoto de 1906, con una magnitud de 8.8 (uno de los más grandes a nivel mundial) [3] o, más recientemente, el terremoto de Pedernales en 2016, con una magnitud de 7.8 [4]; ambos eventos siendo causantes de importantes daños estructurales y numerosas víctimas.

Para comprender este problema, es esencial definir algunos conceptos sismológicos. Un sismo se refiere a cualquier movimiento brusco de la corteza terrestre y se utiliza como un término general para describir cualquier movimiento sísmico, sin importar su magnitud o impacto. Un

temblor es un sismo de menor magnitud que no produce daños perceptibles. Por último, un terremoto describe un sismo de gran magnitud que causa daños significativos [5].1

El Instituto Geofísico de la Escuela Politécnica Nacional (IG-EPN) fue designado en 2003, mediante el decreto oficial No. 3593, como la autoridad nacional encargada del monitoreo sísmico en Ecuador [6]. Este organismo evalúa tanto la magnitud como el grado de daño de los sismos. La magnitud se mide mediante la escala de Richter, que cuantifica la energía liberada durante un evento sísmico. El grado de daño, en cambio, se determina utilizando escalas como la Escala Macrosísmica Europea (EMS), la cual clasifica los efectos observables en las estructuras afectadas. La EMS divide el daño en cinco grados, que van desde fisuras leves (grado 1) hasta la destrucción total (grado 5) [7].

Definición del problema

A pesar de la importancia de las evaluaciones rápidas post-sísmicas, el proceso actual de evaluación del daño de un terremoto se lo realiza mediante una inspección in situ realizada por personal calificado; siendo un proceso lento, costoso y dependiente del desplazamiento de expertos. Esto genera importantes cuellos de botella en la toma de decisiones durante situaciones de emergencia, especialmente en áreas urbanas densamente pobladas y en zonas rurales de difícil acceso, donde la disponibilidad de inspectores cualificados es limitada. Como consecuencia, los retrasos en estas evaluaciones pueden comprometer la seguridad de los habitantes y dificultar la asignación oportuna de recursos para la mitigación de daños.

Objetivos

Objetivo General.

Desarrollar un modelo de clasificación de imágenes utilizando Transformers para evaluar el grado de daño estructural post-sísmico conforme a la Escala Macrosísmica Europea.

Objetivos Específicos.

1. Evaluar el rendimiento de un modelo de Transformers para validar su capacidad en la clasificación de imágenes.
2. Comparar la eficiencia del modelo de Transformers con métodos tradicionales de aprendizaje automático en la clasificación de daños estructurales.

Justificación

Los avances en inteligencia artificial y computer vision ofrecen un panorama prometedor para mejorar la evaluación de daños estructurales después de un terremoto. Los modelos de Transformers, diseñados para procesar lenguaje natural, son altamente eficaces en la clasificación de imágenes gracias a su capacidad para emplear mecanismos de autoatención. Esta funcionalidad les permite analizar simultáneamente diferentes áreas de una imagen y evaluar la relevancia de cada una [8]. Este Proyecto Integrador comparará un modelo de clasificación basado en Transformers con modelos tradicionales de machine learning. Como referencia, se usará la tesis de posgrado de Damaris Tarapues titulada “Modelo de clasificación supervisado de fotografías de fachadas para evaluar el daño estructural ocasionado por sismos de acuerdo con la Escala Macrosísmica Europea para apoyo de toma de decisiones en el Instituto Geofísico-EPN” [9]. Esta investigación empleó técnicas de aprendizaje automático,

específicamente Redes Neuronales Convolucionales (CNN), para clasificar daños en imágenes de fachadas. Para garantizar resultados consistentes, este proyecto utilizará el mismo conjunto de datos. La comparación permitirá determinar si los Transformers tienen mejores métricas que las CNN en la clasificación del daño estructural.

Alcance

El análisis se realizará en edificaciones construidas con mampostería y hormigón armado, utilizando dos notebooks independientes. Esta separación busca mitigar posibles interferencias entre los datos, dado que las propiedades mecánicas y los patrones de daño estructural difieren significativamente entre ambos materiales. El modelo de clasificación seguirá los lineamientos de la Escala Macrosísmica Europea (EMS-98), que categoriza el daño en cinco niveles, desde fisuras leves hasta la destrucción total.

Este estudio se limitará al análisis de datos previamente recolectados y etiquetados en el dataset utilizado en un trabajo previa, sin incluir el desarrollo de una aplicación final. El objetivo principal es determinar si la implementación de un modelo basado en Transformers ofrece un rendimiento superior en comparación con los enfoques tradicionales de machine learning. La evaluación y comparación de los modelos permitirá analizar el impacto de las técnicas basadas en Transformers en la clasificación de imágenes.

ESTADO DEL ARTE

Daños estructurales post-sísmicos

La evaluación precisa y oportuna del daño estructural después de un terremoto es crucial para la gestión eficaz de desastres. Este proceso tradicionalmente se basa en inspecciones visuales in situ, lo que puede ser lento, peligroso y costoso, especialmente en las primeras horas tras un evento sísmico [10]. Los daños en las estructuras varían desde grietas menores hasta el colapso total, dependiendo de factores como la magnitud del sismo, que determina la energía liberada y el movimiento del suelo generado [11]; las distancias del epicentro, las estructuras más cercanas al epicentro experimentarán un movimiento del suelo más fuerte y, por ende, mayor riesgo de daños [11]; las condiciones del suelo, los suelos blandos pueden amplificar las ondas sísmicas, aumentando el daño potencial a las estructuras [12]; el tipo de estructura, materiales como mampostería, hormigón armado o acero presentan diferentes niveles de resistencia a las fuerzas sísmicas [13]; y finalmente, la calidad de la construcción, estructuras mal diseñadas o con deficiencias de diseño son más vulnerables a los daños [14].

Después de un sismo, es crucial evaluar los daños a las estructuras para determinar su seguridad y habitabilidad. Este proceso incluye dos tipos principales de evaluaciones: rápidas y detalladas. Las evaluaciones rápidas se realizan inmediatamente después del sismo para identificar estructuras que representen un peligro inminente para la seguridad pública [15] [16]. Estas se basan en inspecciones visuales y se enfocan en detectar daños evidentes, como grietas grandes, deformaciones o colapsos parciales [16]. Por otro lado, las evaluaciones detalladas implican un análisis más profundo de la estructura, incluyendo el uso de equipos de prueba no destructivos para evaluar la integridad de los elementos estructurales [12]. Estas evaluaciones

permiten determinar el alcance de las reparaciones necesarias o si la estructura debe ser demolida [16].

El comportamiento de una edificación durante un sismo depende en gran medida del material de construcción utilizado. Dos de los materiales más comunes son la mampostería y el hormigón armado. Las estructuras de mampostería, especialmente aquellas no reforzadas, son generalmente más vulnerables a los daños sísmicos [13] [17]. La mampostería es débil frente a las tensiones y puede agrietarse o colapsar fácilmente bajo las fuerzas laterales generadas por un sismo [17]. Además, las edificaciones de mampostería antiguas, comunes en centros históricos, son especialmente propensas a los daños debido a la falta de refuerzo y al deterioro acumulado con el tiempo [16]. En contraste, las estructuras de hormigón armado son más resistentes a los daños sísmico [13]. Tanto la mampostería como el hormigón armado son materiales de construcción ampliamente utilizados, especialmente en países en desarrollo, debido a su disponibilidad, bajo costo y relativa facilidad de construcción [14].

Clasificación de Imágenes: Del Auge de las Redes Convolucionales a la Aparición de los Transformers

La clasificación de imágenes, una tarea esencial en el campo de computer vision, ha avanzado significativamente gracias al desarrollo de modelos basados en deep learning. Durante años, las redes neuronales convolucionales (CNN) han dominado este ámbito, estableciéndose como el estándar para el análisis de imágenes. Las CNN se fundamentan en la operación matemática de convolución, que permite extraer características locales de una imagen con filtros especializados. Estas características son procesadas a través de múltiples capas, que aprenden representaciones progresivamente más abstractas, hasta culminar en una capa de clasificación que asigna la imagen a una categoría específica. Aunque las CNN han demostrado ser altamente efectivas, no están exentas de limitaciones. Por ejemplo, su enfoque en

características locales puede dificultar la captura de relaciones de largo alcance dentro de una imagen. Asimismo, la equivarianza a la traslación inherente a las CNN puede no ser adecuada en tareas donde la posición precisa de los objetos resulta crucial [18].

Aparición de Transformers en Computer Vision.

A diferencia de las redes neuronales convolucionales (CNN), los Transformers no dependen de convoluciones, sino que emplean un mecanismo de autoatención para analizar las relaciones entre diferentes partes de una imagen. Su arquitectura se basa en el concepto de atención, que permite al modelo ponderar la importancia de distintas áreas de la entrada al realizar una predicción [19]. En el contexto de computer vision, un Transformer divide la imagen en parches y calcula las relaciones de atención entre ellos. Una ventaja clave de los Transformers es su capacidad para capturar dependencias globales dentro de una imagen, superando así las limitaciones de las convoluciones locales de las CNN. Además, los Transformers procesan la información en paralelo, lo que los hace más eficientes computacionalmente en comparación con las CNN, especialmente en conjuntos de datos grandes [18].

Vision Transformer (ViT).

Entre los modelos basados en transformer, el Vision Transformer (ViT) destaca como un diseño específico para tareas de clasificación de imágenes. ViT trata los parches de una imagen de manera similar a los tokens (palabras) en un modelo de lenguaje, procesándolos mediante un codificador estándar de Transformer. Para llevar a cabo la clasificación, se añade un token especial de clasificación a la secuencia de parches, cuyo estado final en la salida del codificador se utiliza como representación de la imagen para realizar la predicción [18].

Los experimentos han demostrado que ViT, cuando se entrena a gran escala, supera a las CNN en tareas de clasificación de imágenes. Este modelo no solo es más eficiente computacionalmente que las CNN, sino que también requiere menos recursos para el entrenamiento y, con frecuencia, alcanza una mayor precisión, especialmente al trabajar con grandes volúmenes de datos. Sin embargo, uno de los principales desafíos al entrenar Transformers para tareas de visión por computadora es su alta demanda de datos [18].

Data-efficient Image Transformer (DeiT).

Para abordar la limitación de la alta cantidad de datos requeridos por ViT, se desarrolló el modelo Data-efficient Image Transformer (DeiT), diseñado para lograr un buen rendimiento incluso con conjuntos de datos más pequeños. DeiT utiliza una técnica innovadora denominada distillation token. Este token especial se añade a la secuencia de parches y se entrena para imitar las predicciones de un modelo de CNN preentrenado, permitiendo que DeiT aprenda de la información contenida en el modelo de CNN sin necesidad de acceder al conjunto de datos original utilizado para entrenarlo. Esta técnica de distillation permite a DeiT superar la barrera de la alta demanda de datos, haciéndolo más accesible para tareas de clasificación de imágenes en escenarios donde los datos son limitados [20].

Transfer Learning en Computer Vision

En el ámbito del aprendizaje automático, el transfer learning se define como la técnica de reutilizar un modelo previamente entrenado en una tarea para resolver otra tarea relacionada. Este enfoque se basa en la premisa de que el conocimiento adquirido por el modelo durante su entrenamiento inicial puede ser transferido y adaptado, mejorando así su rendimiento en la

nueva tarea. Esta estrategia resulta especialmente eficiente, ya que aprovecha el conocimiento previo en lugar de comenzar el entrenamiento desde cero [21]. Además, el transfer learning ofrece otras ventajas importantes como: reducción de tiempo de entrenamiento, al reutilizar un modelo preentrenado, se elimina la necesidad de entrenar todas las capas del modelo, lo que disminuye significativamente el tiempo requerido [21]; mejor rendimiento, los modelos preentrenados suelen estar optimizados para extraer características generales, gracias a su entrenamiento en grandes conjuntos de datos, lo que mejora su capacidad de generalización en nuevas tareas.

Transfer Learning y Transformers

Transfer Learning se ha convertido en una técnica fundamental en computer vision, dado que permite aprovechar el conocimiento adquirido por modelos preentrenados en grandes conjuntos de datos, como ImageNet, para mejorar el rendimiento en tareas específicas con conjuntos de datos más pequeños [22].

Se puede combinar las ventajas de Transfer Learning con Transformers. Para ello se selecciona un modelo preentrenado, así como un Transformer en función de la tarea y los recursos computacionales disponibles. Luego, se realiza una modificación de la capa de clasificación, adaptando la capa de salida del modelo preeentrenado para que coincida con el número de clases de la nueva tarea [23], en este proceso los pesos del modelo preeentrenado se actualizan [22].

Estudio Previo

Como se mencionó previamente este estudio parte de un trabajo previo titulado “Modelo de clasificación supervisado de fotografías de fachadas para evaluar el daño estructural ocasionado por sismos de acuerdo con la Escala Macrosísmica Europea para apoyo de toma de

decisiones en el Instituto Geofísico-EPN" [9]. En dicho estudio se emplearon diversas CNN preentrenadas, incluyendo: VGG16, InceptionResnetV2 (híbrida de las arquitecturas Inception y Resnet), MobileNetV2, VGG16, DenseNet121, DenseNet201, Xception, ResNet50, VGG19, InceptionV3, y ResNet152V2.

El conjunto de datos se dividió en subconjuntos de entrenamiento (80%), validación (10%) y prueba (10%). Posteriormente, se aplicaron técnicas de preprocesamiento de imágenes y se realizaron dos experimentos: uno sin aumento de datos y otro con aumento de datos. Para este último, se utilizó el paquete de Python Augmentor, ampliando el conjunto a 1500 imágenes. El entrenamiento de los modelos se configuró con un tamaño de lote (batch size) de 64, 100 épocas, codificación one-hot para las etiquetas, la función de pérdida categorical crossentropy y el optimizador Adam. La evaluación del desempeño se realizó utilizando métricas como accuracy, precisión, balanced accuracy y F1-score [9].

Los resultados indicaron que DenseNet201 presentó el mejor rendimiento general en comparación con los demás modelos. Sin embargo, fue necesario aplicar aumento de datos para mejorar las métricas en clases con pocas muestras. En el primer experimento (sin aumento de datos), MobileNetV2 obtuvo el mejor rendimiento en edificaciones con mampostería (accuracy: 0.86, precision:0.87, accuracy balanced:0.85 y f1-score:0.86) y VGG16 obtuvo el mejor rendimiento en edificaciones de hormigón armado (accuracy: 0.74, precision:0.72, accuracy balanced:0.70 y f1-score:0.69). En el segundo experimento (con aumento de datos), DenseNet201 se destacó, logrando en mapostería (accuracy: 0.91, precision:0.91, accuracy balanced:0.90 y f1-score:0.90) y en hormigón armado (accuracy: 0.95, precision:0.95, accuracy balanced:0.95 y f1-score:0.95) [9].

METODOLOGÍA

La metodología empleada en este proyecto es CRISP-DM (Cross-Industry Standard Process for Data Mining) que es una metodología ampliamente utilizada en proyectos de minería de datos debido a que proporciona una estructura sistemática para guiar el desarrollo de un proyecto [24]. Además, su naturaleza iterativa asegura un enfoque flexible y dinámico [24]. CRISP-DM consta de seis fases principales: comprensión del negocio, donde se define el problema y los objetivos del proyecto; comprensión de los datos, que implica la recopilación y la exploración preliminar de los datos disponibles; preparación de los datos, en la que se procesan, limpian y transforman los datos para que sean adecuados para el modelo; modelado, donde se seleccionan y ajustan los algoritmos de minería de datos más apropiados; evaluación, que verifica la calidad y relevancia de los modelos generados en relación con los objetivos iniciales; y finalmente, despliegue, que abarca la implementación de la solución en el entorno real y su mantenimiento [24].

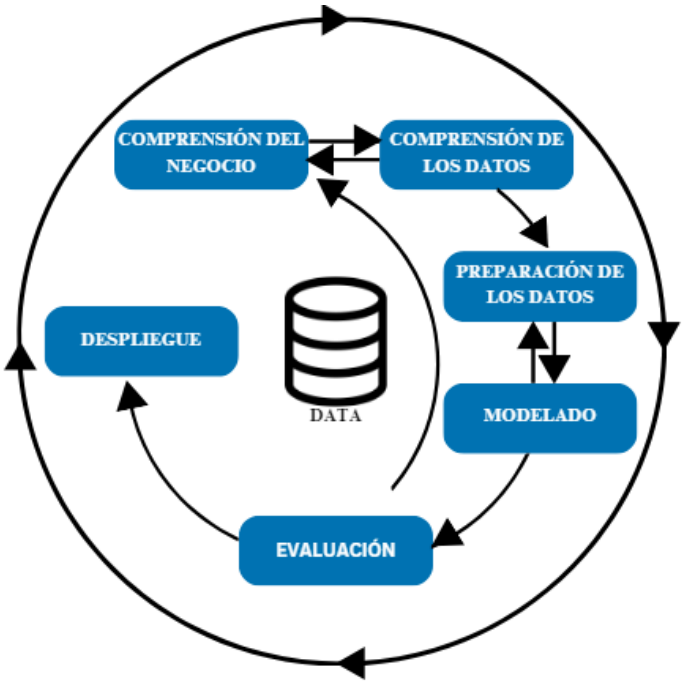

Figura 1. Diagrama de CRISP-DM. Adaptado de [25]

Comprensión del Negocio

Como se menciona en capítulos previos, el objetivo principal de este proyecto es determinar si los Transformers pueden superar el rendimiento de los modelos tradicionales, como las CNN, en la tarea específica de clasificación de imágenes de fachadas de daños estructurales post-sísmicos. Este estudio busca comparar las métricas obtenidas por los Transformers con los resultados presentados en la tesis previa, en la que se utilizaron diferentes arquitecturas de CNN para resolver el mismo problema. A partir de esta comparación, se espera obtener conclusiones claras sobre la eficacia de los Transformers en esta aplicación y evaluar si ofrecen o no ventajas significativas en términos de accuracy, precisión, balanced accuracy y F1-score.

Comprensión de Datos

El conjunto de datos está compuesto por un total de 605 imágenes, divididas en 2 categorías: mampostería (MA) 392 imágenes y hormigón armado (HA) 213 imágenes. Todas las imágenes cuentan con etiquetas asignadas según su grado de daño estructural, determinado a partir de la Escala Macrosísmica Europea (EMS-98). Esta escala clasifica el daño en cinco grados, que van desde fisuras leves (grado 1) hasta destrucción total (grado 5) [9]. Las estructuras representadas en las imágenes corresponden a edificaciones afectadas por sismos registrados en diversos países entre los años 1960 y 2020 [9].

En el caso de las edificaciones de mampostería, se evidencia un desbalance significativo en la distribución de imágenes, con una mayor concentración en el grado 2 (ver Figura 3). Por su parte, las imágenes correspondientes a edificaciones de hormigón armado presentan una distribución más equilibrada, aunque con una cantidad total de datos considerablemente menor (ver Figura 4). Estas particularidades del conjunto de datos, como el desbalance en las etiquetas y la limitada cantidad de información, representan un desafío para el desarrollo de modelos de

aprendizaje automático, ya que pueden afectar la capacidad de los modelos para generalizar correctamente.

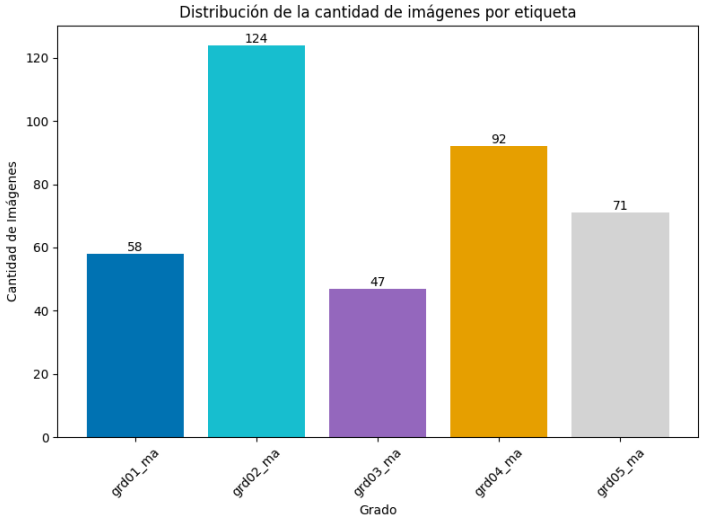

Figura 2. Distribución de la cantidad de imágenes por etiqueta de mampostería

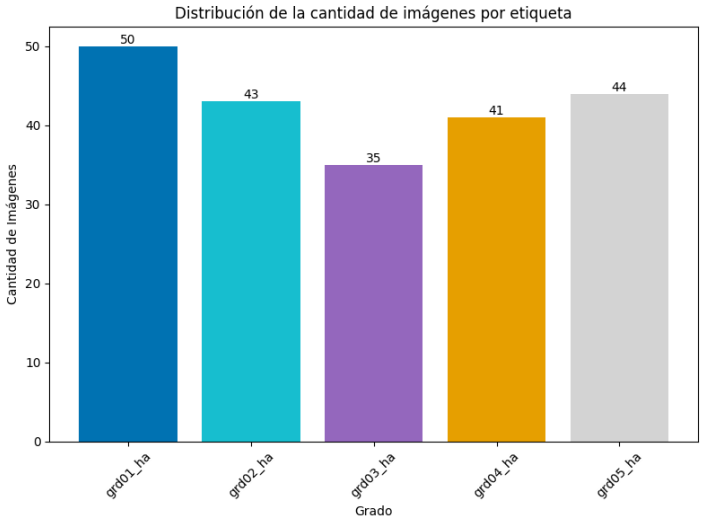

Figura 3. Distribución de la cantidad de imágenes por etiqueta de hormigón armado

Preparación de los datos

La fase de preparación de los datos se enfoca en la manipulación y transformación de los datos brutos para que sean adecuados para el modelado y análisis [26]. En esta etapa, el conjunto de

datos se dividió en entrenamiento, prueba y validación. Para abordar el problema relacionado con la limitada cantidad de datos, se aplicó la técnica de aumento de datos (data augmentation). El aumento de datos es una técnica esencial en el aprendizaje automático, particularmente en escenarios con conjuntos de datos reducidos. Su propósito principal consiste en incrementar la cantidad y diversidad de los datos de entrenamiento sin necesidad de recopilar nuevos datos reales, lo cual suele ser costoso y demandante en tiempo [27]. Para este trabajo, se realizaron las mismas transformaciones aplicadas en la tesis previa, como ecualización del histograma, volteo horizontal y zoom. Sin embargo, en la tesis anterior se utilizó la librería Augmentor, mientras que en este caso se implementaron las transformaciones mediante Pillow y PyTorch debido a su flexibilidad, compatibilidad y eficiencia en el procesamiento de imágenes.

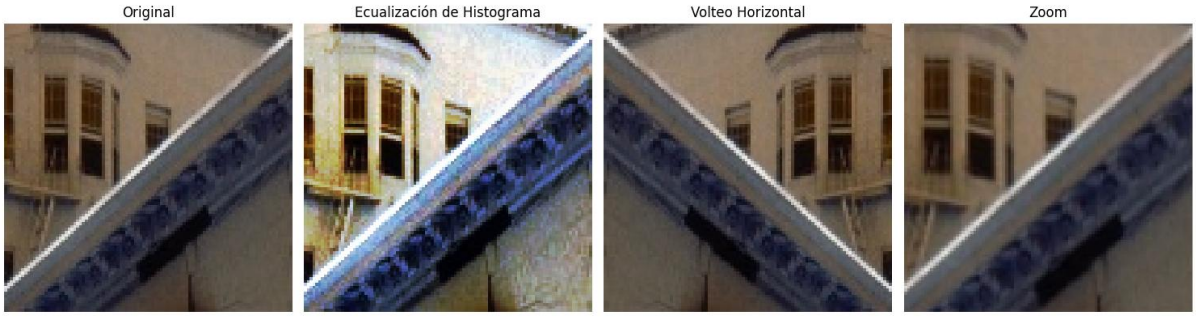

Figura 4. Muestra de imagen con Data Augmentation

Adicionalmente, las imágenes se procesaron utilizando AutoImageProcessor, un módulo especializado en ajustar las características de las imágenes al formato requerido por el modelo Transformers. Este paso es fundamental para garantizar que las imágenes cumplan con los parámetros de entrada específicos del modelo.

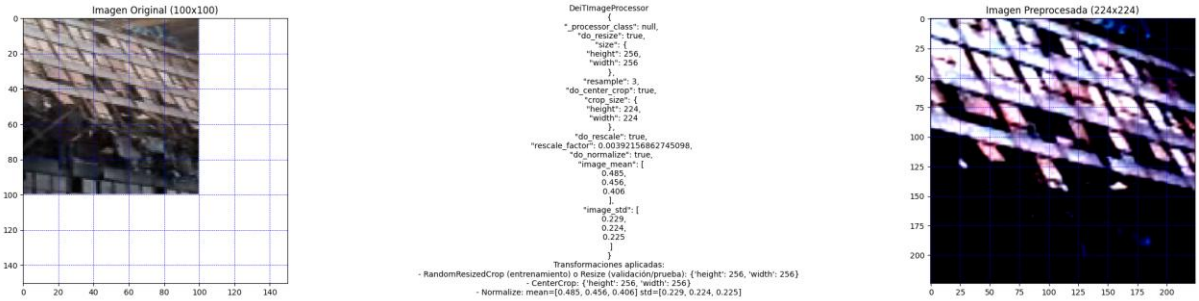

Figura 5. Imagen después de aplicar AutoImageProcessor

Modelado

Para la tarea de clasificación de imágenes se seleccionaron los Vision Transformers (ViT), debido a su notable rendimiento en problemas de computer vision [18]. Estos modelos han demostrado resultados de óptimos en conjuntos de datos de referencia como ImageNet, superando en algunos casos a las redes neuronales convolucionales (CNN) [18]. Su capacidad para capturar dependencias globales mediante mecanismos de autoatención les permite identificar patrones complejos en las imágenes [18]. No obstante, un desafío significativo en el uso de los ViT es su dependencia de grandes cantidades de datos de entrenamiento, lo cual puede ser un problema cuando se trabaja con conjuntos de datos reducidos.

Ante esta limitación, se optó por utilizar los Data-Efficient Image Transformers (DeiT), modelos diseñados para funcionar de manera eficiente con cantidades limitadas de datos [20]. Los modelos seleccionados incluyen DeiT Base Distilled 224, DeiT Base Distilled 384, DeiT Small Distilled y DeiT Tiny Distilled, los cuales permiten mantener la arquitectura fundamental de los ViT, pero incorporan optimizaciones adicionales. Una de estas mejoras es la introducción del distillation token, que aprovecha el conocimiento de un modelo maestro durante el entrenamiento para mejorar el rendimiento, incluso cuando el volumen de datos es reducido [20].

La arquitectura de los ViT y DeiT se basa en el procesamiento de las imágenes mediante su división en parches de tamaño fijo. Posteriormente, estos parches son aplanados y proyectados a un espacio de características mediante una capa de embedding lineal, que transforma cada parche en un vector numérico de dimensión uniforme. A estos vectores se les suman vectores posicionales para preservar la información espacial de cada parche dentro de la imagen original. Los vectores resultantes se introducen en una serie de bloques transformadores, donde se aplican mecanismos de autoatención que permiten capturar relaciones globales entre los parches. El resultado final es un vector de características que se procesa mediante operaciones de pooling y se envía a una red neuronal completamente conectada, encargada de realizar la clasificación (ver Figura 6).

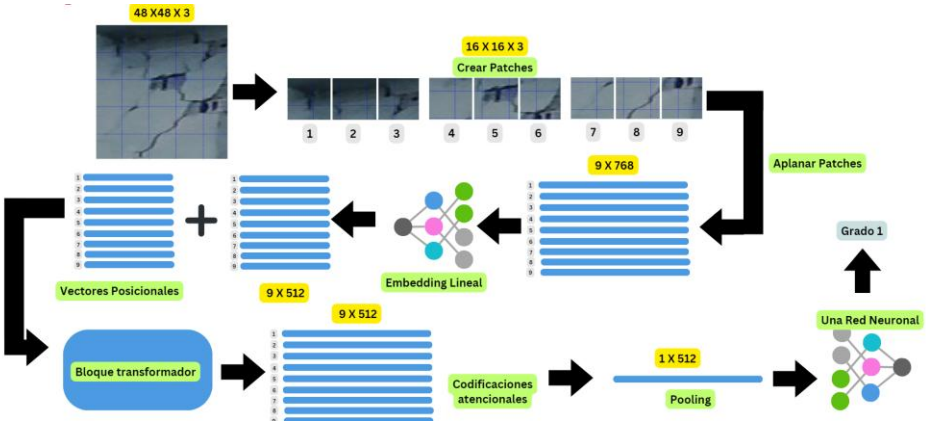

Figura 6. Arquitectura ViT. Adaptado de [28]

Para ajustar estos modelos a la tarea específica, se implementó la técnica de transfer learning. El aprendizaje por transferencia permite reutilizar modelos preentrenados como punto de partida, lo cual acelera el entrenamiento y mejora el rendimiento, especialmente cuando se dispone de un dataset pequeño [27]. Las modificaciones aplicadas incluyeron la congelación de las capas iniciales del modelo, que ya contienen características generales aprendidas durante el preentrenamiento, y la adición de nuevas capas al final de la arquitectura para ajustarse al número de clases específicas de la nueva tarea [27].

Evaluación

Para la fase de evaluación, se seleccionaron las mismas métricas de desempeño utilizadas en la tesis previa con el fin de permitir una comparación directa de los resultados obtenidos. Las métricas elegidas incluyen accuracy, accuracy balanced, F1-score, precision, y matrices de confusión en sus versiones normalizada y desnormalizada. Adicionalmente, se realizó una prueba estadística para evaluar si las diferencias en el desempeño del modelo actual con respecto al modelo previo son estadísticamente significativas.

Despliegue

Aunque el presente proyecto tiene un carácter investigativo y no contempla un despliegue operativo, los resultados obtenidos sientan las bases para una futura implementación práctica. El trabajo realizado proporciona un modelo validado que podría adaptarse y desplegarse en sistemas reales de clasificación de imágenes estructurales, contribuyendo así al análisis automatizado de daños en edificaciones afectadas por sismos.

PIPELINE DEL PROYECTO

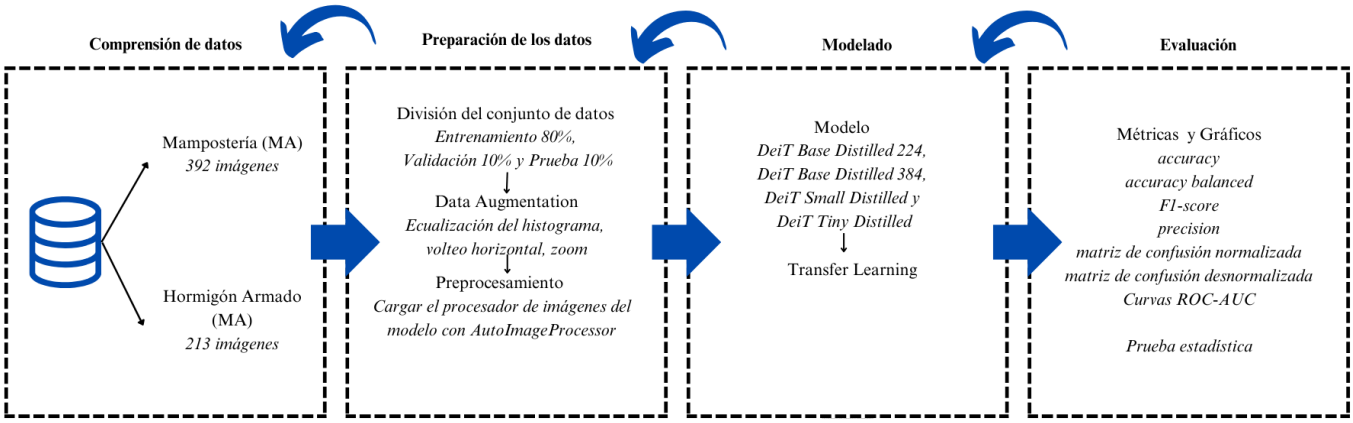

Figura 7. Pipeline del Proyecto. Adaptado de [9]

RESULTADOS

El análisis de los resultados se realizó en dos etapas. En primer lugar, se evaluaron los modelos DeiT aplicados a los conjuntos de datos de mampostería (MA) y hormigón armado (HA). Esta evaluación consideró tres escenarios: (1) utilizando los datos originales, (2) aplicando data augmentation agresivo (donde todas las clases se igualan en cantidad de datos para eliminar sesgos) y (3) aplicando data augmentation moderado, que combina undersampling en la clase mayoritaria y oversampling con data augmentation en las clases minoritarias (con un límite del 30%). Estos tres enfoques permitieron analizar el impacto del balance y desbalance en el desempeño de los modelos.

Tabla 1. Métricas de Evaluación para el Conjunto de datos de Mampostería (MA)

Modelo	**Técnica**		**Métricas**			
			Accuracy	**Accuracy Balanced**	**F1-Score**	**Precision**
DeiT Base Distilled 224	Original		0.83	0.79	0.82	0.85
	Data Augmentation	Agresivo	0.91	0.91	0.91	0.92
		Moderado	0.80	0.80	0.80	0.81
DeiT Base Distilled 384	Original		0.91	0.89	0.91	0.91
	Data Augmentation	Agresivo	0.94	0.94	0.94	0.94
		Moderado	0.90	0.89	0.90	0.91
DeiT Small Distilled 224	Original		0.88	0.85	0.87	0.89
	Data Augmentation	Agresivo	0.94	0.94	0.94	0.94
		Moderado	0.79	0.77	0.78	0.81

DeiT Tiny Distilled 224	Original		0.77	0.75	0.77	0.78
	Data Augmentation	Agresivo	0.84	0.84	0.84	0.85
		Moderado	0.72	0.72	0.71	0.71

La Tabla 1 presenta las métricas obtenidas para el conjunto de datos de mampostería (MA). En el escenario original, el modelo DeiT Base Distilled 384 alcanza un accuracy de 0.91 y un F1-Score de 0.91, lo que lo posiciona como el mejor modelo en esta configuración. Al aplicar data augmentation agresivo, el desempeño mejora significativamente, obteniendo un accuracy y un F1-Score de 0.94 para el mismo modelo. Este incremento demuestra la efectividad del aumento de datos para reducir los sesgos derivados del desbalance de clases. Por otro lado, el data augmentation moderado presenta un desempeño intermedio con un accuracy de 0.90, manteniendo valores competitivos sin necesidad de forzar un balance completo en las clases. Modelos como el DeiT Small Distilled 224 y el DeiT Tiny Distilled 224 muestran resultados similares, aunque ligeramente inferiores en comparación con el modelo Base Distilled 384.

Tabla 2. Métricas de Evaluación para el Conjunto de datos de Hormigón Armado (HA)

Modelo	**Técnica**		**Métricas**			
			Accuracy	**Accuracy Balanced**	**F1-Score**	**Precision**
DeiT Base Distilled 224	Original		0.78	0.79	0.77	0.79
	Data Augmentation	Agresivo	0.94	0.94	0.94	0.94
		Moderado	0.91	0.91	0.91	0.93
DeiT Base Distilled 384	Original		0.67	0.65	0.65	0.68
	Data Augmentation	Agresivo	0.89	0.88	0.88	0.89
		Moderado	0.89	0.89	0.88	0.92

DeiT Small Distilled 224	Original		0.64	0.63	0.60	0.62
	Data Augmentation	Agresivo	0.82	0.82	0.82	0.83
		Moderado	0.91	0.91	0.91	0.93
DeiT Tiny Distilled 224	Original		0.56	0.53	0.54	0.57
	Data Augmentation	Agresivo	0.86	0.86	0.86	0.86
		Moderado	0.71	0.71	0.68	0.79

La Tabla 2 muestra los resultados correspondientes al conjunto de datos de hormigón armado (HA). En este caso, el modelo DeiT Base Distilled 224 logra un accuracy de 0.78 y un F1-Score de 0.77 en el escenario original, lo que refleja un desempeño moderado debido al desbalance presente en los datos. Al aplicar data augmentation agresivo, se observa una mejora considerable, alcanzando un accuracy y un F1-Score de 0.94. El data augmentation moderado, por su parte, también incrementa el rendimiento del modelo, con un accuracy de 0.91, mostrando que la estrategia de balance controlado ofrece resultados sólidos sin llegar a los valores máximos del aumento agresivo.

En una segunda etapa del análisis, se compararon los resultados obtenidos por los modelos DeiT con los modelos CNN tradicionales evaluados en la tesis previa. La Tabla 3 presenta los resultados para el conjunto de datos de mampostería (MA). En el escenario original, el modelo DeiT Base Distilled 384 supera al modelo MobileNetV2 utilizado previamente, alcanzando un accuracy de 0.91 frente a 0.86. Sin embargo, al aplicar data augmentation agresivo, el desempeño del modelo DeiT incrementa a 0.94, superando significativamente al modelo DenseNet121 de la tesis previa, que alcanza un máximo de 0.87.

Tabla 3. Tabla Comparativa CNN vs DeiT en Mampostería (MA)

	Modelo	**Técnica**		**Métricas**			
				Accuracy	**Accuracy Balanced**	**F1-Score**	**Precision**
Tesis Previa	*MobileNetV2*	Original		0.86	0.87	0.85	0.86
Tesis Actual	*DeiT Base Distilled 384*			0.91	0.89	0.91	0.91
Tesis Previa	*DenseNet121*	Data Augmentation	Agresivo	0.86	0.87	0.87	0.86
Tesis Actual	*DeiT Base Distilled 384*			0.94	0.94	0.94	0.94

De manera similar, la Tabla 4 muestra la comparación en el conjunto de datos de hormigón armado (HA). En el escenario original, el modelo DeiT Base Distilled 384 logra un accuracy de 0.78, superando ligeramente al modelo VGG16, que presenta un accuracy de 0.74. No obstante, al aplicar data augmentation agresivo, el modelo MobileNetV2 de la tesis previa alcanza un accuracy de 0.95, ligeramente superior al 0.94 obtenido por el modelo DeiT Base Distilled 384. Estos resultados indican que, si bien los DeiT no superan por completo a todos los modelos CNN en escenarios con data augmentation agresivo, mantienen un desempeño altamente competitivo, especialmente en el manejo de datasets originales y en configuraciones moderadas de aumento de datos.

Tabla 4. Tabla Comparativa CNN vs DeiT en Hormigón Armado (HA)

	Modelo	**Técnica**		**Métricas**			
				Accuracy	**Accuracy Balanced**	**F1-Score**	**Precision**
Tesis Previa	*VGG16*	Original		0.74	0.72	0.70	0.69
Tesis Actual	*DeiT Base Distilled 384*			0.78	0.79	0.77	0.79
Tesis Previa	*MobileNetV2*	Data Augmentation	Agresivo	0.95	0.95	0.95	0.95
Tesis Actual	*DeiT Base Distilled 384*			0.94	0.94	0.94	0.94

Los modelos DeiT, especialmente el DeiT Base Distilled 384, demostraron un rendimiento superior en la mayoría de los escenarios, destacándose su robustez en condiciones de datos originales, así como su capacidad para mejorar significativamente mediante data augmentation agresivo y moderado. La comparación con modelos CNN tradicionales resalta la eficiencia de los Transformers, tanto en conjuntos de datos balanceados como en contextos con desbalance leve.

Este comportamiento puede atribuirse a las diferencias fundamentales en el procesamiento de imágenes entre ambas arquitecturas. Las CNN tradicionales capturan predominantemente características locales mediante operaciones de convolución, lo cual limita su capacidad para aprender dependencias espaciales globales. Por el contrario, los Vision Transformers (ViT) y sus variantes, como DeiT, combinan la captura de relaciones locales y globales a través del mecanismo de autoatención, lo que les permite identificar patrones más complejos y

dependencias a diferentes escalas. Esta capacidad resulta especialmente útil en tareas donde la estructura global de la imagen influye significativamente en el desempeño del modelo, como la clasificación de daños estructurales en edificaciones.

CONCLUSIONES

El presente trabajo ha demostrado la eficacia de los Data-Efficient Image Transformers (DeiT) en la clasificación de imágenes estructurales, particularmente en el análisis de daños en edificaciones de mampostería y hormigón armado. A través de la implementación de diferentes enfoques de data augmentation (agresivo y moderado), se evidenció cómo el balance de las clases influye significativamente en el rendimiento de los modelos. Los resultados obtenidos indican que los modelos DeiT, en especial el DeiT Base Distilled 384, logran un desempeño superior al de modelos CNN tradicionales, como MobileNetV2, DenseNet121 y VGG16, tanto en conjuntos de datos originales como en escenarios con aumento de datos.

La principal diferencia entre ambos enfoques radica en el procesamiento de las imágenes: mientras las CNN capturan características locales, los Vision Transformers tienen la capacidad de identificar tanto relaciones locales como globales gracias al mecanismo de autoatención. Esta ventaja permite a los Transformers adaptarse de manera más eficiente a problemas donde la información espacial a diferentes escalas resulta crítica, como en la clasificación de daños estructurales.

En el contexto nacional, este trabajo contribuye a la automatización de procesos de evaluación estructural, especialmente en países sísmicamente activos, donde la detección y clasificación temprana de daños resulta vital para la toma de decisiones post-desastre. A nivel internacional, los resultados obtenidos se alinean con estudios recientes que demuestran el desempeño superior de los Transformers en tareas de visión por computadora, superando a las arquitecturas tradicionales cuando se aplican estrategias efectivas de entrenamiento y aumento de datos.

Durante el desarrollo de este trabajo, se identificaron desafíos relacionados con la limitada cantidad de datos disponibles y el desbalance de clases, lo que requirió la implementación de

técnicas de data augmentation y el uso de transfer learning para optimizar los modelos. Este proceso permitió no solo mejorar el rendimiento del modelo, sino también explorar metodologías escalables para conjuntos de datos reducidos.

Finalmente, esta investigación ha reafirmado la importancia de seleccionar arquitecturas y técnicas adecuadas a las particularidades del problema abordado. El aprendizaje obtenido proporciona bases sólidas para la implementación futura de modelos Transformers en la evaluación estructural automatizada.

TRABAJOS FUTUROS

A partir de los resultados obtenidos, se identifican diversas líneas de investigación y desarrollos futuros que podrían ampliar y mejorar los aportes del presente trabajo. Un primer enfoque consiste en la ampliación del conjunto de datos, mediante la recopilación de imágenes adicionales que representen edificaciones con distintos niveles de daño, tanto a nivel local como internacional, para fortalecer la capacidad de generalización de los modelos propuestos. Otra línea de investigación relevante es el desarrollo de análisis multimodal, donde se incorpore información complementaria, como datos sísmicos, características estructurales y metadatos de las edificaciones, con el objetivo de crear modelos más robustos en la clasificación de daños. Además, resulta importante evaluar la aplicabilidad de los modelos en tiempo real, mediante su implementación en sistemas automatizados para su uso en drones, dispositivos móviles o plataformas de monitoreo, lo que permitiría realizar evaluaciones rápidas en contextos post-sísmicos. Del mismo modo, se podría extender el estudio a edificaciones construidas con otros materiales, como acero o madera, así como a infraestructuras críticas, como puentes o represas, donde la detección de daños resulta fundamental. Para validar su efectividad, los modelos también deben ser sometidos a pruebas en contextos reales, donde su desempeño pueda medirse en situaciones prácticas tras eventos sísmicos, facilitando su aplicación por profesionales en la evaluación de infraestructuras. Finalmente, se sugiere investigar técnicas avanzadas de data augmentation y regularización que optimicen aún más el desempeño de los modelos en datasets pequeños y desbalanceados, así como explorar enfoques como semi-supervised learning y el few-shot learning, que podrían ofrecer soluciones prometedoras para superar las limitaciones de datos y mejorar la capacidad de los modelos en escenarios similares.

REFERENCIAS BIBLIOGRÁFICAS

[1] T. Menéndez, "70.000 sismos han sacudido a Ecuador en los últimos 10 años," 22 Julio 2022. [Online]. Available: https://www.primicias.ec/noticias/sociedad/sismos-ecuador-ultimos-anos-instituto-geofisico/.

[2] P. Ramón, S. Vallejo, P. Mothes, D. Andrade, F. J. Vásconez, H. Yepes, S. Hidalgo and S. Santamaría, "Instituto Geofísico – Escuela Politécnica Nacional, the Ecuadorian Seismology and Volcanology Service," 01 Noviembre 2021. [Online]. Available: https://doi.org/10.30909/vol.04.S1.93112.

[3] IG-EPN, "Histórico - Instituto Geofísico - EPN," 01 febrero 2012. [Online]. Available: https://www.igepn.edu.ec/servicios/noticias/content/49-historico?start=294#:~:text=A%20las%2010%3A36%20(tiempo,27%20de%20febrero%20de%202010..

[4] INEC, "Reconstruyendo las cifras luego del sismo Memorias," 13 abril 2017. [Online]. Available: https://www.ecuadorencifras.gob.ec/documentos/web-inec/Bibliotecas/Libros/Memorias%2013%20abr%202017.pdf.

[5] J. E. Arteaga Moriano, J. C. Mutis Mamuscay and D. M. Torres Ortiz, "Conceptos preliminares para análisis en ingeniería sísmica," 11 septiembre 2023. [Online]. Available: https://doi.org/10.26507/paper.3353.

[6] EPN, "Monitoreo Permanente," [Online]. Available: https://www.epn.edu.ec/monitoreo-permanente/.

[7] IGN, "ESCALA MACROSÍSMICA EUROPEA," [Online]. Available: https://www.ign.es/web/resources/docs/IGNCnig/SIS-Escala-Intensidad-Macrosismica.pdf.

[8] L. Papa, P. Russo, I. Amerini and L. Zhou, "A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking," 24 abril 2024. [Online]. Available: https://doi.org/10.1109/TPAMI.2024.3392941.

[9] D. Tarapues, "Modelo de clasificación supervisado de fotografías de fachadas para evaluar el daño estructural ocasionado por sismos de acuerdo con la escala Macrosísmica europea para apoyo de toma de decisiones en el Instituto Geofísico – EPN," 15 febrero 2022. [Online]. Available: http://bibdigital.epn.edu.ec/handle/15000/22167.

[10] Z. Reches and J. Fineberg, "Earthquakes as Dynamic Fracture Phenomena," 17 marzo 2023. [Online]. Available: https://doi.org/10.1029/2022JB026295.

[11] S. Ghimire, P. Guéguen, S. Giffard-Roisin and D. Schorlemmer, "Testing machine learning models for seismic damage prediction at a regional scale using building-damage dataset compiled after the 2015 Gorkha Nepal earthquake," 21 julio 2022. [Online]. Available: https://doi.org/10.1177/87552930221106495.

[12] A. Vuoto, J. Ortega, P. Lourenço, F. J. Suárez and A. C. Núñez, "Safety assessment of the Torre de la Vela in la Alhambra, Granada, Spain: The role of on site works," 28 mayo 2022. [Online]. Available: https://doi.org/10.1016/j.engstruct.2022.114443.

[13] S. Del Mese, L. Graziani, F. Meroni, V. Pessina and A. Tertulliani, "Considerations on using MCS and EMS-98 macroseismic scales for the intensity assessment of

contemporary Italian earthquakes," 20 mayo 2023. [Online]. Available: https://doi.org/10.1007/s10518-023-01703-0.

[14] K. S. Ballesteros Salazar, D. G. Caizaguano Montero, A. G. Haro Báez and T. Toulkeridis, "Case Study of the Application of an Innovative Guide for the Seismic Vulnerability Evaluation of Schools Located in Sangolquí, Interandean Valley in Ecuador," 16 septiembre 2022. [Online]. Available: https://doi.org/10.3390/buildings12091471.

[15] P. F. Giordano, C. Lacovino, S. Quqa and M. Pina Limongelli1, "The value of seismic structural health monitoring for post-earthquake building evacuation," 25 marzo 2022. [Online]. Available: https://doi.org/10.1007/s10518-022-01375-2.

[16] I. Hafner, D. Lazarevic, T. Kišicek and M. Stepinac, "Post-Earthquake Assessment of a Historical Masonry after the Zagreb Earthquake—Case Study," 8 marzo 2022. [Online]. Available: https://doi.org/10.3390/buildings12030323.

[17] H. Bilgin, M. Leti, R. Shehu, H. Baytan Özmen and A. Hilmi Deringol, "Reflections from the 2019 Durrës Earthquakes: An Earthquake An Earthquake," 31 agosto 2023. [Online]. Available: https://doi.org/10.3390/buildings13092227.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," 3 junio 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2010.11929.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uzkoreit, L. Jones, A. N. Gomez, L. Kaiser and . I. Polosukhin, "Attention Is All You Need," 12 june 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762.

[20] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles and H. Jégou, "Training data-efficient image transformers & distillation through attention," 15 enero 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2012.12877.

[21] A. Tabbakh and S. Sankar Barpanda, "A Deep Features Extraction Model Based on the Transfer Learning Model and Vision Transformer ‘‘TLMViT’’ for Plant Disease Classification," 05 mayo 2023. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10119138.

[22] J. Hensel Donato, Y. Novanto and Sutrisno, "Mask Usage Recognition using Vision Transformer with Transfer Learning and Data Augmentation," 22 Marzo 2022. [Online]. Available: https://doi.org/10.48550/arXiv. 2203.11542.

[23] U. Mohammed, Z. Tehseen and T. All, "Analyzing Transfer Learning of Vision Transformers for Interpreting Chest Radiography," 11 julio 2022. [Online]. Available: https://doi.org/10.1007/s10278-022-00666-z.

[24] J. Saltz and I. Krasteva, "Current approaches for executing big data science projects—a systematic literature review," 21 febrero 2022. [Online]. Available: https://doi.org/10.7717/peerj-cs.862.

[25] IBM;, "CRISP-DM Help Overview," 2021, 17 july. [Online]. Available: https://www.ibm.com/docs/pt-br/spss-modeler/saas?topic=dm-crisp-help-overview.

[26] F. Martínez Plumed, L. Contreras Ochando, C. Ferri, J. Hernández Orallo, M. Kull, N. Lachiche, P. Flavh and P. Flach, "CRISP-DM Twenty Years Later: From Data Mining

Processes to Data Science Trajectories," 08 august 2021. [Online]. Available: 10.1109/TKDE.2019.2962680.

[27] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," 02 december 2022. [Online]. Available: https://doi.org/10.1016/j.array.2022.100258.

[28] M. Sotaquirá, "El Vision Transformer ¡EXPLICADO!," Codificando Bits, 06 mayo 2024. [Online]. Available: https://www.youtube.com/watch?v=A-6DF9mkDuQ.