



DETECCIÓN DE ANOMALÍAS VISUALES EN TERRENOS PROPENSOS A
DESLIZAMIENTOS MEDIANTE ANÁLISIS MULTITEMPORAL DE IMÁGENES DE
PUNTO FIJO

Visual anomaly detection in landslide-prone terrain based on multitemporal fixed-
camera imagery

FABIAN DAVID SANCHEZ MARTINEZ

Tesis de Maestría

Asesores

PABLO ANDRES SALDARRIAGA ARISTIZABAL
JUAN CARLOS ARBELAEZ ESTRADA

UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA
MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA
MEDELLÍN
2026

Detección de anomalías visuales en terrenos propensos a deslizamientos mediante análisis multitemporal de imágenes de punto fijo

Fabian D. Sánchez M.¹, Pablo A. Saldarriaga A.², Juan C. Arbeláez E.³

¹ Escuela de ciencias aplicadas e ingeniería – Universidad EAFIT, Medellín, Colombia

Email: fdsanchezm@eafit.edu.co

² Escuela de ciencias aplicadas e ingeniería – Universidad EAFIT, Medellín, Colombia

Email: psaldar2@eafit.edu.co

³ Escuela de ciencias aplicadas e ingeniería – Universidad EAFIT, Medellín, Colombia

Email: jarbel16@eafit.edu.co

Resumen. Los movimientos de masa representan un alto riesgo en áreas propensas a deslizamientos y otros fenómenos geotécnicos, afectando tanto la infraestructura (por ejemplo, viviendas, redes viales y construcciones) como la seguridad humana. Sin embargo, los sistemas tradicionales de monitoreo basados únicamente en sensores presentan limitaciones para detectar cambios visuales tempranos en el terreno que podrían preceder a estos eventos. Además, existen desafíos específicos que este proyecto busca abordar, como la falta de datos etiquetados, el fuerte desbalance de clases entre imágenes normales y anómalas, la variabilidad ambiental (debida a condiciones de iluminación, sombra o lluvia) y la desalineación entre las imágenes, causada por ligeros desplazamientos en las cámaras. Estas condiciones dificultan la construcción de modelos supervisados convencionales debido al tiempo, costo y esfuerzo que implica el etiquetar datos. Por este motivo este estudio se orienta hacia el uso de técnicas no supervisadas para la detección de anomalías visuales.

Este estudio se enfoca en analizar imágenes provistas por el Sistema de Alerta Temprana del Valle de Aburrá (SIATA) de diferentes estaciones, con el fin de identificar anomalías o cambios en el suelo y el terreno en general, a fin de que un experto pueda validar si estas constituyen un peligro. A través del análisis de imágenes multitemporales y la aplicación de técnicas de visión por computador, se procesan secuencias de imágenes que permiten detectar cambios inusuales en la textura, el color, la vegetación e incluso la morfología del terreno. Estos cambios se cuantifican mediante métodos estadísticos y modelos de detección de anomalías, lo que permite priorizar las regiones que requieren la verificación de un experto.

Los resultados preliminares muestran que los modelos de visión artificial alcanzaron desempeños consistentes en la detección de anomalías. Entre ellos, DINOv2-ViT-L/14 que obtuvo los mejores valores, con una *Precision* del 0.51, *Recall* del 0.47, F1-Score de 0.49 y un *Accuracy* de 0.82. Estos resultados sugieren que es posible identificar indicios de cambios inminentes en el suelo y generar alertas basadas en cambios específicos observados en las imágenes. Estas conclusiones proporcionan un enfoque innovador y a su vez complementario a los sistemas de monitoreo tradicionales, fortaleciendo la gestión de riesgo y prevención de desastres en zonas previamente marcadas como propensas a deslizamientos.

Palabras clave: Detección de anomalías, visión por computador, imágenes multitemporales, aprendizaje automático, SIATA, gestión de riesgos.

1 Introducción

Los deslizamientos de tierra constituyen uno de los fenómenos naturales más frecuentes y destructivos en regiones montañosas, provocando pérdidas humanas, daños en infraestructura y afectaciones económicas considerables (Dou et al., 2019; Santacana et al., 2003). En Colombia, el Valle de Aburrá, ubicado en el departamento de Antioquia, es una región densamente poblada y con una geografía compleja (Figura 1), caracterizada por fuertes pendientes, suelos inestables y precipitaciones altamente variables. Estas condiciones lo hacen susceptible a movimientos del suelo, deslizamientos de tierra y otros eventos geotécnicos que representan una amenaza significativa para la población, haciendo necesario usar herramientas que permitan detectar de manera oportuna los cambios que preceden estos procesos.



Figura 1. Imagen de la región del Valle de Aburrá, en el departamento de Antioquia, Colombia, región donde se localiza geográficamente este estudio.

Tradicionalmente, los estudios de susceptibilidad y monitoreo de movimientos de masa han sido basados en modelos físicos o estadísticos, apoyados en el uso de sensores geotécnicos, sin embargo, estos sistemas presentan ciertas limitaciones en cobertura, instalación y mantenimiento. Adicionalmente, suelen carecer de datos etiquetados que permitan el entrenamiento de modelos supervisados (Dou et al., 2019; Merghadi et al., 2020). Además, el uso de eventos positivos registrados es muy bajo respecto al volumen de información normal, generando conjuntos de datos altamente desbalanceados, dificultando así la detección de posibles señales tempranas.

En este contexto, el constante monitoreo visual mediante cámaras fijas se ha convertido en una fuente de información complementaria a los sistemas tradicionales basados en sensores (Stumpf y Kerle, 2011). Estas cámaras, instaladas en puntos estratégicos del territorio, permiten capturar imágenes periódicas del entorno que, analizadas en secuencia temporal, pueden revelar transformaciones sutiles en la superficie del suelo, la vegetación o los cuerpos de agua, que podrían ser indicios de procesos de inestabilidad. Estudios recientes de visión por computador han demostrado que los cambios multitemporales en textura, color y cobertura vegetal pueden ser predictores confiables de deslizamientos o erosión superficial (Ghorbanzadeh et al., 2019; Merghadi et al., 2020).

La susceptibilidad a los movimientos del terreno depende de múltiples factores naturales y antrópicos, entre los más importantes se encuentran la pendiente, la litología, la cobertura vegetal, el uso del suelo y las precipitaciones (Santacana et al., 2003; Dou et al., 2019; Ghorbanzadeh et al., 2019; Merghadi et al., 2020), como se observa en la figura 2. Estos factores influyen tanto en la generación de deslizamientos como en los cambios visuales que preceden a su ocurrencia. Por ejemplo, mayores ángulos incrementan las fuerzas de corte y el riesgo de fallas; los suelos arcillosos o altamente meteorizados presentan menor resistencia; la deforestación y urbanización alteran el equilibrio hidrológico; y la saturación del suelo producto de las precipitaciones reduce la cohesión, generando condiciones críticas. Estas condiciones, cuando son observadas a través de imágenes periódicas, pueden reflejarse en cambios sutiles de color, textura o morfología, que sirven como indicadores visuales de posibles procesos de inestabilidad.

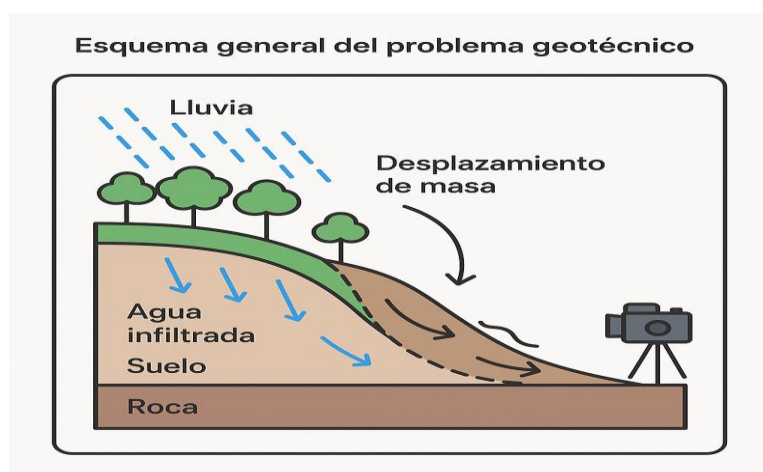


Figura 2. Representación gráfica de la inestabilidad de un terreno y variables que contribuyen la detección visual. *Ilustración original creada para este estudio mediante ChatGPT (OpenAI), 2025.*

A nivel regional, el Sistema de Alerta Temprana del Valle de Aburrá (SIATA) ha desplegado cámaras fijas en puntos estratégicos para el monitoreo del paisaje, complementando así la información de los sensores geotécnicos. Estas imágenes constituyen una fuente de datos visuales para la identificación de cambios graduales o abruptos en el terreno, sin embargo, la ausencia de datos etiquetados y variabilidad en las condiciones de captura como iluminación, sombras o desalineaciones de la cámara, plantean retos significativos para su análisis automatizado. Teniendo en cuenta esta problemática, este estudio propone una estrategia de análisis visual automatizado, que busca mejorar la detección temprana de anomalías en el terreno mediante modelos preentrenados, para que de esta forma se reduzca la dependencia de datos manualmente etiquetados, de la intervención humana y, fortaleciendo la capacidad de monitoreo del SIATA.

Los avances recientes en aprendizaje profundo han permitido el desarrollo de modelos capaces de generar representaciones visuales de alta dimensión (*embeddings*) que capturan información semántica y estructural de las imágenes sin necesidad de entrenamiento específico sobre cada dominio (Xie et al., 2017; Radford et al., 2021; Oquab et al., 2023). A diferencia de los modelos supervisados tradicionales, estos modelos no requieren un proceso de clasificación supervisado, sino que aprenden a través de tareas de contraste o autosupervisión, lo que los hace especialmente útiles en contextos donde no existen etiquetas o eventos conocidos de antemano, como ocurre en la detección de anomalías (Radford et al., 2021; Oquab et al., 2023). En este estudio se utilizaron modelos de visión artificial bajo este paradigma para detectar cambios visuales en el terreno a partir de secuencias temporales de imágenes.

Para ello, se propone una aproximación de procesamiento y análisis visual automatizado, que incluye la selección de regiones de interés (*ROI*), la aplicación de filtros de calidad visual, la extracción de representaciones numéricas o *embeddings* a partir de modelos preentrenados, y el uso de métricas estadísticas para identificar desviaciones o comportamientos anómalos. Este enfoque no busca reemplazar, sino complementar los sistemas de monitoreo tradicionales, permitiendo en un futuro integrar la información visual con otras fuentes de datos para fortalecer las estrategias de prevención y gestión del riesgo en zonas susceptibles a deslizamientos. En conjunto, esta investigación propone un flujo de procesamiento automatizado para la detección de cambios visuales en terrenos mediante imágenes de cámaras fijas, aplicable en regiones con infraestructura de monitoreo similar.

Además, evalúa el uso de modelos de visión artificial y técnicas estadísticas para identificar patrones de cambio sin necesidad de datos etiquetados, abordando una de las principales limitaciones de los sistemas tradicionales, y plantea la integración futura de las observaciones visuales con otros datos físicos. Con ello, se pretende contribuir al desarrollo de sistemas de monitoreo más robustos, eficientes y adaptables a diferentes entornos, promoviendo la protección de la vida y la infraestructura en el Valle de Aburrá y regiones con características similares.

Este documento está organizado de la siguiente manera: la Sección 2 presenta el estado del arte, donde se revisan los estudios previos sobre monitoreo de terreno, detección de deslizamientos y aplicaciones de visión por computador para análisis ambiental. La Sección 3 describe la metodología empleada, que incluye la preparación de las imágenes, la extracción de representaciones visuales y las métricas utilizadas para la detección de anomalías. En la Sección 4 se exponen los resultados y experimentos computacionales, seguidos de la Sección 5, donde se presenta el caso de aplicación en el Valle de Aburrá. Finalmente, la Sección 6 expone las conclusiones y líneas de trabajo futuras, destacando el potencial del modelo para su integración con sistemas de monitoreo físico y su expansión a otras regiones.

2 Estado del arte

En esta sección se detalla de forma general los avances relevantes para este estudio. Primero, se detallan los conceptos fundamentales de detección de anomalías en imágenes y los desarrollos recientes en representaciones visuales basadas en aprendizaje profundo (Sección 2.1). Luego, se presentan los métodos tradicionales empleados para la susceptibilidad a deslizamientos que no usan técnicas de visión por computador (Sección 2.2) y luego se describen los enfoques basados en análisis visual aplicados a la dinámica del terreno (Sección 2.3). Posteriormente, se discuten los desafíos propios del uso de cámaras fijas terrestres en sistemas de monitoreo continuo (Sección

2.4) y finalmente, se introducen los enfoques autosupervisados y no supervisados para detección de anomalías, que motivan la aproximación adoptada en este trabajo (Sección 2.5).

2.1 Detección de anomalías en imágenes

La detección de anomalías en imágenes se ha convertido en una de las ramas más atractivas de la visión por computador, impulsada por el crecimiento de modelos preentrenados capaces de identificar desviaciones sutiles en patrones visuales de normalidad. En términos generales, la detección de anomalías busca reconocer objetos, regiones, comportamientos o cambios en los patrones que se apartan del conjunto mayormente predominante de los datos, sin que esto dependa necesariamente de etiquetas predefinidas. Esta capacidad es relevante en contextos donde los eventos anómalos son poco frecuentes o costosos de registrar, como ocurre en el monitoreo de infraestructuras, control industrial o vigilancia ambiental. (Ruff et al., 2021).

Los recientes avances en aprendizaje profundo han permitido el desarrollo de arquitecturas que aprenden representaciones visuales robustas a partir de grandes volúmenes de datos no etiquetados. En este sentido, enfoques contrastivos como CLIP (Radford et al., 2021) y autosupervisados como DINOv2 (Oquab et al., 2023), han demostrado un notable desempeño al generar *embeddings* (representaciones numéricas de alta dimensión) que capturan la semántica y la estructura de las imágenes, junto con sus características más relevantes. Estos métodos aprenden representaciones visuales robustas, sin requerir de anotaciones manuales: en el caso de CLIP mediante relaciones entre pares imagen-texto, y en el caso de DINOv2 a través de tareas autosupervisadas basadas en reconstrucción y consistencia entre vistas de la misma imagen. De manera complementaria, arquitecturas más tradicionales como ResNeXt (Xie et al., 2017) siguen siendo ampliamente utilizadas como extractores de características por su eficiencia y capacidad para representar patrones visuales complejos.

2.2 Métodos tradicionales para la susceptibilidad a deslizamientos (sin visión por computador)

El aprendizaje automático ha sido ampliamente empleado para la evaluación de la susceptibilidad a deslizamientos y predicción de eventos de remoción en masa. Goetz et al. (2015) utilizaron modelos como *random forest* para mapear zonas propensas a deslizamientos combinando variables topográficas y geológicas, logrando mejorar la precisión respecto a métodos estadísticos clásicos. Dou et al. (2019) evaluaron máquinas de soporte vectorial y modelos de ensamble para estimar la probabilidad espacial de ocurrencia de deslizamientos con buenos niveles de exactitud. Chen et al. (2022) integraron información de diversas fuentes con técnicas de aprendizaje profundo para mejorar la predicción espacial. Merghadi et al. (2020), en uno de los metaanálisis más completos del tema, compararon más de 40 métodos distintos y concluyeron que los modelos basados en árboles, como *random forest*, suelen superar a las aproximaciones estadísticas tradicionales.

A pesar de los avances, a mayoría de estos enfoques requieren bases de datos históricas bien instrumentadas y con registros completos de eventos positivos, condición que limita su aplicabilidad en regiones donde la información geotécnica es escasa, discontinua o altamente desbalanceada, como ocurre en muchas zonas andinas (Goetz et al., 2015; Dou et al., 2019; Merghadi et al., 2020; D'Amato Avanzi et al., 2021; Herrera et al., 2018).

2.3 Visión por computador aplicada a deslizamientos

Diversas investigaciones han explorado la aplicación de modelos de visión por computador para detectar cambios en la superficie de los terrenos, como la cobertura vegetal, desplazamientos en el terreno o incluso erosión superficial. Stumpf y Kerle (2011) demostraron que los métodos basados en análisis de textura y color pueden identificar deslizamientos recientes en imágenes ópticas de alta resolución mediante técnicas de aprendizaje supervisado. Del mismo modo, Ghorbanzadeh et al. (2019) y Merghadi et al. (2020) evaluaron arquitecturas profundas como las redes neuronales convolucionales (*CNN* por sus siglas en inglés) para clasificar ciertas zonas con inestabilidad, mostrando un desempeño superior al de los modelos estadísticos tradicionales. Guo et al. (2021) destacaron la utilidad de los índices de vegetación derivados de imágenes multitemporales para detectar procesos de reptación del terreno, mientras que, recientemente Coluzzi et al. (2025) propusieron un esquema de detección rápida de deslizamientos a partir de imágenes ópticas satelitales libres, combinadas con análisis espectral y segmentación automática.

Otras investigaciones también han incorporado datos de radar e interferometría SAR que capturan micro deformaciones superficiales y desplazamientos milimétricos en el terreno (Li et al., 2024), lo que amplía la cobertura

de la observación, aunque requiere de una infraestructura computacional y de almacenamiento considerablemente amplio y costoso. A pesar de su precisión, estos enfoques basados en teledetección presentan restricciones de temporalidad, pues las imágenes satelitales suelen tener una frecuencia de captura de ciertos días o inclusive de semanas, dificultando la observación continua.

2.4 Cámaras fijas terrestres para monitoreo continuo

El uso de cámaras fijas instaladas en ciertos sistemas de alerta locales, como las que usa el SIATA, ofrecen una fuente de información más densa temporalmente, con capturas minuto a minuto, permitiendo observar la dinámica con mayor detalle de un terreno, pero con desafíos propios. El análisis automatizado de estas imágenes enfrenta retos como variaciones en la iluminación, presencia de sombras y alto contraste, niebla, lluvia, así como cambios ligeros y notorios en el ángulo entre capturas, que afectan la continuidad visual y comparabilidad temporal de las secuencias. Además, la ausencia de etiquetas que indiquen cuándo ocurrió un cambio real impide entrenar modelos supervisados de manera convencional.

2.5 Enfoques autosupervisados y detección no supervisada de anomalías

Frente a estas limitaciones, enfoques basados en modelos de aprendizaje autosupervisado y detección no supervisada de anomalías ofrecen una alternativa prometedora. Bergmann et al. (2019) y Ruff et al. (2021) usan modelos que permiten aprender la distribución de las imágenes normales y detectarían desviaciones significativas sin requerir ejemplos explícitos de eventos anormales, lo que resulta útil especialmente en dominios donde los cambios geotécnicos reales son escasos o difíciles de etiquetar. Este tipo de modelos resulta adecuado para escenarios como el Valle de Aburrá, donde no existe un registro consistente y masivo de deslizamientos visibles en imágenes, pero sí grandes volúmenes de datos normales capturados mediante cámaras fijas.

En resumen, la literatura reciente evidencia un progreso sustancial en la detección automatizada de cambios superficiales en el terreno mediante el uso de técnicas de visión por computador y aprendizaje profundo. Sin embargo, la mayoría de los estudios existentes se centran en imágenes satelitales o aéreas, con frecuencias de muestreo limitadas y conjuntos de datos previamente balanceados. En este contexto, este estudio introduce un enfoque novedoso basado en imágenes de cámaras fijas terrestres, aplicando modelos de aprendizaje no supervisado para detección de anomalías visuales. Dicho enfoque permite un monitoreo más continuo y de menor costo, contribuyendo de esta forma a la identificación temprana de cambios en zonas superficiales, propensas a deslizamientos.

3 Metodología

Este estudio se desarrolló siguiendo la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), ampliamente utilizada en proyectos de ciencia de datos por su estructura iterativa y flexible. Dicha metodología consta de seis fases: (1) Entendimiento del negocio, (2) Entendimiento de los datos, (3) Preparación de los datos, (4) Modelado, (5) Evaluación y (6) Despliegue.

A continuación, se presentan los detalles de la aplicación de cada una de estas fases en el contexto del análisis multitemporal de imágenes de cámara fija para la detección de anomalías visuales en terrenos propensos a deslizamientos.

En las secciones de introducción y estado del arte se presentó el contexto geotécnico del Valle de Aburrá, la relevancia que cobran los deslizamientos en la comunidad y el soporte técnico del uso de visión por computador en este tipo de problemas (p. ej. Santacana et al., 2003; Dou et al., 2019; Ghorbanzadeh et al., 2019; Merghadi et al., 2020; Stumpf & Kerle, 2011). Aquí se retoma ese contexto desde la perspectiva operativa de CRISP-DM.

3.1 Entendimiento del negocio

El objetivo principal de este estudio es identificar cambios inusuales en la superficie de un terreno que puedan estar asociados a procesos de inestabilidad o movimientos de masa, usando imágenes de cámaras de punto fijo del SIATA en estaciones del Valle de Aburrá (Andalucía, Bermejala, Montañita, Playas, Pueblo Viejo, Raya y Vallería), y siguiendo enfoques similares a los usados en la detección de movimientos de masa (Stumpf & Kerle, 2011; Ghorbanzadeh et al., 2019; Behling et al., 2014).

A diferencia de los enfoques tradicionales basados principalmente en sensores geotécnicos e hidrometeorológicos (como lo son acelerómetros, extensómetros, pluviómetros), que pueden presentar limitaciones de cobertura espacial, ser costosos de instalar o incluso de mantener (D'Amato Avanzi et al., 2021), así como también pueden presentar considerable desbalanceo de clases con pocos datos o eventos positivos (Merghadi et al., 2020), este trabajo se apoya en secuencias de imágenes temporales que permiten observar de manera continua la superficie del terreno, la vegetación e incluso algunos cuerpos de agua (Stumpf & Kerle, 2011; Ghorbanzadeh et al., 2019).

La problemática de fondo radica en esta pregunta: ¿En qué momento un cambio visual en una secuencia de imágenes es lo suficientemente relevante para que pueda considerarse como un evento anormal que requiera la revisión de un experto geotécnico? Para responder a esta pregunta tendríamos que transformar un flujo extenso de imágenes minuto a minuto en un conjunto reducido de posibles anomalías, que puedan ser revisados por alguien con conocimiento especializado en geotecnia (Ruff et al., 2021; Bergmann et al., 2019). Desde la óptica de CRISP-DM, el resultado del entendimiento del negocio se traduce así en:

- Un objetivo funcional: priorizando aquellas imágenes que pueden tener cambios relevantes en el terreno (pérdida de vegetación, exposición del suelo, modificación en taludes, etc.).
- Un objetivo técnico: diseñar un flujo automatizado que, tomando imágenes de cámara fija, extraiga características principales de dichas imágenes, calcule métricas de cambio, y genere algún tipo de alerta visual cuando note que una diferencia es significativa.
- Un alcance: que proponga un sistema complementario a sensores físicos existentes, para que posteriormente se integren a los sistemas de información del SIATA.

3.2 Entendimiento de los datos

Durante esta fase se caracterizó el conjunto de datos disponibles y se definió cómo serían utilizados en el estudio.

3.2.1. Origen de las imágenes

Las imágenes provienen de cámaras fijas instaladas por el SIATA, ubicadas en siete estaciones del Valle de Aburrá. Cada cámara genera capturas minuto a minuto del mismo punto de observación durante las 24 horas del día, lo que equivale aproximadamente a 525.600 imágenes por año en un ambiente ideal. Procesar esta cantidad de imágenes por cada una de las estaciones con modelos preentrenados se traduce en un alto costo computacional, y en ciertos casos, representa una redundancia de temporalidad al tratarse de imágenes prácticamente idénticas, por tal razón, se optó por trabajar con una muestra reducida pero que fuera significativa, tomando una imagen por cada día, durante aproximadamente dos años por cada una de las estaciones, lo que en la mayoría de los casos representó alrededor de 730 imágenes por estación, para un total aproximado de 4058 imágenes (antes de aplicar filtros de calidad). Cada una de estas imágenes contiene metadatos de fecha y hora, la cual, se trató sin éxito que fuera en el mismo horario, y están organizadas en carpetas por estación, lo que permite mantener coherencia temporal y agrupación por sitio.

3.2.2. Selección de periodo y condiciones de visibilidad

Con el fin de trabajar con escenas comparables, se seleccionó un periodo temporal común entre todas las cámaras donde existían datos disponibles, el cual, no fue posible de obtener en todos los escenarios, teniendo que tomarse entonces imágenes de diversos horarios. Por otro lado, se descartaron aquellas imágenes con condiciones adversas para el análisis visual, como por ejemplo aquellas que contenían niebla densa, lluvia intensa en el paisaje o incluso sobre el lente de la cámara, oscuridad extrema (falta de iluminación o simplemente un horario nocturno). Este filtrado inicial se basó en observación y métricas de calidad que se definen más adelante en la sección 3.3.

3.2.3. Regiones de interés (*ROI* por sus siglas en inglés)

Como las imágenes contienen un campo de visión muy amplio, se definieron regiones de interés (*ROI*) para centrar el análisis en las zonas del terreno donde se consideró más propenso encontrar cambios relevantes ya mencionados (taludes inestables, áreas con vegetación densamente pobladas y susceptibles a remoción, cauces visibles en la imagen). En este caso, el procedimiento seguido consistió en tomar una muestra representativa de imágenes por estación, sobre las que se dibujaron cuadrantes de forma manual alrededor de zonas de interés, posteriormente, con el uso de herramientas disponibles en la web, se realizó un cálculo estimado de las coordenadas aproximadas de las *ROI* (este cálculo se tomó en píxeles y en fracciones de la imagen) y fueron consolidadas en un archivo formato CSV que contenía información de identificación de la estación (carpeta), nombre del archivo (metadata de la imagen) y coordenadas absolutas o fraccionales de la caja delimitadora (*ROI*). Aquí se definieron dos tipos de *ROI*: específicos por imagen, cuando se requería de precisión más fina, pues en algunas estaciones, los tamaños de las imágenes eran diferentes; por estación, cuando la zona de interés era estable en todas las imágenes y el

tamaño era el mismo. Con este enfoque se buscaba reducir principalmente el ruido que podía generar el cielo, nubes, o incluso la misma vegetación del paisaje que no eran consideradas críticas para el análisis Stumpf & Kerle, 2011; Guo et al., 2021; Coluzzi et al., 2025).

3.2.4. Exploración inicial (EDA)

En esta fase se realizó una exploración preliminar de los datos, que consistió en conteo de imágenes disponibles por estación y por año. En total se dispone de 4058 imágenes, correspondientes a siete estaciones de monitoreo, entre los que se encuentran Andalucía, Bermejala, Montañita, Playas, Pueblo Viejo, Raya y Valeria. Las imágenes fueron tomadas desde el año 2022 en la mayoría de las estaciones y hasta el año 2024. Se distribuyen de la siguiente manera (Tabla 1):

Tabla 1. Distribución de imágenes por estación, año inicio y año final.

Estación	Cant. de Imágenes	Año Inicio	Año Final
Andalucía	707	2022	2023
Bermejala	360	2022	2022
Montanita	719	2023	2024
Playas	170	2022	2022
Pueblo Viejo	720	2023	2024
Raya	665	2023	2024
Valeria	717	2023	2024

En este caso, las imágenes no contenían metadata, por ende, tuvo que ser construido con la forma en que se obtuvieron los datos: un folder por estación, nombre de la imagen indicando la fecha y hora y el tamaño en píxeles de cada una de las imágenes, lo que sirvió para detectar que, en una de las estaciones, los tamaños variaban y que iba a dificultar su posterior análisis (*ROI*). Desde el punto de vista espacial, las imágenes presentan homogeneidad en tamaño: la totalidad de las imágenes de casi todas las estaciones es de 1280×720 px, mientras que en la estación Bermejala, a pesar de tener predominancia este tamaño, también tiene una segunda clase minoritaria de tamaño 800×600 px como se muestra en la figura 3. Esta heterogeneidad puntual en Bermejala es la que puede complicar el proceso de aplicación del *ROI*, mientras que la homogeneidad general facilita la extracción de características principales y reduce la necesidad de operaciones de remuestreo que podrían introducir ruido en las series temporales.

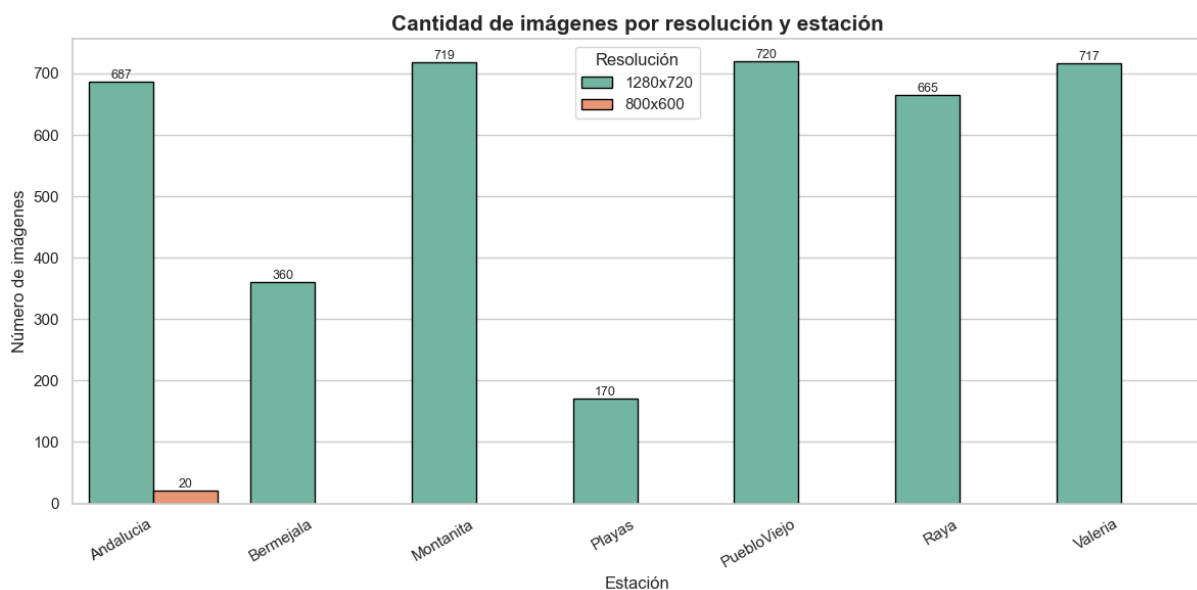


Figura 3. Distribución de imágenes por estación y resolución.

3.2.4.1. Métricas de calidad de imagen

Continuando con la exploración de datos, para cada imagen se calculó el brillo promedio utilizando el canal V (Value) del espacio del color HSV, una métrica robusta que cuantifica la iluminación general. La figura 4 muestra la distribución del brillo por estación, allí se observa que las estaciones presentan valores de iluminación relativamente consistentes, concentrándose entre 110 y 140 unidades, lo que sugiere condiciones de captura relativamente estables en la mayor parte del periodo monitoreado. También se identifican valores atípicos hacia la parte inferior de cada distribución, que correspondería a imágenes con escasa iluminación (amanecer, condiciones nocturnas), así como valores aislados más altos asociados a sobreexposición puntual. Estas observaciones justifican la necesidad de aplicar un filtrado basado en umbrales de brillo para excluir imágenes sin información visual útil.

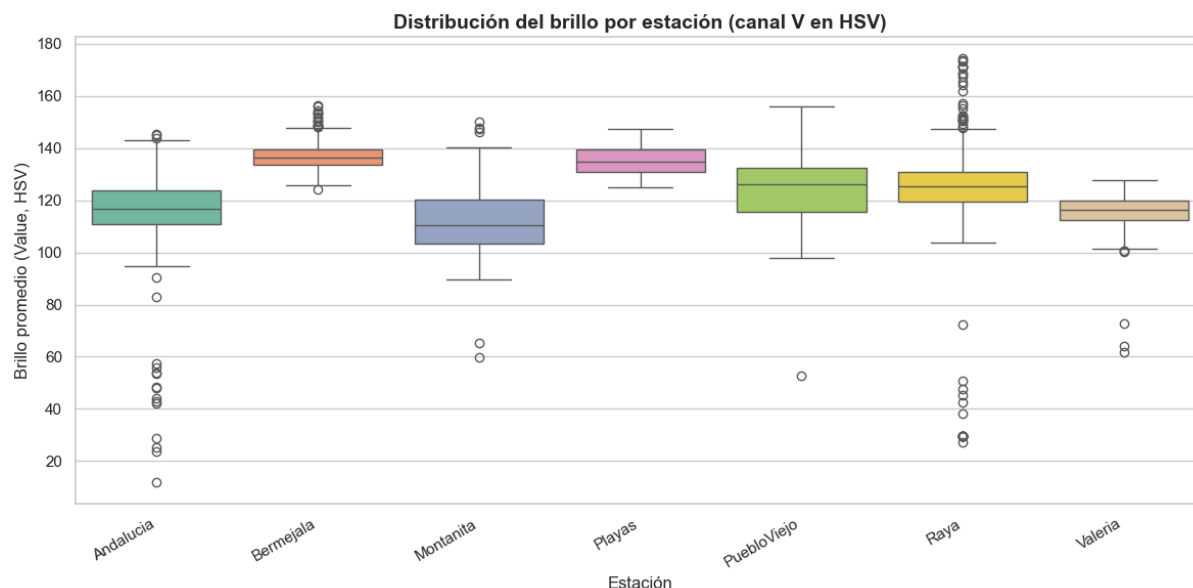


Figura 4. Distribución del brillo por estación (canal V en HSV).

Posteriormente, un análisis de nitidez evaluado mediante la varianza del operador Laplaciano, una métrica ampliamente usada para detectar desenfoque. La figura 5 muestra la distribución de nitidez por estación. En general se observa una variación notable entre cada una de las estaciones, particularmente, Valeria, Raya y Pueblo Viejo presentan las medianas más altas, lo que indica que sus capturas mantienen un nivel de detalle superior de forma consistente. Por otro lado, estaciones como Playas y Bermejala tienen valores de medianas más bajos y rangos comprimidos, sugiriendo una mayor recurrencia de condiciones adversas como humedad, niebla o problemas ópticos de la cámara. Asimismo, todas las estaciones muestran valores atípicos hacia el umbral inferior, lo que corresponde a imágenes con poca nitidez. Estas imágenes pueden hacer correspondencia a gotas en el lente, desenfoque por movimiento, lluvia o condiciones atmosféricas adversas. Su presencia justifica la necesidad de aplicar un filtrado previo para excluir aquellas capturas que no aportan información útil en la representación visual temporal.

A partir del análisis de distribución se definió un umbral de nitidez (Laplacian Variance < 1227, percentil 5) para la identificación automática de imágenes borrosas, lo que constituye parte del proceso de depuración del conjunto de datos antes de la extracción de características mediante modelos de visión por computador.

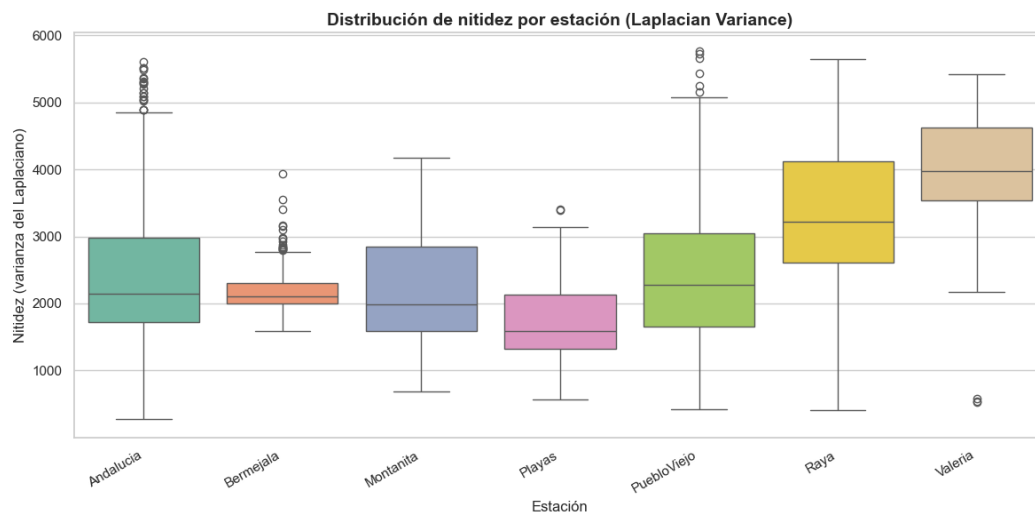


Figura 5. Distribución de la nitidez por estación (Variación Laplaciana)

El contraste de las imágenes se evaluó mediante la desviación estándar de las intensidades en escala de grises. La figura 6 presenta la distribución del contraste para cada estación, mostrando diferencias claras en los sitios monitoreados. Aquí se observa que en estaciones como Andalucía, Raya y Pueblo Viejo presentan las medianas de contraste más altas, es decir, una mayor variabilidad en los tonos, que puede deducirse como una mejor diferenciación entre elementos del terreno, y también la posible presencia de condiciones atmosféricas más favorables o una iluminación más definida entre las capturas.

Por otro lado, en estaciones como Playas y Bermejala, se observan valores más bajos y distribuciones más comprimidas, lo que sería consistente con imágenes afectadas por niebla, nubes densas, humedad o iluminación difusa. Este comportamiento indica que, en estos sitios, pueden aparecer con mayor frecuencia imágenes con detalle visual insuficiente para el análisis estructural del terreno. En este caso, los valores atípicos también están presentes en todas las estaciones en la parte inferior de la distribución, que indica imágenes con contraste muy bajo y que no ofrece información relevante para la extracción de características principales. Nuevamente, este es la justificación que se busca para realizar un filtro basado en un umbral mínimo de contraste, para descartar las imágenes que pueden aportar ruido al modelo.

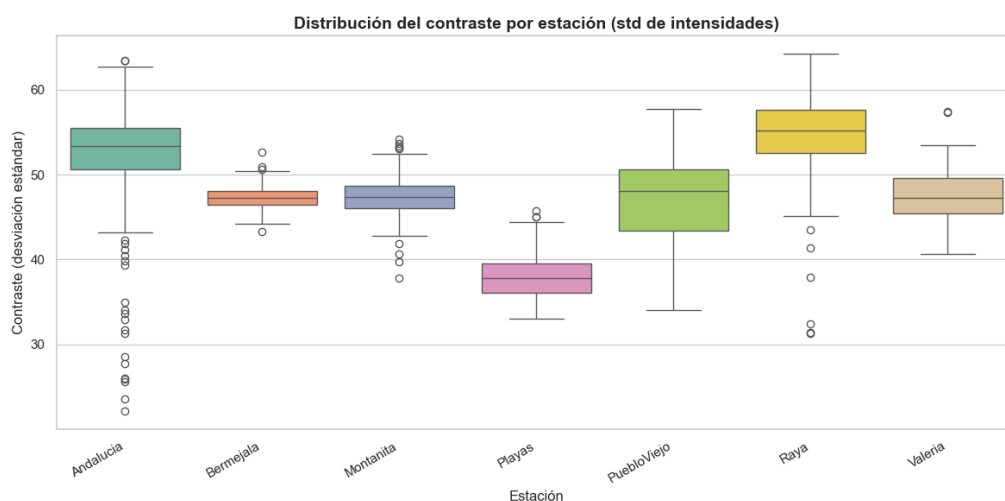


Figura 6. Distribución del contraste por estación (desviación estándar de intensidades)

Los patrones observados en brillo, nitidez y contraste motivan la inclusión de una etapa explícita de filtrado de calidad (sección 3.3), con el fin de asegurar que el modelado posterior se base únicamente en imágenes con información visual suficiente.

3.3 Preparación de los datos

Durante esta etapa se destacan cuatro pasos clave para el análisis de anomalías visuales: 1. Realineación geométrica de las imágenes, para corregir ángulos; 2. Filtrado de calidad de imagen (HQ), donde se descartaron las imágenes con bajo contraste, condiciones climáticas difíciles, entre otras; 3. Aplicación de *ROI* y; 4. Generación y almacenamiento de *embeddings* visuales.

3.3.1. Realineación geométrica mediante ORB + RANSAC

Durante la revisión de las secuencias fotográficas se observó que, en algunos periodos y estaciones, las cámaras presentaban leves cambios de inclinación o ángulo (ocasionadas posiblemente por vibraciones, golpes), que desplazaba el campo de visión y generaba una comparabilidad píxel a píxel desigual.

Con el fin de mitigar este efecto, se implementó en el pipeline un proceso para realinear las imágenes de forma automática: para cada estación se eligió automáticamente una imagen base, una imagen que tuviera mayor nitidez, medida con la varianza de Laplaciano (heurística clásica de autofocus).

Posterior a ello, se detectaron puntos clave y descriptores en la imagen base y en cada imagen a alinear utilizando un detector robusto ORB (Oriented FAST and Rotated BRIEF), método eficiente usado para descripción local de características (Rublee et al., 2011). Se restringió en él la búsqueda de puntos a una máscara interna que excluye parte superior (cielo, logo del SIATA y timestamp), bordes muy cercanos al margen de la imagen.

Luego se emparejaron los descriptores entre la imagen base y la imagen actual mediante un BFMatcher con prueba de razón tipo Lowe (k-NN con ratio test) y chequeo de correspondencia mutua. A partir de los puntos emparejados, se estimó primero una homografía usando RANSAC para rechazar outliers (Fischler & Bolles, 1981); si el número de inliers era suficiente y el resultado geoméricamente estable, se aplicó `cv2.warpPerspective` para alinear cada imagen a la imagen base (Bradski, 2000).

Cuando la homografía no era confiable, se estimó una transformación afín parcial (`estimateAffinePartial2D` con RANSAC) y se aplicó `cv2.warpAffine` como la opción más estable. Para aquellos casos extremos en los que no se lograba una alineación robusta (es decir, pocos puntos válidos, escenas muy diferentes), se conservó la imagen original, pero quedando marcada en los logs como no alineada. Este procedimiento produce, por cada estación, una carpeta de imágenes “aligned” donde la variación y el ángulo se reduce al mínimo, facilitando el recorte *ROI* y análisis multitemporal.

Para ilustrar el efecto de este procedimiento, la figura 7 presenta un ejemplo de una imagen original y su correspondiente versión alineada en la estación Bermejala. Aunque a simple vista las diferencias son casi imperceptibles, al comparar ambas versiones en escala de grises se obtiene un error cuadrático medio (MSE) de aproximadamente 0.46 y un índice de similitud estructural (SSIM) cercano a 0.97. Estos valores indican ligeros desplazamientos geométricos entre capturas consecutivas, y aunque no son evidentes para una persona, sí afectan la comparabilidad píxel a píxel. Esta etapa de realineación permite reducir esas micro variaciones y dejar las imágenes en condiciones más adecuadas para el análisis multitemporal posterior.



Figura 7. Ejemplo de una imagen original y su versión realineada mediante ORB + RANSAC

3.3.2. Filtros de calidad de imagen (HQ)

Una vez las imágenes están alineadas, se someten a un filtro de calidad automática con el objetivo de descartar aquellas capturas que, por condiciones ambientales adversas o problemas de la cámara, no resultan informativas

para el análisis del terreno. Este filtrado está basado en métricas clásicas de calidad de imagen (Bovik, 2013; Hore & Ziou, 2010; Krotkov, 1988):

- Brillo promedio (canal V en espacio HSV): se calcula el valor medio del canal V (Value) tras convertir la imagen de BGR a HSV. Para aquellas imágenes con V por debajo de un umbral (por ejemplo 60) se consideran muy oscuras (noche, sombras extremas) y posteriormente se descartan (Bovik, 2013).
- Contraste (desviación estándar en escala de grises): a partir de la imagen en niveles de gris se calcula la desviación estándar. Valores por debajo de un umbral (por ejemplo 18), indican un bajo contraste y poca diferenciación en detalles (Hore & Ziou, 2010).
- Nitidez (Varianza del operador Laplaciano): se calcula la varianza del Laplaciano a partir de la imagen en escala de grises, esta es una métrica ampliamente utilizada como medida de enfoque en visión por computador (Krotkov, 1988; Pech-Pacheco et al., 2000). Aquellas imágenes con valores por debajo de un umbral (por ejemplo 1200, o que pertenezcan al percentil 5 en nuestro caso), se consideran desenfocadas, borrosas o con pérdida significativa de detalle, asociadas a gotas en el lente, lluvia o condiciones atmosféricas adversas.
- Porcentaje de saturación/subexposición: se calcula el porcentaje de píxeles casi blancos (≥ 245) y casi negros (≤ 10). Si el porcentaje de píxeles muy oscuros supera el 20%, se considera una imagen sobreexpuesta (Bovik, 2013).

Estas métricas se integran en una única evaluación que determina si cada imagen aporta información útil para el análisis. La etiqueta resultante clasifica cada captura como “aceptable” cuando una imagen cumple todos los umbrales o “excluida”, indicando la razón principal. Este enfoque se alinea con la literatura de evaluación de calidad de imagen y enfoque automático, donde combinan medidas de brillo, contraste y gradiente para descartar capturas poco informativas (Bovik, 2013; Hore & Ziou, 2010; Krotkov, 1988).

En total, se descartó aproximadamente el 12% de las imágenes ya alineadas. Las causas predominantes de descarte fueron baja saturación (334 imágenes) y bajo contraste (121 imágenes), seguidas por casos asociados a neblina ligera y presencia de gotas o reflejos en el lente detectados mediante umbrales bajos en nitidez y contraste, mientras que las imágenes completamente oscuras que fueron poco frecuentes, esto es consistente con la naturaleza diurna del monitoreo y con la presencia recurrente de condiciones atmosféricas ligeras (niebla, nubosidad, humedad) en las estaciones analizadas. Algunas de las imágenes descartadas se muestran en la figura 8.



Figura 8. Imágenes descartadas luego de aplicar el filtro de calidad (HQ)

3.3.3. Aplicación de Regiones de Interés (ROI)

Una vez realizado el filtrado de calidad, se aplicaron las Regiones de Interés (ROI) definidas en el archivo de configuración, con el propósito de centrar el análisis en las zonas del terreno donde se esperaba encontrar cambios

visuales relevantes. Estas regiones se especificaron tanto a nivel de estación como de imagen usando coordenadas absolutas o proporcionales:

1. *ROI* específico por imagen, cuando se requería de mayor precisión.
2. *ROI* por estación basado en proporciones, usado cuando las imágenes tenían tamaño diferente.
3. *ROI* por estación basado en coordenadas absolutas, para casos donde existía homogeneidad de resolución.

La asignación siguió una prioridad simple: primero el *ROI* específico de la imagen (si existía), luego el *ROI* fraccional por estación y finalmente el *ROI* absoluto. Cuando una imagen no tenía *ROI* disponible, se optó por omitirla en el análisis, con el fin de evitar que regiones irrelevantes (como cielo, nubes, zonas exteriores al talud) introdujeran ruido en el cálculo de características.

El recorte se implementó mediante funciones que aseguran que las coordenadas estén dentro de los límites de la imagen y devuelven únicamente el subarreglo correspondiente a la zona de interés. De este modo, la información procesada por los modelos preentrenados se limitan exclusivamente a los sectores del terreno más relevantes desde el punto de vista geotécnico (Stumpf & Kerle, 2011; Ghorbanzadeh et al., 2019; Coluzzi et al., 2025). La figura 9 muestra una imagen sometida al recorte por región de interés de la estación Bermejala, correspondiente a 12 de marzo de 2022.

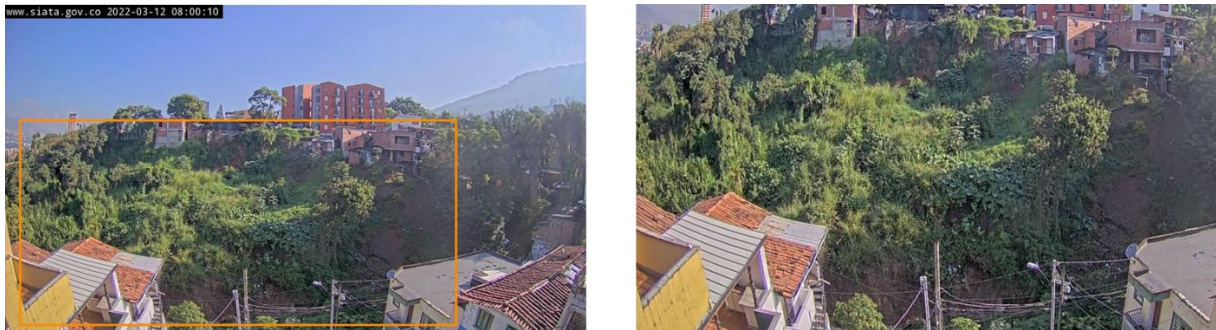


Figura 9. Imágenes estación la Bermejala de 2022-03-12 previo al recorte y posterior al recorte de *ROI*.

3.3.4. Generación y almacenamiento de *embeddings* visuales

Finalmente, las imágenes que se recortaron y pasaron el filtro de calidad se procesaron mediante cuatro modelos preentrenados, seleccionados por su capacidad para detectar cambios precisos en la superficie, y que actúan como extractores de características (Xie et al., 2017; Oquab et al., 2023; Radford et al., 2021):

- ResNeXt-101 64x4D: arquitectura basada en agregación de transformaciones residuales, propuesta para mejorar la capacidad de representación incrementando la cardinalidad del modelo. Implementada a partir de la versión disponible en torchvision/timm, entrenada previamente en ImageNet-1K. Se utilizó la salida de la penúltima capa (antes de la capa de clasificación) como un vector de características globales de dimensión 2048 (Xie et al., 2017).
- DINOv2 ViT-L/14: Modelo de auto-supervisión basado en Vision Transformers, entrenado sobre grandes colecciones de imágenes no etiquetadas para aprender representaciones visuales generalistas y robustas. A partir del token de clase (CLS) se obtiene un embedding global de dimensión 1024 para cada imagen (Oquab et al., 2023).
- CLIP ViT-L/14@336px: Modelo contrastivo multimodal que alinea representaciones de imágenes y texto en un espacio común, entrenado maximizando la similitud entre pares imagen-texto. Para este trabajo se utilizó únicamente el encoder visual, extrayendo para cada imagen un vector global de dimensión 768 (Radford et al., 2021).
- CLIP ViT-H/14: variante de mayor capacidad del modelo CLIP, también basada en Vision Transformers, que incrementa la profundidad y el ancho del modelo para capturar patrones visuales más complejos. De forma análoga, se empleó únicamente el encoder visual para generar un embedding global por imagen, con dimensión 1024 (Radford et al., 2021)

Cada imagen del conjunto de datos limpio se procesa por los cuatro modelos, generando así cuatro *embeddings* por imagen:

- $\text{resnext_emb} \in \mathbb{R}^{2048}$
- $\text{dinov2_emb} \in \mathbb{R}^{1024}$
- $\text{clip_114_336_emb} \in \mathbb{R}^{768}$

- $\text{clip_h14_emb} \in \mathbb{R}^{1024}$

Estas representaciones se almacenan conservando información asociada a cada imagen (estación, fecha, nombre de archivo, ruta de la imagen recortada, métricas de calidad utilizadas en el filtrado). Esto permite reutilizar las representaciones numéricas sin necesidad de reprocesar nuevamente las imágenes originales en etapas posteriores.

El resultado de la fase de preparación es, por tanto, un conjunto de datos estructurado de *embeddings* multitemporales, listo para calcular distancias y métricas de cambio entre imágenes (fase de modelado, sección 3.4), integrarse con etiquetas manuales y anotaciones de expertos (fase de evaluación, sección 3.5), alimentar un flujo automatizado de generación de alertas visuales (fase de despliegue, sección 3.6).

3.4 Modelado

En la etapa de modelado, el objetivo fue detectar anomalías visuales en las secuencias temporales de imágenes a partir de los *embeddings* generados por los 4 modelos de visión por computador descritos en la sección 3.3.4. Aunque se cuenta con imágenes etiquetadas manualmente para la fase de evaluación, no existe un volumen suficiente, representativo y diverso de anomalías reales que permita entrenar un modelo supervisado, debido a que los eventos de deslizamiento o inestabilidad son escasos, heterogéneos y en muchos casos, no presentan un patrón visual repetible-

Por esta razón, siguiendo múltiples trabajos en detección de anomalías visuales, se adopta un enfoque usando modelos no supervisados, donde la referencia principal es el comportamiento normal del terreno. La premisa es que los deslizamientos o cualquier otro fenómeno de inestabilidad producirán cambios inusuales en el espacio de características frente al patrón dominante en cada estación. Este planteamiento es consistente con la definición general de anomalía como una observación que se desvía considerablemente de la noción de normalidad y con la recomendación de modelar explícitamente la distribución de patrones normales para luego hallar desviaciones respecto a ella (Ruff et al., 2021).

A diferencia de estudios tradicionales, donde se entrenan modelos supervisados (SVM, árboles de decisión o redes neuronales), el modelo propuesto opera en tres capas:

1. Cálculo de distancias en el espacio de *embeddings* por estación.
2. Normalización robusta y fusión multimodelo.
3. Selección de posibles eventos anómalos mediante umbrales basados en cuantiles por estación.

3.4.1. Distancias temporales basadas en la similitud coseno

Sea $\mathbf{e}_{t,m} \in \mathbb{R}^d$ el *embedding* de la imagen en el instante t para el modelo $m \in \{\text{ResNeXt-101}, \text{DINOv2}, \text{CLIP-L/14-336}, \text{CLIP-H/14}\}$, luego de aplicar alineación, filtros de calidad y *ROI*. Para cada estación (folder) y modelo se ordenan las imágenes por timestamp y se construyen dos métricas:

1. Distancia coseno a la imagen anterior (cambio abrupto entre $t-1$ y t): para cada par consecutivo $(t-1, t)$:

$$d_{\text{prev}}^{(m)}(t) = 1 - \frac{\mathbf{e}_{t-1,m} \cdot \mathbf{e}_{t,m}}{\|\mathbf{e}_{t-1,m}\|_2 \|\mathbf{e}_{t,m}\|_2}.$$

La primera imagen de cada secuencia no tiene predecesora y se asigna como N. Esta métrica captura cambios bruscos entre imágenes consecutivas (por ejemplo, deslizamientos repentinos, aparición de barro, pérdida súbita de vegetación).

2. Distancia coseno a un baseline local (cambio respecto a un comportamiento promedio): Para cada instante t , se define un *baseline* temporal:

$$b_{t,m} = \frac{1}{|W_t|} \sum_{k \in W_t} \mathbf{e}_{k,m},$$

donde W_t es una ventana centrada en t de tamaño $\pm \text{WINDOW_BASELINE} = \pm 10$ imágenes, excluyendo la propia imagen.

La distancia respecto a ese baseline es:

$$d_{\text{base}}^{(m)}(t) = 1 - \frac{e_{t,m} \cdot b_{t,m}}{\|e_{t,m}\|_2 \|b_{t,m}\|_2}.$$

Esta métrica cuantifica el grado de desviación de una imagen respecto al comportamiento típico de su entorno temporal inmediato.

El uso de distancias en el espacio de *embeddings* sigue la línea de métodos recientes que detectan anomalías midiendo la discrepancia entre features de redes preentrenadas y una distribución de referencia. Por ejemplo, Schulthess y Konukoglu (2025) ajustan una mezcla de Gaussianas sobre *embeddings* de DINOv2 y definen el score de anomalía en función de la distancia coseno al centro de componentes normales.

En este trabajo, a diferencia de propuestas que comparan varias métricas (likelihood, distancia euclidiana, similitud coseno) no existe un consenso universal sobre una métrica superior a las demás. En nuestro caso, se optó por utilizar únicamente la distancia coseno, apoyándose en su buena estabilidad numérica en espacios de alta dimensión, su invariancia al escalado de los *embeddings*, su desempeño en tareas de “*change-detection*” con modelos tipo ViT y porque permite mantener el flujo de trabajo más interpretable y homogéneo entre modelos.

3.4.2. Normalización por estación y scores tipo z-score robusto

Las distribuciones de la métrica $d_{\text{base}}^{(m)}(t)$ (distancia al baseline local definida en la sección anterior) pueden variar significativamente entre estaciones debido a diferencias geométricas, de textura o de condiciones locales de captura. Para hacer comparables los valores dentro de cada estación y modelo, se aplica una normalización tipo *z-score* calculada de forma independiente para cada estación.

Sea $d_{\text{base}}^{(m)}(t)$ la distancia al baseline local para el modelo m en el instante t . Se calcula:

$$\mu_m = \frac{1}{N} \sum_{t=1}^N d_{\text{base}}^{(m)}(t),$$

$$\sigma_m = \sqrt{\frac{1}{N} \sum_{t=1}^N (d_{\text{base}}^{(m)}(t) - \mu_m)^2},$$

y el score estandarizado:

$$z_{\text{base}}^{(m)}(t) = \frac{d_{\text{base}}^{(m)}(t) - \mu_m}{\sigma_m}.$$

Este valor expresa cuántas desviaciones estándar se aleja una imagen del patrón promedio de su estación para un modelo específico. Esta normalización reduce la variabilidad intra-normal atribuible a factores como iluminación o textura, evitando que estas variaciones dominen los valores de distancia e impidan identificar anomalías relevantes (Ruff et al., 2021).

3.4.3. Fusión multimodelo mediante rankings robustos

Con el fin de explotar la complementariedad de los cuatro modelos (ResNeXt, DINOv2, CLIP-L, CLIP-H), e construyó un esquema de fusión multimodelo basado en dos tipos de información por imagen:

- i) los puntajes estandarizados $z_{\text{base}}^{(m)}(t)$ de cada modelo, y
- ii) la posición relativa (*rank*) de cada imagen dentro de su estación según dichos puntajes.

a) Integración de los puntajes por modelo.

Para cada imagen se agruparon los cuatro valores $z_{\text{base}}^{(m)}(t)$, uno por cada modelo m . Esta consolidación permite disponer, en un mismo vector, de la estimación multifuente de qué tan atípica resulta una imagen respecto al comportamiento normal de su estación. (Obs.: la distancia previa, d_{prev} , también fue calculada, pero en esta etapa no se incluyó en la fusión porque captura un fenómeno distinto (cambios abruptos entre imágenes consecutivas, que se analiza por separado).

b) Construcción de rankings locales.

Para cada estación y modelo, las imágenes se ordenaron de menor a mayor según su valor $z_{\text{base}}^{(m)}(t)$. El *rank* obtenido para cada imagen refleja su posición relativa dentro de la distribución de su propia estación: imágenes muy anómalas quedan consistentemente en los primeros lugares del ranking independientemente de su escala numérica.

Este *rank* no hace referencia a la posición temporal, sino a la posición relativa dentro de la distribución estadística del modelo.

c) Normalización robusta y fusión de scores.

Dado que los rangos y dispersión de los *rank*s pueden diferir entre modelos, se aplicó una normalización robusta basada en la mediana y la desviación absoluta mediana (*MAD*). Esto reduce el efecto de valores extremos y homogeniza la escala entre modelos. Con estas versiones normalizadas se definieron dos puntajes fusionados:

- $z_{\text{fusion_mean}}$: promedio de los cuatro puntajes $z_{\text{base}}^{(m)}$. Representa el nivel de consenso entre modelos en términos de desviación respecto al baseline.
- $z_{\text{fusion_rmean}}$: promedio robusto de los *rank*s normalizados. Resume la coincidencia entre modelos en cuanto a la posición relativa de una imagen dentro de su estación.

La estrategia de combinar múltiples extractores de características está respaldada por estudios recientes que muestran que la fusión de descriptores provenientes de redes preentrenadas puede aumentar la sensibilidad a cambios sutiles en comparación con el uso de un solo modelo (Bergmann et al., 2021). En el análisis posterior se utilizó principalmente $z_{\text{fusion_mean}}$, dejando $z_{\text{fusion_rmean}}$ como alternativa para análisis de sensibilidad.

3.4.4. Selección de eventos candidatos mediante umbrales por cuantiles

A partir del conjunto de puntajes fusionados, se identifican las imágenes con mayor probabilidad de corresponder a eventos anómalos mediante un umbral por cuantiles aplicado de forma independiente a cada estación. Para cada estación se calcula el umbral τ como el percentil 99.5 de la distribución de los *scores* previamente calculados y aquellas imágenes con $\text{score} \geq \tau$ se marcan como posibles eventos anómalos y se almacenan. Este enfoque es consistente con prácticas en detección de anomalías no supervisada, donde una vez definido un score continuo de anomalía, la obtención de predicciones binarias se realiza mediante selección de umbrales sobre cuantiles, tasas de falso positivo o métricas relacionadas (Bergmann et al., 2021).

En términos de hiperparámetros, el modelo se caracteriza por tener una ventana de baseline temporal ($\text{window_baseline}=10$); el uso de $z_{\text{fusion_mean}}$ como *score* principal y; cuantil de selección de eventos = 0.995. Estos valores se seleccionaron buscando una relación entre sensibilidad a cambios relevantes (picos altos en el score) y un número manejable de posibles eventos anómalos para revisión experta y posterior anotación manual. No obstante, es importante mencionar que esta estimación de umbrales se realizó utilizando la distribución completa del conjunto de datos, lo que es válido para un análisis retrospectivo.

En un posterior escenario de despliegue en tiempo real, la información disponible para calcular umbrales y *baselines* sería únicamente la correspondiente al historial previo del sistema, es decir, a los frames anteriores al instante actual, por lo que dichos umbrales deberían recalibrarse de forma incremental o mediante enfoques de actualización continua basado en estadísticas móviles.

3.4.5. Conexión con la etapa de evaluación

Los scores fusionados y la lista de eventos candidatos sirven como entrada para la fase de evaluación (Sección 3.5), donde se construye un subconjunto etiquetado mediante la combinación de:

- Imágenes con scores muy altos por modelo y estación (potenciales anomalías), y
- Un conjunto de imágenes normales seleccionadas aleatoriamente en la cola inferior de la distribución de scores.

Ese diseño de muestra estratificada permite cuantificar, con apoyo de expertos, qué proporción de los picos de score corresponden efectivamente a cambios relevantes en el terreno y qué fracción se asocia a ruido residual (condiciones ambientales, errores residuales de alineación, etc.), alineándose con la idea general de que la detección de anomalías se valida comparando el modelo con anotaciones humanas o señales externas de referencia, tal como discuten Ruff et al. (2021) y Bergmann et al. (2021) en el contexto más amplio de métodos no supervisados.

3.5 Evaluación

La evaluación del modelo se realizó siguiendo un esquema que fuera coherente con el problema de desbalanceo de clases (pocas anomalías frente a gran cantidad de imágenes normales), tal como se recomienda en la literatura reciente de detección de anomalías (Rachwal et al., 2025). En esta sección se describe entonces la estrategia del muestreo, el proceso de etiquetado manual y cada una de las métricas empleadas, mientras que los valores numéricos, tablas comparativas y figuras se presentarán en la sección de resultados.

3.5.1. Construcción de la muestra para evaluación

El punto de partida para la evaluación fue el conjunto de datos que contiene, para cada imagen que superó los filtros de calidad de la sección 3.3.2, los puntajes de anomalía por modelo y las puntuaciones fusionadas (z_{fusion_mean} y z_{fusion_rmean}). Teniendo este conjunto se construyó una muestra estratificada por estación siguiendo estos pasos:

1. Cálculo de z-scores robustos por modelo y estación: para cada modelo (CLIP ViT-H/14, CLIP ViT-L/14@336, DINOv2 ViT-L/14 y ResNeXt-101 64x4D) se calculó, en cada estación, un z-score robusto z_m utilizando la mediana y la desviación absoluta mediana (MAD) sobre el puntaje de anomalía del modelo. Este enfoque reduce la sensibilidad a valores extremos, tal como recomiendan estudios de detección de anomalías y series temporales donde la presencia de outliers severos es esperada (Rachwal et al., 2025).
2. Score combinado del “peor caso” por imagen: para cada imagen se definió

$$z_{any} = \max_m z_m,$$

de modo que una imagen se considere posiblemente como anómala si algún modelo la percibe como muy atípica.

3. Muestreo de imágenes normales: para cada estación se ordenaron las imágenes por z_{any} de menor a mayor (los valores más “normales” primero) y se seleccionaron hasta 80 imágenes al azar de esa cola inferior. Estas muestras funcionan como muestra de referencia para el comportamiento habitual.
4. Muestreo de imágenes anómalas por modelo: en cada estación se seleccionaron las 5 imágenes con mayor z-score z_m (top-K por modelo), etiquetadas como anomalías. Este esquema garantiza que cada modelo aporte un conjunto de posibles anomalías, lo que permite evaluar su capacidad de priorizar imágenes verdaderamente relevantes.
5. Muestra final por estación: las imágenes “normales” y “anómalas” se concatenaron para obtener una muestra de 100 imágenes por estación (80 normales + 4x5 anómalas), utilizada posteriormente en la fase de etiquetado manual.

La intención final de este diseño de muestreo busca cubrir dos objetivos: representar de forma adecuada el comportamiento normal de cada estación y; enfocar la revisión humana en la cola extrema de los modelos, donde se esperaría se hayan ubicado las imágenes “más anómalas”, en línea con recientes enfoques de detección de anomalías, donde se enfatiza en la inspección de casos de mayor score (Bergmann et al., 2019).

3.5.2. Protocolo de etiquetado manual y depuración de etiquetas

La muestra anterior fue sometida a un etiquetado manual, en la que se inspeccionó cada imagen (ya alineada y recortada al *ROI*) y se asignó una etiqueta binaria: “Y” como presencia de un cambio visual potencialmente relevante para análisis geotécnico (por ejemplo, deslizamientos, pérdida de vegetación, cambios morfológicos); “N” ausencia de dichos cambios relevantes.

3.5.3. Evaluación del esquema de candidatos extremos por modelo

En una etapa inicial se analizó, para cada modelo, qué tan bien se seleccionan las imágenes anómalas con la etiqueta manual binaria. Para ello, sobre la muestra etiquetada:

Se definió

$$y_{true} \in \{0,1\}$$

donde 0 corresponde a imágenes sin cambios relevantes (N) y 1 a imágenes con presencia de cambios relevantes (Y), según la etiqueta manual. Luego, para cada modelo m , se generó un vector de predicciones binarias y_{pred} siguiendo la convención utilizada en esta evaluación: una imagen se marca como positiva (1) únicamente si proviene del conjunto de candidatos extremos seleccionados por ese modelo en la etapa previa; en caso contrario se marca como negativa (0).

Con estos vectores y_{true} y y_{pred} se construyó la matriz de confusión para cada modelo y se calcularon las métricas clásicas, incluyendo: Precisión, *Recall*, F1-score y *Accuracy*. Estas métricas, ampliamente utilizadas en la evaluación de sistemas de detección de anomalías y clasificación binaria bajo desbalance de clases (Rachwal et al., 2025), permiten cuantificar distintos aspectos del desempeño: la pureza de la lista de candidatos (*precision*); la capacidad de recuperar verdaderas anomalías (*recall*) y; el compromiso global entre ambas (F1-score). Los valores numéricos por modelo se presentan en una tabla resumen y mediante matrices de confusión gráficas en el capítulo de resultados.

3.5.4. Evaluación basada en distancias en el espacio de *embeddings*

En una segunda etapa se evaluó cada modelo de manera más directa a partir de sus *embeddings* completos, con el fin de analizar su capacidad para separar imágenes normales y anómalas, a partir de medidas de distancia. Esta evaluación se realizó únicamente sobre las imágenes que contaban con etiqueta manual.

1. Normalización y distancia al centroide de normales: para cada modelo y por cada estación se estandarizaron los *embeddings* para eliminar efectos de escala entre dimensiones. Posteriormente, se calculó un centroide de referencia μ_f usando únicamente las imágenes etiquetadas como normales, siempre que hubiera al menos 5; el centroide representa el comportamiento visual normal del terreno. Para cada imagen se calculó la distancia euclidiana al centroide:

$$d_i = \|x_i - \mu_f\|_2.$$

de manera que valores altos de d_i indican desviaciones significativas respecto al patrón normal.

2. Umbralización por cuantiles: para evaluar qué tan bien separa ambas clases cada modelo, se aplicó un esquema de umbralización basado en cuantiles. Para un cuantil q (por ejemplo, $q = 0,90$), se estimó para cada estación un umbral τ_q definido como:

$$\tau_q = Q_q(\{d_i: y_{\text{true}} = 0\}).$$

Una imagen se considera predicha como anómala cuando:

$$y_{\text{pred}} = \begin{cases} 1, & \text{si } d_i \geq \tau_q, \\ 0, & \text{en otro caso.} \end{cases}$$

Este esquema permite evaluar el desempeño del modelo bajo diferentes niveles de sensibilidad, siguiendo prácticas habituales en detección de anomalías donde los umbrales se controlan explícitamente a través de cuantiles (Bergmann et al., 2019).

3. Métricas de evaluación: Con los vectores y_{true} y y_{pred} se calcularon nuevamente precisión, *recall*, F1-score y exactitud. Adicionalmente, se utilizaron: i) Área bajo la curva ROC (AUC-ROC), calculada a partir de las distancias d_i como scores continuos. Esta métrica resume la capacidad del modelo para separar normales y anómalas en todo el rango de umbrales (Rachwal et al., 2025); ii) Lift@10: se ordenaron las imágenes por d_i de mayor a menor y se calculó el porcentaje de anomalías reales en el 10 % superior de la lista, dividido entre la tasa global de anomalías, como se define en los gráficos Lift usados en minería de datos (University of Regina, s.f.). iii) Ganancia acumulada al 10 % (Gain@10): proporción de todas las anomalías verdaderas que se capturan dentro del 10 % superior de la lista ordenada por score (H2O.ai, s.f.).

Los resultados para cada modelo (CLIP-H/224, CLIP-L/336, DINOv2 ViT-L/14 y ResNeXt-101 64x4D) se resumen en tablas con F1, AUC-ROC, Lift@10 y Gain@10, que se presentan en la sección de resultados.

3.5.5. Análisis de sensibilidad al cuantil de umbralización

Dado que la elección del cuantil q tiene un impacto directo en el equilibrio entre *recall* (cantidad de anomalías detectadas) y precisión (proporción de verdaderos positivos entre los detectados), se realizó un análisis de sensibilidad sistemático:

- Se recalcularon, para cada modelo, las predicciones y_{pred} usando cuantiles $q \in \{0,80, 0,85, 0,90, 0,95\}$ definidos por estación.
- Para cada par (modelo, q) se obtuvieron la matriz de confusión y las métricas *Accuracy*, precisión, *recall*, F1, AUC-ROC y Lift@10.
- Estas métricas se representaron gráficamente como función de q , permitiendo visualizar la evolución de F1, *recall*, AUC-ROC y Lift@10 y seleccionar rangos de cuantiles que representen el mejor compromiso entre sensibilidad y robustez a falsos positivos, tal como se recomienda en problemas de detección de anomalías con fuerte desbalance y alta variabilidad (Rachwal et al., 2025; Mehta et al., 2024).

Nuevamente, los valores concretos y las curvas de desempeño por cuantil se presentan en el capítulo de resultados; aquí se documenta únicamente el protocolo de evaluación seguido.

3.6 Despliegue

El despliegue de la solución se concibe como un flujo automatizado que pueda integrarse en los sistemas de información del SIATA, en donde el objetivo principal es que, a partir de las nuevas imágenes capturadas periódicamente por las cámaras fijas, el sistema se pueda actualizar de forma continua con los indicadores de cambio y generar alertas visuales cuando se detecten patrones anómalos en el terreno. El flujo del pipeline se propone de la siguiente manera, como se observa en la figura 10:

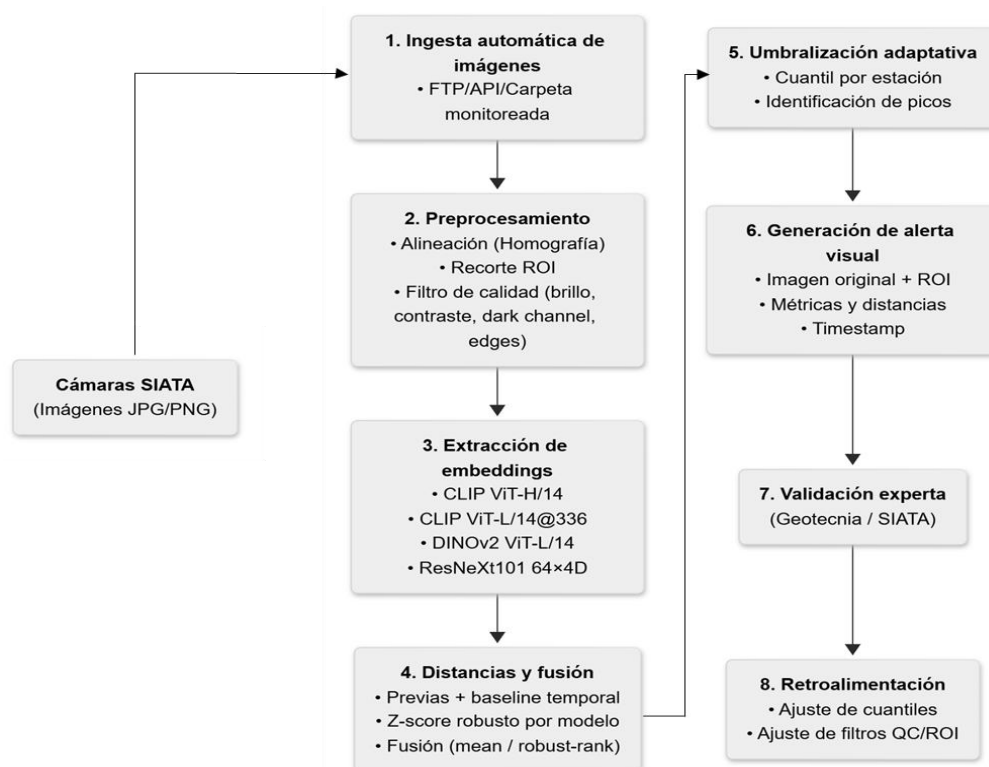


Figura 10. Diagrama de solución propuesto para la implementación futura por parte del SIATA. Este flujo propuesto es flexible y permite una integración futura en módulos existentes en el SIATA dedicados al monitoreo geotécnico. La solución es escalable a nuevas estaciones o regiones, requiriendo únicamente ajustar los umbrales estadísticos según la variabilidad local de cada cámara.

4 Resultados

En esta etapa se presentan los resultados obtenidos al aplicar los modelos preentrenados de detección de anomalías a las imágenes procesadas, siguiendo los pasos descritos previamente en la metodología (Sección 3). El objetivo principal es comparar el desempeño de cada una de las arquitecturas de *embeddings* (CLIP ViT-H/14, CLIP ViT-L/14@336, DINOv2 ViT-L/14 y ResNeXt101 64×4d), bajo dos esquemas de predicción que fueron introducidos en la sección 3.5:

1. Evaluación por selección directa según el modelo de origen, en la que cada arquitectura aporta sus casos más extremos según su propio score (Sección 3.5.3).
2. Evaluación basada en la distancia en el espacio de *embeddings* y umbralización de cuantiles, que examina qué tan bien cada modelo separa imágenes normales y anómalas en función de su distancia al centroide de referencia (Sección 3.5.4).

Ambos esquemas de evaluación fueron definidos previamente y se emplean aquí para analizar la capacidad de cada modelo para identificar anomalías visuales bajo distintos criterios.

Dentro de los resultados se encuentran: matrices de confusión, métricas agregadas, curvas (*Recall/F1/AUC/Lift*), y ejemplos cualitativos de las imágenes que fueron correctamente detectadas y etiquetadas como verdaderos positivos, falsos positivos y falsos negativos. Todos los criterios de evaluación fueron descritos previamente en la sección 3.5; aquí se presentan exclusivamente los hallazgos detectados.

4.1 Resultados selección directa por modelo

Esta fase inicial evalúa si cada modelo, por sí solo, tiene la capacidad de priorizar imágenes consideradas anomalías por la revisión humana durante el muestreo inicial. Este procedimiento corresponde al enfoque denominado selección directa por modelo, que consiste en analizar el desempeño de cada arquitectura únicamente sobre las imágenes que ella misma seleccionó como sus casos más extremos durante la etapa de muestreo (es decir, su *top-K*). La regla que aquí se aplicó es:

$$y_{\text{pred}} = \begin{cases} 1 & \text{si la imagen proviene del modelo } m, \\ 0 & \text{en otro caso.} \end{cases}$$

donde m representa una de las cuatro arquitecturas evaluadas (ResNeXt-101, DINOv2, CLIP-L/336 o CLIP-H/224). En otras palabras, la idea es evaluar qué tan bien funcionó la intuición del propio modelo al escoger su top-K , y evaluar si coinciden con las etiquetas humanas.

4.1.1 Métricas globales por modelo

Para evaluar de forma preliminar la capacidad de cada uno de los modelos para seleccionar imágenes anómalas, se construyeron matrices de confusión aplicando el criterio de selección directa, tomando como predicción positiva aquellas imágenes seleccionadas por el modelo en el muestreo inicial ($\text{modelo_origen} = m$). Las etiquetas manuales binarias (0 o 1) se utilizan como referencia de verdad. La tabla 2 muestra las métricas de evaluación globales, las matrices comparativas se muestran en el Anexo (A).

Tabla 2. Métricas por modelo bajo la regla $y_{\text{pred}} = (\text{modelo_origen} = m)$.

Modelo	TP	FP	TN	FN	Accuracy	Precision	Recall	F1
clip_vit_l14_336_clean	24	11	567	98	0.8443	0.6857	0.1967	0.3057
dino_vitl14_clean	24	11	567	98	0.8443	0.6857	0.1967	0.3057
clip_vit_h14_clean	22	13	565	100	0.8386	0.6286	0.1803	0.2803
resnext101_64x4d_clean	17	18	560	105	0.8243	0.4857	0.1393	0.2166

En términos de F1 score, los mejores resultados se obtienen con el modelo CLIP_ViT-L/14@336 y DINOv2 ViT-L/14, ambos con un F1 $\approx 0,31$, precisión cercana a 0,69 y un *recall* de 0,20. Esto indica que cuando estos modelos marcan una imagen como anómala, alrededor de 0.68 de los casos corresponden efectivamente a cambios relevantes según el etiquetado manual, pero solo recuperan aproximadamente una de cada cinco anomalías presentes en la muestra.

El modelo CLIP ViT_H/14 por su parte, presenta un comportamiento similar, con una ligera disminución tanto en precisión ($\approx 0,63$) como en *recall* ($\approx 0,18$), lo que se traduce en un F1 alrededor de ($\approx 0,28$). Por su parte, ResNeXt-101 64×4D es el modelo con menor desempeño en este esquema de selección, con precisión $\approx 0,49$, *recall* $\approx 0,14$ y F1 $\approx 0,22$, lo que indica que tiende a proponer más falsos positivos y a dejar pasar una mayor proporción de anomalías reales.

En todos los casos la exactitud global (*Accuracy*) se mantiene alta (entre 0.82 y 0.84), lo que es esperable por el desbalanceo de clases (la mayoría de las imágenes de la muestra son normales). De esta forma, estas métricas se pueden interpretar más como un diagnóstico inicial del criterio de selección de posibles anomalías por cada modelo, que como una evaluación definitiva del sistema, la cual se realiza posteriormente en el espacio de *embeddings* utilizando distancias y umbrales basados en cuantiles.

4.2 Resultados Distancias en *embeddings* + umbralización por cuantiles

Esta sección presenta los resultados del segundo esquema de predicción, distinto al enfoque previo de selección directa por modelo. Aquí cada modelo se evalúa en función a sus distancias en el espacio de *embeddings*, aplicando una umbralización por cuantiles. A diferencia del método anterior, que solo examinaba qué tan bien funcionaba la intuición del modelo al seleccionar su propio top-K durante el muestreo, este enfoque utiliza: distancias al centroide de imágenes normales, normalización por estación y umbrales basados en cuantiles, lo que permite medir la separación real entre patrones normales y anómalos en todo el conjunto de *embeddings*.

Este esquema muestra un desempeño más estable, pues reduce la variabilidad entre estaciones y captura mejor los cambios sutiles que no necesariamente aparecen entre las imágenes más extremas del top-K inicial.

Los valores obtenidos por modelo se resumen en la Tabla 3. En general, el modelo DINOv2 ViT-L/14 suele alcanzar los mejores resultados (ver figura 11): alcanza precisiones cercanas a 0,51, *recall* alrededor de 0,47 y un F1 $\approx 0,49$, junto con la mayor AUC-ROC ($\approx 0,78$) (ver figura 12) y los mayores valores de Lift@10 y Gain@10 ($\approx 3,66$ y 0,37 respectivamente). Esto indica que no solo separa mejor las imágenes normales de las anómalas en

todo el rango de umbrales, sino que además concentra una fracción importante de las anomalías reales en el 10 % superior del ranking de revisión.

El modelo CLIP ViT-L/14@336 por su parte, se ubica como el segundo mejor, con un $F1 \approx 0,47$, $AUC-ROC \approx 0,74$ y $Lift@10$ cercano a 3,0, lo que sugiere un comportamiento muy competitivo y ligeramente inferior al de DINOv2 en términos de separación global y capacidad de priorización. CLIP ViT-H/14 muestra un desempeño similar, aunque un poco más conservador ($F1 \approx 0,46$, $AUC-ROC \approx 0,73$, $Lift@10 \approx 2,6$), mientras que ResNeXt-101 64×4D es el modelo con menor desempeño en este esquema ($F1 \approx 0,39$ y $AUC-ROC \approx 0,63$), lo que indica que, aunque aporta información complementaria, resulta menos efectivo que los vision transformers para discriminar cambios sutiles en el terreno.

Tabla 3. Métricas globales en el espacio de *embeddings* para cada modelo (umbral $q = 0,90$ por estación).

Model	Precision	Recall	F1	Accuracy	AUC_ROC	Lift@10	Gain@10
CLIP_H_224	0.483	0.435	0.458	0.807	0.733	2.595	0.260
CLIP_L_336	0.492	0.458	0.474	0.810	0.739	2.977	0.298
DINOv2_ViT_L14	0.508	0.473	0.490	0.816	0.780	3.664	0.366
ResNeXt101_64x4d	0.434	0.351	0.388	0.793	0.625	2.214	0.221

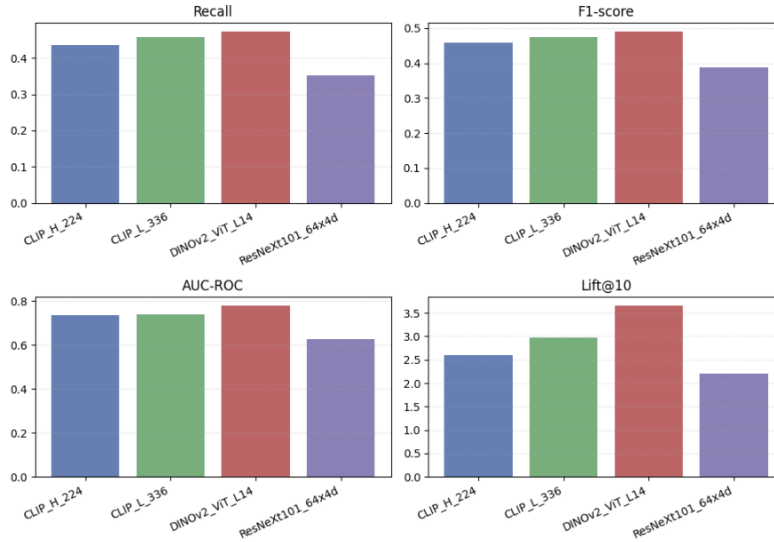


Figura 11. Comparación de *Recall*, *F1*, *AUC-ROC* y *Lift@10* para los cuatro modelos en el espacio de *embeddings*.

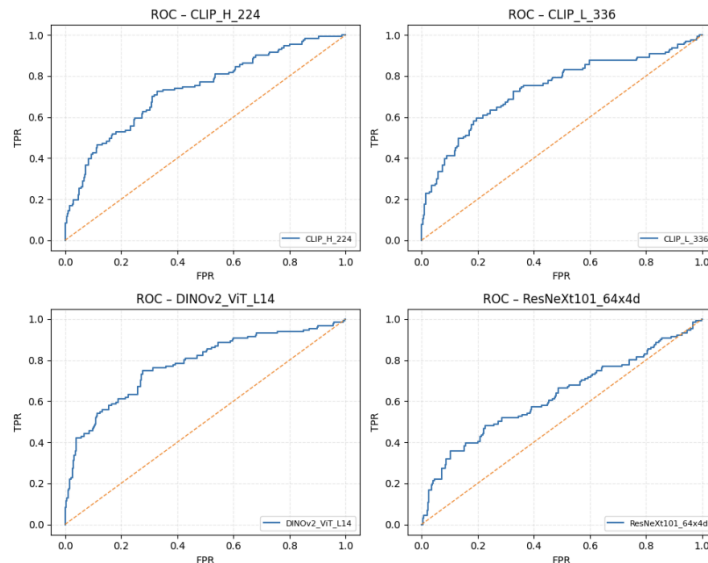


Figura 12. Curvas ROC por modelo usando la distancia al centroide de normales como score.

De este modo, y siendo consistente con la literatura de detección de anomalías bajo fuerte desbalance de clases, las métricas basadas en ranking (Lift@10 y Gain@10) muestran que los mejores modelos, particularmente DINOv2 y CLIP-L/336, logran aumentar entre ~ 3 y $\sim 3,7$ veces la concentración de anomalías en las primeras posiciones del listado respecto a una selección aleatoria (Bergmann et al., 2019; Rachwał et al., 2025).

Estos resultados globales se observan en la sección de resultados por estación y ejemplos visuales de eventos detectados.

4.2.2. Análisis por estación

Además del desempeño agregado por modelo, se realizó un análisis detallado por estación con el fin de evaluar si las distancias en el espacio de *embeddings* preservan patrones espaciales coherentes con la dinámica geotécnica observada en cada sitio. El análisis consistió en tres verificaciones principales:

1. Revisar si las distribuciones de distancias normalizadas por estación reflejan la estabilidad o variabilidad típica del sitio, por ejemplo, estaciones con mayor humedad, niebla o iluminación variable deberían mostrar distribuciones más dispersas.
2. Inspeccionar los casos de las anomalías más extremas para determinar si los cambios detectados corresponden a modificaciones reales en el talud (humedad superficial, erosión, pérdida de vegetación, exposición de material).
3. Comparación de ejemplos visuales representativos “antes/después” para verificar que los eventos detectados no son fallas residuales de la alineación/ROI/HQ, sino alteraciones consistentes con procesos superficiales en el terreno.

La Figura 13 muestra la distribución de los valores de distancia para cada estación bajo el modelo DINOv2 ViT-L/14, el cual obtuvo el mejor desempeño global según la sección 4.2.1. que todas las estaciones presentan medianas en un rango relativamente acotado (en torno a 28–32 unidades), pero con diferencias claras en la dispersión: Playas y Andalucía muestran colas más largas y valores máximos más altos, mientras que Montañita, Pueblo Viejo y Valeria exhiben distribuciones algo más compactas. Estas diferencias son coherentes con las condiciones de captura descritas en el EDA (sección 3.2.4): las estaciones más expuestas a niebla, brillos o variaciones atmosféricas tienden a producir distancias más dispersas, mientras que aquellas con escenas más estables presentan distribuciones más concentradas.

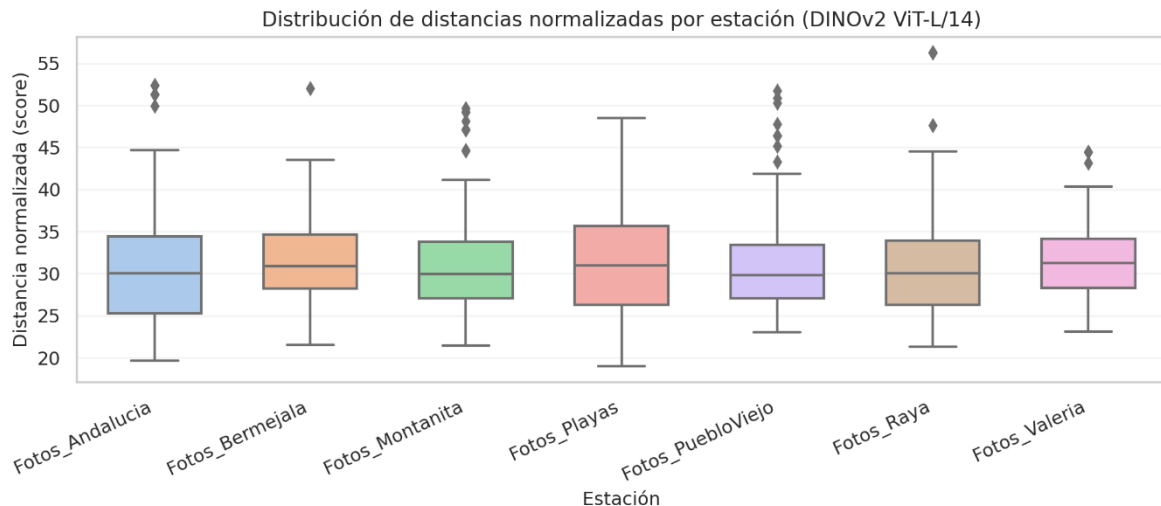


Figura 13. Distribución de distancias normalizadas en el espacio de *embeddings* para cada estación.

Top-10 anomalías por estación. La figura 14 muestra, para cada estación, las 10 imágenes con mayor score de anomalía según DINOv2 ViT-L/14. En particular, la estación Pueblo Viejo resulta ilustrativa, ya que el ROI corresponde casi por completo al talud monitoreado y la serie presenta condiciones atmosféricas relativamente estables. Esto hace que las anomalías seleccionadas por el modelo reflejen principalmente cambios geomorfológicos reales, como variaciones en humedad superficial, zonas de remoción reciente, pérdida de vegetación o exposición de material del subsuelo. Estas transiciones pueden observarse como outliers claros en la serie temporal, lo que

refuerza que los *embeddings* capturan patrones relevantes para el seguimiento del talud y no solamente efectos ópticos o meteorológicos.



Figura 14. Imágenes top 10 anómalas detectadas en la estación Pueblo Viejo.

Comparación “antes/después” de eventos anómalos. Para una interpretación geotécnica, se seleccionaron los casos con mayor score de anomalía en cada estación y se compararon con la imagen válida anterior dentro de la serie, utilizando únicamente capturas que superaron el alineamiento, el recorte por *ROI* y el filtrado estricto por calidad (sección 3.3.4). Esta comparación permite identificar cambios visuales reales asociados a la dinámica superficial del talud, sin interferencias de condiciones atmosféricas adversas.

La Figura 15 presenta el caso más representativo en la estación Pueblo Viejo, donde el modelo DINOv2 ViT-L/14 asignó un score elevado (50.31). Entre la imagen previa y el evento anómalo se observa un cambio claro en la superficie expuesta del terreno:

- Aumento del área de suelo desnudo,
- Reducción localizada de vegetación,
- Ampliación de manchas de material fino,
- Variaciones en textura que sugieren procesos de erosión o lavado superficial.

Estas transiciones, visibles aun después de aplicar el recorte por *ROI* y el filtrado de calidad, indican que el modelo detecta no solo diferencias radiométricas residuales, sino cambios geomorfológicos consistentes con procesos de remoción superficial y degradación del talud.

Fotos_PuebloViejo – comparación antes/después (DINOv2 ViT-L/14)

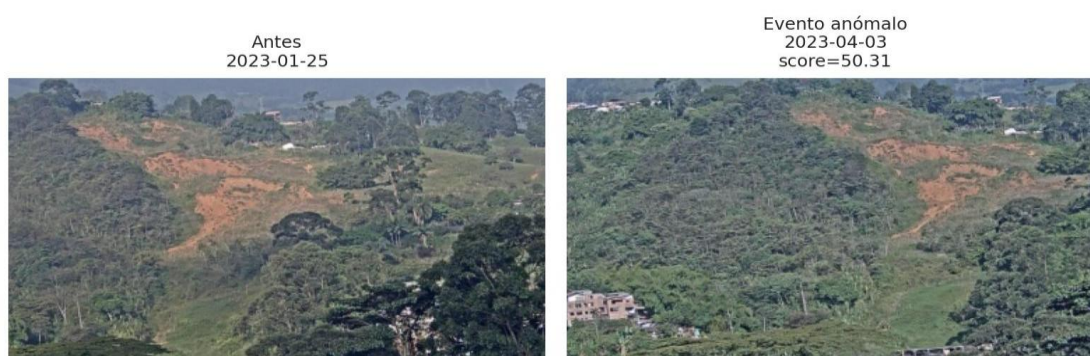


Figura 15. Comparación visual entre una imagen anómala (día D) y la captura temporalmente previa (día D–1 o imagen base más cercana disponible).

4.3 Análisis de sensibilidad al cuantil (q)

Con el fin de analizar la sensibilidad del sistema frente a la elección del cuantil de umbralización, se re-evaluaron los cuatro modelos variando $q \in \{0,80, 0,85, 0,90, 0,95\}$ por estación. Para cada par (modelo, q) se recalcularon las matrices de confusión y las métricas de evaluación (precisión, *recall*, F1, *accuracy*, AUC-ROC y Lift@10),

manteniendo fija la definición del score como distancia al centroide de normales en el espacio de *embeddings* (sección 3.5.4). La Tabla 4 resume presenta métricas por modelo y cuantil, estos resultados permiten observar cómo la elección de q influye directamente en el balance entre sensibilidad y especificidad del sistema.

Tabla 4. Métricas por modelo y cuantil q en el espacio de *embeddings*.

Model	Q	TP	FP	TN	FN	Precision	Recall	F1	Accuracy	AUC ROC	Lift@10
CLIP_H_224	0.80	70	116	453	61	0.376	0.534	0.442	0.747	0.733	2.595
CLIP_H_224	0.85	65	88	481	66	0.425	0.496	0.458	0.780	0.733	2.595
CLIP_H_224	0.90	57	61	508	74	0.483	0.435	0.458	0.807	0.733	2.595
CLIP_H_224	0.95	32	32	537	99	0.500	0.244	0.328	0.813	0.733	2.595
CLIP_L_336	0.80	74	116	453	57	0.389	0.565	0.461	0.753	0.739	2.977
CLIP_L_336	0.85	68	90	479	63	0.430	0.519	0.471	0.781	0.739	2.977
CLIP_L_336	0.90	60	62	507	71	0.492	0.458	0.474	0.810	0.739	2.977
CLIP_L_336	0.95	46	35	534	85	0.568	0.351	0.434	0.829	0.739	2.977
DINOv2_ViT_L14	0.80	78	118	451	53	0.398	0.595	0.477	0.756	0.780	3.664
DINOv2_ViT_L14	0.85	74	88	481	57	0.457	0.565	0.505	0.793	0.780	3.664
DINOv2_ViT_L14	0.90	62	60	509	69	0.508	0.473	0.490	0.816	0.780	3.664
DINOv2_ViT_L14	0.95	57	33	536	74	0.633	0.435	0.516	0.847	0.780	3.664
ResNeXt101_64x4d	0.80	55	116	453	76	0.322	0.420	0.364	0.726	0.625	2.214
ResNeXt101_64x4d	0.85	49	88	481	82	0.358	0.374	0.366	0.757	0.625	2.214
ResNeXt101_64x4d	0.90	46	60	509	85	0.434	0.351	0.388	0.793	0.625	2.214
ResNeXt101_64x4d	0.95	31	33	536	100	0.484	0.237	0.318	0.810	0.625	2.214

En general, se observa que el modelo DINOv2 ViT-L/14 es el más estable y con mejor desempeño global: para $q = 0,85$ y $q = 0,90$ alcanzando los valores más altos de F1 ($\approx 0,49$) con *recall* entre 0,55 y 0,47, manteniendo además los AUC-ROC y Lift@10 más altos entre los cuatro modelos. Esto indica que DINOv2 no solo separa bien normales y anómalas en todo el rango de umbrales, sino que también concentra una fracción importante de las anomalías reales en el 10 % superior del ranking, lo cual es especialmente relevante en contextos de fuerte desbalanceo (Rachwal et al., 2025; Mehta et al., 2020).

Por su parte, el modelo CLIP ViT-L/14@336 muestra un comportamiento bastante cercano: sus mejores F1 se dan también alrededor de $q = 0,85 - 0,90$, con valores de *recall* ligeramente inferiores a DINOv2 pero con precisiones similares. CLIP ViT-H/14 mantiene un desempeño intermedio: mejora su F1 al pasar de $q = 0,80$ a $q = 0,85$, pero a partir de $q = 0,90$ el *recall* cae más rápido. Finalmente, ResNeXt-101 64×4D es el modelo con menor F1 en todos los cuantiles, confirmando que, en este problema, la familia de vision transformers resulta más adecuada que la arquitectura convolucional tradicional.

Analizando los efectos de los cuantiles, los resultados muestran el patrón esperado en detección de anomalías: cuantiles bajos ($q = 0,80$) favorecen el *recall* (marcando así más imágenes como anómalas), pero a costa de un incremento en falsos positivos y una ligera caída en F1; cuantiles intermedios ($q = 0,85 - 0,90$) ofrecen el mejor compromiso entre precisión y *recall*, maximizando F1 para la mayoría de modelos y; cuantiles altos ($q = 0,95$) reducen de forma notable el *recall* sin una ganancia proporcional en precisión, por lo que resultan demasiado conservadores en este contexto.

Basado en la literatura sobre detección de anomalías con desbalanceo de clases y selección basada en cuantiles, (Rachwal et al., 2025; Bergmann et al., 2019; Mehta et al., 2020), se adopta $q = 0,90$ como valor de referencia para el resto del análisis, aplicándolo de forma independiente por estación.

Otro aspecto importante en este estudio, es que, tal como muestra la tabla 3, AUC-ROC y Lift@10 se mantienen prácticamente constantes para cada modelo al variar el cuantil, esto se debe a que el cálculo de AUC-ROC utiliza el score continuo (o distancia al centroide) y es invariante al umbral específico elegido, por tanto cambiar q no altera el área bajo la curva y; Lift@10 depende del ranking de scores solamente, más no del umbral, es decir que

mientras el orden de las imágenes por score se mantenga, el número de anomalías en el 10% superior permanecen iguales.

4.4 Resultados Cualitativos

Además de las métricas cuantitativas presentadas previamente, se realizó un análisis cualitativo para evaluar visualmente la naturaleza de las decisiones realizadas por el modelo. El análisis está comprendido por 3 categorías: verdaderos positivos (TP), que son las anomalías reales correctamente identificadas por el modelo; falsos positivos (FP) que son imágenes sin cambios relevantes que el modelo etiquetó como anomalías y; falsos negativos (FN), que son las imágenes donde había cambios visuales detectables, pero el modelo no los contempló como una anomalía.

El objetivo de este análisis es comprender mejor qué tipo de cambios en el terreno es capaz de reconocer el modelo, en qué situaciones tiende a fallar y qué patrones visuales puede estar confundiendo con anomalías. Los verdaderos negativos (TN) no se incluyen en este apartado debido a que, por definición, corresponden a imágenes sin cambios visuales relevantes, correctamente clasificadas como normales, por lo que no aportan información cualitativa adicional sobre el comportamiento del modelo.

4.4.1. Ejemplos de anomalías correctamente detectadas (TP)

Los verdaderos positivos corresponden a imágenes donde se observan cambios geomorfológicos evidentes, tales como: aumento del área de suelo expuesto, pérdida localizada de vegetación, ampliación de manchas de erosión, aparición de cicatrices superficiales recientes, alteraciones en texturas del terreno asociadas a remoción fina de material. La Figura 16 presenta ejemplos representativos detectados por el modelo DINOv2 ViT-L/14. En estas imágenes se evidencia que el modelo captura transiciones físicas relevantes del talud, incluso bajo variabilidad atmosférica mínima, lo que valida su utilidad para el monitoreo geotécnico basado en imágenes.

Ejemplos de Verdaderos Positivos (TP) – DINOv2 ViT-L/14



Figura 16. Verdaderos positivos detectados por el modelo DINOv2 ViT-L/14.

4.4.2. Falsos positivos (FP)

Los falsos positivos corresponden a imágenes clasificadas como anómalas por el modelo pese a no existir un cambio geomorfológico relevante en el terreno. En los ejemplos de la Figura 17, se observa que las tres escenas corresponden a taludes estables, con vegetación densa y sin evidencias claras de remoción reciente. En estos casos, el modelo asignó puntuaciones altas debido a patrones que se asemejan visualmente a zonas de suelo expuesto, pero que en realidad forman parte de la cobertura vegetal o de inclinaciones naturales del talud. Entre los factores que inducen estas falsas alarmas se identifican: tonos de vegetación seca que el modelo interpreta como material removido, contrastes locales fuertes en la iluminación que resaltan ciertas franjas del terreno y las hacen parecer cicatrices de deslizamiento, pendientes amplias con escasa vegetación, que pueden confundirse con superficies inestables aun cuando corresponden a laderas estables.

Ejemplos de Falsos Positivos (FP) – DINOv2 ViT-L/14

**Figura 17.** Ejemplos de falsos positivos generados por el modelo DINOv2 ViT-L/14.

4.4.3. Falsos negativos (FN)

Los falsos negativos ocurren cuando existen cambios sutiles en el talud, pero el modelo no los detecta como anomalías. Entre los patrones más frecuentes se encuentran: transiciones de baja magnitud (pequeños aumentos en la superficie expuesta), cambios lentos y graduales, alteraciones muy localizadas en vegetación, eventos con baja señal visual respecto al ruido. En la Figura 18 se muestran ejemplos representativos de este tipo de error.

Ejemplos de Falsos Negativos (FN) – DINOv2 ViT-L/14

**Figura 18.** Falsos negativos del modelo DINOv2 ViT-L/14.

En resumen, los resultados muestran que DINOv2 ViT-L/14 se comporta como el modelo más equilibrado para la tarea de priorización de anomalías geotécnicas. Aunque CLIP-L/336 obtiene valores ligeramente superiores de F1 y Lift@10 en ciertos conjuntos, DINOv2 mantiene una consistencia inter-estaciones más alta, una AUC-ROC superior, y una estabilidad más robusta frente a variaciones radiométricas y condiciones climáticas. Este comportamiento es coherente con la literatura reciente, donde los modelos auto supervisados basados en ViT muestran mayor invariancia frente a cambios de iluminación, textura y escala (Caron et al., 2021; Oquab et al., 2023). En nuestros experimentos, esta robustez se refleja en una menor dispersión de los puntajes entre estaciones, una reducción de falsos positivos asociados a variabilidad natural del talud, y una mejor preservación de la separabilidad entre casos verdaderamente atípicos y escenas normales.

Por su parte, CLIP-L/336, a pesar de su excelente Lift@10, mostró mayor sensibilidad a artefactos atmosféricos y a variaciones abruptas de iluminación, lo que incrementó los FP en estaciones con fuerte heterogeneidad radiométrica. Los modelos ResNeXt101 y CLIP-H/224 resultaron útiles como referencia, pero exhibieron menor discriminación y mayor inestabilidad entre estaciones, consistente con lo reportado por Radford et al. (2021) y estudios de cambio multitemporal en entornos naturales. En consecuencia, y considerando tanto desempeño cuantitativo como estabilidad operacional, DINOv2 ViT-L/14 se recomienda como la arquitectura principal para la etapa de priorización automática, mientras que CLIP-L/336 puede utilizarse como apoyo complementario en un esquema de ensemble o para análisis exploratorios.

5 Caso de aplicación

Para ilustrar el funcionamiento operativo del sistema se seleccionó un caso de estudio en la estación Andalucía, asociado a un episodio de inestabilidad superficial ocurrido a mediados de noviembre de 2023. La Figura 19 muestra la evolución inmediata del talud en una ventana de cinco días: dos días antes, el día del evento y dos días después. Todas las imágenes corresponden al conjunto de datos alineado y recortado por ROI, sobre el cual se aplicó el filtrado estricto de calidad y el cálculo de distancias en el espacio de *embeddings* con el modelo DINOv2 ViT-L/14.

En la imagen objetivo (15/11/2023, 08:00:30) se observa el evento que el modelo marca como anomalía ($y_{pred} = 1$, $y_{true} = 1$, $score \approx 51.3$, percentil > 98 dentro de la estación). A pesar de la presencia de gotas en el lente, se aprecia claramente la quebrada desbordada, con un flujo de agua turbia que remueve material a lo largo del cauce y genera un contraste muy marcado entre el suelo desnudo y la vegetación circundante. Las imágenes de los dos días previos (13 y 14 de noviembre) muestran un talud visualmente más estable: el cauce aparece menos definido, con menor huella de erosión reciente y sin evidencia clara de desbordamiento. En los dos días posteriores (16 y 17 de noviembre) el flujo superficial disminuye y el cauce vuelve a su estado “normal”, pero persiste una franja de suelo expuesto y sedimentos depositados que no estaba presente en los días anteriores, coherente con un episodio de mayor descarga en la quebrada.

Evolución inmediata alrededor del evento – Fotos_Andalucia (DINOv2 ViT-L/14)



Figura 19. Caso de uso del modelo DINOv2 ViT-L/14 para la estación Andalucía 2 días antes y después de un evento clasificado como anómalo.

Este caso ejemplifica, no de manera hipotética sino como un ejemplo real, el tipo de señal que el sistema es capaz de priorizar en un escenario operativo: el evento del 15 de noviembre aparece entre las imágenes con mayor score de anomalía de toda la serie para Andalucía, mientras que las capturas inmediatamente anteriores y posteriores, aunque comparten condiciones de iluminación similares, no son clasificadas como anómalas. En otras palabras, el modelo no reacciona solo a cambios radiométricos o a la presencia de gotas en la cámara, sino a la combinación de geometría del cauce, extensión del suelo desnudo y textura asociada al material recientemente movilizado. Esto sugiere que, en la práctica, revisar de forma rutinaria el top-k de anomalías recientes por estación (por ejemplo, las 10 mayores de un periodo de tiempo definido) permitiría a un analista identificar de forma temprana episodios de saturación y erosión en el talud, incluso en condiciones meteorológicas adversas.

6 Conclusiones

Los resultados obtenidos en este estudio demuestran que la aproximación para detección de anomalías visuales basado en *embeddings* es viable para el monitoreo multitemporal de imágenes de punto fijo en terrenos propensos a deslizamientos. Más que identificar deslizamientos en sentido estricto, el sistema demuestra capacidad para resaltar cambios visuales relevantes como por ejemplo la exposición de suelo, modificaciones abruptas en textura, alteraciones en vegetación o cuerpos de agua, que funcionan como indicadores tempranos de procesos geomorfológicos anómalos. Este comportamiento coincide con la literatura reciente que subraya la potencia de las representaciones auto supervisadas (Caron et al., 2021; Oquab et al., 2023) para capturar variaciones espaciales complejas incluso en contextos con etiquetas limitadas.

Los modelos de tipo Vision Transformer, particularmente DINOv2 ViT-L/14 y CLIP-L/336, lograron ordenar las imágenes según su nivel de rareza visual, priorizando escenas asociadas a perturbaciones reales y manteniendo estabilidad ante variaciones de iluminación propias de cámaras en campo. El caso de uso mostró, además, que el

sistema responde de forma coherente en el tiempo: cuando se produce una alteración significativa, como la crecida abrupta del cauce en Andalucía (noviembre de 2023), el modelo incrementa sus puntajes y distingue el evento frente a días previos y posteriores, como se evidenció en la secuencia temporal de ± 2 días.

Aun con limitaciones inherentes al desbalance extremo de clases, artefactos meteorológicos y heterogeneidad entre cámaras, los resultados respaldan que este enfoque es adecuado como primera etapa de priorización dentro de un sistema de monitoreo continuo. En conjunto, los hallazgos permiten concluir que la anomalía visual es un indicador práctico, sensible y computacionalmente eficiente para el seguimiento de laderas, y que el análisis multitemporal mediante *embeddings* ofrece una base sólida para los módulos posteriores para continuar construyendo sistemas de alerta temprana orientados a identificar procesos de inestabilidad superficial.

Trabajo futuro

Como proyección de este estudio se propone:

- Incorporar reglas temporales de persistencia para distinguir fluctuaciones breves de cambios sostenidos,
- Evaluar arquitecturas más recientes de visión por computador y modelos fundacionales,
- Explorar esquemas semi-supervisados que aprovechen etiquetas parciales o débiles y,
- Avanzar hacia la integración operativa de este pipeline en la plataforma de monitoreo geotécnico del SIATA, de forma que pueda analizar imágenes en tiempo real y priorizar eventos para revisión experta.

7 Anexos

Anexo A: Matrices de confusión por modelo

A continuación, en la figura 20 se presentan las matrices de confusión completas para cada uno de los modelos evaluados en la etapa de selección directa por modelo. Estas figuras complementan la Tabla 1 del capítulo de resultados, donde se resumen las métricas globales de desempeño (*Precision*, *Recall*, *F1-Score*, *Accuracy*).

Las matrices se incluyen aquí con el propósito de mostrar el detalle de las distribuciones de *TP/FP/FN/TN* por modelo, permitiendo una inspección visual más precisa de los patrones de error. No se introduce interpretación adicional en este anexo, debido a que el análisis cuantitativo y comparativo se desarrolló en la Sección 4.1.1 del documento.

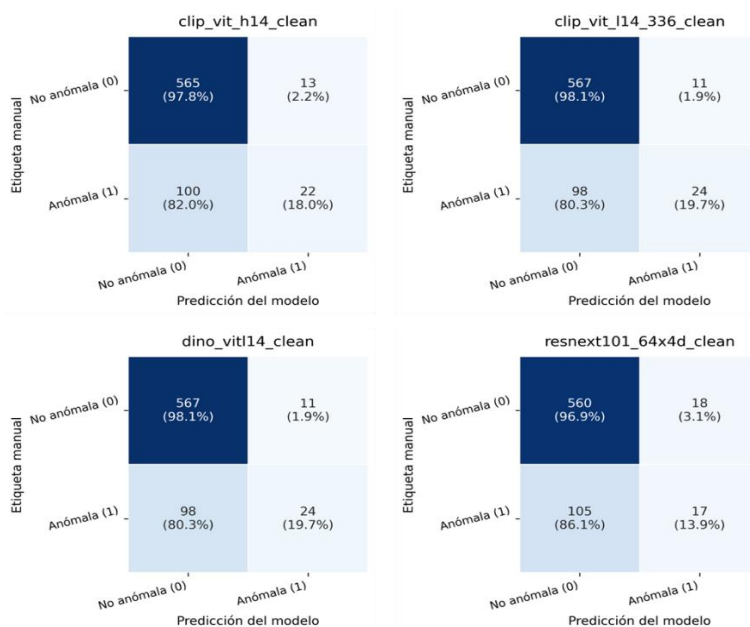


Figura A20. Matrices de confusión por modelo (CLIP L/14-336, DINOv2 L/14, CLIP H/14 y ResNeXt).

Anexo B: Evolución de las métricas en función del cuantil

La Figura 21 presenta la evolución del *Recall*, *F1*, *AUC-ROC* y *Lift@10* para los cuatro modelos evaluados al variar el cuantil q . Esta visualización complementa los valores numéricos de la Tabla 3, permitiendo apreciar con claridad tendencias globales como: el incremento del *recall* en cuantiles bajos, la estabilización y máxima eficiencia alrededor de $q = 0.85$ – 0.90 , y la caída del *recall* para cuantiles altos.

Esta figura sirve como soporte visual al análisis presentado en la sección 4.3 del documento principal.

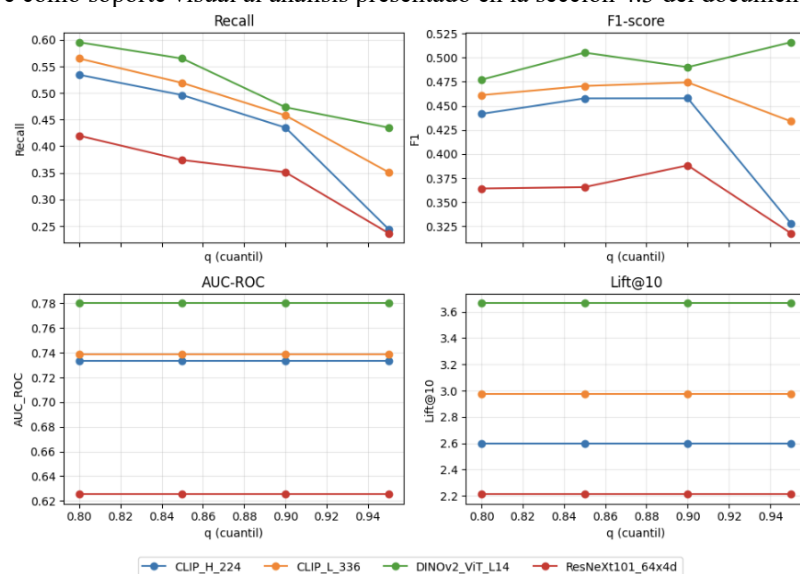


Figura B21. Comportamiento de las métricas de evaluación en función del cuantil q para cada modelo

En la figura 22 se presentan las matrices de confusión generadas para distintos valores de q en cada uno de los modelos, donde se puede evidenciar cómo el incremento de q reduce FP (falsos positivos) a costa de aumentar FN (falsos negativos). Este comportamiento confirma las tendencias descritas en el cuerpo principal del documento sobre el impacto del cuantil en la sensibilidad del sistema.

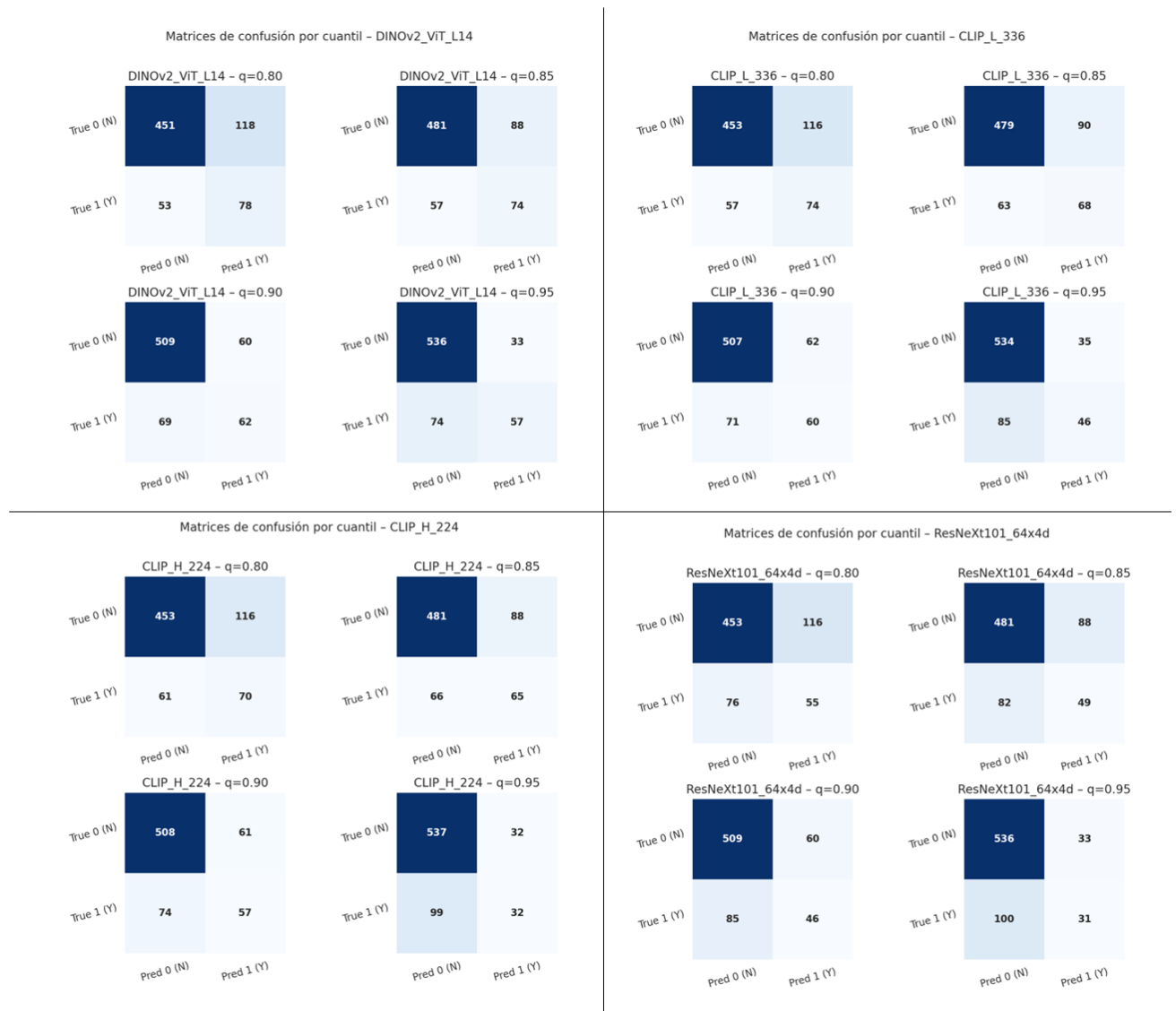


Figura B22. Matrices de confusión por modelo para distintos cuantiles q en el espacio de *embeddings*.

Referencias

1. Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection in industrial inspection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9592–9600
2. Bovik, A. C. (2013). *The essential guide to image quality assessment*. Academic Press.
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9650–9660.
4. Chen, W., Hong, H., Li, S., & Wang, Y. (2022). Integration of remote sensing and machine learning for landslide early warning systems: A review. *Earth-Science Reviews*, 232, 104076. <https://doi.org/10.1016/j.earscirev.2022.104076>
5. Coluzzi, R., et al. (2025). Rapid landslide detection from free optical satellite imagery. *Scientific Reports*, 15(2), 1124.
6. D'Amato Avanzi, G., Intrieri, E., & Gigli, G. (2021). Sensor networks for landslide monitoring and early warning: State of the art and perspectives. *Sensors*, 21(3), 680. <https://doi.org/10.3390/s21030680>

7. Dou, J., Yunus, A. P., Tien Bui, D., Merghadi, A., Sahana, M., Zhu, Z., Chen, C. W., Khosravi, K., Yang, Y., & Pham, B. T. (2019). Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Science of the Total Environment*, 662, 332–346. <https://doi.org/10.1016/j.scitotenv.2019.01.221>
8. Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S. R., Tiede, D., & Aryal, J. (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing*, 11(2), 196. <https://doi.org/10.3390/rs11020196>
9. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation (R-CNN). *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580–587. <https://doi.org/10.1109/CVPR.2014.81>
10. Goetz, J. N., Brenning, A., Petschko, H., & Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & Geosciences*, 81, 1–11.
11. Guo, X., et al. (2021). Detecting the vegetation change related to the creep of landslides. *Frontiers in Earth Science*, 9, 706998.
12. Herrera, G., Mateos, R. M., García-Davalillo, J. C., Grandjean, G., Poyiadji, E., Maftai, R., & Van Westen, C. J. (2018). The European Landslide Early Warning Network (E-LENS): Advances and challenges. *Landslides*, 15(7), 1323–1334. <https://doi.org/10.1007/s10346-018-0979-0>
13. Hore, A., & Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. *2010 20th International Conference on Pattern Recognition*, 2366–2369. <https://doi.org/10.1109/ICPR.2010.579>
14. H2O.ai. (s.f.). Cumulative Gains and Lift Charts. H2O Documentation. Recuperado de <https://h2oai.github.io/model-docs/latest/cumulative-gains-lift-charts.html>
15. Krotkov, E. (1988). Focusing. *International Journal of Computer Vision*, 1(3), 223–237. <https://doi.org/10.1007/BF00127947>
16. Li, N., et al. (2024). A deep-learning-based algorithm for landslide detection over wide areas using InSAR images considering topographic features. *Remote Sensing*, 16(4), 755.
17. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
18. Merghadi, A., Yunus, A. P., Dou, J., Whiteley, J., Thai Pham, B., Bui, D. T., Avtar, R., & Abderrahmane, B. (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Science Reviews*, 207, 103225. <https://doi.org/10.1016/j.earscirev.2020.103225>
19. Mehta, S., Mehendale, S., Fernandes, N., Sarkar, J., Sarkar, S., & Saha, S. (2024). Benchmarking anomaly detection algorithms: Deep learning and beyond. *IEEE Transactions on Artificial Intelligence*. arXiv:2402.07281.
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernández, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P. Y., Li, S. W., Misra, I., Rabbat, M., Joulin, A. (2023). DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*. <https://doi.org/10.48550/arXiv.2304.07193>
21. Rachwał, A., Banaś, P., Dobrzański, P., & Dadas, S. (2025). Aggregating evaluation metrics for anomaly detection. *Advances in Science and Technology Research Journal*, 19(1), 402–419. <https://doi.org/10.12913/22998624/183113>
22. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, 8748–8763. <https://arxiv.org/abs/2103.00020>
23. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., & Kloft, M. (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5), 756–795
24. Santacana, N., Baeza, B., Corominas, J., De Paz, A., & Marturiá, J. (2003). A GIS-based multivariate statistical analysis for shallow landslide susceptibility mapping in La Pobra de Lillet area (Eastern Pyrenees, Spain). *Natural Hazards*, 30(3), 281–295. <https://doi.org/10.1023/B:NHAZ.0000007169.28860.80>

25. Stumpf, A., & Kerle, N. (2011). Object-oriented mapping of landslides using Random Forests. *Remote Sensing of Environment*, 115(10), 2564–2577. <https://doi.org/10.1016/j.rse.2011.05.013>
26. University of Regina. (s.f.). Lift Charts. Pattern Discovery Lab. Recuperado de <https://www2.cs.uregina.ca/~dbd/cs831/notes/liftchart/liftchart.html>
27. Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective Search for Object Recognition. *International Journal of Computer Vision*, 104, 154–171. <https://doi.org/10.1007/s11263-013-0620-5>
28. Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 5987–5995. <https://doi.org/10.1109/CVPR.2017.634>