AIMS Environmental Science, 12(6): 936–957.
DOI: 10.3934/environsci.2025041
Received: 27 June 2025
Revised: 13 August 2025
Accepted: 03 September 2025
Published: 30 October 2025

https://www.aimspress.com/journal/environmental

Research article

Explainable Transfer Learning with Attention Mechanisms for Landslide Crack Classification

Chenxi Zhang[1,2], Qi Ge[1,2,*], Wei Wei[2], Wei Zhan[1], Xin Yan[1] and Jin Li[3]

[1] Zhejiang Scientific Research Institute of Transport, Hangzhou, 310005, China

[2] College of Civil Engineering, Nanjing Forestry University, Nanjing, 210037, China

[3] College of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, 210044, China

* **Correspondence:** Email: geqi@njfu.edu.cn.

Abstract: Landslide crack identification is crucial for risk management and mitigation, yet challenges like limited data and the lack of model interpretability hinder progress. This study proposes a deep learning-based solution for classifying landslide cracks using the ResNet50 architecture integrated with a squeeze-and-excitation (SE) attention mechanism, the transfer learning (TL) strategy, and a model interpretability approach. We introduced the SE module to enhance the model's ability to capture subtle crack features, and, in the TL framework, we first pretrained the model on a large concrete crack dataset to learn general crack characteristics, then fine-tuned it on a landslide crack dataset. To further improve interpretability, we applied gradient-weighted class activation mapping (Grad-CAM) to visualize the areas of the image most influential to the model's decisions. Our results demonstrated that ResNet50+SE outperforms both the standard convolutional neural network (CNN) and residual network with 50 layers (ResNet50), particularly in recall and F1 score, highlighting its superior ability to detect challenging cracks and improve overall landslide crack identification accuracy. Additionally, TL boosts performance, even with the limited availability of landslide crack data. Grad-CAM heatmaps provide valuable insights into the model's focus and decision-making, enhancing transparency. This study also tackled data imbalance challenges. Overall, the proposed approach offers an effective, interpretable solution for landslide crack identification, enhancing accuracy and transparency for landslide early warning and risk assessment.

Keywords: machine learning; landslide; risk assessment; crack identification; model interpretability

1. Introduction

Landslides are widespread natural disasters that cause significant loss of life, property, and environmental damage [1–3]. While engineering solutions like slope reinforcement are important, monitoring and predicting landslides using deformation data offers key advantages [4–7]. This approach enables early warning and real-time tracking of slope instability [8–10], which is crucial for disaster prevention and risk reduction [11–13].

Ground cracks are direct indicators of rock and soil fractures that occur during landslide deformation, reflecting stress adjustments and displacement concentration within the landslide mass [14, 15]. The geometric shapes, extension directions, and patterns of these cracks serve as key markers for identifying the different stages of landslide evolution [16, 17]. In the early phase, isolated cracks dominate, gradually connecting to form a crack network as deformation intensifies. On one hand, the continuous expansion of these cracks compromises the structural integrity of the slope, accelerating the penetration of the sliding surface, which may eventually trigger a catastrophic landslide [18, 19]. On the other hand, these cracks provide a pathway for surface water to rapidly infiltrate, further softening the sliding surface soil and increasing pore water pressure [20]. This creates a vicious cycle that could ultimately result in landslide failure. Therefore, accurately identifying landslide cracks is crucial for determining the current slope stability and deploying monitoring equipment effectively.

Currently, manual inspection is the most commonly used method for crack identification [21, 22]. However, this approach presents several challenges, such as blind spots during the inspection process, high fieldwork intensity, low efficiency, subjective biases affecting results, and overall low accuracy [23]. To overcome these limitations, various image processing-based identification methods have been developed [24, 25]. While these techniques offer a more automated solution, they are still highly susceptible to environmental factors, such as lighting conditions, shadows, and background interference. Additionally, the computational processes involved in image processing-based methods are often complex, leading to reduced efficiency [26].

In recent years, computer vision (CV) technology, powered by deep learning (DL), has made rapid advancements, offering substantial benefits for detecting visible issues like cracks. These advantages include high accuracy, long-range detection, ease of implementation, and cost-effectiveness. With the development of hardware technologies such as drones [27, 28] and continuous improvements in detection algorithms, CV holds great potential to replace manual visual inspections in landslide crack identification. Many researchers have proposed diverse methods for crack identification by integrating mechanical image acquisition with CV techniques. For example, [29] employed CV for the health monitoring of shield tunnels, using it to detect defects in tunnel linings. [30] utilized feature learning for the identification and width quantification of bridge cracks. [31] proposed a method for detecting trailing edge cracks in landslides through image processing technology, employing a custom comparison algorithm to calculate crack motion parameters. [32] designed an automated recognition model for slope cracks using RetinaNet, applying it to map cracks in landslide-prone areas.

However, research on using deep learning models to detect landslide cracks remains limited. This is primarily because developing such models requires a substantial amount of training data, which is difficult to obtain for landslide cracks compared to more readily available concrete crack images. As a result, building intelligent identification models for landslide surface cracks using deep learning is challenging. Additionally, a significant concern is the interpretability of these models. While deep

learning can greatly enhance efficiency compared to manual inspections, its "black-box" nature makes it difficult for researchers to understand how the model arrives at its conclusions [33, 34]. This lack of transparency presents challenges in high-risk fields like landslide risk assessment, where understanding the decision-making process is vital. In these situations, the inability to explain the model's underlying reasoning can erode trust in its results and limit its practical application in critical decision-making.

To overcome the limitations mentioned above, this paper proposes a deep learning-based classification model for landslide crack identification. The model utilizes the residual network with 50 layers (ResNet50) architecture combined with an attention mechanism. It is pre-trained on concrete crack images, which share visual similarities with landslide crack images, and then fine-tuned on a landslide crack dataset using a transfer learning (TL) strategy [35]. This approach effectively addresses the challenge of limited landslide crack data. Furthermore, the study employs gradient-weighted class activation mapping (Grad-CAM) to generate attention heatmaps, which visualize gradient information from the final layer. This provides an intuitive representation of the key features influencing the model's decisions, thereby enhancing the interpretability and reliability of the identification results. Additionally, the impact of data imbalance on landslide crack classification is considered to ensure more accurate and reliable outcomes.

2. Methodology

2.1. Workflow overview

The flowchart in Figure 1 outlines the implementation procedure of the research, which can be summarized as follows:

- **Step 1:** Collect a large public dataset of concrete crack images and establish a small landslide crack image dataset, followed by data augmentation.
- **Step 2:** Pretrain models on the concrete dataset using CNN, ResNet50, and the proposed ResNet50 with a squeeze and excitation (SE) model.
- **Step 3:** Use the landslide crack dataset to fine-tune the pretrained models or train the models that were not pretrained directly.
- **Step 4:** Evaluate and compare the output results of models with and without the transfer learning strategy.
- **Step 5:** Use the Grad-CAM algorithm to generate heatmaps, visualize the results, and compare the interpretability of the different models.

2.2. Baseline networks

To ensure efficient and accurate landslide crack detection in data-limited scenarios, CNN and ResNet50 were chosen as baseline models for their simplicity, computational efficiency, and effectiveness with small datasets. However, this approach is also applicable to other advanced backbones (e.g., EfficientNet, MobileNet, Vision Transformer) in future work when more data and resources are available.

(1) CNN: The typical CNN architecture comprises convolutional layers, pooling layers, and fully

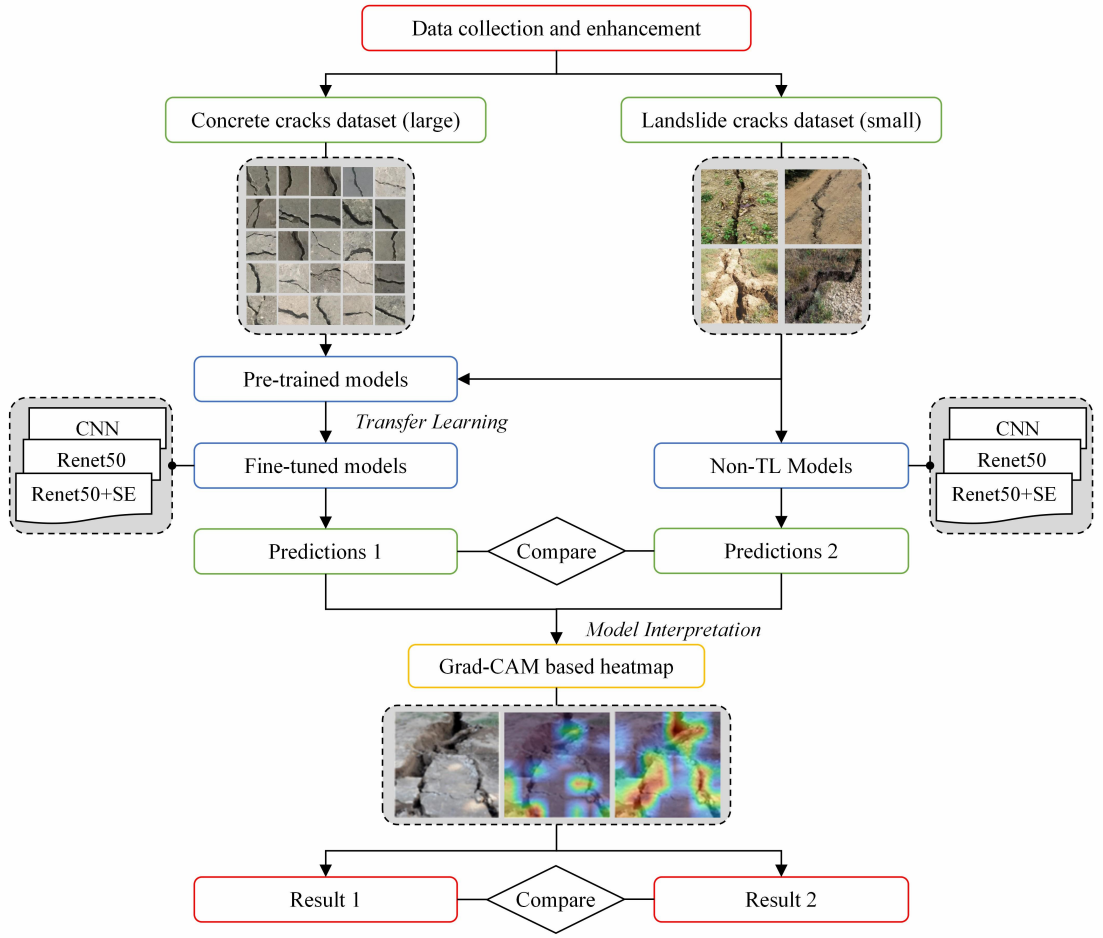

Figure 1. The flowchart of the research

connected layers [36–39]. The convolutional layers extract meaningful features from the input data by capitalizing on local correlations within images. Meanwhile, the pooling layers perform downsampling on the feature maps, which helps reduce the number of parameters and enhances the model's robustness to image transformations, such as rotation, translation, and scaling. Common pooling techniques include max pooling and average pooling. The CNN architecture used in this study consists of three convolutional layers and two fully connected layers. Each convolutional layer is followed by a ReLU activation function and a 2×2 max pooling operation, which reduces the spatial dimensions of the feature maps by half, thereby decreasing computational complexity [40]. The formulas for the max pooling, and activation functions are provided as follows:

$$y = \text{Maxpool}(x; k, l, s) \tag{2.1}$$

where y represents the output feature map obtained after the operation of maximum pooling, k denotes the window height, and l denotes the width. The operation moves these windows across the input feature map x with a step size of s.

$$\text{ReLU}(x) = \begin{cases} x & , x \geq 0 \\ 0 & , x \leq 0 \end{cases} \tag{2.2}$$

The output of the final fully connected layer consists of two classes, "crack" and "no crack", providing a binary classification output, as depicted in Figure 2(a). The input images are in RGB format, consisting of three channels: red, green, and blue. The depth of the convolutional kernels matches the number of channels in the input images. The initial convolutional layers primarily capture low-level features, such as crack edges, texture, gradient direction, and color. As the number of convolutional layers increases, the model becomes more adept at extracting high-level features, such as the distinct shapes of cracks. This enables the neural network to gain a deeper understanding of the images within the dataset.

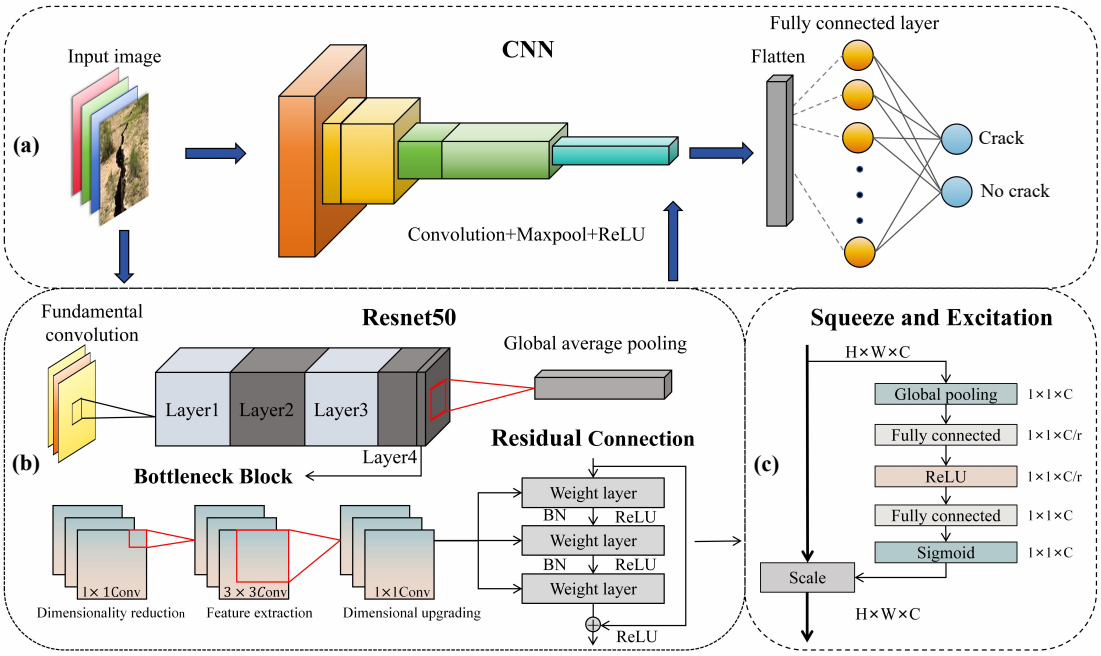

Figure 2. The network architectures of (a) CNN, (b) ResNet50, and (c) ResNet50+SE.

(2) ResNet50: a pretrained model based on a CNN, has proven highly effective in numerous image classification tasks [41]. However, its deeper architecture increases training complexity [42]. As the network depth increases, gradient backpropagation becomes unstable, risking vanishing or exploding gradients. Although deeper networks are theoretically expected to improve training, they often result in longer training times, slower convergence, and reduced accuracy once saturation is reached. To mitigate these issues, ResNet50 incorporates residual connections, ensuring that model performance does not degrade while preserving the advantages of deep networks.

The structure of ResNet50 is illustrated in Figure 2(b). Initially, images augmented through data enhancement are fed into the first convolutional layer. The features then pass through four stages, each consisting of multiple bottleneck blocks. Each bottleneck block includes 1x1, 3x3, and 1x1 convolutional layers. Residual connections allow the network to effectively integrate and process information. After each convolutional layer, batch normalization and ReLU activation functions are applied. The features are then passed through a global average pooling layer before being input into two fully connected layers, ultimately producing the binary output: "crack" or "no crack."

2.3. Feature attention enhancement with SE

The baseline networks may perform poorly in certain situations, particularly when they fail to capture fine details. For example, when identifying multiple complex cracks, factors such as weak color contrast, indistinct crack edges, and large crack widths can significantly impact the model's accuracy. To address this issue, this paper proposes an improved ResNet50 model for landslide crack classification, incorporating the SE module. An SE module is added to each bottleneck in the final layer, along with a dropout layer, which allows the model to better capture unclear edges of tensile cracks, large crack terminations, and small cracks. Furthermore, the dropout layer helps prevent overfitting, thus optimizing the model's performance.

The SE module is a lightweight attention mechanism that can be seamlessly integrated into various network architectures. This attention mechanism dynamically learns from the global context, prioritizing channels with higher information richness. Ultimately, the model adaptively adjusts the learned channel weights to recalibrate the feature responses across the channel dimensions. Moreover, it simplifies insertion and usage, and significantly enhances performance with minimal computational cost. The SE module consists of three components: squeezing, excitation, and scaling, as illustrated in Figure 2(c). Initially, the SE module compresses the two-dimensional features using global average pooling, as shown in the following mathematical expression:

$$z_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,c} \tag{2.3}$$

The input feature map tensor of the SE module is denoted as $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H and W represent the height and width of the feature map, respectively, and C denotes the number of channels. $z_c$represents the global average pooling result of the input feature map on channel c. $X_{i,j,c}$ refers to the value at spatial location (i, j) in the c-th channel, while X_c represents the entire $H \times W$ feature map corresponding to the c-th channel.

Next, a small two-layer fully connected network is used to learn the nonlinear relationships between channels. The first layer, W_1, performs compression and is followed by the ReLU activation function. The second layer, W_2, restores the dimensionality, with the sigmoid activation function, ensuring that the output range is (0, 1):

$$\mathbf{s} = \sigma(\mathbf{W}_2 \cdot \delta(\mathbf{W}_1 \cdot \mathbf{z})) \tag{2.4}$$

where $\delta(\cdot)$ denotes the ReLU function, and $\sigma(\cdot)$ denotes the sigmoid function.

Finally, by applying the channel weight s to the original input, each channel is multiplied by its corresponding coefficient, effectively re-weighting the features:

$$\tilde{X}_{i,j,c} = s_c \cdot X_{i,j,c} \tag{2.5}$$

The SE module was chosen due to its computational efficiency and ability to refine feature responses through channel-wise attention. This is particularly beneficial for landslide crack identification, where subtle and irregular features need to be captured. Unlike other attention mechanisms like the convolutional block attention module (CBAM), SE blocks provide a simpler and effective solution, which is well-suited for the small and complex landslide crack dataset. Moreover, the SE module was inserted after each bottleneck block in layer 4. Layers 1 to 3 primarily extract low-level features

such as edges and textures, which are crucial for crack detection. Adding the SE module to these layers could lead to overemphasis on certain features, potentially disrupting the pretrained model's performance. Thus, the SE module was applied only to layer 4, and layers 1–3 were frozen to preserve feature extraction efficiency.

2.4. Domain adaptation with TL

A major limitation of traditional deep learning is its reliance on large amounts of data and considerable effort to achieve satisfactory results. While obtaining a large volume of landslide crack images is challenging due to the constraints of the natural environment, concrete crack images are more readily available, with numerous publicly accessible datasets. The TL strategy employed in this study utilizes pretraining on a concrete crack dataset due to the visual similarities between concrete and landslide cracks, such as edges and textures. This allows the model to learn general crack features that are then fine-tuned on the landslide crack dataset, effectively overcoming the domain shift. This approach is more efficient than other TL strategies, as it directly adapts the model to landslide cracks without the complexity of domain adaptation techniques. Therefore, we incorporated TL during the model training process.

The core idea behind TL is to leverage knowledge gained from previous tasks to improve performance in new, related scenarios [43,44]. In the transferred model, as shown in Figure 3, the lower and middle layers are frozen to retain the weights associated with general crack feature recognition learned during pretraining. The higher layers are then specifically adjusted by unfreezing the weights of layer 4, adding the SE module, and unfreezing the final fully connected layer, which is modified to output two classes for binary classification. This partial unfreezing strategy was adopted to balance generalization and task-specific adaptation: fully fine-tuning the network or unfreezing earlier layers could lead to overfitting and training instability due to limited data, while freezing too much would restrict the model's ability to adapt to landslide-specific features.

2.5. Model interpretability with Grad-CAM

Deep learning models like CNN and ResNet50 have complex structures and many parameters, making them "black-box" models that are hard to interpret. This opacity complicates understanding their decision-making process. Interpretability in machine learning refers to the ability to explain the model's decisions and key influencing features. Class activation mapping (CAM) [45], an advanced interpretability method, reveals pixel contributions through the final convolutional layer. However, it requires removing the fully connected layer and adding global average pooling, which can reduce model performance. In contrast, Grad-CAM [46, 47] overcomes these limitations, allowing for application without modifying the model. Grad-CAM is widely used in CNN architectures to explain model performance by identifying activations in the last convolutional layer during inference and generating an activation map.

As shown in Figure 4, in Grad-CAM, forward propagation is first performed to compute the score y^c for class c in the network's output. Then, backpropagation is applied to calculate the gradient of the feature activation A^K in the convolutional layer. These gradients are globally averaged over width and height to obtain the neuron importance weights α_K^c. The weight matrix is then multiplied by the continuous activation gradients until the last convolutional layer. Finally, the forward activation map

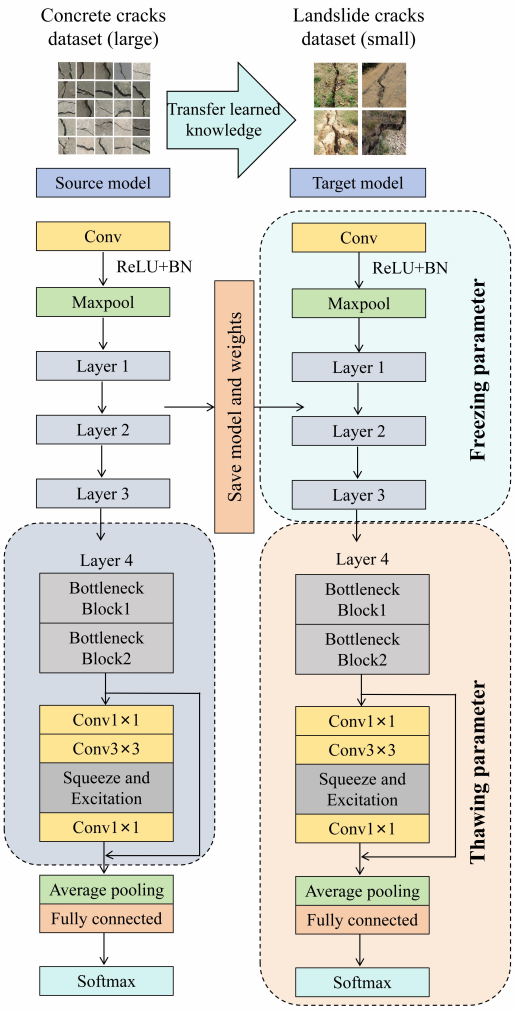

Figure 3. Illustrations of the landslide crack classification transfer learning.

is combined with the weights, followed by a ReLU operation [48]. This process is mathematically expressed as follows:

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}\right) \tag{2.6}$$

$$\alpha^{c}_{k} = \frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial y^{c}}{\partial A^{k}_{i,j}} \tag{2.7}$$

The resulting heatmap, which matches the size of the convolutional feature map, can be analyzed alongside the color contrast in the output image to evaluate the model's performance.

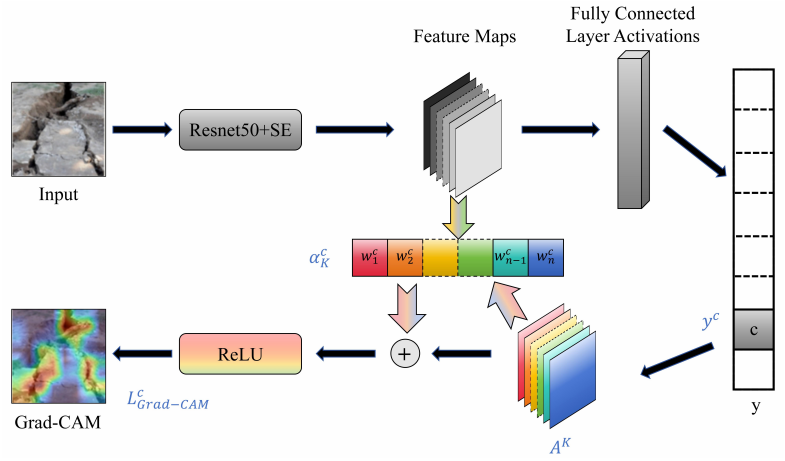

Figure 4. The mechanism of Grad-CAM.

3. Data and materials

3.1. Data acquisition and preprocessing

The large concrete crack dataset used for model pretraining is publicly available (https://www.kaggle.com/) and contains 10,000 positive samples (with cracks) and 10,000 negative samples (without cracks) in RGB format. As illustrated in Figure 5, the concrete cracks in the dataset exhibit clear edges, minimal interference, and distinct colors, making it well-suited for machine learning to extract high-level features of surface cracks. The datasets were split into training, validation, and test sets in the proportions of 0.7, 0.2, and 0.1, respectively.

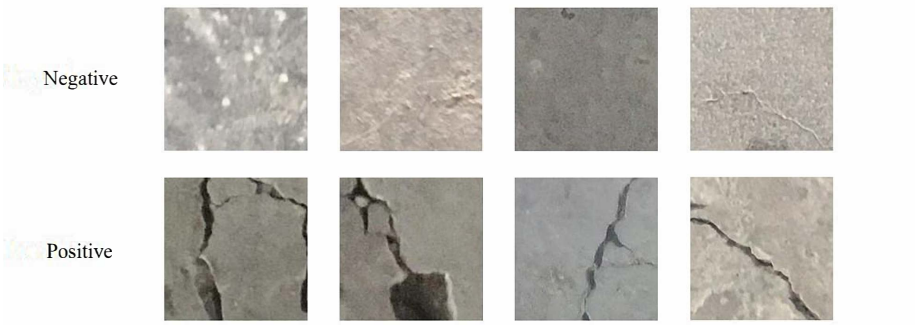

Figure 5. Examples of concrete crack images in the public dataset.

The landslide crack images used in this study were manually collected from online sources due to the absence of publicly available datasets. After conducting a manual review and excluding low-quality images, the final dataset consists of 100 positive images (with cracks) and 100 negative images (without cracks), as shown in Figure 6. All images were resized to a consistent resolution of 227×227 pixels, matching the resolution of the concrete crack dataset used in this study. The landslide crack images have the following characteristics: 1) The crack edges are often irregular or obscured by vegetation. 2) Some cracks are highly branched, with their lower ends filled with loose soil, and the color variation is minimal. 3) In the negative images, dark shadow areas resemble fine cracks, which may interfere with the model's decision-making. Additionally, landslide cracks exhibit complex patterns (e.g., tensile,

shear, and bulging cracks). Different patterns of landslide cracks exhibit distinct characteristics in images. To mitigate the potential impact of these variations on model training and prediction when the dataset is randomly split, the positive images in this study are visually categorized into two groups: single simple cracks (57 images) and multiple complex cracks (43 images). Both simple and complex crack types are divided into training, validation, and test sets using a stratified split in a 7:2:1 ratio. This approach helps preserve the diversity of crack patterns in each subset and avoids underrepresentation of minority classes, which is especially important given the limited dataset size. As a result, the final test set of the landslide crack image dataset contains a total of 22 images, including 11 images with landslide cracks and 11 images without landslide cracks.

Figure 6. Examples of landslide crack images in the dataset.

Data augmentation involves applying a series of transformation techniques to a limited dataset to create new data, reduce network overfitting, and enhance the generalization ability of training models. Due to the limited number of landslide images in the dataset, this study applied data augmentation methods to expand the training data by generating additional variations of existing samples. The data augmentation techniques employed include:

- **Random Rotation:** Rotation angles are multiples of 90 degrees, simulating multi-directionality while avoiding false features from interpolation artifacts in non-90-degree rotations.
- **Random Horizontal Flip:** Images are flipped horizontally with 50% probability, doubling the dataset size without bias and reducing storage needs.
- **Random Color Jitter:** Brightness, contrast, and saturation are adjusted within 20%, and hue is shifted by $[-0.1, 0.1]$, preventing overreliance on color features and mitigating environmental lighting effects.

As shown in Figure 7, these image augmentation techniques significantly expand the dataset while preserving the original features of the images, without introducing irrelevant details into the augmented data.

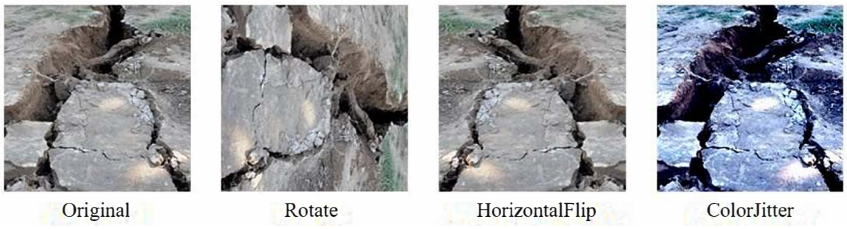

Figure 7. Demonstrations of the data augmentation of landslide cracks.

3.2. Training and evaluations

Model training was conducted on an Ubuntu system featuring an NVIDIA RTX4090/24G GPU, an Intel Xeon 8365A 12-core CPU, and 48GB of memory. PyTorch, along with CUDA 11.8 for GPU acceleration, was used, while TensorFlow served as the backend framework.

During training, the Adam optimizer was employed with automatic learning rate adjustment. The initial learning rate was set to 0.01 to optimize performance in the early stages. Additionally, cross-entropy loss (CEL) and a reduced learning rate on plateau (RLRP) were used alongside Adam. CEL measures the difference between predicted and actual probability distributions, providing strong gradient signals for early-stage convergence. When combined with a high learning rate, it helps achieve rapid model convergence. The RLRP dynamically adjusts the learning rate based on model performance, reducing it when improvement stalls to help escape local optima and find the global optimum. In this study, the decay factor for the RLRP was set to 0.1, with a patience of 5. The number of epochs was initially set to 500, but early stopping was applied based on performance metrics, and we observed that the model stabilized after approximately 100 epochs. Thus, we used 100 epochs for the final training. Hyperparameters such as batch size and learning rate were tuned using a grid search. We tested batch sizes of 8, 16, 32, and 64, and learning rates of 0.1, 0.01, 0.001, and 0.0001. The final hyperparameters were: for the CNN model, batch size = 8 and learning rate = 0.1; for the ResNet50 model, batch size = 16 and learning rate = 0.01.

The evaluation metrics used in this experiment include accuracy (*Acc.*), recall (*Rec.*), precision (*Pre.*), and F1 score ($F1$). *Acc.* represents the proportion of correctly classified samples out of the total, reflecting the model's overall prediction accuracy. *Rec.* assesses the model's ability to correctly identify positive samples, while *Pre.* measures the proportion of true positives among those predicted as positive. The $F1$ is the harmonic mean of accuracy and recall, providing a comprehensive evaluation of both metrics. The equations for these metrics are as follows:

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$Rec. = \frac{TP}{TP + FN} \tag{3.2}$$

$$Pre. = \frac{TP}{TP + FP} \tag{3.3}$$

$$F1 = 2 \times \frac{Pre. \times Rec.}{Pre. + Rec.} \tag{3.4}$$

where TP represents the number of correctly classified positive samples, TN denotes the number of correctly classified negative samples, FP is the number of negative samples misclassified as positive,

and *FN* is the number of positive samples misclassified as negative. Higher values of *TP* and *TN*, and lower values of *FP* and *FN*, indicate better model performance.

Additionally, receiver operating characteristic curves (ROC), area under the curve (AUC), and confusion matrices are used to assess the model's performance. The ROC curve illustrates the model's performance across all possible thresholds, while the AUC value indicates the probability that a randomly selected positive sample will receive a higher prediction score than a randomly selected negative sample. The confusion matrix provides insight into the model's prediction accuracy for different categories.

4. Results

4.1. Non-TL models

We initially trained the models directly on the landslide crack dataset, without employing thr TL strategy. This subsection evaluates the models' ability to perform the landslide crack classification task without any prior pretraining. The evaluation results for the non-TL models are presented in Table 1.

As shown in Table 1, the two models from the ResNet series significantly outperform the CNN model, highlighting the crucial role of structural improvements and increased model depth in enhancing the ability to identify landslide cracks. Deeper networks are better equipped to capture intricate features of the data, while residual connections help mitigate the vanishing gradient problem to some extent, facilitating more effective learning in deeper layers.

Table 1. The evaluation results of the non-TL models (subscript values indicate percentage improvement over the CNN baseline).

Model	*Rec.*	*Pre.*	*Acc.*	*F*1
CNN	0.667	0.889	0.773	0.762
ResNet50	0.833 $_{\uparrow 24.9\%}$	1.000 $_{\uparrow 12.5\%}$	0.909 $_{\uparrow 17.6\%}$	0.909 $_{\uparrow 19.3\%}$
ResNet50+SE	0.917 $_{\uparrow 37.5\%}$	1.000 $_{\uparrow 12.5\%}$	0.955 $_{\uparrow 23.5\%}$	0.957 $_{\uparrow 25.6\%}$

Both ResNet50 and ResNet50+SE models achieve a precision (*Pre.*) of 1.000, indicating perfect classification of images containing landslide cracks. However, the inclusion of the SE module increases the recall (*Rec.*) and F1 score of ResNet50 from 0.909 to 0.955 and from 0.909 to 0.957, respectively, significantly enhancing overall performance. The improvement in recall suggests that the SE module helps ResNet50 better capture subtle crack features, improving its ability to classify difficult-to-detect cracks and enhancing overall accuracy in landslide crack recognition.

The ROC curves and confusion matrices for the CNN, ResNet50, and ResNet50+SE models are presented in Figure 8. Among the three models, ResNet50+SE achieved the highest area under the ROC curve (AUC = 0.98), indicating a clear performance advantage with increased model complexity and the integration of attention mechanisms. Notably, the maximum discrepancy between TPR and FPR increases with model complexity: 0.65 for CNN, 0.83 for ResNet50, and 0.92 for ResNet50+SE. A higher discrepancy between true positive rate (TPR) and false positive rate (FPR) means the model is better at distinguishing true positives (landslide cracks) from false positives (non-crack images). This shows the model's ability to classify landslide cracks more accurately.

The confusion matrices further support this interpretation. The ResNet50+SE model correctly identified all Class 0 images (without cracks) and 11 out of 12 Class 1 images (with cracks), resulting in only a single false negative. In contrast, ResNet50 misclassified two Class 1 images, and the CNN model showed the weakest performance, with four false negatives and one false positive. These results demonstrate that the ResNet50+SE model, enhanced by the SE module, not only improves sensitivity to complex crack patterns but also maintains a high level of specificity, thus offering the most robust and reliable performance for landslide crack classification.

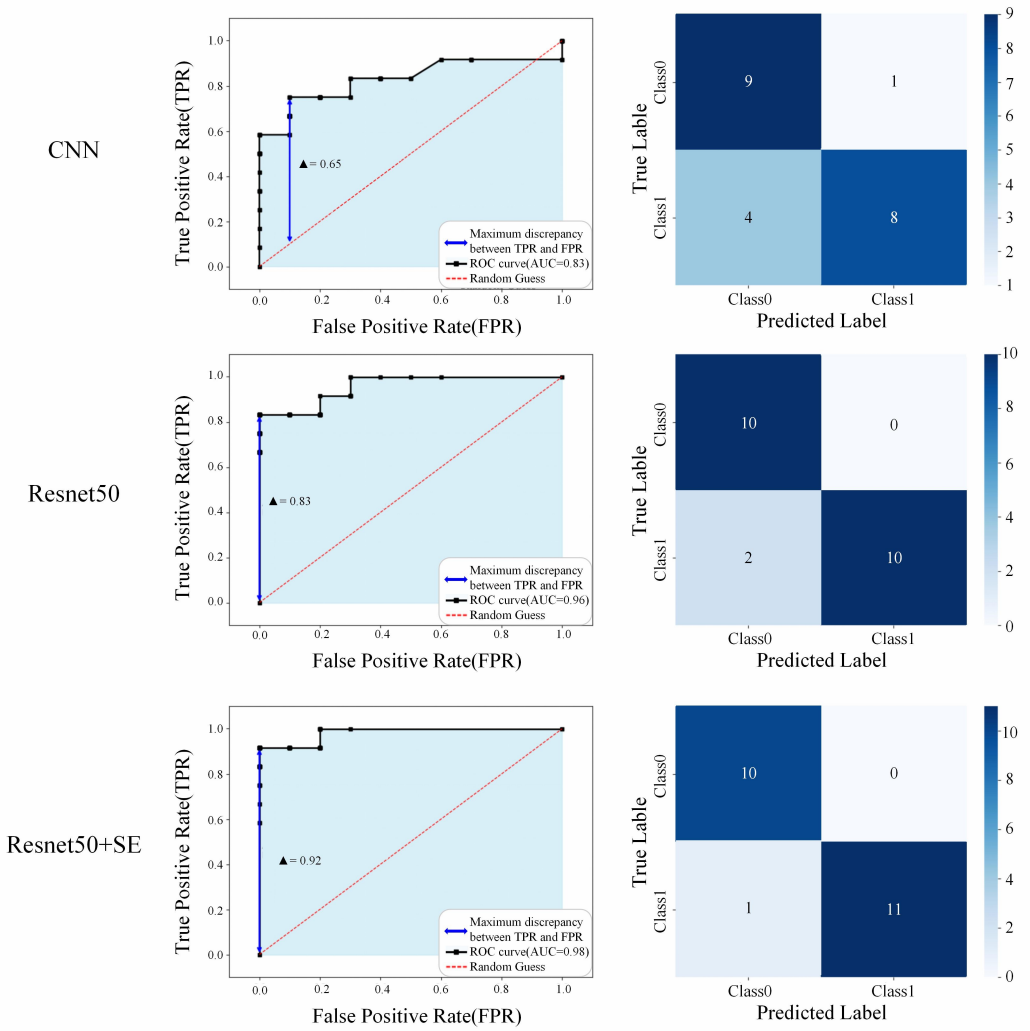

Figure 8. The ROC curve and confusion matrices of the non-TL models.

4.2. Models using TL

The performance of the three models is presented in Table 2.

As presented in Table 2, for CNN, TL slightly improves precision by helping the model learn from similar images, thus reducing false positives. However, recall remains low due to CNN's limited ability to detect complex features, causing many landslide crack images to be missed. ResNet50 with TL maintains high recall and perfect precision, with scores identical to the non-TL model, indicating

Table 2. The evaluation results of the TL models (subscript values indicate percentage improvement over the CNN baseline.)

Model	*Rec.*	*Pre.*	*Acc.*	*F*1
CNN	0.667	1.000	0.818	0.800
ResNet50	0.833 $_{\uparrow 24.9\%}$	1.000 $_{\uparrow 0.0\%}$	0.909 $_{\uparrow 11.1\%}$	0.909 $_{\uparrow 13.6\%}$
ResNet50+SE	1.000 $_{\uparrow 50.0\%}$	0.923 $_{\downarrow 7.7\%}$	0.955 $_{\uparrow 16.7\%}$	0.960 $_{\uparrow 20.0\%}$

that TL has minimal impact on its performance. ResNet50+SE with TL achieves perfect recall (1.000) but a slight decrease in precision, suggesting the model may classify some non-crack images as cracks. Nevertheless, the improvement in recall significantly enhances the model's ability to detect landslide cracks. In this study, recall is prioritized to ensure the model captures as many true positives (landslide cracks) as possible, which is critical for a safer landslide risk analysis.

The results in Figure 9 demonstrate a clear improvement in model performance when TL was applied. All models showed increases in AUC scores, with the CNN model improving from 0.83 to 0.93, Resnet50 from 0.96 to 0.97, and Resnet50+SE achieving a perfect score of 1.00, up from 0.98. Confusion matrix analysis revealed fewer misclassifications in the CNN and ResNet50+SE with TL strategies, particularly in Resnet50+SE, which successfully distinguished all the images with landslide cracks.

4.3. Model interpretability

We calculated attention heatmaps for different TL models based on Grad-CAM. The results, as illustrated in Figure 10, reveal notable differences in the model's ability to focus on the cracks in both single and multiple crack scenarios. The CNN model demonstrated a fundamental level of crack identification. The heatmaps generated by this model showed its ability to identify the general location of cracks, but with limited accuracy. Specifically, the heatmaps lacked precision, often overlooking critical details and even displaying diffuse attention areas that failed to delineate the boundaries of the cracks. This suggests that the CNN model is highly susceptible to environmental factors, such as lighting conditions. In scenarios where the lighting is dim or the contrast between the cracks and the surrounding terrain is subtle, the model tends to misinterpret non-cracked dark regions as cracks, leading to misidentifications and overlapping attention areas.

In contrast, the ResNet50 model, a deeper architecture with residual connections, outperformed the CNN in crack identification. The heatmaps generated by ResNet50 showed more concentrated attention on the cracks, with clearer demarcations of crack regions. This model demonstrated a more effective focus on single cracks, marking an improvement over the CNN in crack localization. However, despite its improved performance, ResNet50 still exhibited limitations in capturing the intricate details of mutiple cracks. Specifically, the attention focused on key areas was smaller compared to that of ResNet50+SE. The high-attention points, marked in red, were less concentrated in critical regions, indicating that ResNet50's ability to emphasize the most important areas was not as refined as that of the ResNet50+SE model.

The ResNet50+SE model further enhanced crack identification by incorporating SE blocks, which assigned higher weights to feature channels associated with cracks and edges. The resulting heatmaps

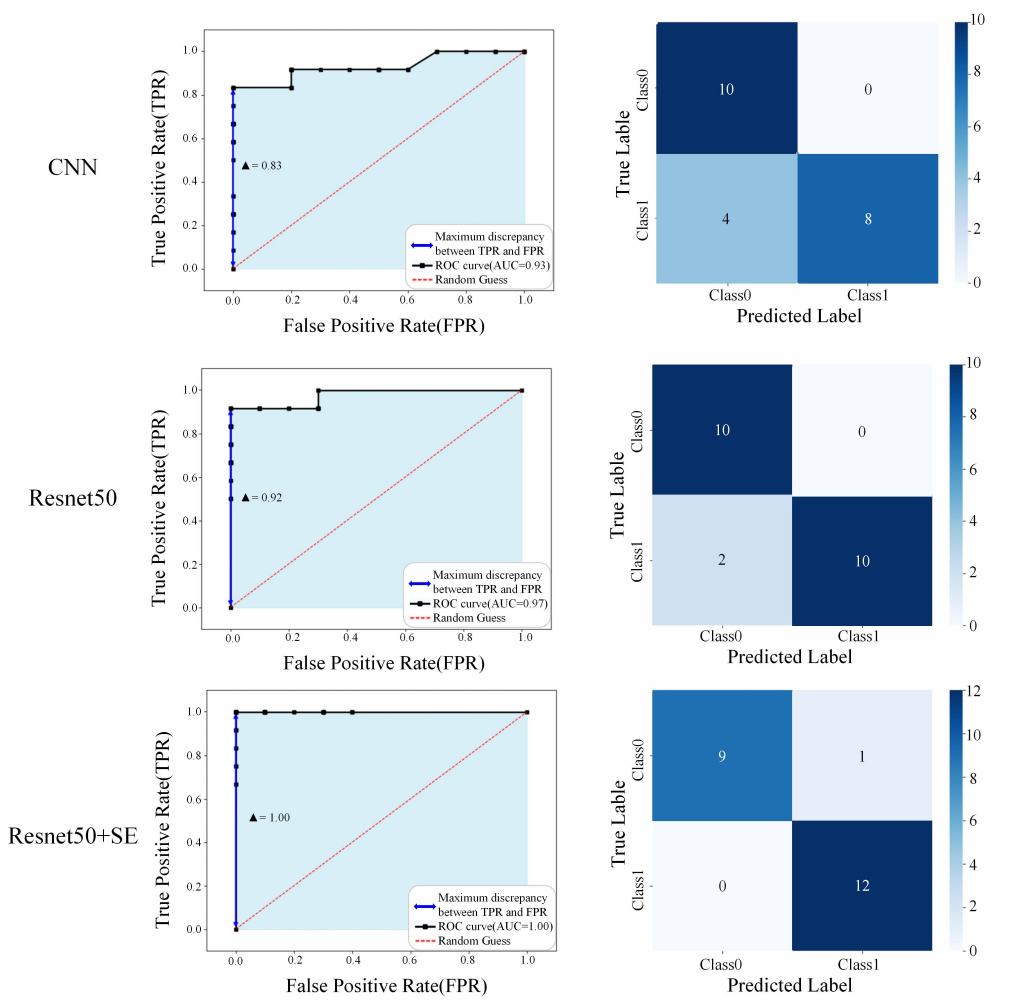

Figure 9. The ROC curve and confusion matrices of the TL models.

were the most precise, with the red areas nearly entirely covering single cracks, and the blue areas extending to the crack edges. This indicates that ResNet50+SE exhibits a strong ability to recognize single, clearly defined cracks. Additionally, in the case of multiple crack images, where fine crack branches—difficult for the previous two models to detect—are present, ResNet50+SE effectively identifies these finer cracks and allocates focused attention to them. In both single and multiple crack scenarios, ResNet50+SE demonstrated superior crack localization, showcasing its enhanced precision and ability to handle complex crack structures.

5. Discussion

In the context of landslide crack classification, data imbalance presents a significant challenge, particularly with regard to the practical application of machine learning models. While our previous research developed a relatively balanced dataset consisting of 100 images containing landslide cracks and 100 images without, the real-world process of image acquisition often results in substantial class imbalance. For example, when large-scale landslide surface images captured by drones are

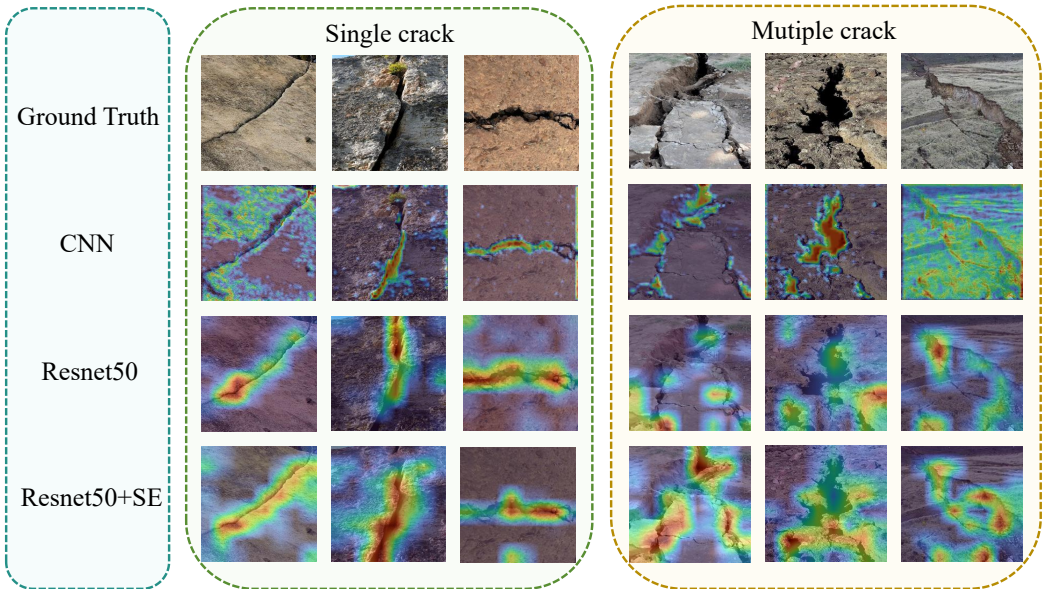

Figure 10. The heatmap of the TL models based on Grad-CAM.

segmented into smaller patches, the majority of these resulting images tend to lack cracks, making crack-containing images a rare occurrence within the dataset.

This imbalance, wherein images containing cracks become the minority class, creates a major obstacle for classification tasks. The over-representation of the majority class in such datasets can dominate the learning process, leading to models that exhibit a bias toward predicting the majority class. Consequently, the model's ability to accurately identify the minority class—images containing cracks—becomes compromised, resulting in a higher rate of misclassification. This issue is particularly critical in landslide risk analysis, where failure to correctly identify cracks could have serious safety and preventive implications. Thus, this section further explores the impact of data imbalance on the classification of landslide cracks, examining how this imbalance affects model performance and underscoring the importance of addressing it to ensure more reliable and accurate outcomes.

Figure 11 illustrates the performance metrics of the ResNet50+SE model when trained on datasets with varying proportions of positive samples (images containing cracks) and negative samples (images without cracks). The datasets are increasingly imbalanced, with the percentage of positive samples reduced from 100% to 75%, 50%, and 25%. The figure evaluates the model's performance across three metrics: F1 score, precision, and recall.

As shown in Figure 11, the proportion of positive samples significantly impacts the model's recall. As the percentage of positive samples (images containing cracks) decreases, recall decreases sharply. This decline suggests that with fewer positive samples, the model becomes more prone to misclassifying images that contain cracks as crack-free. This issue is particularly concerning in landslide risk analysis, where the accurate identification of cracks is critical. The reduction in positive samples complicates the model's ability to learn and distinguish the features of the minority class, resulting in an increase in missed identifications. The problem becomes particularly pronounced when the proportion of positive samples drops to 25%, with recall falling below 0.6.

Interestingly, as the proportion of positive samples decreases (from 75% to 25%), precision increases from 0.8 to 1. When the proportion of positive samples is reduced to 25%, the majority of

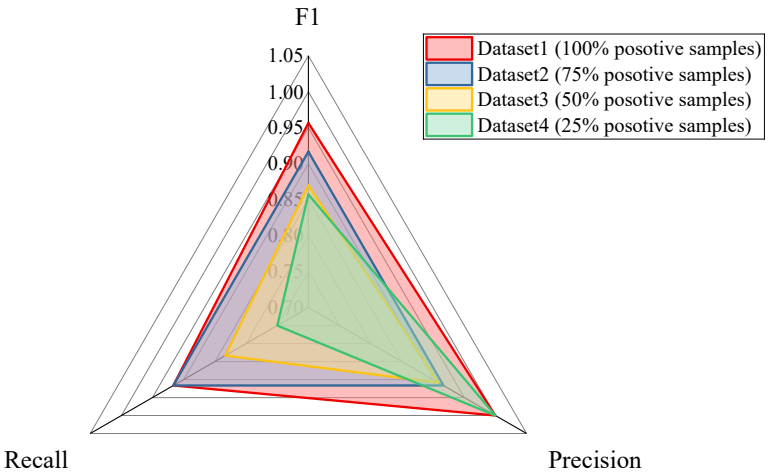

Figure 11. The evaluation results of the models using 100%, 75%, 50%, and 25% positive samples.

the dataset consists of negative samples. Consequently, the model tends to adopt a more conservative approach when classifying positives, such as raising the classification threshold. This adjustment helps the model reduce the number of false positives (i.e., incorrectly predicting a negative sample as positive), which leads to an increase in precision, but at the cost of a decrease in recall. This rise in precision is a result of the model shifting its decision boundary due to the scarcity of positive samples, making it more heavily influenced by the statistical patterns of the majority class. It is crucial to note that this increase in precision does not signify an actual improvement in model performance; rather, it reflects a change in the model's behavior driven by the dataset imbalance.

The F1 score, which represents the harmonic mean of precision and recall, provides a comprehensive measure of the model's overall performance. As the proportion of positive samples decreases, the F1 score also declines, signaling a deterioration in overall model performance. However, the compensatory effect of precision helps to alleviate the decline in F1 to some extent, making the reduction in F1 less pronounced than that of recall. Nonetheless, the overall trend underscores the significant challenge posed by class imbalance and emphasizes the need for strategies that address this issue. It should be noted that imbalance mitigation techniques such as the synthetic minority oversampling technique (SMOTE) and focal loss could improve model performance [49], particularly by enhancing recall, which is crucial for practical applications in landslide crack detection. Future work could explore these techniques and their impact on the recall/precision trade-off.

Moreover, the relatively small size of the landslide crack dataset may constrain the generalization ability of the proposed model. Future research could address this by expanding the dataset through further data collection, synthetic image generation techniques (e.g., generative adversarial networks, GANs), or enhanced data augmentation to improve diversity. Additionally, although transfer learning from a concrete crack dataset helps mitigate data scarcity, a domain gap remains due to differences in texture, background, and morphology between concrete and landslide cracks. This gap may influence feature transferability and model performance, suggesting the need for domain adaptation strategies in future work. While we adopted a fixed stratified split to preserve crack type diversity across subsets, we acknowledge that k-fold cross-validation could offer stronger statistical robustness and plan to explore

it in future work as larger datasets become available. Additionally, due to the limited dataset size, we adopted a partial unfreezing strategy for transfer learning instead of full fine-tuning. Future studies could explore alternative unfreezing schemes as more data becomes available. Last but not least, the impact of the compression ratio (r-value) in the SE module showed minimal effect on key performance metrics in our study. However, in other similar scenarios, the r-value could be an important factor worth exploring further.

Beyond performance metrics, the proposed landslide crack classification method offers practical value for hazard mitigation. Accurate and timely crack detection can trigger early warnings, enabling targeted monitoring and preventive measures to reduce risks. Real-world cases, such as the Shaziba landslide in China [50], show that early intervention can prevent fatalities and economic losses. Integrating our method into monitoring frameworks, especially in data-scarce regions, could enhance proactive risk management and public safety.

6. Conclusion

In this study, we proposed a deep learning-based approach to classify landslide crack images by integrating the SE attention mechanism, TL strategy, and a model interpretability method based on Grad-CAM.

Our results demonstrate that the inclusion of the SE module significantly enhanced the model's ability to focus on subtle features of landslide cracks, leading to an overall improvement in performance. The proposed ResNet50+SE model outperformed both the CNN and ResNet50 architectures, particularly in terms of recall, which is essential for comprehensive identification of landslide cracks. Additionally, the integration of TL further optimized performance by allowing the model to leverage knowledge from a pretrained concrete crack dataset. This strategy improved classification accuracy despite the limited availability of landslide crack images.

The analysis of model interpretability through Grad-CAM provided valuable insights into the decision-making process of the models. The heatmaps generated by ResNet50+SE effectively highlighted the crack regions, offering transparency in the model's focus areas. This interpretability is vital for practical applications, where understanding the reasoning behind classification decisions is crucial for ensuring safety and reliability in landslide risk analysis.

In conclusion, the ResNet50+SE model, enhanced by TL, attention mechanisms, and interpretability, offers a robust and effective solution for landslide crack classification, improving both performance and transparency. Future work will focus on optimizing the model's ability to handle imbalanced datasets, exploring domain adaptation and multi-spectral imaging to mitigate environmental factors, and enhancing robustness for real-world applications, particularly in overcoming limitations such as sensitivity to lighting, shadows, vegetation occlusion, and generalization to unseen terrains or imaging modalities.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 424007081), the Natural Science Foundation of Jiangsu Province (No. BK20220421), the Jiangsu Province College Students Innovation Training Program (No. 202410298142Y), and the Open Foundation from the Research and Development Center of Transport Industry of New Generation of Artificial Intelligence Technology (No. 202403H).

Conflict of interest

None of the authors have any relevant financial or non-financial competing interests.

References

1. Tianzheng Li, Limin Zhang, Wenping Gong, et al. (2024) Initiation mechanism of landslides in cold regions: role of freeze-thaw cycles. *Int J Rock Mech Min* 183: 105906.
2. Faming Huang, Ronghui Li, Filippo Catani, et al. (2024) Uncertainties in landslide susceptibility prediction: Influence rule of different levels of errors in landslide spatial position. *J Rock Mech Geotech Eng* 16: 4177–4191.
3. Haijia Wen, Fangyi Yan, Junhao Huang, et al. (2025) Interpretable machine learning models and decision-making mechanisms for landslide hazard assessment under different rainfall conditions. *Expert Syst Appl* page 126582.
4. Qi Ge, Hongyue Sun, Zhongqiang Liu, et al. (2022) A novel approach for displacement interval forecasting of landslides with step-like displacement pattern. *Georisk* 16: 489–503.
5. Yu Lei, Jinsong Huang, Yifei Cui, et al. (2023) Time capsule for landslide risk assessment. *Georisk* 17: 613–634.
6. Guotao Ma, Mohammad Rezania, Mohaddeseh Mousavi Nezhad, et al. (2024) Multivariate copula-based framework for stochastic analysis of landslide runout distance. *Reliab Eng Syst Saf* 250: 110270.
7. Te Xiao and Li-Min Zhang (2023) Data-driven landslide forecasting: Methods, data completeness, and real-time warning. *Eng Geol* 317: 107068.
8. Desheng Zhu, DV Griffiths, Gordon A Fenton, et al. (2025) Probabilistic stability analyses of two-layer undrained slopes. *Comput Geotech* 182: 107178.
9. Himanshu Rana and GL Sivakumar (2024) Probabilistic back analysis for rainfall-induced slope failure using mls-svr and bayesian analysis. *Georisk* 18: 107–120.
10. ShuiHua Jiang, Hong-Hu Jie, Jiawei Xie, et al. (2024) Probabilistic back-analysis of rainfall-induced landslides for slope reliability prediction with multi-source information. *J Rock Mech Geotech Eng* 16: 3575–3594.
11. Junrong Zhang, Huiming Tang, Qinwen Tan, et al. (2024) A generalized early warning criterion for the landslide risk assessment: deformation probability index (dpi). *Acta Geotechnica* 19: 2607–2627.

12.Fausto Guzzetti, Massimo Melillo, Michele Calvello, et al. (2024) Independent demonstration of a deep-learning system for rainfall-induced landslide forecasting in italy. *Landslides* 21: 2171–2178.

13.Qi Cui, Lulu Zhang, Xiangyu Chen, et al. (2022) Quantitative risk assessment of landslides with direct simulation of pre-failure to post-failure behaviors. *Acta Geotechnica* 17: 4497–4514.

14.Xuanmei Fan, Qiang Xu, Jie Liu, et al. (2019) Successful early warning and emergency response of a disastrous rockslide in guizhou province, china. *Landslides* 16: 2445–2457.

15.Minghao Miao, Huiming Tang, Kun Fang, et al. (2025) Influence of tensile crack development on the deformation behavior and failure mode of reservoir-induced landslides: insights from model tests. *Landslides* 1–16.

16.Li Wang, Yushan Chen, Shimei Wang, et al. (2025) Response of landslide deformation to rainfall based on multi-index monitoring: a case of the tanjiawan landslide in the three gorges reservoir. *B Eng Geol Environ* 81: 231.

17.Qi Ge, Jin Li, Suzanne Lacasse, et al. (2024) Data-augmented landslide displacement prediction using generative adversarial network. *J Rock Mech Geotech Eng* 16: 4017–4033.

18.Jia-Xing Chen, Han-Dong Liu, Zhi-Fei Guo, et al. (2024) Research on failure mechanism of landslide with retaining-wall-like locked segment and instability prediction by inverse velocity method. *Sci Rep* 14: 21359.

19.Zhongqiang Liu, Suzanne Lacasse, Luqi Wang, et al. (2024) Deformation triggers and stability evolution of landslide. *Spatial Modelling and Failure Analysis of Natural and Engineering Disasters through Data-Based Methods, volume III* 29.

20.Yueming Yin, Qinglu Deng, Weibo Li, et al. (2023) Insight into the crack characteristics and mechanisms of retrogressive slope failures: A large-scale model test. *Eng Geol* 327: 107360.

21.Yifei Gong, Aijun Yao, Yanlin Li, et al. (2022) Classification and distribution of large-scale high-position landslides in southeastern edge of the qinghai–tibet plateau, china. *Environ Earth Sci* 81: 311.

22.Laxman Kafle, Wen-Jie Xu, Shu-Yuan Zeng, et al. (2022) A numerical investigation of slope stability influenced by the combined effects of reservoir water level fluctuations and precipitation: A case study of the bianjiazhai landslide in china. *Eng Geol* 297: 106508.

23.Yuting Yang and Gang Mei (2021) Deep transfer learning approach for identifying slope surface cracks. *Appl Sci* 11: 11193.

24.Minh-Vuong Pham, Yong-Soo Ha, Yun-Tae Kim (2023). Automatic detection and measurement of ground crack propagation using deep learning networks and an image processing technique. *Measurement* 215: 112832.

25.Bo Deng, Qiang Xu, Xiujun Dong, et al. (2024) Automatic method for detecting deformation cracks in landslides based on multidimensional information fusion. *Remote Sens* 16: 4075.

26.Ionut Sandric, Zenaida Chitu, Viorel Ilinca, et al. (2024) Using high-resolution uav imagery and artificial intelligence to detect and map landslide cracks automatically. *Landslides* 21: 2535–2543.

27.Remzi Eker, Abdurrahim Aydın, Tolga Görüm (2024) Tracking deformation velocity via psi and sbas as a sign of landslide failure: an open-pit mine-induced landslide in himmetoğlu (bolu, nw turkey). *Nat Hazards* 120: 7701–7724.

28. Konstantinos G Nikolakopoulos, Aggeliki Kyriou, Ioannis K Koukouvelas, (2023) Uav, gnss, and insar data analyses for landslide monitoring in a mountainous village in western greece. *Remote Sens* 15: 2870.

29. Hongwei Huang, Shuai Zhao, Dongming Zhang, et al. (2022). Deep learning-based instance segmentation of cracks from shield tunnel lining images. *Struct Infrastruct Eng* 18: 183–196.

30. Xiong Peng, Xingu Zhong, Chao Zhao, et al. (2021) A uav-based machine vision method for bridge crack recognition and width quantification through hybrid feature learning. *Constr Build Mater* 299: 123896.

31. Honghui Wang, Donglin Nie, Xianguo Tuo, et al. (2020). Research on crack monitoring at the trailing edge of landslides based on image processing. *Landslides* 17: 985–1007.

32. Zhan Cheng, Wenping Gong, Huiming Tang, et al. (2021) Uav photogrammetry-based remote sensing and preliminary assessment of the behavior of a landslide in guizhou, china. *Eng Geol* 289: 106172.

33. Qi Ge, Jin Li, Xiaohong Wang, et al. (2024). Litetransnet: An interpretable approach for landslide displacement prediction using transformer model with attention mechanism. *Eng Geol* 331: 107446.

34. Husam AH Al-Najjar, Biswajeet Pradhan, Ghassan Beydoun, et al. (2023) A novel method using explainable artificial intelligence (xai)-based shapley additive explanations for spatial landslide prediction using time-series sar dataset. *Gondwana Res* 123: 107–124.

35. Haojie Wang, Lin Wang, Limin Zhang (2023). Transfer learning improves landslide susceptibility assessment. *Gondwana Res* 123: 238–254.

36. Zeyang Zhao, Tao Chen, Jie Dou, et al. (2024) Landslide susceptibility mapping considering landslide local-global features based on cnn and transformer. *IEEE J Sel Top Appl Earth Observ Remote Sens*.

37. Teruyuki Kikuchi, Koki Sakita, Satoshi Nishiyama, et al. (2023). Landslide susceptibility mapping using automatically constructed cnn architectures with pre-slide topographic dem of deep-seated catastrophic landslides caused by typhoon talas. *Nat Hazards* 117: 339–364.

38. Luqi Wang, Lin Wang, Wengang Zhang, et al. (2024). Time series prediction of reservoir bank landslide failure probability considering the spatial variability of soil properties. *J Rock Mech Geotech Eng* 16: 3951–3960.

39. Chongzhi Wu, Li Hong, Lin Wang, et al. (2023). Prediction of wall deflection induced by braced excavation in spatially variable soils via convolutional neural network. *Gondwana Res* 123: 184–197.

40. Mohammad Razavi, Samira Mavaddati, Hamidreza Koohi (2024). ResNet deep models and transfer learning technique for classification and quality detection of rice cultivars. *Expert Syst Appl* 247: 123276.

41. Bilal Aslam, Adeel Zafar, Umer Khalil (2023). Comparative analysis of multiple conventional neural networks for landslide susceptibility mapping. *Nat Hazards* 115: 673–707.

42. Masoud Mahdianpari, Bahram Salehi, Mohammad Rezaee, et al. (2018) Very Deep Convolutional Neural Networks for Complex Land Cover Mapping Using Multispectral Remote Sensing Imagery. *Remote Sens* 10: 1119.

43.Zhiyong Fu, Li Changdong, Yao Wenmin (2023) Landslide susceptibility assessment through tradaboost transfer learning models using two landslide inventories. *Catena* 222: 106799.

44.Zhihao Wang, Alexander Brenning (2023). Unsupervised active–transfer learning for automated landslide mapping. *Comput Geosci* 181: 105457.

45.Hyungsik Jung, Youngrock Oh (2021) Towards better explanations of class activation mapping. In *Proc IEEE/CVF international conference on computer vision* 1336–1344.

46.Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, et al. (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* 618–626.

47.Shuihua Wang, Yudong Zhang (2023). Grad-cam: understanding ai models. *Comput Mater Contin* 76: 1321–1324.

48.Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626.

49.Junhao Huang, Haijia Wen, Jiwei Hu, et al. (2025) Deciphering decision-making mechanisms for the susceptibility of different slope geohazards: A case study on a smote-rf-shap hybrid model. *J Rock Mech Geotech Eng* 17: 1612–1630.

50.Jian Wang, Xinli Hu, Hongchao Zheng, et al. (2024) Energy transfer mechanisms of mobility alteration in landslide-debris flows controlled by entrainment and runout-path terrain: A case study. *Landslides* 21: 1189–1206.

AIMS Press