Journal of Al-Turath University College (JATUC)
Open Access Journal
ISSN: 2074-5621
DOI: https://doi.org/10.63964/JATUC.43.2.2026.46
https://jatuc.uoturath.edu.iq/

Early Crack Detection in Civil Infrastructure Using Multi-Threshold Evaluation of Fine-Tuned ResNet Backbones on SDNET2018

Hadeel Talib Mohammed Jawad[1*], Muhanad Tahrir Younis[2]

[1] Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, Iraq.
[2] Department of Artificial Intelligence, College of Science, Mustansiriyah University, Baghdad, Iraq.

[*]*Correspondence email*: hadeeltalib.altalib@uomustansiriyah.edu.iq

KEYWORDS

Crack Detection; SDNET2018; ResNet; Fine-Tuning; Multi-Threshold Evaluation; PR-AUC; Gate-to-Segmentation.

ABSTRACT

Early detection of cracks on civil infrastructure is vital for implementing preventive maintenance measures. In this paper, five different ImageNet pretrained models of the residual architecture ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNetRS-101 are benchmarked for their performance in classifying crack images within the SDNET2018 dataset. Each model is trained using two different training procedures, namely training with a frozen backbone (NOFT) and training with end-to-end fine-tuning (FT). Due to the need for false positive and false negative rates to be considered in addition to the classification accuracy, a multi-threshold analysis for both metrics was conducted. The results of this study indicate that fine-tuning improves the classification performance of all five model backbones. Furthermore, the fine-tuned ResNet-34 model exhibits the best overall performance. At a threshold of 0.60 on the fixed test set of the SDNET2018 dataset (n = 8414), the model achieves an accuracy of 96.39%, precision of 87.58%, recall of 88.68%, and an F1-score of 88.13%, in addition to area under the ROC curve and area under the precision-recall curve values of 0.9805 and 0.9411, respectively. Furthermore, the model indicates 160 false positives and 144 false negatives at this threshold. Overall, then, the fine-tuned ResNet-34 model is selected as an efficient model to be utilized as a classification gate for early crack detection prior to performing crack segmentation.

1. INTRODUCTION

Cracking is one of the earliest signs of deterioration in many types of civil infrastructure built with concrete and asphalt-based materials. Inspecting these infrastructures for cracks is a manual process that is inherently slow and subjective. Thus, automated methods for detecting and analyzing cracks in civil infrastructure are of interest and growing in importance [1-4]. Figure 1 shows some of the different ways that cracks can appear within civil infrastructure, as well as where they are some of the most common locations. These representative examples include some of the different textures and backgrounds of these infrastructures.

Fig. 1. Representative crack morphologies and common locations in civil infrastructure (roads/pavements, bridges/decks, and building walls/columns)

In this work, the SDNET2018 data set is adopted as a controlled benchmark for binary crack classification across different concrete surfaces. Several different deep learning models, with and without fine-tuning, are assessed on the benchmark using threshold-free and threshold-dependent measures. Since the occurrence of missed cracks is more costly than false crack detections, false negative (FN) counts are of particular interest, as are the trade-offs between FN and false positive (FP) counts [5]. Deep convolutional neural network models have the potential to learn effective representations of crack features and avoid the need for hand-engineered features. In practice, however, the available crack images can contain a variety of different features such as illumination variations, scale differences, different background appearances, and class distribution issues. Therefore, it is necessary to use fixed splits and other preprocessing steps for the models to be fair and meaningful to evaluate beyond a single operating point [1], [5-9]. Furthermore, the problem of detecting cracks in concrete can be difficult in that there are various appearances in concrete that can resemble cracks. Recent deep learning approaches to classification have found success using transfer learning, adaptation techniques, and preprocessing steps to improve the robustness of the models in the presence of these issues, particularly when using models adapted from a general domain like concrete cracking, rather than using the models without domain-specific tuning [6]- [9], [12-15]. Thus, the SDNET2018 benchmark is used in this paper as a reference point for stage-1 crack screening. More specifically, the study is focused upon a fair and reproducible protocol for assessing various deep learning models for crack detection, and reports upon the performance of a fine-tuned ResNet-34 model as a suitable candidate for the stage-1 screening process prior to more detailed stage-2 analysis of the detected cracks. This model choice is based upon the evidence that deep learning models based upon residual blocks have competitive performance in the problem of crack detection when fine-tuned using transfer learning approaches [10], [11]. The remainder of this paper is organized as follows. Section 2 reviews the related work on crack classification and deep learning-based crack analysis. Section 3 presents the problem formulation and the contributions of the present work. Section 4 describes the SDNET2018 data set and the preprocessing steps performed on the data. Section 5 describes the research methodology and experiments performed. Section 6 introduces the various ResNet backbone models evaluated in this work. Section 7 describes the training protocol for fine-tuning the ResNet backbone models. Section 8 presents the procedure for evaluating the models at various thresholds. Section 9 defines the evaluation metrics used in the study. Section 10 presents the results and discussion of the experiment. Section 11 presents the analysis of the model selection. Finally, Section 12 concludes the paper and presents possible future work.

2. RELATED WORK

While there is considerable literature on the topic of pixel-level crack segmentation, image-level classification of crack and non-crack images remains a valuable first-line screening tool for rejecting obvious non-crack images early in the process. Transfer learning strategies using pre-trained deep learning models, including AlexNet [6], EfficientNet [7], MobileNet [8], Inception [9], VGG [10], ResNet [11], and, more recently, CNN-transformer models [12-16] have been explored for the detection of crack and non-crack images, as well as for the determination of crack orientation. Similarly, models that are specifically built with the intention of segmenting cracks from full concrete images have also been explored and published in the literature [17-25]. Furthermore, models that are specifically built with the intention of monitoring for crack progression over time, such as those based on multiresolution analysis, have also been explored [26]. Despite the growing literature on these topics, however,

there are still some limitations to existing methods. For instance, many studies that investigate deep learning models for crack detection employ different deep learning models, datasets, and settings for their studies, making it difficult to directly compare the results of different studies. Furthermore, deep learning model outputs are often evaluated with a focus on determining the overall accuracy and F1-score of the model, without analyzing the false positive and false negative rates of the model. Finally, few studies have evaluated the performance of multiple residual deep learning backbone models (with and without fine-tuning) on the same dataset. In this paper, therefore, we aim to address these knowledge gaps by presenting a study that evaluates different types of deep learning residual models under the same settings and protocols on the SDNET2018 dataset.

3. PROBLEM FORMULATION AND CONTRIBUTIONS

In this paper, the problem of early crack screening is modeled as a binary classification problem, where the classifier has to determine whether the given input image x contains a crack or not. Let f(x) be the output of the classifier for input image x. The classifier estimates the probability that the image contains a crack: $p = Pr(y = 1 | x)$. The class label y takes the value 1 if the image contains a crack and 0 if it does not contain a crack. In practice, however, it is often difficult to obtain images that contain cracks due to the low contrast between the crack and the remainder of the infrastructure, the irregularity of the crack, shadows, stains on the infrastructure, cracks in the background of the infrastructure, and various other factors that may be likely to confuse the classifier. Furthermore, existing benchmark datasets tend to be unbalanced between the number of images that contain cracks and those that do not. Finally, in practice, it is more important to avoid false negatives than false positives in crack detection: false negative results would mean that maintenance on the infrastructure is not performed when it would be needed due to the presence of a crack, while false positive results merely lead to unnecessary maintenance efforts. Thus, the classification problem cannot be assessed using only a single score of the classifier at a threshold. Rather, the trade-off between the number of false positives and false negatives should be examined in order to determine the strategy that can be used for screening the infrastructure in a way that is both reliable and practical. The main contributions of this paper are as follows:

- A standardized evaluation protocol is established for the binary crack classification task on the SDNET2018 dataset.
- Multiple types of residual neural network backbones are evaluated under two training regimes to study the impact of domain adaptation on crack screening performance.
- A method for evaluating models across multiple classification thresholds is developed to assess the trade-off between false-positive and false-negative predictions.
- Based on the experiments conducted, a suitable backbone architecture is identified for stage-1 crack screening before performing stage-2 crack analysis.

4. DATASET AND PREPROCESSING

SDNET2018 is a widely used dataset for the detection and classification of cracks in concrete. The dataset was developed for use with image-based learning methods and comprises image patches collected from three different categories of concrete surfaces: reinforced bridge decks, walls, and unreinforced pavements [5]. Each of these surface types contains both cracked and non-cracked samples of concrete.

Table 1. Original SDNET2018 dataset composition by surface type

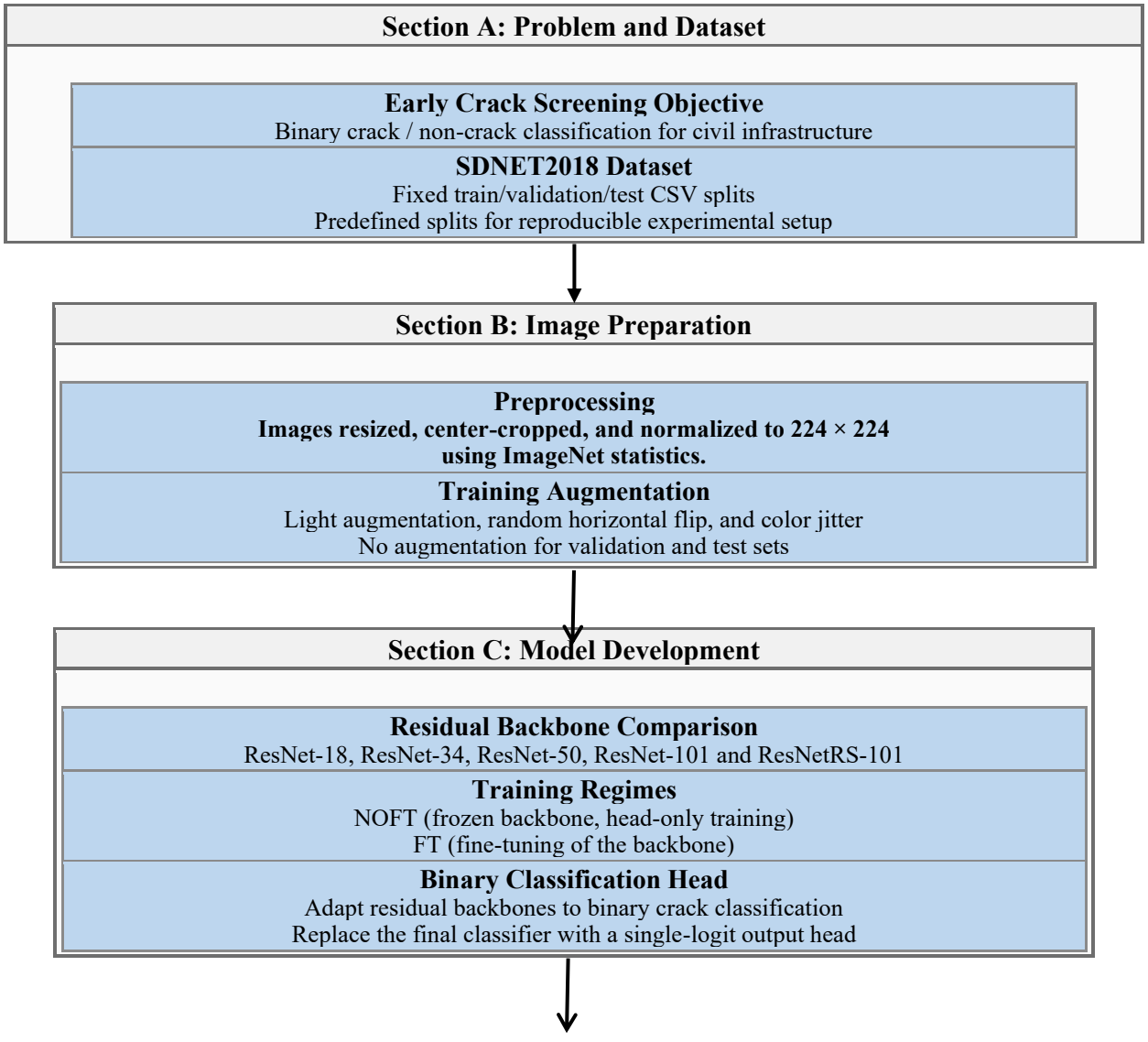
Surface type	Cracked	Non-cracked	Total
Reinforced Bridge Deck	2,025	11,595	13,620
Wall	3,851	14,287	18,138
Unreinforced Pavement	2,608	21,726	24,334
Total	**8,484**	**47,608**	**56,092**

Table 1 illustrates the statistics of the SDNET2018 dataset. The dataset contains 56,092 image patches in total, including 13,620 bridge-deck images, 18,138 wall images, and 24,334 pavement images. These images include 8,484 cracked patches and 47,608 non-cracked patches. These images were taken under realistic conditions and divided into 256 × 256 patches with crack widths between 0.06 and 25 mm. The dataset includes common visual challenges to crack detection, such as shadows, surface roughness, scaling, edges, holes, and debris in the background of the patches [5]. A test set of 8414 samples was used in this study. All images were resized to 224 × 224 pixel images and normalized using ImageNet statistics. During training, the images were randomly cropped to

224 × 224, horizontally flipped with probability 0.5, and color jittered with probability 0.5. For evaluation, deterministic preprocessing was used.

5. RESEARCH METHODOLOGY

This study presents an experimental study that evaluates the performance of deep residual backbones for stage-1 crack screening on the SDNET2018 dataset. The purpose of this study is to compare the performance of deep learning models that use different deep residual backbones for the task of identifying whether an asphalt pavement sample contains a crack or not. Furthermore, the study investigates the effects of fine-tuning the deep residual backbones and using different thresholds in the classification of asphalt samples. The experimental conditions that are used for the evaluation of each model are all the same in order to enable a fair comparison between their performances. Each of the deep learning models that are assessed in this study uses the same data splits, preprocessing steps, and evaluation procedure. The use of deep residual backbones that are pretrained on other datasets and that are fine-tuned to the SDNET2018 dataset follows the general principle of transfer learning that has been established in many computer vision tasks [10], [11]. The overall methodology that is employed in this study is illustrated in Fig. 2. The proposed method was implemented as a complete end-to-end screening system rather than just as a single experiment to evaluate deep learning models with various deep residual backbones. The framework for the method is shown in the figure and consists of several connected stages from the initial preparation of the dataset to the final selection of the best deep learning model for stage-1 crack screening.



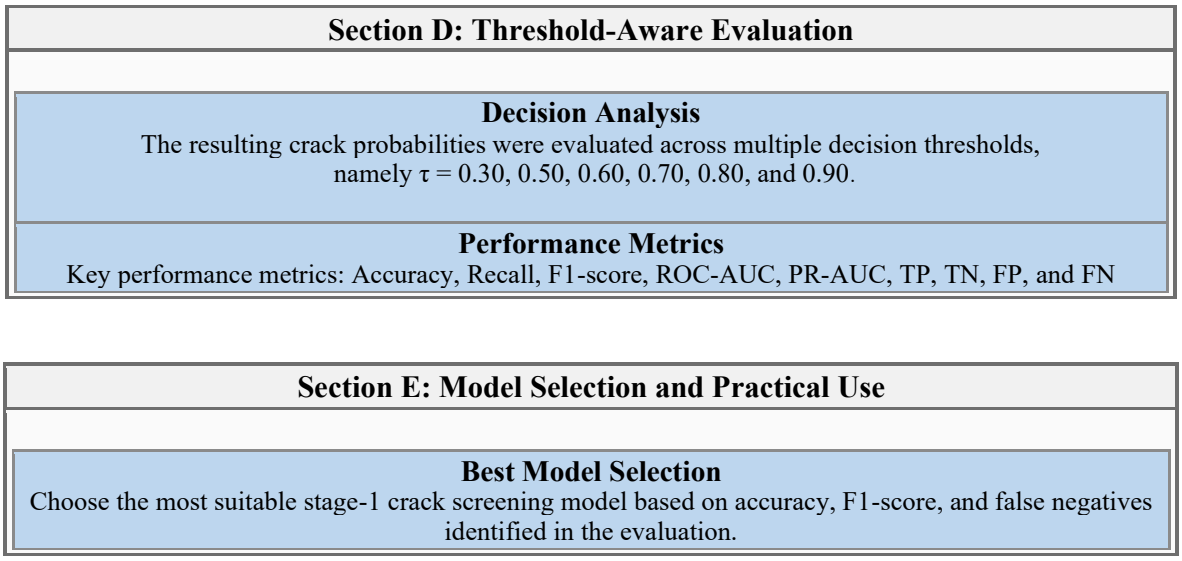


Fig. 2. General workflow of the proposed threshold-aware crack classification framework.

As shown in Fig. 2, the workflow includes steps to prepare the SDNET2018 images as per the predefined splits of the train, validation, and test CSV files. All images will be resized to 224 × 224 pixels and normalized using ImageNet statistics. During training, only light augmentations will be used on the training images, and no augmentations will be applied to the validation and test images. Several different types of ImageNet-pretrained residual networks will be employed to classify the presence or absence of cracks in the concrete images. Each of these models will be trained in two different settings: frozen-backbone training (NOFT), where only the classification layer of the pretrained networks will be trained, and end-to-end training with fine-tuning (FT) of the pretrained backbones to adapt them to the classification task of detecting cracks. For each of the pretrained networks, crack probabilities will be determined both at a fixed threshold and at several different shared thresholds to enable the evaluation of the false negatives and false positive rates of each network. Based on these evaluations, the most suitable of the models will be selected for use as a practical stage-1 process for crack classification before proceeding to later stages of analysis of the cracks. Thus, this method is a general approach to utilizing transfer learning with residual networks for crack analysis, and it can be applied to develop a classification stage before performing a subsequent stage focused on the localization of the cracks within the concrete samples [10], [11], [24].

Following the description of the general workflow of the study, Fig. 3 presents the detailed inference pipeline of the finally selected model. The finally selected model is the residual network with 34 layers that has undergone fine-tuning. This figure presents the detailed steps that will be followed in transforming the input concrete samples to crack or non-crack classifications.

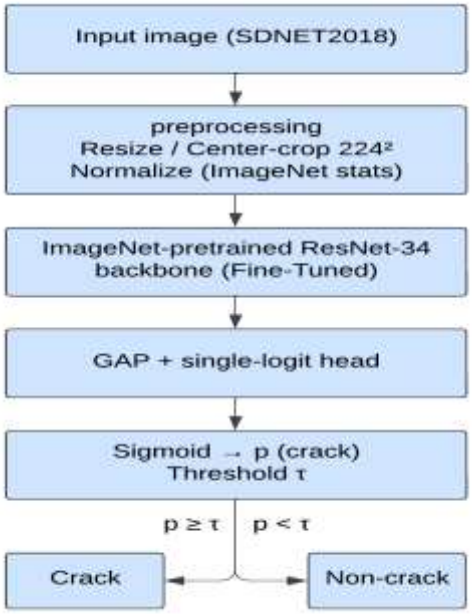


Fig. 3. Detailed inference workflow of the selected fine-tuned ResNet-34 crack classification model.

As illustrated in Fig. 3, the input image is first resized and normalized in the same way as described before. Then the preprocessed image is passed through the ImageNet-pretrained ResNet-34 model, followed by a global average pooling and a logit classification head. The output logit is passed through a sigmoid activation function to yield a crack probability, which is then thresholded to obtain the binary crack/non-crack classification result.

Table 2. Common experimental configuration used across the evaluated backbones on SDNET2018.

Item	Setting
Dataset	SDNET2018 for binary crack classification (Crack vs. Non-crack) using fixed CSV splits (train/val/test).
Input resolution	224 × 224
Normalization	ImageNet mean [0.485, 0.456, 0.406], std [0.229, 0.224, 0.225]
Splits	Loaded from sdnet_train.csv / sdnet_val.csv / sdnet_test.csv
Random seed	42
Training augmentation	Random horizontal flip (p=0.5) + color jitter (brightness=0.15, contrast=0.15, saturation=0.10, hue=0.02)
Validation/test preprocessing	Resize + ImageNet normalization (no augmentation)
Backbone (selected)	ResNet-34 (ImageNet-pretrained, torch vision), fine-tuned
Classification head	Single-logit binary head: Linear(in features, 1) with sigmoid at inference
Training strategy	2-phase FT: (i) backbone frozen warm-up (1 epoch), (ii) end-to-end fine-tuning (remaining epochs)
Total epochs	Up to 15 with early stopping (patience=4) on validation
Batch size	64
Optimizer	AdamW
Learning rate	3×10^{-4}
Weight decay	5×10^{-4}
Loss	BCE With Logits Loss
Imbalance handling	Weighted Random Sampler (inverse class-frequency sampling; no pos weight in loss)
Mixed precision	AMP is enabled when CUDA is available
Checkpointing and logs	Save best.pth and last.pth + train_log.xlsx + evaluation tables

*, ** This setup was kept consistent across experiments to support fair comparison and reliable reporting of the results.

To reduce the effect of class imbalance, the training mini-batches are balanced using a Weighted Random Sampler according to the inverse class frequency. The model is trained using BCE with LogitsLoss with the AdamW optimizer, with mixed-precision training enabled if the CUDA GPU is available. Training is performed in two stages during fine-tuning: first, with the backbone weights frozen, then without freezing. Training with early stopping and model checkpointing enables training stability and reproducibility. The trained model is evaluated on the fixed test dataset using multiple metrics. The threshold-free evaluation metrics include the area under the ROC curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC). Additionally, the threshold-dependent evaluation metrics include accuracy, precision, recall, the F1-score, and confusion matrix counts. In addition to the threshold $\tau = 0.50$ that is used for fair comparison between the different backbones, the thresholds that were selected during model validation are also presented in order to demonstrate the trade-off between false positives and false negatives [1], [14], [15].

6. MODELS

In addition to the four ResNet models described above, evaluation also includes a fifth model based on the residual model ResNetRS-101 [10], [11]. Each of these models determines a crack probability for each image and is evaluated on the SDNET2018 data set using the same splits and testing protocols. These residual models have been used in previous studies on crack detection [8]-[10], [13], [16].

7. TRAINING PROTOCOL AND FINE-TUNING

In addition to training the models with the standard protocol described above with the pre-trained ImageNet weights, each model can also be fine-tuned. More specifically, the weights of the trained models can be further trained with the SDNET2018 data instead of just training the classification head of the models [6], [7], [12]. In this way, the model learns about the features of the cracks in the data set. Thus, the fine-tuned models are the models that are used to evaluate crack probabilities.

Within the baseline model, NOFT, the weights of the trained models are held frozen, and only the classification heads of the models are trained [5], [6]. For the fine-tuning models, FT, the models are trained end-to-end with the same training splits as the baseline model. More specifically, for the fine-tuned models, training begins with training only the classification head of the model (head-only warm-up), followed by training the entire model for a few epochs [6], [8], [12]. This training protocol allows the models to better learn features that describe cracks, but not those of the background textures that are similar to crack patterns.

8. MULTI-THRESHOLD EVALUATION

Evaluation is performed for each model at a variety of decision thresholds τ [1], [14], [15]. In addition to the overall performance scores for each model, a confusion matrix is also reported to provide a more interpretable description of the performance of each model. The confusion matrix for each model displays the following:

- TP: the number of crack images correctly predicted as crack
- TN: the number of non-crack images correctly predicted as non-crack
- FP: the number of non-crack images incorrectly predicted as crack
- FN: the number of crack images incorrectly predicted as non-crack

9. METRICS

The classification performance will be evaluated using accuracy (Acc), precision (Prec), recall (Rec; sensitivity), F1-score, ROC-AUC, and PR-AUC, as these five metrics provide a complete description of the classification model performance under class imbalance [1], [14]. These same measures have been used in other deep learning studies on the classification and detection of defects in concrete [18].

The classification model's predictions are summarized in a confusion matrix, which captures the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) predictions made by the model. In this study, TP represents the number of crack images that were correctly classified as being crack images. Similarly, TN represents the number of non-crack images that were correctly classified as non-crack images. FP represents the number of non-crack images that were incorrectly classified as crack images. Finally, FN represents the number of crack images that were incorrectly classified as non-crack images. Based on these predictions, the threshold-dependent classification metrics can be calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP, and FN are the confusion-matrix entries defined above. Accuracy measures the proportion of all correctly classified images among all evaluated images [18].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

where TP is the number of correctly predicted crack images, and FP is the number of non-crack images incorrectly predicted as crack. Precision measures how often a positive crack prediction is actually correct [18].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

where TP is the number of correctly detected crack images, and FN is the number of actual crack images missed by the classifier. Recall measures the proportion of real crack images that are successfully identified [18].

$$\text{F1} = \frac{2.Precision\,.Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{4}$$

where Precision and Recall are the values defined in Equations (2) and (3), respectively. The F1-score is the harmonic meaning of Precision and Recall and is especially useful in situations where the data is imbalanced between correct and false alarms regarding the presence of cracks in the bridge beams [1], [14], [18].

$$\text{PR} - \text{AUC} = \int_0^1 Precision(Recall)\, d(Recall), AP \approx \sum_{k=1}^{m} (Recall_k - Recall_{k-1}) Precision_k \tag{5}$$

In addition to the evaluation at the threshold τ, the area under the ROC and precision-recall curves (ROC-AUC and PR-AUC, respectively) are also reported. The ROC-AUC is a metric that summarizes the area under the ROC curve calculated for all thresholds of the predicted scores. The PR-AUC is similarly a metric that summarizes the precision-recall curve, but is especially useful in cases where the crack class itself is a minority class. Thus, the PR-AUC is useful in evaluating the crack-screening model's abilities, especially due to the nature of the crack class within the dataset [1], [14].

In addition to these metrics, the false negative (FN) and false positive (FP) rates are also of interest to the inspection systems. False negatives in the context of crack detection are especially critical to evaluate, as false negatives may indicate the presence of potential cracks that would otherwise require maintenance; false positives would only increase the effort required of the manual crack inspection process. Thus, in addition to evaluating each model at the threshold of $\tau = 0.50$, the models are also evaluated at a range of different thresholds [1], [15].

10. RESULTS AND DISCUSSION

The results of the evaluated models on the fixed SDNET2018 test set are presented in this section. These results not only provide an overview of the performance of these models on this test set but also offer an analysis of the suitability of each model under different thresholds. Such an analysis allows for an understanding of the different trade-offs that must be made in the deployment of each of these models for early crack detection. In this section, three main parts will be presented: an overview of the evaluation, the analysis of each model at the same thresholds, and an interpretation of the results that may assist in the selection of the best model for stage-1 crack screening in practice.

10.1. Comparative Evaluation Overview

All of the models were trained and tested under the same splits of the data in order to ensure a fair comparison between the models. The results presented are therefore a reflection of the differences between the backbones of the models, rather than any differences that may have existed within the experimentation of the models themselves. Furthermore, because it is important for the models to exhibit both accuracy in detecting cracks, as well as a balance between sensitivity to those cracks and false alarms, the thresholds of the models were also of particular interest in the observations of this experiment.

10.2. Threshold-Based Performance Analysis

Tables 3-7 report the performance of the evaluated backbones at the five shared operating thresholds. Each table considers both the metrics and the confusion matrix behavior of each model at that threshold. Table 8 summarizes the best operating point for each model, as determined by the F1-score, at each of the shared thresholds.

Table 3 examines the behavior of each backbone at a relatively low operating threshold, $\tau = 0.30$. At this threshold, the system aims to maximize the number of detected cracks in the network, even at the cost of potentially false positive alerts. Thus, Table 3 provides a view of the behavior of each backbone under relatively aggressive screening of the SDNET2018 test set.

Table 3. Threshold-wise comparison of the evaluated backbones at the shared operating threshold $\tau = 0.30$ on the SDNET2018 test set**.**

Model	Stage	thr	Acc	Prec	Rec	F1	TN	FP	FN	TP	N	ROC-AUC	PR- AUC
ResNet- 18	NOFT	0.30	0.5917	0.2601	0.9221	0.4058	3805	3336	99	1173	8414	0.8819	0.7157
ResNet- 18	FT	0.30	0.9114	0.6767	0.7932	0.7303	6660	482	263	1009	8414	0.9333	0.8425
ResNet- 34	NOFT	0.30	0.7836	0.4064	0.9355	0.5666	5404	1738	82	1190	8414	0.9518	0.8734
ResNet- 34	FT	0.30	0.9469	0.7724	0.9205	0.8400	6797	345	101	1171	8414	0.9804	0.9411
ResNet- 50	FT	0.30	0.8131	0.4408	0.8797	0.5874	5723	1419	153	1119	8414	0.9351	0.8502
ResNet-101	NOFT	0.30	0.7245	0.3474	0.9363	0.5068	4905	2237	81	1191	8414	0.9351	0.8334
ResNet-101	FT	0.30	0.9271	0.7501	0.7767	0.7632	6813	329	284	988	8414	0.9356	0.8531
ResNet RS-101	FT	0.30	0.8635	0.5287	0.8970	0.6653	6125	1017	131	1141	8414	0.9531	0.8845

Table 3 illustrates the benefits of lowering the decision threshold to $\tau = 0.30$. Such a threshold adjustment increases the recall of all models, which is a desired attribute in the context of crack detection. However, lowering the threshold also increases the number of false positive detections for several backbone models. The best performing model is ResNet-34 (FT) with an F1-score of 0.8400, accuracy of 0.9469, and recall of 0.9205. ResNet-34 (FT) has a relatively high count of false positives at 345 but a low count of false negatives at 101. Therefore, ResNet-34 (FT) is the more suitable model for screening for the presence of early-stage cracks. The NOFT models have high recall counts but generate a large number of false positives that make them less useful in detecting cracks. Thus, Table 3 shows that lowering the decision threshold is beneficial to the detection process. However, the benefit is only truly realized when fine-tuning of the models is applied.

To provide a fair and standardized comparison between each of these backbones, Table 4 presents the results at each threshold on the same test set, at the same threshold value of \tau = 0.50. This threshold is important in that it is the threshold that is typically employed for binary classification models as a means of balancing the costs of false positives and false negatives. Thus, this threshold is the main reference point for comparing the models prior to considering their performance at different thresholds.

Table 4. Threshold-wise comparison of the evaluated backbones at the shared operating threshold $\tau = 0.50$ on the SDNET2018 test set.

Model	Stage	thr	Acc	Prec	Rec	F1	TN	FP	FN	TP	N	ROC-AUC	PR-AUC
ResNet-18	NOFT	0.50	0.8592	0.5250	0.7240	0.6087	6308	833	351	921	8414	0.8819	0.7157
ResNet-18	FT	0.50	0.9259	0.7557	0.7539	0.7548	6832	310	313	959	8414	0.9333	0.8425

ResNet-34	NOFT	0.50	0.8869	0.5844	0.8734	0.7002	6352	790	161	1111	8414	0.9518	0.8734
ResNet-34	FT	0.50	0.9610	0.8501	0.9009	0.8748	6940	202	126	1146	8414	0.9804	0.9411
ResNet-50	FT	0.50	0.9345	0.8105	0.7397	0.7735	6922	220	331	941	8414	0.9351	0.8502
ResNet101	NOFT	0.50	0.8678	0.5400	0.8490	0.6601	6222	920	192	1080	8414	0.9351	0.8334
ResNet101	FT	0.50	0.9353	0.8058	0.7539	0.7790	6911	231	313	959	8414	0.9356	0.8531
ResNetRS101	FT	0.50	0.9176	0.6805	0.8577	0.7589	6630	512	181	1091	8414	0.9531	0.8845

Table 4 presents the main results of the comparison of the backbones at the operating threshold of $\tau = 0.50$. The effect of fine-tuning is visible at this threshold, with each FT model outperforming the corresponding NOFT model in each case. ResNet-34 (FT) achieves the highest accuracy (0.9610), precision (0.8501), recall (0.9009), and F1-score (0.8748). Furthermore, the values in the confusion matrix for this model are more favorable than those of the other candidates for each metric; there are fewer false positives (FP = 202) and fewer false negatives (FN = 126) with this model. The NOFT models achieve lower precision and F1-score values than the FT models with each backbone. Thus, Table 4 provides confirmation of the fair comparison of the different backbones at a shared threshold of $\tau = 0.50$, and further supports the selection of the ResNet-34 (FT) model as the most suitable classifier for detecting stage-1 cracks.

To investigate the effect of threshold adjustment beyond the default threshold setting of $\tau = 0.50$, Table 5 reports the model comparison at threshold $\tau = 0.60$ on the same fixed test set (SDNET2018). Such a threshold is more conservative than the default threshold of 0.50; it is shifted towards the region of lower predicted probabilities to reduce the occurrence of false alarms while maintaining strong sensitivity to crack detections. Thus, the results of this higher threshold help to indicate whether such a threshold leads to better results for the goal of early crack screening.

Table 5. Threshold-wise comparison of the evaluated backbones at the shared operating threshold $\tau = 0.60$ on the SDNET2018 test set.

Model	Stage	thr	Acc	Prec	Rec	F1	TN	FP	FN	TP	N	ROC-AUC	PR- AUC
ResNet-18	NOFT	0.60	0.8952	0.6663	0.6155	0.6399	6749	392	489	783	8414	0.8819	0.7157
ResNet- 18	FT	0.60	0.9310	0.7907	0.7397	0.7644	6893	249	331	941	8414	0.9333	0.8425
ResNet-34	NOFT	0.60	0.9100	0.6582	0.8419	0.7388	6586	556	201	1071	8414	0.9518	0.8734
ResNet- 34	FT	0.60	0.9639	0.8757	0.8867	0.8813	6982	160	144	1128	8414	0.9804	0.9411
ResNet- 50	FT	0.60	0.9404	0.8978	0.6839	0.7764	7043	99	402	870	8414	0.9351	0.8502
ResNet-101	NOFT	0.60	0.8993	0.6312	0.8034	0.7070	6545	597	250	1022	8414	0.9351	0.8334

ResNet-101	FT	0.60	0.9373	0.8270	0.7405	0.7814	6945	197	330	942	8414	0.9356	0.8531
ResNet RS- 101	FT	0.60	0.9305	0.7402	0.8333	0.7840	6770	372	212	1060	8414	0.9531	0.8845

Table 5 also demonstrates that setting the threshold to $\tau = 0.60$ provides a better balance between sensitivity and the false-alarm rate for several of the models. False positives decrease more noticeably at these higher thresholds while maintaining a high rate of recalled positives. ResNet-34 (FT) achieves the highest values for the F1-score (0.8812), accuracy (0.9638), precision (0.8757), and recall (0.8867). Furthermore, the confusion matrix for this model at this threshold reveals false positives (FP) and false negatives (FN) rates that are more balanced than they are at the lower thresholds. Thus, $\tau = 0.60$ is a better threshold value for the model under consideration. Overall, then, not only does this table support the choice of ResNet-34 as the backbone for the model, but it also demonstrates that adjusting the threshold to reduce false alarms will enhance the model's performance during the screening process.

In addition to the figures presented in Table 5 at the same operating threshold, Fig. 4 further presents a visualization of the relationship between accuracy and the F1-score for each of the model backbones. Such a visualization is presented to assess whether high accuracy rates are correlated with good F1 scores as well, as the F1 score provides a measure of the balance between precision and recall of the model - two metrics that provide more informative measurements of the model's performance at screening for cracks than accuracy alone.

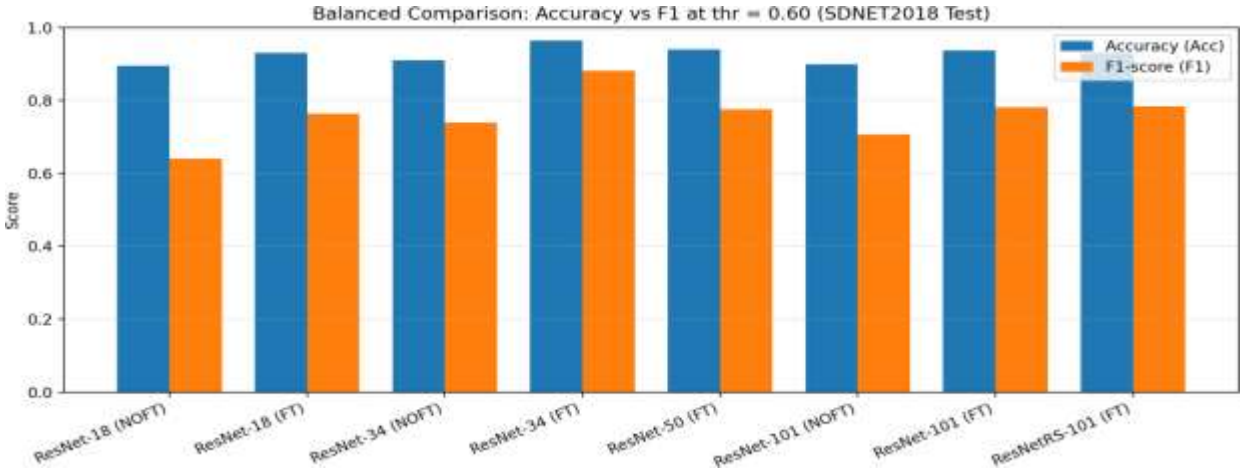


Fig. 4. Visual comparison of Accuracy and F1-score for the evaluated backbones at the shared operating threshold $\tau = 0.60$ on the SDNET2018 test set.

Fig. 4 demonstrates that accuracy is insufficient to assess the effectiveness of the screening models. Several models have high accuracy values, some of which are even similar to each other. However, the F1-score values for these models highlight the fact that accuracy alone is not a suitable metric for evaluating the effectiveness of these models. Fine-tuning helps to reduce the gap between the accuracy and F1-score of the models. The ResNet-34 model with fine-tuning achieves the highest accuracy (0.9639) and F1-score (0.8813) among all evaluated models with $\tau = 0.60$. Thus, this model is chosen as the main model that will be used as the classification gate for the crack segmentation stage and as the best-suited model for stage-1 crack screening.

Given the importance of false negatives in the context of safety-oriented crack screening, Fig. 5 presents the false-negative count for each of the evaluated backbone networks at the shared operating threshold of $\tau = 0.60$ on the SDNET2018 dataset. This figure is presented to allow for the visual comparison of the false-negative rates of the different backbone networks, which is an important factor to consider when selecting a backbone network to utilize as a stage-1 screening model.

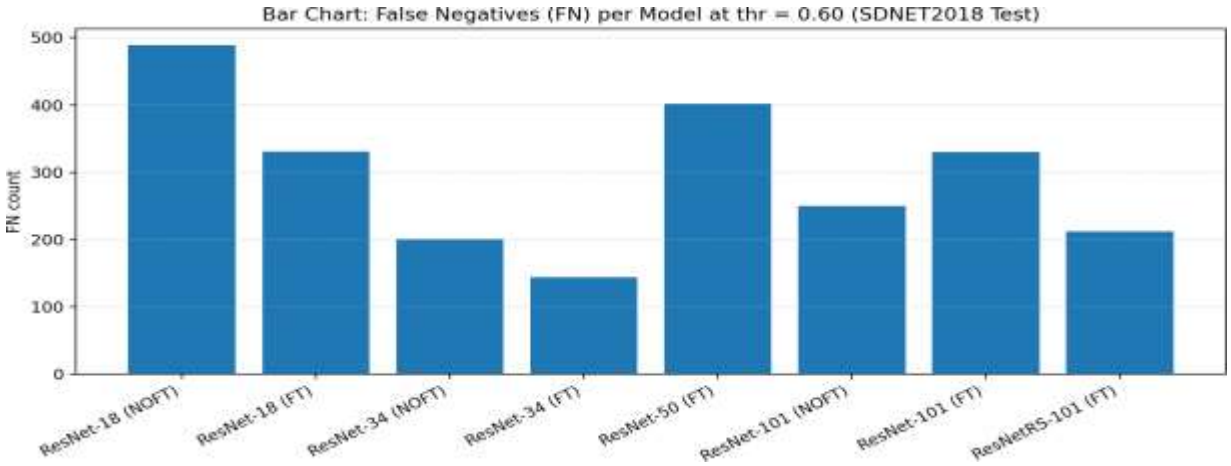


Fig. 5. Visual comparison of false-negative (FN) counts across the evaluated backbones at the shared operating threshold $\tau = 0.60$ on the SDNET2018 test set.

Fig. 5 illustrates the importance of false negatives in the context of detecting cracks in infrastructure. The different representations of missed cracks in the illustrated backbones at 0.60 thresholds demonstrate the different impacts of each backbone on the detection of these cracks. ResNet-34 (FT) has the lowest count of false negatives among all models yet maintains a high F1-score. Thus, it is the least likely to fail to detect crack-containing images. The reduction of false negatives presents benefits to the stage of screening of these images prior to the segmentation of the cracks within those images. Therefore, based upon these findings, ResNet-34 (FT) can be selected as the stage-1 model for crack detection. Table 6 presents the results of the same models at a threshold of 0.70 on the same test dataset of SDNET2018. This threshold introduces a stricter standard for the detection of crack-containing images. Thus, this table can reveal whether it is possible to reduce the false positives at the cost of only a small reduction in the detection of crack-containing images.

Table 6. Threshold-wise comparison of the evaluated backbones at the shared operating threshold $\tau = 0.70$ on the SDNET2018 test set**.**

Model	Stage	thr	Acc	Prec	Rec	F1	TN	FP	FN	TP	N	ROC-AUC	PR-AUC
ResNet-18	NOFT	0.70	0.9052	0.7752	0.5259	0.6266	6947	194	603	669	8414	0.8819	0.7157
ResNet-18	FT	0.70	0.9339	0.8213	0.7193	0.7669	6943	199	357	915	8414	0.9333	0.8425
ResNet-34	NOFT	0.70	0.9269	0.7302	0.8191	0.7721	6757	385	230	1042	8414	0.9518	0.8734
ResNet-34	FT	0.70	0.9643	0.8983	0.8616	0.8796	7018	124	176	1096	8414	0.9804	0.9411
ResNet-101	NOFT	0.70	0.9170	0.7161	0.7476	0.7315	6765	377	321	951	8414	0.9351	0.8334
ResNet-101	FT	0.70	0.9386	0.8487	0.7232	0.7809	6978	164	352	920	8414	0.9356	0.8531
ResNet-50	FT	0.70	0.9390	0.9566	0.6250	0.7560	7106	36	477	795	8414	0.9351	0.8502

As presented in Table 6, increasing the threshold to $\tau = 0.70$ leads to a reduction in the number of false positive outcomes identified by the majority of deep learning models. However, there is a reduction in the recall metric for the majority of models, especially those that are less well-calibrated. The model that performs the best overall at this threshold is the ResNet-34 (FT) model, which attains the highest F1-score (0.8796) as well as accuracy (0.9643). Furthermore, it also has high precision (0.8983) and recall (0.8616) scores. The values within its confusion

matrix are also favorable, with a false positive value of 124 and a false negative value of 176. Other models have lower false positive rates, but at the cost of a significant drop in the recall metric. Thus, this table indicates that the threshold of $\tau = 0.70$ is competitive with the ResNet-34 (FT) deep learning model, but increasing the threshold beyond this value should be done with care due to the drop in the recall metric becoming more evident as the threshold increases.

To assess the behavior of the models under such a conservative setting for the decision threshold, Table 7 reports the results of the models at the threshold of $\tau = 0.90$ on the same fixed test set. This threshold is much more conservative than the threshold of $\tau = 0.50$, which is utilized in the remaining tables; at $\tau = 0.90$, the model will be much more restrictive in its assignment of the label "crack". This table is included to determine if it is possible to significantly reduce the number of false positive detections while at the same time not sacrificing too many of the actual cracks that were present in the images.

Table 7. Threshold-wise comparison of the evaluated backbones at the shared operating threshold $\tau = 0.90$ on the SDNET2018 test set.

Model	Stage	thr	Acc	Prec	Rec	F1	TN	FP	FN	TP	N	ROC-AUC	PR-AUC
ResNet-18	NOFT	0.90	0.8978	0.9056	0.3624	0.5176	7093	48	811	461	8414	0.8819	0.7157
ResNet-18	FT	0.90	0.9372	0.8974	0.6603	0.7608	7046	96	432	840	8414	0.9333	0.8425
ResNet-34	NOFT	0.90	0.9433	0.8825	0.7209	0.7935	7020	122	355	917	8414	0.9518	0.8734
ResNet-34	FT	0.90	0.9603	0.9518	0.7767	0.8554	7092	50	284	988	8414	0.9804	0.9411
ResNet-50	FT	0.90	0.9113	0.9906	0.4174	0.5873	7137	5	741	531	8414	0.9351	0.8502
ResNet-101	NOFT	0.90	0.9307	0.9057	0.6045	0.7251	7062	80	503	769	8414	0.9351	0.8334
ResNet-101	FT	0.90	0.9415	0.9157	0.6753	0.7773	7063	79	413	859	8414	0.9356	0.8531
ResNet RS-101	FT	0.90	0.9474	0.9036	0.7303	0.8078	7043	99	343	929	8414	0.9531	0.8845

As Table 7 illustrates, increasing the threshold to $\tau = 0.90$ results in a false positive reduction at the cost of recall for several models. This demonstrates that attempting to increase the threshold for false positives will result in increasing the threshold for the detection of cracks. The model with the best results at this threshold is the ResNet-34 (FT) model. This model attains the highest F1-score (0.8554), accuracy (0.9603), and precision (0.9518). Furthermore, the values within its confusion matrix exhibit fewer false positives (FP = 50) and have a higher recall than the other models at this threshold. However, the false negative rate for this model is more significant compared to the other models at $\tau = 0.90$, indicating that this threshold may be too conservative for the task of detecting cracks for safety purposes. Thus, although false positives can be reduced by increasing the threshold, the threshold for detecting the cracks limits its usefulness for detecting these flaws in the steel.

11. MODEL SELECTION

Following the individual analysis of each threshold, Table 8 presents a comparison of the selected best operating points based on the F1-score. The purpose of this table is to provide a selection of the best operating points for each of the backbone models, based on these same test conditions and thresholds.

Table 8. Selected Best Operating Points Based on F1-Score

Model	Stage	thr	Acc	Prec	Rec	F1	TN	FP	FN	TP	N	ROC-AUC	PR-AUC
ResNet-34	FT	0.60	0.9639	0.8757	0.8867	0.8813	6982	160	144	1128	8414	0.9804	0.9411

ResNet RS-101	FT	0.90	0.9474	0.9036	0.7303	0.8078	7043	99	343	929	8414	0.9531	0.8845
ResNet- 34	NOFT	0.90	0.9433	0.8825	0.7209	0.7935	7020	122	355	917	8414	0.9518	0.8734
ResNet-101	FT	0.60	0.9373	0.8270	0.7405	0.7814	6945	197	330	942	8414	0.9356	0.8531
ResNet- 50	FT	0.60	0.9404	0.8978	0.6839	0.7764	7043	99	402	870	8414	0.9351	0.8502

Table 8 indicates that ResNet-34 (FT) has the strongest overall operating point. At $\tau = 0.60$, ResNet-34 (FT) achieves the highest F1-score (0.8813) and accuracy (0.9639), while also having the most favorable balance between false positives and false negatives. Thus, ResNet-34 (FT) at $\tau = 0.60$ is selected as the model to employ as the stage-1 crack detection model.

12. CONCLUSIONS

This paper presents a threshold-aware benchmark for the classification of concrete cracks on the SDNET2018 dataset using residual backbone networks. Fine-tuning was shown to be not just beneficial, but important for reducing false responses from the networks. Furthermore, ResNet-34 with fine-tuning was the best-performing network in terms of accuracy and F1 score on this specific benchmark. Thus, the importance of this study lies in the fact that it provides a means of comparing the performance of these models in a fair and relevant way to their potential deployment, rather than simply reporting the accuracy of each network. Therefore, this benchmark can be helpful in model selection in real-world scenarios, and also in guiding future studies on crack detection and segmentation.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Consent for publications

All authors have read and approved the final manuscript for publication.

Availability of data and material

All data supporting the findings of this study are included within the paper.

Authors' contributions

H.T.M. designed the main idea of the study, prepared and organized the datasets, conducted the experiments, analyzed the results, and wrote the initial draft of the manuscript.
M.T.Y. supervised the research, reviewed the methodology and experimental design, contributed to the interpretation of results, and participated in the revision and final editing of the manuscript.

Funding

This research received no external funding.

13. REFERENCES

[1] L. Ali, F. Alnajjar, H. Al Jassmi, M. Gocho, W. Khan, and M. A. Serhani, “Performance evaluation of deep CNN-based crack detection and localization techniques for concrete structures,” Sensors, vol. 21, no. 5, art. no. 1688, 2021, doi: 10.3390/s21051688.

[2] E. W. C. Chen and M. Jahanshahi, “A satellite image-based convolutional neural network classifier for detecting and classifying the surface cracks in concrete infrastructure,” Sensors, vol. 18, no. 12, art. no. 4097, 2018, doi: 10.3390/s18124097.

[3] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, “Road crack detection using deep convolutional neural network,” in Proc. IEEE Int. Conf. Image Processing (ICIP), Phoenix, AZ, USA, 2016, pp. 3708–3712, doi: 10.1109/ICIP.2016.7533052.

[4] R.-S. Rajadurai and S.-T. Kang, "Automated vision-based crack detection on concrete surfaces using deep learning," Applied Sciences, vol. 11, no. 11, art. no. 5229, 2021, doi: 10.3390/app11115229.

[5] M. Dorafshan, R. J. Thomas, and M. Maguire, "SDNET2018: An annotated image dataset for non-contact concrete crack detection using deep convolutional neural networks," Data in Brief, vol. 21, pp. 1664–1668, 2018, doi: 10.1016/j.dib.2018.11.015.

[6] C. Su and W. Wang, "Concrete cracks detection using convolutional neural network based on transfer learning," Mathematical Problems in Engineering, vol. 2020, art. no. 7240129, 2020, doi: 10.1155/2020/7240129.

[7] R. E. Philip, A. D. Andrushia, A. Nammalvar, B. G. A. Gurupatham, and K. Roy, "A comparative study on crack detection in concrete walls using transfer learning techniques," J. Compos. Sci., vol. 7, no. 4, art. no. 169, 2023, doi: 10.3390/jcs7040169.

[8] W. Qayyum, R. Ehtisham, A. Bahrami, C. Camp, J. Mir, and A. Ahmad, "Assessment of convolutional neural network pre-trained models for detection and orientation of cracks," Materials, vol. 16, no. 2, art. no. 826, 2023, doi: 10.3390/ma16020826.

[9] R. Wang, X. Zhou, Y. Liu, D. Liu, Y. Lu, and M. Su, "Identification of the surface cracks of concrete based on ResNet-18 depth residual network," Applied Sciences, vol. 14, no. 8, art. no. 3142, 2024, doi: 10.3390/app14083142.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[11] I. Bello, W. Fedus, X. Du, E. D. Cubuk, A. Srinivas, T.-Y. Lin, J. Shlens, and B. Zoph, "Revisiting ResNets: Improved training and scaling strategies," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, 2021.

[12] J. Li, J. Dong, J. Xie, J. Wu, and L. Yang, "Intelligent detection method for concrete dam surface cracks based on two-stage transfer learning," Water, vol. 15, no. 11, art. no. 2082, 2023, doi: 10.3390/w15112082.

[13] D. P. Yadav, B. Sharma, S. Chauhan, and I. B. Dhaou, "Bridging convolutional neural networks and transformers for efficient crack detection in concrete building structures," Sensors, vol. 24, no. 13, art. no. 4257, 2024, doi: 10.3390/s24134257.

[14] Z. Fang, X. Wang, J. Gao, and B. Eskandarpour, "Optimizing concrete crack detection: An echo state network approach with improved fish migration optimization," Scientific Reports, vol. 15, art. no. 40, 2025, doi: 10.1038/s41598-024-84458-1.

[15] S. Egodawela, A. Khodadadian Gostar, H. A. D. S. Buddika, A. J. Dammika, N. Harischandra, S. Navaratnam, and M. Mahmoodian, "A deep learning approach for surface crack classification and segmentation in unmanned aerial vehicle assisted infrastructure inspections," Sensors, vol. 24, no. 6, art. no. 1936, 2024, doi: 10.3390/s24061936.

[16] D. Li, S. Wu, D. Zhai, and J. Zhang, "Intelligent crack detection method based on GM-ResNet," Sensors, vol. 23, no. 20, art. no. 8369, 2023, doi: 10.3390/s23208369.

[17] R. Pu, G. Ren, H. Li, W. Jiang, J. Zhang, and H. Qin, "Autonomous concrete crack semantic segmentation using deep fully convolutional encoder–decoder network in concrete structures inspection," Buildings, vol. 12, no. 11, art. no. 2019, 2022, doi: 10.3390/buildings12112019.

[18] P. Arafin, A. H. M. Billah, and A. Issa, "Deep learning-based concrete defects classification and detection using semantic segmentation," Structural Health Monitoring, vol. 23, no. 1, pp. 383–409, 2024, doi: 10.1177/14759217231168212.

[19] H. Hacıefendioğlu, A. Akbulut, and N. U. K. Unal, "Road surface crack segmentation based on transformer-enhanced encoder–decoder networks," Buildings, vol. 13, no. 12, art. no. 3113, 2023, doi: 10.3390/buildings13123113.

[20] K. Feng and W. Xu, "Lightweight dual-attention network for concrete crack segmentation," Sensors, vol. 25, no. 14, art. no. 4436, 2025, doi: 10.3390/s25144436.

[21] E. Elhariri, N. El-Bendary, and S. A. Taie, "Automated pixel-level deep crack segmentation on historical surfaces using U-Net models," Algorithms, vol. 15, no. 8, art. no. 281, 2022, doi: 10.3390/a15080281.

[22] S. Kang, X. Zhang, L. Yang, and Y. Yang, "Pavement crack detection with DeepLabv3+ based on residual network," in Proc. Int. Conf. Intelligent Systems Design and Applications, 2019, pp. 1227–1237.

[23] D. Xiao, G. Wang, K. Wang, S. Liu, G. Shang, Q.-A. Wang, X. Fan, M. Hu, R. Liu, G. Chen, and Z. Chen, "Intelligent measurement of concrete crack width based on U-Net deep learning and binocular vision 3D reconstruction," Applied Sciences, vol. 16, art. no. 2355, 2026, doi: 10.3390/app16052355.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[25] J. Zhang, H. Xia, P. Li, K. Zhang, W. Hong, and R. Guo, "A pavement crack detection method via deep learning and a binocular-vision-based unmanned aerial vehicle," Applied Sciences, vol. 14, no. 5, art. no. 1778, 2024, doi: 10.3390/app14051778.

[26] A. Arbaoui, A. Ouahabi, S. Jacques, and M. Hamiane, "Concrete cracks detection and monitoring using deep learning-based multiresolution analysis," Electronics, vol. 10, no. 15, art. no. 1772, 2021, doi: 10.3390/electronics10151772.

Journal of Al-Turath University College (JATUC)
Open Access Journal
ISSN : 2074-5621
DOI: https://doi.org/10.63964/JATUC.43.2.2026.46
https://jatuc.uoturath.edu.iq/

الكشف المبكر عن الشقوق في البنية التحتية المدنية باستخدام التقييم متعدد العتبات لنماذج ResNet المضبوطة بدقة على مجموعة بيانات SDNET2018

هديل طالب محمد جواد[1]*, مهند تحرير يونس[2]

[1]قسم علوم الحاسبات, كلية العلوم, الجامعة المستنصرية, بغداد, العراق
[2]قسم الذكاء الاصطناعي, كلية العلوم, الجامعة المستنصرية, بغداد, العراق

* *البريد الإلكتروني للتواصل* : hadeeltalib.altalib@uomustansiriyah.edu.iq

الملخص

يُعدّ الاكتشاف المبكر للتشققات عنصرًا أساسيًا في الصيانة الوقائية وضمان السلامة في البنى التحتية المدنية. يقدّم هذا البحث مقارنة معيارية لمهمة التصنيف الثنائي للتشققات على مجموعة بيانات SDNET2018 باستخدام خمسة نماذج متبقية مدرّبة مسبقًا علىImageNet ، وهي ResNet-18 :و ResNet-34وResNet-50 و ResNet-101وResNetRS-101. وقد جرى تقييم نظامين للتدريب، هما: تدريب رأس التصنيف مع تجميد العمود الفقري(NOFT) ، والضبط الدقيق الشامل من طرف إلى طرف .(FT) ولدعم اتخاذ القرار العملي في سياق الفحص الهندسي، تم اعتماد بروتوكول تقييم متعدد العتبات لتحليل المفاضلة بين الإنذارات الكاذبة (FP) والتشققات الفائتة (FN) عند نقاط تشغيل مشتركة. أظهرت النتائج أن الضبط الدقيق حسّن بصورة متسقة القدرة التمييزية العملية لجميع النماذج المقيمة. ومن بين جميع النماذج، حقق النموذج ResNet-34 بعد الضبط الدقيق أفضل توازن إجمالي في الأداء. فعند عتبة القرار المختارة 0.60 على مجموعة الاختبار الثابتة من SDNET2018والتي تضم 8414 صورة، حقق Accuracy بنسبة 96.39% و Precisionبنسبة 87.58% و Recallبنسبة 88.68% و F1-scoreبنسبة 88.13%، مع ROC-AUC = 0.9805 و PR-AUC = 0.9411.وعند هذه العتبة، سجل النموذج 160 حالة FP و144 حالةFN ، مما يعكس توازنًا عمليًا مناسبًا بين حساسية الفحص وتقليل العبء على المراحل اللاحقة. وبناءً على هذه النتائج، تم اعتماد النموذج ResNet-34 بعد الضبط الدقيق كبوابة تصنيف فعالة للفحص المبكر للتشققات قبل مرحلة التقسيم اللاحقة.

الكلمات المفتاحية

الكشف عن الشقوق؛ SDNET2018؛ ResNet؛ الضبط الدقيق؛ التقييم متعدد العتبات؛ PR-AUC؛ البوابة إلى التجزئة.