

CLASSIFICATION OF BUILDING ASSET DAMAGE LEVELS BASED ON CRACK DETECTION USING CONVOLUTIONAL NEURAL NETWORK AND SWIN TRANSFORMER

KLASIFIKASI TINGKAT KERUSAKAN ASET BANGUNAN BERDASARKAN DETEKSI RETAKAN MENGGUNAKAN CONVOLUTIONAL NEURAL NETWORK (CNN) DAN SWIN TRANSFORMER

Alfandi Ahmad^{1‡}, Eneng Tita Tosida², and Yusma Yanti³

^{1,2,3}Department Computer Science, Pakuan University, Indonesia

[‡]corresponding author: 065122105@student.unpak.ac.id

Copyright © 2026 Alfandi Ahmad, Eneng Tita Tosida, and Yusma Yanti. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Structural damage caused by cracks is a serious problem that requires early identification to prevent a decline in structural stability. Conventional inspections are considered inefficient and prone to human subjectivity. This study proposes a comparison of two deep learning models, a MobileNetV2-based Convolutional Neural Network (CNN) and a Swin Transformer to automatically classify the severity of building cracks. The dataset consists of 3,000 images of concrete cracks from Kaggle, divided into three classes: hairline cracks, small cracks, and moderate cracks, with 1,000 images in each class. Training was conducted using three variations of the data split ratio (70:15:15, 80:10:10, 90:5:5). Evaluation results using the 90:5:5 split for both models showed that the Swin Transformer model achieved the highest accuracy of 98%, outperforming MobileNetV2, which achieved an accuracy of 93%. This study confirms that Transformer-based approaches are superior to conventional CNNs in capturing global relationships in crack images.

Keywords: Classification, CNN, Crack, Deep Learning, Swin Transformer.

Abstrak

Kerusakan bangunan akibat retakan merupakan permasalahan serius yang memerlukan identifikasi dini untuk mencegah penurunan stabilitas struktur. Inspeksi konvensional dinilai kurang efisien dan rentan terhadap subjektivitas manusia. Penelitian ini mengusulkan komparasi dua model deep learning, yaitu Convolutional Neural Network (CNN) berbasis MobileNetV2 dan Swin Transformer untuk mengklasifikasikan tingkat kerusakan retakan bangunan secara otomatis. Dataset yang digunakan terdiri dari 3.000 citra retakan beton dari Kaggle yang dibagi ke dalam tiga kelas, yaitu hairline crack, small crack, dan moderate crack, masing-masing sebanyak 1.000 citra. Data diproses melalui tahapan resize ke 224×224 piksel, normalisasi, serta augmentasi meliputi rotasi, zoom, flip horizontal, dan penyesuaian kecerahan. Pelatihan dilakukan dengan tiga variasi rasio pembagian data (70:15:15, 80:10:10, 90:5:5). Hasil evaluasi dengan proporsi 90:5:5 pada keduanya menunjukkan bahwa model Swin Transformer mencapai akurasi tertinggi sebesar 98%, mengungguli MobileNetV2 yang memperoleh akurasi 93%. Penelitian ini mengonfirmasi bahwa pendekatan berbasis Transformer lebih unggul dalam menangkap hubungan global pada citra retakan dibandingkan CNN konvensional.

Kata Kunci: CNN, *Deep Learning*, Klasifikasi, Retakan, Swin Transformer.

1. Pendahuluan

Retakan (crack) merupakan salah satu kerusakan struktur beton yang paling umum dan dapat mengindikasikan penurunan performa struktural (Novigu *et al.*, 2026). Fenomena retakan dapat berkembang baik sebelum maupun setelah beton mencapai kematangan struktural, dipicu oleh beban eksternal, penyusutan (shrinkage), settlement, dan fluktuasi kondisi lingkungan. Deteksi dini dan evaluasi sistem tingkat keparahan retakan sangat penting dalam menjaga integritas struktural dan keamanan bangunan serta memperpanjang umur layanan konstruksi.

Inspeksi retakan secara manual memiliki keterbatasan signifikan, meliputi ketergantungan pada subjektivitas pengamat, durasi pemeriksaan yang panjang, dan risiko inkonsistensi dalam penilaian. Untuk mengatasi keterbatasan tersebut, pendekatan deep learning menjadi solusi yang menjanjikan. Deep learning merupakan teknologi yang memungkinkan sistem komputer mempelajari representasi data yang semakin abstrak dan kompleks melalui arsitektur jaringan saraf berlapis. Kemampuan ini memfasilitasi sistem dalam mengenali dan mengklasifikasi pola visual, seperti jenis dan tingkat keparahan retakan, secara otomatis dengan akurasi tinggi tanpa bergantung pada penilaian manusia (Yahfizham, 2024).

Suherman *et al.*, (2023) melaporkan bahwa model berbasis Convolutional Neural Network (CNN) mampu mencapai akurasi sekitar

94% dalam tugas klasifikasi citra retakan beton, dengan mekanisme operasi konvolusi yang memfasilitasi ekstraksi representasi fitur lokal berupa karakteristik tepi dan tekstur. (Sarhadi *et al.*, 2024) menunjukkan keunggulan arsitektur SWIN U-Net dalam melakukan deteksi retakan beton pada skala piksel dengan performa yang sangat tinggi akurasi 99%. Keunggulan model ini didasari oleh mekanisme self-attention Swin Transformer yang memungkinkan pemrosesan citra dengan berbagai resolusi secara efisien (Liu *et al.*, 2021)

Studi komparatif yang mendalam antara arsitektur CNN dan Swin Transformer untuk klasifikasi retakan beton masih belum banyak dilakukan dan masih terbatas. Penelitian ini mengusulkan evaluasi yang membandingkan performa kedua arsitektur secara sistematis untuk memberikan panduan pemilihan model yang lebih optimal dalam implementasi sistem inspeksi otomatis bangunan.

2. Metodologi

2.1 Bahan dan Data

Dataset yang digunakan merupakan data sekunder dari platform Kaggle Concrete & Pavement Crack Dataset, terdiri atas 30.000 citra permukaan beton dan aspal yang terdiri dari dua kelas *crack* dan *non-crack*, yang diakuisisi di Nigerian Army University Bui, Nigeria menggunakan DJI Mavic 2 Enterprise drone dan smartphone, kemudian disimpan dalam format RGB JPEG dengan dimensi 227×227 piksel. Gambar kedua kelas bisa dilihat pada Gambar 1.



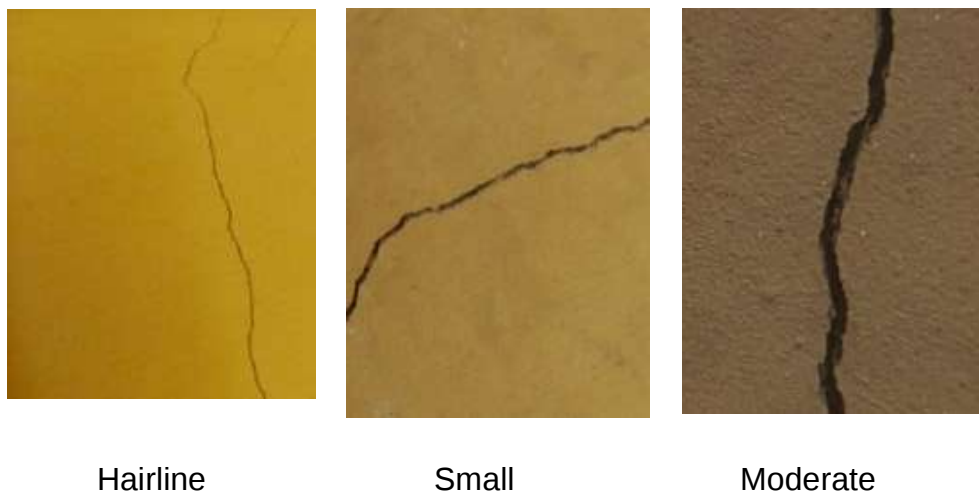
(a) (b)
Gambar 1: crack (a) dan non crack (b).

2.2 Metode Penelitian

Penelitian ini menggunakan pendekatan *Knowledge Discovery in Databases* (KDD) karena menyediakan tahapan sistematis dalam pengolahan data hingga evaluasi klasifikasi.

2.2.1. Data Selection

Tahapan awal dilakukan dengan memilih dan mengelompokkan dataset Concrete & Pavement Crack Dataset yang terdiri dari dua kelas. crack dan non-crack. Dipilih secara manual mengacu pada karakteristik visual dari standar ACI (American Concrete Institute) yang mendefinisikan hairline crack (lebar $< 0,1$ mm), small crack ($0,1-0,3$ mm), dan moderate crack ($> 0,3$ mm). Ciri crack ditunjukkan pada Gambar 2, pengelompokan dilakukan secara visual dengan mempertimbangkan ketebalan garis, kontras, dan kejelasan retakan pada citra. Pemilihan tiga kelas juga didasarkan dengan pertimbangan efisiensi komputasi, keseimbangan distribusi data, serta kemudahan interpretasi hasil klasifikasi. Dari keseluruhan dataset, dipilih sebanyak 3.000 citra secara merata dengan masing-masing kelas sebanyak 1.000 citra. Dataset selanjutnya dibagi menjadi data pelatihan, validasi, dan pengujian menggunakan tiga skenario rasio, yaitu 70:15:15, 80:10:10, dan 90:5:5 untuk mengevaluasi performa model pada berbagai kondisi pembagian data.



Gambar 2: Kategori Retakan.

2.2.2. Data Preprocessing

Untuk memastikan kesesuaian format input, maka dilakukan praproses. Proses pertama meliputi *resize* citra dari ukuran awal 227×227 piksel menjadi 224×224 piksel agar sesuai dengan kebutuhan input model MobileNetV2 dan Swin Transformer. Selain itu, dilakukan normalisasi nilai piksel dengan mengubah rentang intensitas dari 0–255 menjadi rentang yang lebih terstandarisasi. Proses normalisasi bertujuan untuk meningkatkan stabilitas pelatihan, serta membantu model dalam mempelajari pola citra secara lebih optimal.

2.2.3. Data Augmentation

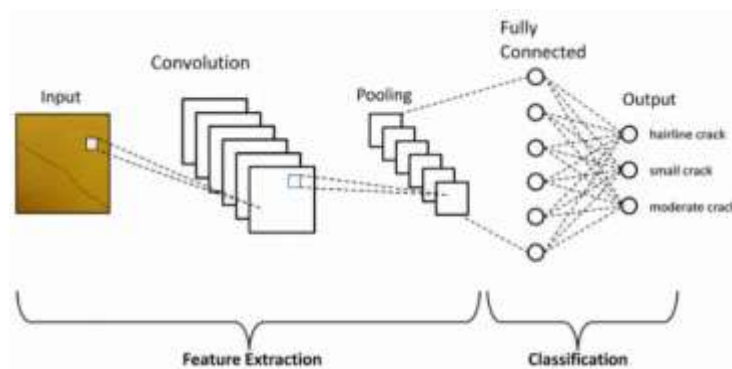
Teknik augmentasi dilakukan meliputi *rotation* hingga $\pm 30^\circ$ untuk membantu model mengenali retakan dari berbagai sudut pandang, *zoom* hingga 30% untuk mendeteksi retakan pada berbagai skala, serta horizontal flip untuk menambah variasi orientasi citra. Selain itu, diterapkan shear transformation untuk mensimulasikan perubahan perspektif pengambilan gambar dan *brightness adjustment* untuk meningkatkan ketahanan model terhadap variasi pencahayaan. Teknik *random resized crop* juga digunakan agar model mampu fokus pada area retakan yang relevan.

2.2.4. Skenario Pemodelan dan Komparasi

Dua skenario dibandingkan pada pemodelan arsitektur yang berbeda dalam cara memproses dan mengekstraksi fitur dari citra retakan. Skenario pertama menggunakan arsitektur *Convolutional Neural Network* (CNN) berbasis MobileNetV2 yang bekerja melalui operasi konvolusi, sedangkan skenario kedua menggunakan arsitektur *Swin Transformer* yang menggunakan mekanisme *self-attention* untuk menangkap hubungan fitur secara lokal maupun global.

Convolutional Neural Network (MobileNetV2)

Skenario pertama menggunakan arsitektur CNN berbasis MobileNetV2 dengan pendekatan fine-tuning menggunakan pretrained weights dari ImageNet. Model menerima input citra berukuran $224 \times 224 \times 3$ yang diproses melalui convolutional layer, pooling layer, dan fully connected layer untuk klasifikasi ke dalam tiga kelas: hairline crack, small crack, dan moderate crack. Lapisan awal dibekukan guna mempertahankan fitur umum, sementara lapisan akhir yang terdiri dari Global Average Pooling, fully connected layer dengan aktivasi ReLU, dropout, dan softmax dilatih untuk menyesuaikan karakteristik dataset retakan beton. Alur proses klasifikasi dapat dilihat pada Gambar 3.



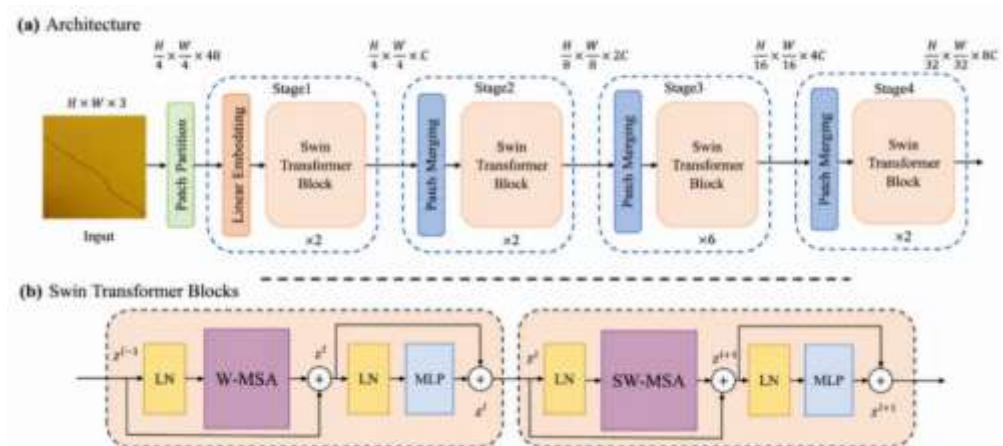
Gambar 3: Alur Proses CNN.

Perumusan konvolusi pada CNN secara matematis dapat dilihat pada persamaan(1), dimana $S(i, j)$ merupakan nilai output feature map pada posisi (i, j) , $(i + m, j + n)$ adalah nilai piksel citra input, dan $K(m, n)$ merupakan kernel atau filter konvolusi yang digunakan untuk mengekstraksi fitur pada citra.

$$S(i, j) = \sum_m \sum_n I(i + m, j + n) \cdot K(m, n) \quad (1)$$

Swin Transformer

Skenario kedua menggunakan arsitektur Swin Transformer varian Swin-Tiny (Swin-T) yang diimplementasikan melalui library timm dengan pretrained weights dari ImageNet. Seluruh parameter model dilatih secara end-to-end tanpa pembekuan lapisan guna mengoptimalkan representasi fitur terhadap karakteristik dataset retakan beton. Alur kerja diawali dengan *patch partition* yang membagi citra masukan menjadi patch-patch kecil, kemudian diekstraksi menggunakan mekanisme Window-based Self-Attention (W-MSA) dan Shifted Window Self-Attention (SW-MSA) untuk menangkap relasi fitur secara lokal dan global. Representasi hierarkis dibentuk melalui *patch merging*, lalu direduksi menggunakan *Global Average Pooling* sebelum diteruskan ke lapisan klasifikasi dengan fungsi aktivasi softmax untuk menghasilkan prediksi. Workflow arsitektur Swin Transformer ditunjukkan pada Gambar 4.



Gambar 4: Alur Proses Swin Transformer.

Dasar mekanisme self-attention Swin Transformer dirumuskan pada persamaan (2), dimana Q (Query), K (Key), dan V (Value). merupakan representasi fitur yang digunakan untuk menghitung hubungan antar bagian citra. Nilai QK^T , menunjukkan tingkat kesamaan antara query dan key. kemudian dinormalisasi dengan $\sqrt{d}$ untuk menjaga stabilitas perhitungan.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (2)$$

2.2.5. Evaluasi Model

Kedua model dievaluasi menggunakan *Confusion Matrix*. Terdapat empat metrik utama pada evaluasi kuantitatif *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). *Accuracy* untuk mengukur proporsi prediksi yang benar secara keseluruhan sebagaimana ditunjukkan pada persamaan (3). *Precision* untuk mengukur ketepatan prediksi kelas positif yang ditunjukkan pada persamaan (4). *Recall* untuk mengukur kemampuan model mendeteksi seluruh sampel positif yang dapat dilihat pada persamaan (5). Persamaan (6) *F1-Score* sebagai rata-rata harmonik antara *Precision* dan *Recall*.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

3. Hasil dan Pembahasan

Hasil evaluasi menggunakan tiga variasi rasio pembagian, yaitu 70:15:15, 80:10:10, dan 90:5:5. Ringkasan akurasi kedua model disajikan pada Tabel 1. Secara umum, kedua model menunjukkan peningkatan performa seiring bertambahnya proporsi data pelatihan. Secara umum, kedua model menunjukkan peningkatan performa seiring bertambahnya proporsi data pelatihan. MobileNetV2 memperoleh akurasi berturut-turut sebesar 90%, 92%, dan 93%, sementara Swin Transformer mencapai 95%, 97%, dan 98%. Rasio 90:5:5 menghasilkan performa tertinggi pada kedua model sehingga dipilih sebagai dasar analisis lebih lanjut.

Tabel 1: Perbandingan Akurasi Model pada Berbagai Proporsi.

Proporsi	MobileNetV2	Swin Transformer
70:15:15	90%	95%
80:10:10	92%	97%
90:5:5	93%	98%

Hasil Model CNN (MobileNetV2)

Model CNN yang digunakan berbasis arsitektur *MobileNetV2* dengan pendekatan *fine-tuning*. Pengujian dilakukan terhadap 150 citra data uji pada proporsi 90:5:5. *Confusion matrix* yang dihasilkan pada Tabel 2.

Tabel 2: Confusion Matrix model CNN (MobileNetV2).

		Prediksi		
		Hairline	Small	Moderate
Aktual	Hairline	45	5	0
	Small	3	47	0
	Moderate	2	0	48

Akurasi yang ditunjukkan pada model CNN secara keseluruhan sebesar 93% dengan nilai *precision*, *recall*, dan *F1-score* per kelas sebagaimana ditampilkan pada Tabel 3.

Tabel 3: Metrik Evaluasi Model CNN per Kelas.

Kelas	Precision	Recall	F1-Score
Hairline	0,90	0,90	0,90
Small	0,90	0,94	0,92
Moderate	1,00	0,96	0,98
Rata-rata Makro	0,93	0,93	0,93

Hasil Model Swin Transformer

Model Swin Transformer yang digunakan adalah varian Swin-Tiny (Swin-T) yang diimplementasikan melalui *library* timm. Pengujian juga dilakukan terhadap 150 citra data uji pada proporsi 90:5:5. Hasil dari confusion matrix disajikan pada Tabel 4.

Tabel 4: Confusion Matrix Model Swin Transformer.

		Prediksi		
		Hairline	Small	Moderate
Aktual	Hairline	48	2	0
	Small	1	49	0
	Moderate	0	0	50

Model Swin Transformer mencapai akurasi keseluruhan sebesar 98%. Metrik evaluasi per kelas dirangkum pada Tabel 5.

Tabel 5: Metrik Evaluasi Model Swin Transformer.

Kelas	Precision	Recall	F1-Score
Hairline	0,98	0,96	0,97
Small	0,96	0,98	0,97
Moderate	1,00	1,00	1,00
Rata-rata Makro	0,98	0,98	0,98

Perbandingan Kedua Model

Swin Transformer secara konsisten menghasilkan performa yang lebih baik dibandingkan MobileNetV2 pada seluruh proporsi pembagian data. Model

ini memperoleh akurasi sebesar 95%, 97%, dan 98% pada rasio 70:15:15, 80:10:10, dan 90:5:5, sementara MobileNetV2 mencapai 90%, 92%, dan 93% pada rasio yang sama. Kesalahan klasifikasi pada kedua model umumnya terjadi antara kelas *hairline crack* dan *small crack* yang memiliki kemiripan karakteristik visual, sedangkan kelas *moderate crack* cenderung lebih mudah dikenali. Pada proporsi terbaik (90:5:5), Swin Transformer hanya menghasilkan 3 kesalahan klasifikasi dibandingkan 10 kesalahan pada MobileNetV2. Hasil ini menunjukkan bahwa meskipun kedua model mampu mengklasifikasikan tingkat kerusakan retakan dengan baik, Swin Transformer lebih unggul dalam membedakan kelas-kelas dengan kemiripan visual yang tinggi.

Analisis dan pembahasan

Kedua model menunjukkan performa yang baik dalam mengklasifikasikan tingkat kerusakan retakan beton, meski dengan tingkat keunggulan yang berbeda-beda. Berdasarkan confusion matrix, kesalahan klasifikasi pada kedua model lebih banyak terjadi antara kelas *hairline crack* dan *small crack*, sementara kelas *moderate crack* cenderung lebih mudah dikenali dengan tingkat kesalahan yang lebih rendah. Hal ini menunjukkan bahwa kedua kelas pertama memiliki kemiripan visual yang lebih tinggi sehingga lebih sulit dibedakan oleh model, sebagaimana terlihat pada Gambar 5 yang menampilkan beberapa contoh citra yang salah diklasifikasikan oleh model MobileNetV2, di mana citra *hairline crack* yang diprediksi sebagai *small crack* memiliki karakteristik visual yang serupa, terutama dalam hal ketebalan garis retakan dan tingkat kontras terhadap permukaan beton.



Gambar 5: Aktual dan prediksi citra.

4. Simpulan dan Saran

Perbandingan dua arsitektur deep learning yaitu CNN berbasis MobileNetV2 dan Swin Transformer, untuk klasifikasi tingkat kerusakan retakan bangunan ke dalam tiga kelas, *hairline crack*, *small crack*, dan *moderate crack*. Berdasarkan hasil evaluasi pada rasio pembagian data terbaik (90:5:5) terhadap 150 citra uji, model Swin Transformer (Swin-T) mencapai akurasi 98% mengungguli model CNN MobileNetV2 yang

mendapatkan hasil 93%. Keunggulan Swin Transformer disebabkan oleh mekanisme *Window-based Self-Attention* dan *Shifted Window Self-Attention* yang mampu menangkap hubungan fitur secara lokal maupun global, sehingga lebih efektif dalam mengenali pola retakan yang bervariasi dibandingkan CNN yang terbatas pada *receptive field* lokal.

Daftar Pustaka

- Bukaita, W., Vankudothu, K., & Khan, J. (2025). *Automated Multi-Class Concrete Crack Detection and Severity Classification Using CNN-Based Deep Learning*. 13(4), 197–210. <https://doi.org/10.11648/j.ajce.20251304.12>.
- Firnando, R., Buchari, M. A., & Marjusalinah, A. D. (2025). *Performance analysis of MobileNetV2 based automatic waste classification using transfer learning*. 14(1), 257–268. <https://doi.org/10.35335/mandiri.v14i1.451>.
- Gowda, P. et al. (2025) 'Adaptive Swin Transformer V2-Tiny Based Model for Classification of Bacteria , Fungus , Virus , and Healthy Fruit and Leaf Images', (August), pp. 1–17. <https://doi.org/10.47852/bonviewAIA52026081>.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. <https://doi.org/https://doi.org/10.48550/arXiv.2103.14030>.
- Lumbanraja, F.R. and Ekananda, R. (2022) 'Classification of Cracked Concrete Images Using Convolutional Neural Algorithm', 2563(October), pp. 040006-1-040006–7. <https://doi.org/10.1063/5.0103114>.
- Novigu, C., Buulolo, D., Sipil, T., Sains, F., Teknologi, D., Nias, U., & Utara, S. (2026). *ANALISIS KERUSAKAN PADA STRUKTUR BANGUNAN*. 02, 36–42. <https://doi.org/10.70134/satedik.v2i1.1268>.
- Sardika, R. P., & Widhiarso, W. (2025). *Klasifikasi Otomatis Tingkat Kerusakan Retak Bangunan pada Citra Digital Menggunakan MobileNetV2 dan Augmentasi Data*. 5(1), 108–124. <https://doi.org/10.29240/arcitech.v5i1.13938>.
- Sarhadi, A., Ravanshadnia, M., Ghanbari, M., & Monirabbasi, A. (2024). *Optimizing Concrete Crack Detection : An Attention-Based SWIN U-Net Approach*. 12(June), 77575–77585. <https://doi.org/https://doi.org/10.1109/ACCESS.2024.3403389>.
- Suherman, N. P., Palar, P. S., & Zuhal, L. R. (2023). *Convolutional Neural Network and Interpretable Deep Learning for Concrete Crack Image Classification*. 3, 1–10. <https://gcs.itb.ac.id/proceeding-igsc/index.php/igsc/article/view/128>.
- Yahfizham. (2024). Studi Kepustakaan: Pengenalan 4 Algoritma Pada

Pembelajaran Deep Learning Beserta Implikasinya. *Pendekar: Jurnal Pendidikan Berkarakter*, 2(1), 270–281. <https://doi.org/10.51903/pendekar.v2i1.598>.

Ye, G., Dai, W., Tao, J., Qu, J., Zhu, L., & Jin, Q. (2024). An improved transformer - based concrete crack classification method. *Scientific Reports*, 1–11. <https://doi.org/10.1038/s41598-024-54835-x>.