VOL. 3. NO. 03 MEI 2026

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

EXPLAINABLE DEEP LEARNING MODELS FOR ENHANCED HANDWRITTEN TEXT RECOGNITION ACCURACY

Septia Dewi Rahayu[1], Nana Suarna[2], Agus Bahtiar[3], Nining Rahaningsih[4], Willy Prihartono[5].

Program Studi Teknik Informatika[123]
Program Studi Komputerisasi Akuntansi[45]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
septiadewi.rahayu09@gmail.com

(*) Corresponding Author : septiadewi.rahayu09@gmail.com
Published : 30 Mei 2026

Abstract—*This study aims to evaluate the performance and interpretability of two deep learning models for handwriting recognition, namely the Convolutional Recurrent Neural Network and the CNN-Transformer. The main problem addressed is the low accuracy in word-level prediction and the limited understanding of which image regions contribute most to model decisions. A quantitative approach was employed by measuring character error rate and word error rate, while a qualitative approach was conducted using sensitivity map visualization to assess the consistency of each model's attention to character structures. The dataset consisted of handwriting images that underwent systematic preprocessing, including normalization and size adjustment. The results show that the Convolutional Recurrent Neural Network demonstrates more stable performance and stronger generalization than the CNN-Transformer, achieving lower character-level errors despite both models exhibiting substantial differences in word-level errors. Explainable visualization indicates that the Convolutional Recurrent Neural Network consistently highlights relevant character strokes, while the CNN-Transformer displays a more dispersed focus pattern. These findings suggest that the ability to capture structural details of handwritten characters plays a crucial role in improving recognition accuracy. This research contributes an integrated analysis combining quantitative evaluation and visual interpretability, providing a foundation for developing more transparent and accurate handwriting recognition models.*

Keywords: *Handwriting Recognition, Deep Learning, Interpretability, Model Accuracy*

Abstrak—Penelitian ini bertujuan untuk mengevaluasi kinerja dan interpretabilitas dua model *deep learning* untuk pengenalan tulisan tangan, yaitu *Convolutional Recurrent Neural Network* (CRNN) dan CNN-Transformer. Permasalahan utama yang dibahas adalah rendahnya akurasi prediksi pada tingkat kata serta keterbatasan pemahaman mengenai bagian citra yang paling berkontribusi terhadap keputusan model. Pendekatan kuantitatif dilakukan dengan mengukur *Character Error Rate* (CER) dan *Word Error Rate* (WER), sedangkan pendekatan kualitatif dilakukan melalui visualisasi sensitivity map untuk menilai konsistensi perhatian masing-masing model terhadap struktur karakter. Dataset yang digunakan terdiri atas citra tulisan tangan yang melalui tahap praproses secara sistematis, termasuk normalisasi dan penyesuaian ukuran. Hasil penelitian menunjukkan bahwa CRNN memiliki performa yang lebih stabil dan kemampuan generalisasi yang lebih baik dibandingkan CNN–Transformer, dengan tingkat kesalahan karakter yang lebih rendah meskipun kedua model menunjukkan perbedaan yang cukup besar pada kesalahan tingkat kata. Visualisasi interpretabilitas menunjukkan bahwa CRNN secara konsisten menyoroti goresan karakter yang relevan, sedangkan CNN–Transformer menampilkan pola fokus yang lebih tersebar. Temuan ini menunjukkan bahwa kemampuan dalam menangkap detail struktural karakter tulisan tangan berperan penting dalam meningkatkan akurasi pengenalan. Penelitian ini memberikan kontribusi berupa analisis terintegrasi yang menggabungkan evaluasi kuantitatif dan interpretabilitas visual, sehingga dapat menjadi dasar bagi pengembangan model pengenalan tulisan tangan yang lebih transparan dan akurat.

Kata Kunci: Pengenalan Tulisan Tangan, Deep Learning, Interpretabilitas, Akurasi Model

INTRODUCTION

Di tengah perkembangan teknologi saat ini, proses digitalisasi dokumen mulai dari arsip

berbasis kertas, catatan manual, hingga naskah historis menjadi kebutuhan yang semakin penting untuk mendukung pengelolaan informasi modern. Sistem Handwritten Text Recognition (HTR) hadir sebagai solusi yang mampu mengubah tulisan tangan ke dalam bentuk teks digital secara otomatis, sehingga proses pencarian, penyimpanan, dan analisis data dapat dilakukan dengan lebih cepat dan terstruktur [1]. Meskipun demikian, karakteristik tulisan tangan yang sangat beragam, seperti variasi gaya penulisan, perbedaan ukuran huruf, tingkat kemiringan, hingga tekanan pena yang tidak konsisten, memunculkan tantangan signifikan dalam mengidentifikasi pola visual dan menyusun urutan karakter secara akurat [2]. Kompleksitas ini membuat sistem HTR memerlukan pendekatan komputasional yang canggih, termasuk kemampuan untuk menangkap detail halus pada bentuk huruf, menyesuaikan model dengan dinamika goresan individu, serta mengatasi noise yang sering muncul akibat kualitas dokumen yang menurun. Dengan demikian, keberhasilan digitalisasi berbasis HTR tidak hanya bergantung pada algoritma pengenalan yang kuat, tetapi juga strategi pra-pemrosesan dan pelatihan model yang mampu mengakomodasi keragaman tulisan manusia.

Pendekatan konvensional berbasis *feature engineering* terbukti belum mampu menangani tingkat variasi dan kompleksitas bentuk tulisan tangan yang sangat tinggi. Kehadiran metode deep learning, khususnya kombinasi CNN–RNN seperti pada arsitektur CRNN, memberikan peningkatan signifikan dalam akurasi HTR karena kemampuannya mengekstraksi pola spasial sekaligus memodelkan urutan karakter secara berkelanjutan [3]. Walaupun demikian, arsitektur tersebut masih menemui kendala ketika berhadapan dengan tulisan tangan yang sangat bervariasi, tidak konsisten, atau memiliki panjang urutan karakter yang besar, sehingga performanya menjadi kurang stabil. Di sisi lain, model berbasis Transformer mulai menunjukkan hasil yang menjanjikan berkat mekanisme self-attention yang mampu memahami keterhubungan global antar karakter dalam satu rangkaian tulisan [4][5]. Meski potensinya besar, efektivitas Transformer pada kondisi dataset yang terbatas atau tanpa proses pretraining masih belum teruji secara komprehensif dalam berbagai studi empiris. Dengan mempertimbangkan batasan masing-masing pendekatan, diperlukan analisis komparatif yang lebih sistematis antara CRNN dan arsitektur Transformer khususnya dalam skema *writer-independent*, yaitu ketika model dilatih menggunakan tulisan dari sekelompok penulis tertentu dan diuji pada penulis yang tidak pernah terlihat sebelumnya. Evaluasi semacam ini penting untuk menilai sejauh mana model benar-benar mampu melakukan generalisasi terhadap keragaman tulisan tangan yang muncul di dunia nyata.

Berbagai kajian terdahulu menunjukkan bahwa arsitektur CRNN masih menjadi salah satu pendekatan yang kompetitif dalam tugas HTR, terutama karena kombinasi modul *convolutional* dan *recurrent* yang memungkinkan model mengekstraksi pola visual lokal sekaligus menangkap urutan karakter secara lebih stabil [6]. Di sisi lain, kemunculan model berbasis Transformer seperti TrOCR memperlihatkan kemampuan yang lebih unggul ketika dilatih pada korpus berskala besar, terutama berkat proses *pretraining* yang memperkaya representasi teks dan visual [7]. Ulasan mendalam oleh Rakesh at al. (2024) juga menekankan bahwa performa model HTR tidak hanya ditentukan oleh arsitekturnya, tetapi sangat bergantung pada kualitas preprocessing, ketersediaan data pelatihan, serta penyesuaian parameter yang digunakan. Namun demikian, kesenjangan penelitian masih terlihat karena hanya sedikit studi yang membandingkan CRNN dan CNN-Transformer secara langsung dalam kondisi eksperimental yang benar-benar setara, khususnya pada dataset CVL.

Penelitian ini bertujuan memberikan evaluasi komprehensif terhadap performa CRNN dan CNN-Transformer melalui desain komparatif pada dataset CVL *writer-independent* yang memuat 996 citra tulisan tangan. Analisis dilakukan dengan mengacu pada metrik evaluasi utama seperti Character Error Rate (CER) dan Word Error Rate (WER), yang kemudian dilengkapi dengan telaah dinamika proses pembelajaran melalui *kurva training* dan *validation loss* untuk melihat stabilitas konvergensi masing-masing model. Selain itu, penelitian ini memanfaatkan *saliency map* sebagai pendekatan Explainable HTR untuk menginterpretasi fokus perhatian model terhadap elemen visual yang relevan selama proses pengenalan. Kombinasi metode evaluasi ini dirancang agar mampu memberikan gambaran yang lebih menyeluruh mengenai performa, ketahanan, dan kecenderungan kesalahan pada kedua arsitektur.

Kontribusi utama penelitian terletak pada penyatuan evaluasi kuantitatif, analisis konvergensi, dan interpretasi visual dalam satu kerangka studi yang koheren untuk membandingkan dua arsitektur HTR modern dalam

skenario writer-independent. Aspek kebaruan penelitian tampak melalui pemanfaatan *saliency map* untuk menilai konsistensi perhatian model terhadap pola stroke tulisan tangan, sehingga memberi wawasan mengenai sejauh mana model mengekstraksi fitur visual yang benar-benar bermakna. Selain itu, penelitian ini menyajikan evaluasi empiris terhadap CRNN dan CNN-Transformer tanpa mengandalkan proses *pretraining*, sehingga dapat menggambarkan performa murni kedua arsitektur dalam kondisi yang lebih terkontrol. Pendekatan tersebut diharapkan dapat menjadi dasar metodologis bagi pengembangan sistem HTR yang lebih transparan, adaptif, dan relevan untuk implementasi di berbagai konteks dunia nyata.

MATERIALS AND METHODS

Penelitian ini menggunakan CVL Handwriting Dataset yang terdiri dari 996 citra tulisan tangan beserta transkripsi ground-truth. Berikut tahapan perancangan penelitian.

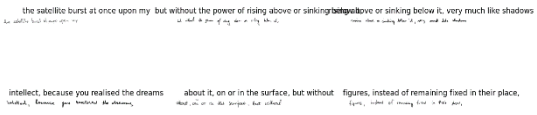
1. Preprocessing
   Seluruh citra melalui tahap praproses yang meliputi *resizing*, binarisasi, dan augmentasi data untuk menstabilkan kualitas visual serta meningkatkan variasi sampel pelatihan.
2. Prosedur Pelatihan
   Proses pelatihan dilakukan untuk membandingkan performa CRNN dan CNN–Transformer dalam mengenali tulisan tangan. Pada CRNN, fitur visual yang dihasilkan CNN diproses oleh Bidirectional LSTM untuk menangkap ketergantungan sekuensial dua arah antar karakter. Sebaliknya, CNN–Transformer menggantikan komponen rekuren dengan mekanisme *self-attention* yang mampu memodelkan konteks global secara paralel.
3. Evaluasi
   Kinerja model dievaluasi menggunakan Character Error Rate (CER) dan Word Error Rate (WER) untuk mengukur tingkat kesalahan pada level karakter dan kata. CER menghitung proporsi kesalahan karakter terhadap label acuan, sedangkan WER menilai kesalahan pada tingkat kata.

4. Analisis Interpretabilitas
   Analisis interpretabilitas dilakukan dengan menggunakan saliency map dan Grad-CAM untuk menyoroti area citra yang paling berpengaruh terhadap prediksi model.

RESULTS AND DISCUSSION

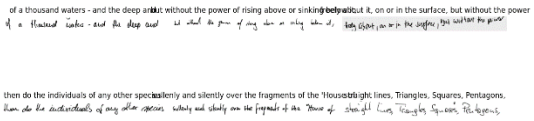
1. Hasil Preprocessing Data

Tahap preprocessing diterapkan untuk memastikan seluruh citra berada pada kondisi yang seragam sebelum diproses oleh model CRNN maupun CNN–Transformer.



Gambar 1 Hasil Preprocesing CRNN

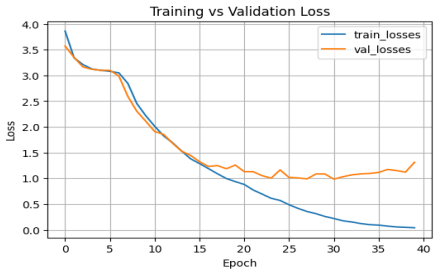
Gambar tersebut menampilkan hasil preprocessing model CRNN terhadap tulisan tangan di mana teks cetak dan tulisan tangan diproses dalam satu alur terintegrasi dengan susunan baris yang tetap linier. Hal ini mencerminkan kemampuan CRNN dalam mengekstraksi fitur spasial melalui lapisan konvolusi serta memodelkan keterkaitan antar karakter secara berurutan melalui lapisan rekuren.



Gambar 2 Hasil Preprocessing CNN-Transformer

Gambar tersebut menampilkan hasil preprocessing model CNN-Transformer yang menonjolkan pemodelan fitur secara spasial dan kontekstual yang lebih luas. Struktur teks dalam gambar terlihat diproses dengan mekanisme yang memungkinkan analisis hubungan antarbagian teks secara menyeluruh, tidak terbatas pada pola linier.

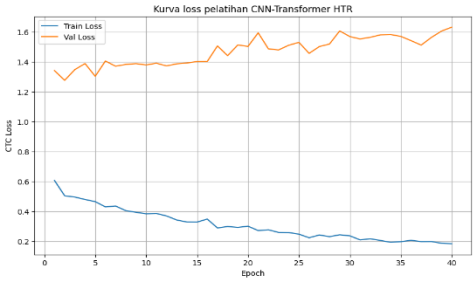
2. Hasil Pelatihan Model



Gambar 3. Kurva Loss CRNN

Gambar tersebut menunjukkan bahwa pelatihan model CRNN mengalami penurunan *training loss* yang stabil dari sekitar 3,8 hingga mendekati 0,2 pada epoch akhir. *Validation loss* yang awalnya menurun sejalan dengan *training loss* mulai

berfluktuasi sejak sekitar epoch ke-15, yang mengindikasikan gejala *overfitting* moderat. Kondisi ini menunjukkan bahwa model mulai lebih menyesuaikan diri dengan data latih dibandingkan data validasi, meskipun masih dalam batas yang dapat diterima dan sejalan dengan karakteristik arsitektur CRNN yang stabil namun sensitif terhadap ukuran dan keberagaman dataset [8].



Gambar 4. Kurva Loss CNN-Transformer

Hasil pada gambar menunjukkan bahwa arsitektur CNN–Transformer mengalami penurunan *training loss* secara konsisten hingga sekitar 0,2, yang menandakan kemampuan model dalam mempelajari data latih dengan baik. Namun, *validation loss* tidak mengikuti tren tersebut dan tetap berada pada kisaran 1,3–1,6 serta mengalami beberapa peningkatan selama pelatihan. Perbedaan yang cukup besar antara *training loss* dan *validation loss* ini mengindikasikan terjadinya *overfitting* yang signifikan, di mana model cenderung menyesuaikan diri dengan data latih tetapi kurang mampu melakukan generalisasi terhadap data baru. Kondisi ini sejalan dengan literatur yang menyebutkan bahwa arsitektur berbasis Transformer umumnya memerlukan jumlah data pelatihan yang lebih besar serta strategi regularisasi yang lebih kuat untuk mencapai konvergensi dan generalisasi yang optimal [7], [9].

3. Evaluasi Model

	metric	value
0	cer	0.295086
1	wer	0.746584

Gambar 5. Evaluasi CRNN

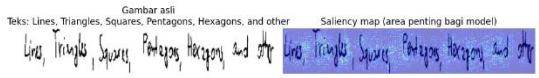
Model CRNN memperoleh nilai CER sebesar 0,2951 dan WER sebesar 0,7466. Nilai CER yang lebih rendah menunjukkan bahwa sebagian besar karakter dapat dikenali dengan baik, sedangkan WER yang masih tinggi mengindikasikan kesulitan model dalam menyusun karakter menjadi kata yang utuh. Hal ini menunjukkan keterbatasan dalam memodelkan dependensi konteks yang lebih panjang, terutama pada variasi tulisan tangan yang kompleks, sebagaimana juga dilaporkan pada penelitian sebelumnya [10].

	metric	value
0	CER	0.296037
1	WER	0.776650

Gambar 6. Evaluasi CNN-Transformer

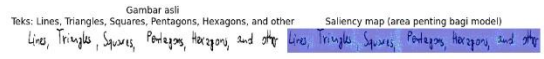
Model CNN–Transformer mencatat CER sebesar 0,2960 dan WER sebesar 0,7767. Nilai CER yang hampir sama dengan CRNN menunjukkan kemampuan pengenalan karakter yang sebanding, namun WER yang lebih tinggi menandakan kesulitan dalam mempelajari koherensi urutan karakter pada tingkat kata. Kondisi ini menunjukkan bahwa mekanisme *self-attention* pada Transformer sangat bergantung pada jumlah dan kualitas data pelatihan, sehingga pada data terbatas performa generalisasi cenderung menurun [11].

4. Analisis Explainable



Gambar 7. Explainable CRNN

Hasil visualisasi menunjukkan perbandingan antara citra tulisan tangan asli dan peta *saliency* yang dihasilkan model CRNN untuk mengidentifikasi area yang paling berpengaruh dalam prediksi. Peta *saliency* memperlihatkan perhatian model yang mengikuti alur karakter tulisan, terutama pada goresan huruf utama, meskipun masih terdapat aktivasi lemah pada latar belakang. Pola ini menunjukkan bahwa CRNN cukup efektif dalam menangkap fitur sekuensial teks serta mempertahankan akurasi pada tingkat karakter dan kata.



Gambar 8. Explainable CNN-Tranformer

Visualisasi *saliency map* pada CNN–Transformer terhadap citra yang sama menunjukkan perhatian yang lebih terfokus pada bagian huruf yang informatif dan relatif minim gangguan latar belakang. Hal ini mencerminkan kemampuan mekanisme *self-attention* dalam menangkap konteks global. Namun demikian, meskipun fokus visual lebih tajam, performa prediksi pada tingkat

kata masih belum melampaui model CRNN, sehingga ketajaman perhatian tidak selalu berbanding lurus dengan peningkatan metrik pengenalan.

CONCLUSION

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut.

1. Penelitian ini berhasil melakukan evaluasi komparatif antara model Convolutional Recurrent Neural Network (CRNN) dan CNN–Transformer pada tugas *Handwritten Text Recognition (HTR)* menggunakan CVL dataset dengan skema *writer-independent*. Hasil analisis kuantitatif menggunakan metrik Character Error Rate (CER) dan Word Error Rate (WER) serta pengamatan terhadap dinamika konvergensi pelatihan menunjukkan bahwa model CRNN memiliki proses pembelajaran yang lebih stabil dan kemampuan generalisasi yang lebih baik dibandingkan CNN–Transformer.
2. Model CRNN memperoleh nilai CER sebesar 0,2951 dan WER sebesar 0,7466**,** sedangkan CNN–Transformer menghasilkan nilai CER yang hampir sama namun memiliki WER yang lebih tinggi**.** Hal ini menunjukkan bahwa kedua model memiliki kemampuan pengenalan karakter yang sebanding, tetapi CNN–Transformer masih mengalami kesulitan dalam memahami struktur kata secara utuh.
3. Analisis saliency map menunjukkan bahwa model CRNN memiliki fokus visual yang lebih terarah pada stroke huruf yang relevan, sedangkan CNN–Transformer memperlihatkan perhatian yang lebih tersebar dan kurang selektif. Pola ini membantu menjelaskan perbedaan kinerja kedua model dalam hal generalisasi pada tingkat kata serta menunjukkan bahwa CRNN lebih responsif terhadap fitur struktural tulisan tangan.
4. Secara keseluruhan, penelitian ini menunjukkan bahwa CRNN merupakan arsitektur yang lebih efektif dan relatif tangguh pada kondisi data terbatas**,** sedangkan CNN–Transformer masih memiliki keterbatasan tanpa dukungan *pretraining* berskala besar. Pengembangan penelitian selanjutnya dapat dilakukan melalui penerapan *self-supervised learning* atau *large-scale pretraining* untuk meningkatkan stabilitas Transformer, penggunaan *language model* untuk menurunkan nilai WER, penerapan teknik augmentasi lanjutan guna memperkaya variasi data, serta eksplorasi metode interpretabilitas yang lebih lanjut seperti Integrated Gradients dan Attention Rollout guna memperoleh pemahaman yang lebih mendalam terhadap mekanisme pengambilan keputusan model.

REFERENCE

[1] S. Rakesh, P. K. Reddy, V. Prashanth, and K. S. Reddy, “Handwritten text recognition using deep learning techniques : a survey,” vol. 01126, 2024.

[2] R. Li, “A review of neural networks in handwritten character recognition,” vol. 0, pp. 169–174, 2024, doi: 10.54254/2755-2721/92/20241736.

[3] K. Chaudhary, “Easter 2 . 0 : IMPROVING CONVOLUTIONAL MODELS FOR HANDWRITTEN,” 2022.

[4] D. Kass and E. Vats, “AttentionHTR : Handwritten Text Recognition Based on Attention Encoder-Decoder Networks,” 2022.

[5] M. Hamdan, A. Rahiche, M. Cheriet, and S. Member, “HTR-JAND : Handwritten Text Recognition with Joint Attention Network and Knowledge Distillation,” pp. 1–13, 2025.

[6] E. M. J. Calvo-zaragoza and E. Mas-candela, “Exploring recursive neural networks for compact handwritten text recognition models,” *Int. J. Doc. Anal. Recognit.*, vol. 27, no. 3, pp. 213–223, 2024, doi: 10.1007/s10032-024-00481-y.

[7] M. Li *et al.*, “TrOCR : Transformer-based Optical Character Recognition with Pre-trained Models,” 2022.