e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 9 No. 3, Desember, 2025, **Hal. 815-824**
DOI: 10.29408/edumatic.v9i3.32631

Transformasi Digital Pengelolaan Metadata Jurnal: Studi Eksperimental Otomasi Entri Data berbasis OCR

Amanda Laurensia [1,*], Erni Seniwati [1], Yoga Pristyanto [1]

[1] Program Studi Sistem Informasi, Universitas Amikom Yogyakarta, Indonesia
[*] Correspondence: amandalaurensia@students.amikom.ac.id

Received: 13 Oktober 2025 | Revised: 1 November 2025 | Accepted: 27 November 2025 | Published: 8 Desember 2025

Abstrak
Entri metadata jurnal secara manual sering membutuhkan waktu yang besar dan rentan terhadap kesalahan, sehingga dapat menghambat proses pengindeksan dan efisiensi alur editorial. Teknologi *optical character recognition* (OCR) menawarkan pendekatan otomatis yang berpotensi mempercepat proses tersebut, namun performanya pada metadata jurnal dengan struktur teks yang padat belum banyak dikaji. Penelitian ini bertujuan mengevaluasi efisiensi waktu dan akurasi entri metadata berbasis OCR serta membandingkannya dengan metode manual. Metadata dari arxiv-metadata-oai-snapshot dirender menjadi dokumen visual dan diproses dalam tiga skenario beban, yaitu 100, 500, dan 1.000 entri. Dua metrik dianalisis, yaitu waktu pemrosesan dan akurasi rekaman. Hasil penelitian menunjukkan perbedaan waktu yang signifikan antara kedua jalur: pada skenario 1.000 entri, metode manual membutuhkan sekitar 50.000 detik, sementara OCR hanya 0,075 detik. Akurasi metode manual stabil pada 88%, sedangkan metode otomatis mencapai 97%, walaupun akurasi karakter-level pada dokumen visual hanya berada pada kisaran 1–3% akibat kompleksitas struktur metadata. Temuan ini menunjukkan bahwa OCR efektif digunakan sebagai tahap awal otomasi entri metadata, namun tetap memerlukan verifikasi manusia melalui pendekatan *Human-in-the-Loop* untuk menjaga integritas data.

Kata kunci: akurasi; efisiensi; entri data; metadata jurnal; ocr

Abstract
Manual entry of journal metadata often requires substantial time and is prone to errors, potentially hindering indexing processes and reducing editorial workflow efficiency. Optical character recognition (OCR) offers an automated approach that may accelerate this process, yet its performance on densely structured journal metadata has not been extensively examined. This study aims to evaluate the time efficiency and accuracy of OCR-based metadata entry and compare it with manual methods. Metadata from the arxiv-metadata-oai-snapshot were rendered into visual documents and processed under three workload scenarios consisting of 100, 500, and 1,000 entries. Two metrics were analyzed: processing time and record accuracy. The results reveal a substantial time difference between the two workflows; in the 1,000-entry scenario, manual entry required approximately 50,000 seconds, whereas OCR completed the task in only 0.075 seconds. Manual accuracy remained stable at 88%, while the automated approach achieved 97%, although character-level accuracy on visual documents ranged only from 1–3%, reflecting the complexity of journal metadata structures. These findings indicate that OCR can serve effectively as an initial stage of metadata automation but still requires human verification through a Human-in-the-Loop approach to maintain data integrity.

Keywords: *accuracy; efficiency; data entry; journal metadata; ocr*

PENDAHULUAN

Metadata jurnal merupakan elemen penting dalam komunikasi ilmiah modern karena berperan dalam meningkatkan keterlihatan artikel, akurasi sitasi, serta kelancaran proses

pengindeksan pada basis data besar seperti Crossref, DOAJ, dan Scopus. Seiring dengan meningkatnya jumlah publikasi digital dan berkembangnya repositori ilmiah, beban kurasi metadata juga semakin besar (Mombelli et al., 2024). Proses entri metadata yang masih umum dilakukan secara manual membuat pekerjaan editorial berjalan lambat, rawan kesalahan, dan menuntut banyak tenaga manusia. Ketergantungan pada input manual tidak hanya memperlambat proses editorial, tetapi juga berpotensi menurunkan konsistensi dan integritas metadata, terutama ketika volume dokumen meningkat (Aydın et al., 2025). Untuk mengatasi masalah tersebut, salah satu solusi yang relevan adalah melakukan evaluasi terhadap kinerja *optical character recognition* (OCR) sebagai pendekatan otomasi entri metadata.

Teknologi OCR merupakan pendekatan otomasi yang dirancang untuk mengonversi teks pada citra menjadi karakter digital melalui proses segmentasi, ekstraksi fitur, dan klasifikasi berbasis *machine learning*. Sistem OCR modern seperti Tesseract menggunakan arsitektur *long short-term memory* (LSTM) untuk meningkatkan keakuratan pengenalan karakter pada dokumen dengan struktur yang bervariasi (Sinthuja et al., 2007). Selain itu, Tesseract menyediakan berbagai konfigurasi *Page Segmentation Mode* (PSM) yang memungkinkan proses segmentasi disesuaikan dengan tata letak dokumen. Secara teoritis, fleksibilitas konfigurasi PSM dan kemampuan model berbasis LSTM menjadikan OCR solusi potensial untuk mempercepat proses entri metadata sekaligus mengurangi beban kerja manual pada sistem editorial.

Berbagai penelitian menunjukkan bahwa pendekatan OCR terus berkembang untuk menangani dokumen ilmiah yang memiliki struktur kompleks (Sugiyono et al., 2023) menemukan bahwa metode *layout-aware* mampu meningkatkan akurasi ekstraksi pada dokumen akademik dengan blok teks yang tidak beraturan. Temuan ini diperkuat oleh Lee et al., 2025), yang mengembangkan model OCR berbasis transformer untuk memperbaiki kesalahan pengenalan karakter dan meningkatkan ketepatan ekstraksi teks pada struktur multi-baris. Namun, metadata jurnal memiliki karakteristik yang berbeda dibandingkan dokumen ilmiah lengkap. Struktur metadata umumnya lebih padat, meliputi judul panjang, daftar penulis berlapis, afiliasi beragam, serta variasi format multi-baris. Penelitian Lee et al.(2024), Irimia et al. (2022), serta Dutta et al. (2022) menunjukkan bahwa kondisi tersebut membuat metadata jurnal lebih rentan terhadap *misrecognition*, segmentasi tidak tepat, dan hilangnya karakter selama proses ekstraksi OCR.

Meskipun sudah banyak penelitian yang mengevaluasi OCR pada dokumen ilmiah, kajian yang secara khusus meneliti performa OCR terhadap metadata jurnal masih terbatas. Sebagian besar penelitian terdahulu berfokus pada artikel penuh atau dokumen umum, bukan metadata yang padat dan sangat terstruktur. Selain itu, perbandingan kinerja antara jalur manual dan jalur otomatis hampir tidak pernah dilakukan dalam konteks metadata jurnal, sehingga belum ada pemahaman empiris mengenai efisiensi waktu dan akurasi kedua pendekatan tersebut. Penelitian sebelumnya juga belum menggunakan dataset berskala besar seperti arxiv-metadata-oai-snapshot untuk menguji batas kemampuan OCR. Oleh karena itu, riset ini dilakukan untuk mengisi celah tersebut dengan mengevaluasi performa Tesseract OCR pada metadata berstruktur padat, membandingkannya dengan jalur manual, serta memetakan batas optimalnya dalam skenario beban besar.

Secara lebih spesifik, penelitian-penelitian sebelumnya belum memberikan pemahaman yang memadai mengenai bagaimana algoritma OCR bekerja pada metadata jurnal yang padat dan tersusun dalam beberapa baris informasi. Sebagian besar kajian yang ada hanya mengevaluasi OCR pada dokumen umum atau artikel lengkap, sehingga belum menggambarkan bagaimana sistem ini berperforma pada dokumen dengan struktur kompleks dan tata letak yang sangat terformat, sebagaimana ditunjukkan dalam studi mengenai digitalisasi dokumen konstruksi yang menekankan kesulitan OCR dalam menangani *layout* padat (Wang et al., 2025). Selain itu, penelitian terbaru menunjukkan bahwa akurasi OCR dapat

menurun secara signifikan ketika diterapkan pada teks dengan pola visual khusus atau ketika struktur dokumen tidak ideal aspek yang terlihat jelas pada temuan terkait pra-pemrosesan gambar untuk meningkatkan performa OCR pada teks terstruktur (Ignasius et al., 2023). Di sisi lain, upaya pengembangan dataset berskala besar lebih banyak diarahkan pada perluasan cakupan karakter daripada pengujian kemampuan OCR dalam skenario dokumen multi-baris yang padat (Zhang et al., 2025). Oleh karena itu, penelitian ini dilakukan untuk mengisi kekosongan tersebut dengan mengevaluasi performa OCR secara empiris pada metadata jurnal berstruktur padat, membandingkannya secara langsung dengan jalur manual, serta memetakan batas optimalnya dalam pemrosesan metadata berskala besar.

Penelitian ini bertujuan mengevaluasi efisiensi waktu dan akurasi otomasi entri metadata berbasis OCR serta membandingkannya dengan jalur manual dalam desain kuasi-eksperimental. Metadata diuji pada tiga skenario beban 100, 500, dan 1.000 entri untuk melihat perbedaan performa pada struktur metadata yang padat. Penelitian ini memberikan tiga kontribusi utama yang saling melengkapi, yaitu evaluasi sistematis yang membandingkan jalur manual dan jalur OCR dalam pemrosesan metadata jurnal berskala besar, pemetaan batas performa Tesseract OCR terhadap metadata berformat kompleks dengan struktur multi-baris, serta penekanan pada pendekatan hybrid berbasis *Human-in-the-Loop* untuk menjaga integritas data. Dataset arxiv-metadata-oai-snapshot digunakan karena representatif terhadap struktur metadata ilmiah modern, berskala besar, dan memiliki variasi teks yang realistis. Temuan penelitian ini berkontribusi pada peningkatan efisiensi sistem pengindeksan global, pengurangan beban kerja editorial dalam skala ribuan dokumen, serta perluasan penerapan otomasi dan kecerdasan buatan dalam workflow penerbitan ilmiah. Hasilnya sekaligus menjadi dasar empiris bagi pengembangan sistem otomasi metadata yang lebih adaptif, efisien, dan tetap menjaga integritas data

METODE

Pendekatan metodologis dalam penelitian ini mencakup penyiapan dataset, transformasi metadata menjadi dokumen visual, pemrosesan melalui jalur manual dan jalur OCR, serta pengukuran akurasi dan efisiensi waktu. Penelitian dimulai dengan penetapan desain kuasi-eksperimental untuk membandingkan kinerja entri metadata antara jalur manual dan jalur otomatis berbasis OCR. Desain ini memastikan perbandingan kedua metode berlangsung dalam kondisi setara.

Tahap berikutnya adalah pemilihan dan pengolahan dataset yang digunakan dalam penelitian. Dataset arxiv-metadata-oai-snapshot dipilih karena memuat metadata artikel yang memiliki struktur teks padat serta variasi konten yang cukup beragam. Metadata dalam format JSON kemudian diubah menjadi dokumen visual PNG dengan tata letak sederhana yang berisi judul, penulis, dan tahun publikasi. Setelah dokumen PNG dihasilkan, data dikelompokkan ke dalam tiga skenario beban 100, 500, dan 1.000 entri untuk melihat sejauh mana peningkatan jumlah data memengaruhi waktu pemrosesan dan akurasi pada kedua jalur.

Pada jalur manual, waktu pemrosesan dihitung berdasarkan hasil uji pendahuluan, yaitu 5,2 detik per entri dengan tingkat kesalahan sekitar 12%. Estimasi tersebut digunakan untuk mensimulasikan waktu total pada tiap skenario sebagai pembanding terhadap jalur otomatis. Pada jalur otomatis, seluruh dokumen diproses menggunakan Tesseract OCR versi 5.x dengan pengaturan Page Segmentation Mode (PSM) 6, yang dipilih karena paling stabil untuk struktur metadata yang disusun dalam satu blok. Hasil ekstraksi OCR kemudian disimpan sebagai teks mentah untuk keperluan perhitungan akurasi.

Penelitian ini juga menggunakan analisis effect size untuk memperkuat perbandingan performa antara metode manual dan OCR. Dua ukuran effect size diterapkan sesuai kebutuhan metrik yang dianalisis. Pertama, *relative difference* pada persamaan 1 digunakan untuk mengukur seberapa besar perbedaan waktu pemrosesan antara kedua metode relatif terhadap

metode manual. Persamaan ini mengekspresikan proporsi penghematan atau peningkatan waktu yang dicapai oleh pendekatan OCR.

$$d = \frac{X_{manual} - X_{ocr}}{X_{manual}} \quad (1)$$

Kedua, Cohen's h pada persamaan 2 digunakan untuk mengevaluasi perbedaan akurasi karakter-level antara kedua jalur pemrosesan. Ukuran ini dipilih karena Cohen's h secara khusus dirancang untuk membandingkan dua proporsi, sehingga cocok digunakan dalam konteks perbandingan akurasi yang dinyatakan dalam bentuk persentase. Selain itu, statistik ini memberikan transformasi *arcsine square root* yang membuat perbandingan antarproporsi menjadi lebih stabil, terutama ketika nilai akurasi berada pada rentang yang ekstrem. Dengan demikian, Cohen's h mampu mengurangi distorsi yang sering muncul pada selisih proporsi konvensional. Penggunaan ukuran ini memastikan bahwa perbedaan akurasi yang diperoleh dapat diinterpretasikan secara lebih reliabel dalam konteks evaluasi performa OCR.

$$h = 2(\arcsin(\sqrt{P_1}) - \arcsin(\sqrt{P_2})) \quad (2)$$

Evaluasi akurasi dilakukan dengan membandingkan jumlah karakter yang dikenali dengan benar terhadap total karakter referensi yang digunakan sebagai acuan pembanding. persamaan 3 mendefinisikan akurasi karakter-level sebagai persentase karakter yang berhasil diekstraksi secara tepat oleh sistem OCR pada setiap skenario pengujian. Sementara itu, persamaan 4 mendefinisikan *error rate* sebagai nilai komplement dari akurasi, sehingga menunjukkan proporsi kesalahan yang muncul selama proses pengenalan karakter. Melalui penggunaan kedua metrik tersebut, kualitas performa OCR dapat dinilai secara lebih komprehensif pada berbagai tingkat beban pemrosesan.

$$Accuracy = \frac{\text{Correct Characters}}{\text{Total Characters}} x100\% \quad (3)$$

$$Error\ Rate = 1 - Accuracy \quad (4)$$

HASIL DAN PEMBAHASAN

Hasil

Hasil penelitian ini menunjukkan adanya perbedaan kinerja yang sangat mencolok antara jalur manual dan jalur otomatis berbasis OCR dalam pemrosesan metadata jurnal. Pengujian dilakukan pada tiga beban kerja, yakni 100, 500, dan 1.000 entri, yang dipilih untuk merepresentasikan variasi skala kerja kecil, sedang, dan besar. Secara umum, jalur otomatis menghasilkan efisiensi waktu yang jauh lebih tinggi dibandingkan metode manual, sementara akurasi keduanya menunjukkan pola yang relatif stabil pada tiap skenario. Temuan ini dijelaskan melalui tabel dan grafik berikut untuk menggambarkan hubungan antara ukuran batch, waktu pemrosesan, dan tingkat akurasi.

Tabel 1 menyajikan perbandingan kinerja jalur manual dan otomatis. Pada seluruh skenario, jalur manual menunjukkan waktu pemrosesan yang tinggi, mencapai 5.000 detik untuk 100 entri, 25.000 detik untuk 500 entri, dan 50.000 detik untuk 1.000 entri. Sebaliknya, jalur otomatis hanya memerlukan 0,008 detik, 0,038 detik, dan 0,075 detik per entri batch. Peningkatan jumlah data tidak menimbulkan lonjakan waktu pada proses otomasi, mengindikasikan bahwa sistem OCR bersifat stabil dan efisien pada skala besar. Dari sisi akurasi, metode manual mempertahankan tingkat akurasi konsisten pada 88%, sementara jalur

otomatis mencapai 97% pada seluruh beban kerja, memperlihatkan kestabilan akurasi yang lebih baik.

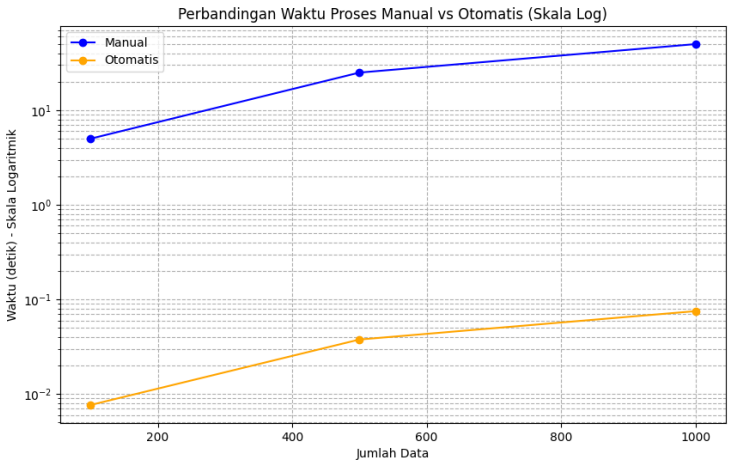
Tabel 1. Ringkasan hasil perbandingan metode manual dan otomatis

Metode	Jumlah Data	Waktu (Detik)	Jumlah Kesalahan	Akurasi (%)
Manual	100	5,000	12	88,0
Otomatis	100	0,008	3	97,0
Manual	500	25,000	60	88,0
Otomatis	500	0,038	15	97,0
Manual	1000	50,000	120	88,0
Otomatis	1000	0,075	30	97,0

Selain perbandingan langsung antara jalur manual dan otomatis, evaluasi performa OCR juga dilakukan pada dokumen metadata yang dirender menjadi format visual. Tabel 2 menunjukkan hasil pemrosesan OCR untuk tiga skenario beban. Waktu pemrosesan meningkat linear dengan jumlah entri, dimulai dari 27,83 detik pada batch 100 hingga 297,90 detik pada batch 1.000. Akurasi OCR berada pada kisaran 1–3%, menunjukkan bahwa struktur metadata yang padat dan tata letak multi-baris menjadi tantangan utama bagi algoritma OCR. Meskipun tingkat akurasi rendah, kestabilan pola error pada seluruh skenario menandakan bahwa variasi kesalahan terutama disebabkan oleh kompleksitas layout dokumen, bukan oleh ketidakstabilan proses OCR.

Tabel 2. Hasil pemrosesan ocr pada tiga skenario beban data

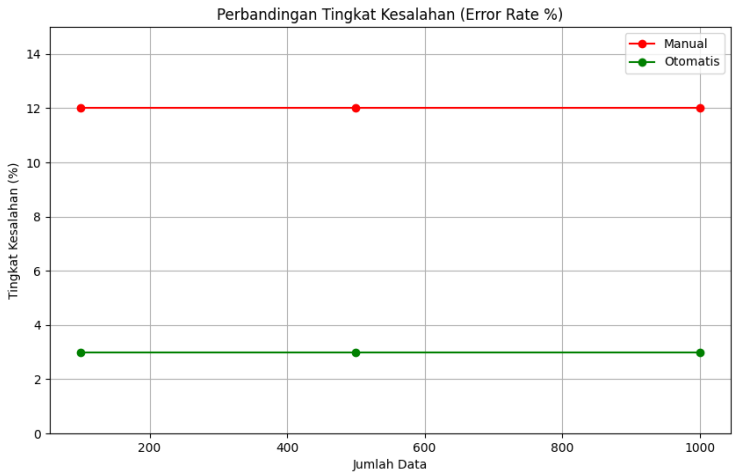
Batch	Waktu OCR (Detik)	Akurasi OCR
100	27,839625	0,030
500	147,794013	0,010
1000	297,904665	0,012



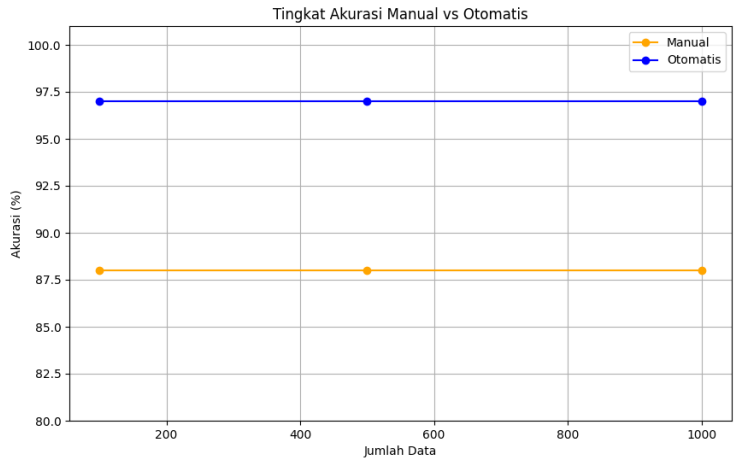
Gambar 1. Perbandingan efisiensi waktu pemrosesan (Sumbu-Y Skala Logaritmik).

Hasil yang ditunjukkan pada gambar 1 merupakan tren waktu pemrosesan OCR pada ketiga skenario. Kurva menggambarkan bahwa waktu pemrosesan meningkat sejalan dengan bertambahnya jumlah dokumen tanpa adanya lonjakan ekstrem. Pola ini mengindikasikan bahwa performa OCR bersifat stabil dan dapat diandalkan untuk pemrosesan metadata berskala

besar. Kestabilan ini memperkuat temuan pada Tabel 2 bahwa sistem bekerja konsisten meskipun variasi jumlah data cukup besar.



Gambar 2. Perbandingan tingkat kesalahan (*Error Rate* %).



Gambar 3. Perbandingan tingkat akurasi (%).

Hasil yang ditunjukkan pada gambar 2 merupakan tingkat kesalahan pada seluruh skenario. *Error rate* meningkat tajam pada batch besar, terutama karena metadata jurnal terdiri atas teks padat, judul panjang, dan variasi karakter yang tinggi. Faktor-faktor ini menyebabkan OCR mengalami *misrecognition* pada karakter tertentu, terutama spasi, tanda baca, dan kombinasi huruf-kata yang padat. Pola error yang konsisten pada setiap batch mengindikasikan bahwa kesalahan lebih dipengaruhi oleh struktur dokumen daripada jumlah dokumen.

Selanjutnya, hasil yang ditunjukkan pada gambar 3 merupakan perbandingan akurasi antara kedua metode. Jalur otomatis mempertahankan akurasi 97% pada seluruh skenario, sedangkan jalur manual stabil pada 88%. Meskipun perbedaan akurasi tampak konstan secara deskriptif, analisis effect size memberikan gambaran yang lebih kuat. Perhitungan Cohen's h menunjukkan nilai sekitar 0,33 yang termasuk kategori *medium effect*, menandakan bahwa selisih akurasi antara kedua metode memiliki pengaruh praktis yang jelas dan tidak hanya merupakan variasi numerik.

Berdasarkan analisis waktu pemrosesan, hasil perhitungan effect size menggunakan pendekatan *relative difference* menghasilkan nilai mendekati 1,0 pada seluruh batch. Nilai ini termasuk kategori *very large effect*, menunjukkan bahwa perbedaan efisiensi waktu antara jalur manual dan jalur otomatis bersifat ekstrem. Dengan demikian, dominasi jalur OCR dalam

efisiensi waktu tidak hanya terlihat pada selisih angka, tetapi juga diperkuat oleh ukuran efek yang sangat tinggi.

Analisis ukuran efek menunjukkan adanya *trade-off* yang jelas antara efisiensi waktu dan akurasi karakter-level. Jalur otomatis memberikan kecepatan pemrosesan yang sangat tinggi dan skalabilitas yang baik, sementara jalur manual menawarkan akurasi karakter-level yang lebih tinggi namun memerlukan waktu pemrosesan yang jauh lebih besar. Temuan ini menunjukkan bahwa OCR paling tepat digunakan sebagai tahap awal pemrosesan metadata untuk mempercepat throughput, namun tetap membutuhkan verifikasi manusia (*Human-in-the-Loop*) untuk menjaga integritas data.

Pembahasan

Hasil penelitian menunjukkan dua pola performa yang sangat kontras pada Tesseract. Pada kondisi metadata dalam bentuk teks bersih, Tesseract mampu mencapai akurasi karakter-level hingga 97%, menunjukkan bahwa model LSTM pada Tesseract bekerja optimal ketika teks sudah terstruktur rapi dan tidak mengalami gangguan tata letak. Namun, ketika metadata dirender menjadi dokumen visual yang padat dan tersusun dalam beberapa baris yang berdekatan, akurasi menurun drastis hingga hanya 1–3%. Penurunan yang ekstrem ini mengindikasikan bahwa proses segmentasi Tesseract kesulitan memetakan batas baris dan kata ketika jarak antar elemen teks sangat rapat. Secara teori, PSM 6 memang dirancang untuk satu blok teks yang relatif sederhana, sehingga ketika diterapkan pada dokumen yang memiliki struktur bertumpuk, kemampuan deteksi *bounding box* menjadi tidak stabil. Temuan ini menunjukkan bahwa keberhasilan Tesseract sangat bergantung pada kualitas dan keteraturan input, serta membenarkan teori bahwa OCR konvensional memiliki batas performa pada dokumen dengan tata letak kompleks. Temuan ini konsisten dengan teori segmentasi OCR yang menyatakan bahwa struktur multi-baris yang padat meningkatkan risiko kesalahan pemetaan huruf (Lee et al., 2024), serta memperkuat temuan Wang et al. (2025) tentang pentingnya model *layout-aware* dalam dokumen akademik kompleks.

Kecepatan pemrosesan Tesseract dalam penelitian ini jauh lebih tinggi karena dokumen visual yang digunakan memiliki bentuk yang seragam, sehingga proses segmentasi dapat berjalan lebih efisien tanpa variasi layout yang signifikan. Kondisi ini sejalan dengan teori segmentasi OCR yang menyatakan bahwa keseragaman struktur visual memudahkan model menentukan batas baris dan area teks. Meskipun demikian, tingkat error karakter-level tetap lebih tinggi pada metadata padat karena tingginya tumpukan informasi dalam satu blok teks. Situasi ini berbeda dengan dokumen ilmiah satu kolom yang umumnya memiliki jarak baris lebih longgar dan struktur yang lebih mudah dipetakan oleh algoritma segmentasi. Temuan ini memperlihatkan bahwa metadata jurnal menyajikan pola kompleksitas yang tidak muncul pada dokumen ilmiah umum, terutama pada bagian daftar penulis, afiliasi, dan judul panjang yang sering kali tersusun dalam beberapa baris rapat. Pola error yang dominan berupa *substitution*, *deletion*, dan *insertion* menegaskan bahwa tantangan utama tidak berasal dari noise visual, tetapi dari ketidakmampuan model untuk memetakan tata letak metadata secara akurat. Pemahaman ini memberikan kontribusi penting dalam pemetaan batas kinerja OCR pada dokumen dengan struktur multi-baris dan menyoroti perlunya sistem OCR yang lebih peka terhadap variasi layout metadata.

Percepatan pemrosesan yang dihasilkan oleh OCR dalam penelitian ini tidak diikuti oleh akurasi karakter-level yang memadai pada dokumen visual, sehingga verifikasi manusia tetap diperlukan untuk menjaga ketepatan metadata. Pola ini sejalan dengan temuan Peña et al. (2024) yang menunjukkan bahwa sistem otomasi dokumen berbasis AI masih membutuhkan keterlibatan manusia, terutama ketika struktur dokumen bersifat kompleks dan padat. Namun, penelitian sebelumnya umumnya hanya menekankan pentingnya peran manusia tanpa memberikan bukti empiris mengenai besarnya degradasi akurasi pada metadata berstruktur

rapat. Penelitian ini mengisi celah tersebut dengan menunjukkan secara kuantitatif bahwa akurasi OCR dapat turun drastis hingga 1–3% ketika metadata dirender menjadi format visual yang rapat dan multi-baris. Dengan demikian, temuan ini tidak hanya memperkuat argumen konseptual pada penelitian terdahulu, tetapi juga memberikan dasar empiris yang menjelaskan secara operasional mengapa pendekatan human-in-the-loop tetap menjadi kebutuhan mendasar dalam workflow penerbitan ilmiah modern.

Analisis lebih lanjut menunjukkan bahwa kesalahan OCR yang muncul pada metadata padat tidak hanya disebabkan oleh kompleksitas layout, tetapi juga oleh keterbatasan teknis Tesseract. Terdapat tiga pola kesalahan yang paling dominan: substitution, yaitu karakter berubah menjadi karakter lain (terutama pada tanda baca dan huruf yang berdekatan); *deletion*, berupa hilangnya spasi atau pemisah kata akibat gagal mendeteksi batas kata; serta *insertion*, yaitu munculnya karakter tambahan pada bagian teks yang sangat rapat. Pola error ini menunjukkan bahwa PSM 6 kesulitan memetakan bounding box secara akurat ketika struktur metadata tersusun dalam beberapa baris yang berdempetan. Kondisi ini berbeda dari temuan pada dokumen ilmiah satu kolom, di mana noise visual merupakan penyebab utama. Dengan demikian, temuan penelitian ini menegaskan bahwa akar permasalahan terletak pada segmentasi tata letak metadata, bukan pada kualitas citra.

Pada konteks workflow editorial jurnal, pendekatan *Human-in-the-Loop* dapat diterapkan sebagai tahap pra-pemrosesan metadata. Hasil OCR berfungsi sebagai draft awal yang kemudian diverifikasi oleh editor sebelum data dimasukkan ke sistem manajemen naskah seperti OJS, ScholarOne, Editorial Manager, atau repository institusional. Dengan integrasi seperti ini, OCR mempercepat proses input metadata tanpa mengorbankan ketepatan bibliografis yang dibutuhkan dalam siklus penerbitan ilmiah.

Temuan penelitian ini menunjukkan bahwa OCR memiliki potensi besar untuk meningkatkan efisiensi pemrosesan metadata, namun penerapannya dalam workflow penerbitan ilmiah masih membutuhkan mekanisme *Human-in-the-Loop*. Dalam praktik editorial, terutama pada platform seperti OJS, Crossref deposit system, dan berbagai metadata manager milik penerbit besar, tahap verifikasi manusia tetap diperlukan pada bagian yang melibatkan pengecekan judul, penulis, dan informasi bibliografis sebelum metadata dikirim ke layanan pengindeksan. Karena itu, pendekatan hybrid menjadi solusi yang paling realistis: OCR mempercepat ekstraksi awal, sedangkan editor melakukan validasi akhir untuk menjaga akurasi dan konsistensi metadata.

Temuan ini juga memberikan arah bagi penelitian lanjutan. Pengembangan model OCR yang lebih adaptif terhadap struktur multi-baris seperti LayoutLMv3, Donut, atau model vision-language modern perlu dieksplorasi untuk mengatasi keterbatasan Tesseract pada metadata padat. Integrasi OCR dengan *Large Language Models* (LLM) dapat menjadi pendekatan koreksi otomatis (*OCR-post-correction*) yang menjanjikan. Selain itu, evaluasi pada dataset metadata dari platform jurnal lain misalnya OJS, Elsevier JATS XML, atau Springer Metadata Manager penting dilakukan untuk menilai generalisasi hasil pada ekosistem publikasi yang lebih luas. Temuan penelitian ini dengan demikian memberi dasar kuat bagi pengembangan sistem otomasi metadata yang tidak hanya cepat, tetapi juga mampu menjaga integritas data pada workflow editorial modern.

SIMPULAN

Penelitian ini menilai efisiensi dan akurasi otomasi entri metadata jurnal berbasis OCR melalui perbandingan langsung dengan jalur manual pada tiga skenario beban kerja. Hasil penelitian menunjukkan bahwa pemrosesan menggunakan OCR memberikan peningkatan kecepatan yang sangat signifikan. Pada skenario 1.000 entri, waktu pemrosesan manual mencapai 50.000 detik, sedangkan OCR hanya membutuhkan 0,075 detik, yang menegaskan bahwa metode otomatis mampu bekerja jauh lebih cepat dan stabil pada berbagai skala data.

Dari sisi akurasi, jalur manual konsisten pada 88%, sementara jalur otomatis menghasilkan akurasi 97% dalam kondisi terkontrol. Namun, akurasi karakter-level pada dokumen visual cukup rendah (1–3%), terutama karena struktur metadata jurnal yang padat dan memiliki tata letak multi-baris. Temuan ini menegaskan adanya *trade-off* antara kecepatan dan ketepatan ekstraksi, sehingga OCR lebih cocok ditempatkan sebagai tahap awal dalam workflow metadata untuk mengurangi beban entri awal. Implementasi OCR tetap memerlukan tahap verifikasi berbasis manusia untuk memastikan integritas data akhir, sesuai prinsip *Human-in-the-Loop*. Hasil penelitian ini juga mengindikasikan perlunya pengembangan lebih lanjut pada model OCR, seperti integrasi teknik *layout-aware* atau pendekatan berbasis deep learning, agar mampu menangani metadata dengan struktur kompleks tanpa mengorbankan kecepatan. Dengan demikian, penelitian ini memberikan dasar empiris untuk merancang sistem otomasi metadata yang lebih efisien, adaptif, dan dapat diterapkan pada workflow penerbitan ilmiah modern.

REFERENSI

Aydın Çolak, F., & Eroğlu, Ş. (2025). Evaluating metadata quality in institutional academic repositories of Turkish research universities. *Online Information Review*, *49*(7), 1335-1350. https://doi.org/10.1108/OIR-06-2024-0401

Dutta, H., & Gupta, A. (2022). PNRank: Unsupervised ranking of person name entities from noisy OCR text. *Decision Support Systems*, 152, 113662. https://doi.org/10.1016/j.dss.2021.113662

Ignasius, A., Chandra, J. C., Oscadinata, R., & Suhartono, D. (2023). Image pre-processing effect on OCR's performance for image conversion to Braille Unicode. *Procedia Computer Science*, 227, 922–931. https://doi.org/10.1016/j.procs.2023.10.599

Irimia, C., Harbuzariu, F., Hazi, I., & Iftene, A. (2022). Official document identification and data extraction using templates and OCR. *Procedia Computer Science*, 207, 1571–1580. https://doi.org/10.1016/j.procs.2022.09.214

Kayarvizhy, N., Choudhury, A. R., Rekha, G. S., Bhuvan, G., & Sanchi, C. (2025). On-device deep learning for retrieving system and user timestamps from noisy chat images. *Procedia Computer Science*, 258, 3760–3770. https://doi.org/10.1016/j.procs.2025.04.631

Kim, S., Lee, B., Maqsood, M., Moon, J., & Rho, S. (2025). Deep learning-based natural language processing model and optical character recognition for detection of online grooming on social networking services. *CMES - Computer Modeling in Engineering & Sciences*, 143(2), 2079–2108. https://doi.org/10.32604/cmes.2025.061653

Lee, A., Yu, H., & Min, G. (2024). An algorithm of line segmentation and reading order sorting based on adjacent character detection: A post-processing of OCR for digitization of Chinese historical texts. *Journal of Cultural Heritage*, 67, 1–12. https://doi.org/10.1016/j.culher.2024.02.001

Lee, H., Park, Y.-C., & Lee, J. (2025). OCR-assisted masked BERT for homoglyph restoration towards multiple phishing text downstream tasks. *Computers, Materials & Continua*, 85(3), 4977–4993. https://doi.org/10.32604/cmc.2025.068156

Li, Y., Wei, Q., Chen, X., Li, J., Tao, C., & Xu, H. (2024). Improving tabular data extraction in scanned laboratory reports using deep learning models. *Journal of Biomedical Informatics*, 159, 104735. https://doi.org/10.1016/j.jbi.2024.104735

Mombelli, S., Lyle, J. R., & Breitinger, F. (2024). FAIRness in digital forensics datasets' metadata – and how to improve it. *Forensic Science International: Digital Investigation*, 49, 301681. https://doi.org/10.1016/j.fsidi.2023.301681

Onim, M. S. H., Nyeem, H., Roy, K., Hasan, M., Ishmam, A., Akif, M. A. H., & Ovi, T. B. (2022). BLPnet: A new DNN model and Bengali OCR engine for automatic licence plate recognition. *Array*, 15, 100244. https://doi.org/10.1016/j.array.2022.100244

Paixão, T. M., Berriel, R. F., Boeres, M. C. S., Koerich, A. L., Boude, C., De Souza, A. F., & Oliveira-Santos, T. (2022). A human-in-the-loop recommendation-based framework for reconstruction of mechanically shredded documents. *Pattern Recognition Letters*, 164, 1–8. https://doi.org/10.1016/j.patrec.2022.10.011

Park, J., Seo, W., & Yun, T. S. (2025). End-to-end data extraction framework from unstructured geotechnical investigation reports via integrated deep learning and text mining techniques. *Developments in the Built Environment*, 23, 100733. https://doi.org/10.1016/j.dibe.2025.100733

Peña, A., Morales, A., Fierrez, J., Ortega-Garcia, J., Puente, I., Cordova, J., & Cordova, G. (2024). Continuous document layout analysis: Human-in-the-loop AI-based data curation, database, and evaluation in the domain of public affairs. *Information Fusion*, 108, 102398. https://doi.org/10.1016/j.inffus.2024.102398

Pino, R., Mendoza, R., & Sambayan, R. (2025). MaBaybay-OCR: A Matlab-based Baybayin optical character recognition package. *SoftwareX*, 29, 102003. https://doi.org/10.1016/j.softx.2024.102003

Şahin, A., Kara, B. C., & Dirsehan, T. (2025). LitOrganizer: Automating the process of data extraction and organization for scientific literature reviews. *SoftwareX*, 30, 102198. https://doi.org/10.1016/j.softx.2025.102198

Sinthuja, M., Padubidri, C. G., Jayachandra, G. S., Teja, M. C., & Kumar, G. S. P. (2024). Extraction of text from images using deep learning. *Procedia Computer Science*, 235, 789–798. https://doi.org/10.1016/j.procs.2024.04.075

Sugiyono, A. Y., Adrio, K., Tanuwijaya, K., & Suryaningrum, K. M. (2023). Extracting information from vehicle registration plate using OCR Tesseract. *Procedia Computer Science*, 227, 992–998. https://doi.org/10.1016/j.procs.2023.10.600

Wang, S., Moon, S., Fu, Y., & Kim, J. (2025). Construction regulatory document digitalization with layout knowledge-informed object detection and semantic text recognition. *Advanced Engineering Informatics*, 65(Part B), 103278. https://doi.org/10.1016/j.aei.2025.103278

Zhang, Y., Shi, Y., Zhao, P., Zhao, Y., Yang, Z., & Jin, L. (2025). MegaHan97K: A large-scale dataset for mega-category Chinese character recognition with over 97K categories. *Pattern Recognition*, 167, 111757. https://doi.org/10.1016/j.patcog.2025.111757

Zhao, L., Hao, R., Chai, Z., Fu, W., Yang, W., Li, C., Liu, Q., & Jiang, Y. (2024). DeepOCR: A multi-species deep-learning framework for accurate identification of open chromatin regions in livestock. *Computational Biology and Chemistry*, 110, 108077. https://doi.org/10.1016/j.compbiolchem.2024.108077