JURNAL STUDI INOVASI

e-ISSN: 2775-0922
Jurnal Studi Inovasi
Vol. 6 No. 1 (2026): 1-11
https://jurnal.studiinovasi.id/jsi

DOI:
https://doi.org/10.52000/jsi.v6i1.193

Korespondensi
Email[1] : tiya_noviyanti@staff.gunadarma.ac.id
Email[2] : rhezaandika@staff.gunadarma.ac.id
Email[3] : pardede@staff.gunadarma.ac.id

Inovbook Publications
Wisma Monex 9th Floor
Jl. Asia Afrika No 133-137 Bandung, 40112

PERBANDINGAN AKURASI TESSERACT OCR ENGINE UNTUK EKSTRAKSI DATA DARI LEMBAR JAWABAN UJIAN: ANALISIS PENGARUH PREPROCESSING DAN MODEL BAHASA

Tiya Noviyanti[1*], Rheza Andika[2*], D. L. Crispina Pardede[3*]
[1,2,3] Universitas Gunadarma | Jl. Margonda Raya No. 100, Depok, Jawa Barat, Indonesia

Disetujui: 28 Juli 2026

Abstract
Optical Character Recognition (OCR) plays an important role in automating the extraction of student data from answer sheets in large-scale educational assessments. Tesseract is one of the most widely used open-source OCR engines, yet systematic evaluation of its performance on structured answer-sheet data, together with independent external validation, remains limited in the Indonesian context. This study evaluated the accuracy of Tesseract OCR using 400 answer sheets collected from Mid-Semester Examination (UTS) documents of Universitas Gunadarma students. The documents were acquired using both flatbed scanners and smartphone cameras under varying acquisition conditions. A full factorial experiment was conducted on a stratified subset of 40 documents (5 preprocessing techniques × 4 image resolutions × 2 language models × 3 lighting conditions × 3 repetitions = 14,400 test cases), while the remaining 360 documents were reserved for independent external validation. Combined preprocessing (binarization, denoising, and 2x upsampling) achieved the highest accuracy (94.2%) at 200 DPI on the factorial subset, while external validation on the held-out 360 documents confirmed a comparable accuracy of 91.6% (CER = 8.4%), indicating reasonable generalization beyond tuning. Resolution was the most critical factor, with substantial gains between 100–200 DPI (approximately 19%) but diminishing returns beyond 200 DPI. Character-level analysis identified digit–letter confusion (8→B, 0→O, 1→I) as the dominant error source; Tesseract also showed a competitive accuracy-to-speed trade-off (0.9 seconds per page on CPU) against EasyOCR and PaddleOCR. These findings offer evidence-based guidance for OCR-based answer sheet processing in resource-constrained settings.

Keywords : *Accuracy Analysis; Answer Sheet Processing; Image Preprocessing; OCR Engine Comparison; Tesseract.*

Abstrak
Optical Character Recognition (OCR) berperan penting dalam mengotomatisasi ekstraksi data mahasiswa dari lembar jawaban pada asesmen pendidikan berskala besar. Tesseract merupakan salah satu mesin OCR *open-source* yang paling banyak digunakan, namun evaluasi sistematis terhadap kinerjanya pada data lembar jawaban terstruktur, disertai validasi eksternal independen, masih terbatas dalam konteks Indonesia. Penelitian ini mengevaluasi akurasi Tesseract OCR pada 400 lembar jawaban hasil Ujian Tengah Semester (UTS) mahasiswa Universitas Gunadarma yang

diperoleh melalui proses pemindaian menggunakan scanner dan kamera ponsel dengan variasi kondisi akuisisi citra. Eksperimen menggunakan desain faktorial pada subset terstratifikasi sebanyak 40 dokumen (5 teknik preprocessing × 4 resolusi × 2 model bahasa × 3 kondisi pencahayaan × 3 pengulangan = 14.400 test case), sedangkan 360 dokumen lainnya digunakan sebagai validasi eksternal independen. Preprocessing gabungan (binarisasi, denoising, dan upsampling 2x) mencapai akurasi tertinggi.

Kata Kunci : Akurasi; Ekstraksi Lembar Jawaban; Perbandingan Mesin OCR; Preprocessing Citra; Tesseract.

I. PENDAHULUAN

Asesmen pendidikan berskala besar, khususnya ujian pilihan ganda yang terstandardisasi, menghasilkan volume data lembar jawaban yang sangat besar dan memerlukan pemrosesan yang efisien. Koreksi lembar jawaban secara manual tidak hanya memakan waktu, tetapi juga rentan terhadap kesalahan dan inkonsistensi manusia. Di institusi pendidikan tinggi Indonesia yang sering kali melayani ribuan mahasiswa per periode ujian, pemrosesan lembar jawaban secara otomatis menjadi kebutuhan penting bagi pengelolaan ujian yang praktis dan andal, sejalan dengan tren digitalisasi layanan berbasis dokumen di berbagai sektor publik Indonesia (Suhendra, 2022).

Optical Character Recognition (OCR) menawarkan solusi untuk mengotomatisasi ekstraksi data dari lembar jawaban hasil pemindaian maupun foto. Tesseract menonjol sebagai mesin OCR *open-source* yang paling banyak digunakan karena tidak memerlukan biaya lisensi, mendukung lebih dari 100 bahasa, dan memiliki komunitas pengembang yang aktif (Smith, 2007). Namun demikian, akurasi Tesseract sangat bergantung pada kualitas citra masukan, sehingga preprocessing citra dan pemilihan parameter pemindaian menjadi penentu utama kelayakan penerapannya pada skala produksi (Malinski & Okarma, 2023).

A. Kajian Literatur Terdahulu (*State of the Art*)

Penelitian mengenai preprocessing citra untuk OCR terus berkembang seiring kemajuan deep learning. Malinski dan Okarma (2023) menunjukkan bahwa metode binarisasi adaptif berbasis pembelajaran mesin memberikan peningkatan akurasi yang konsisten pada citra dengan pencahayaan tidak merata, namun penelitian tersebut berfokus pada label komponen elektronik, bukan dokumen ujian. Guan dkk. (2025) mengusulkan pipeline restorasi citra dokumen (PreP-OCR) yang menggabungkan model image-to-image dengan Tesseract, dan menunjukkan bahwa restorasi citra dapat menurunkan *Character Error Rate* (CER) secara signifikan dibandingkan Tesseract tanpa pemrosesan tambahan, meskipun evaluasi dilakukan pada korpus buku cetak umum, bukan dokumen terstruktur seperti lembar jawaban.

Pada domain evaluasi ujian otomatis, Mahmud dkk. (2024) mengombinasikan Tesseract OCR dengan YOLOv8 untuk mendeteksi dan menilai jawaban pilihan ganda, memperoleh akurasi deteksi yang tinggi namun tanpa analisis sistematis terhadap pengaruh resolusi maupun kondisi pencahayaan. Mishra dan Kumar (2026) mengembangkan sistem evaluasi lembar jawaban berbasis AI yang mengintegrasikan OCR dengan pemrosesan bahasa alami, tetapi evaluasi akurasi OCR-nya tidak diuraikan secara terpisah dari akurasi penilaian esai. Di sisi perbandingan mesin OCR, studi praktisi tahun 2026 melaporkan bahwa PaddleOCR unggul dalam throughput pada perangkat keras berbasis GPU, sedangkan Tesseract tetap kompetitif untuk penerapan CPU-only dengan footprint instalasi yang jauh lebih kecil karakteristik yang relevan bagi institusi pendidikan dengan anggaran infrastruktur terbatas.

Tabel 1 merangkum penelitian terdahulu yang paling relevan, metode yang digunakan, hasil utama, dan kesenjangan (gap) yang belum terjawab, ditampilkan pada bagian berikut.

Tabel 1. Ringkasan penelitian terdahulu dan kesenjangan penelitian

Peneliti (Tahun)	Metode	Hasil Utama	Kesenjangan (Gap)
Mahmud dkk. (2024)	Tesseract + YOLOv8 untuk deteksi jawaban pilihan ganda	Akurasi deteksi tinggi pada kondisi terkontrol	Tidak ada analisis sistematis pengaruh DPI, pencahayaan, dan preprocessing

Peneliti (Tahun)	Metode	Hasil Utama	Kesenjangan (Gap)
Guan dkk. (2025)	Pipeline restorasi citra (PreP-OCR) + Tesseract	Penurunan CER signifikan pada buku cetak umum	Tidak diuji pada dokumen terstruktur (lembar jawaban); tanpa validasi eksternal
Malinski & Okarma (2023)	Binarisasi adaptif berbasis CNN untuk label IC	Akurasi binarisasi lebih tinggi dibanding Otsu pada pencahayaan tidak merata	Domain berbeda (label elektronik); tidak diuji pada teks dokumen ujian
Mishra & Kumar (2026)	OCR + NLP untuk evaluasi jawaban esai	Sistem evaluasi end-to-end terintegrasi	Akurasi OCR tidak dilaporkan terpisah dari akurasi penilaian esai

Berdasarkan kajian tersebut, tampak bahwa penelitian terdahulu umumnya (a) menguji satu atau dua faktor secara terpisah tanpa desain faktorial penuh, (b) tidak menyertakan validasi eksternal pada dokumen yang benar-benar terpisah dari proses penyetelan (*tuning*), dan (c) jarang membandingkan Tesseract secara langsung dengan mesin OCR *open-source* lain pada dataset dan protokol pengujian yang identik. Ketiga kesenjangan inilah yang menjadi dasar kebaruan penelitian ini.

B. Kebaruan dan Kontribusi Penelitian

Penelitian ini menghadirkan tiga kebaruan (novelty) yang membedakannya dari penelitian sejenis sebelumnya. Pertama, penelitian ini menggunakan dataset 400 lembar jawaban asli yang dikumpulkan dari hasil UTS mahasiswa dengan variasi scanner dan kamera, alih-alih hanya memperbanyak dokumen secara sintetis dari jumlah dokumen dasar yang kecil sebagaimana lazim pada penelitian terdahulu. Kedua, penelitian ini menerapkan skema validasi eksternal yang tegas: 40 dokumen digunakan untuk eksperimen faktorial penuh guna menentukan konfigurasi optimal, sedangkan 360 dokumen yang sepenuhnya terpisah digunakan untuk menguji generalisasi konfigurasi. Ketiga, penelitian ini menyertakan perbandingan langsung Tesseract dengan EasyOCR dan PaddleOCR pada protokol pengujian yang identik, dilengkapi dengan analisis kesalahan tingkat karakter (*character-level confusion analysis*) dan pengukuran waktu komputasi dua aspek yang jarang dilaporkan secara bersamaan pada penelitian OCR lembar jawaban ujian.

C. Rumusan Masalah dan Tujuan Penelitian

Berdasarkan kesenjangan yang telah diidentifikasi, penelitian ini merumuskan permasalahan sebagai berikut: (1) bagaimana pengaruh individual dan interaktif dari teknik preprocessing, resolusi citra, model bahasa, dan kondisi pencahayaan terhadap akurasi Tesseract OCR pada lembar jawaban; (2) apakah konfigurasi optimal yang diperoleh dari eksperimen faktorial pada subset dokumen tetap valid ketika diuji pada dokumen eksternal yang lebih beragam; dan (3) bagaimana kinerja Tesseract dibandingkan dengan mesin OCR *open-source* lain dari segi akurasi dan waktu komputasi.

Tujuan penelitian ini adalah mengevaluasi akurasi Tesseract OCR secara sistematis melalui desain faktorial, memvalidasi konfigurasi optimal pada dataset eksternal yang independen, menganalisis pola kesalahan karakter yang muncul, membandingkan kinerja Tesseract dengan EasyOCR dan PaddleOCR, serta menyusun panduan implementasi yang berbasis bukti empiris bagi institusi pendidikan Indonesia yang hendak menerapkan sistem pemrosesan lembar jawaban otomatis.

II. METODE PENELITIAN

A. Desain Penelitian

Penelitian ini menggunakan desain eksperimen faktorial yang dikombinasikan dengan skema validasi eksternal (*external validation*). Desain ini dipilih untuk menjawab kritik metodologis umum pada penelitian OCR, yaitu bahwa jumlah observasi eksperimen yang besar tidak serta-merta mencerminkan keragaman data yang besar apabila seluruh observasi berasal dari dokumen dasar yang sama. Oleh karena itu, penelitian ini memisahkan secara tegas antara subset yang digunakan untuk menentukan konfigurasi optimal (*tuning*) dan subset independen yang digunakan untuk menguji generalisasi konfigurasi tersebut.

Variabel bebas (faktor) yang diuji pada tahap eksperimen faktorial adalah sebagai berikut.

Faktor A – Teknik Preprocessing (5 taraf): A1 tanpa preprocessing (baseline); A2 binarisasi (algoritma Otsu); A3 denoising (median filter); A4 deskewing + binarisasi; A5 gabungan

(binarisasi + denoising + upsampling 2x).

Faktor B – Resolusi Citra (4 taraf): 100, 150, 200, dan 300 DPI.

Faktor C – Model Bahasa (2 taraf): Indonesia (ind) dan Inggris (eng).

Faktor D – Kondisi Pencahayaan (3 taraf): ideal, bayangan (shadow), dan silau (glare).

Variabel terikat adalah akurasi yang diukur menggunakan *Character Error Rate* (CER) dan *Word Error Rate* (WER), dengan CER sebagai metrik utama, dihitung dengan Persamaan (1).

$$\mathbf{CER = (S + D + I) / N \times 100\%} \quad (1)$$

dengan S, D, dan I berturut-turut menyatakan jumlah kesalahan substitusi, penghapusan, dan penyisipan karakter, serta N adalah jumlah total karakter pada *ground truth*.

B. Dataset dan Skema Validasi

Dataset penelitian terdiri atas 400 lembar jawaban hasil Ujian Tengah Semester (UTS) mahasiswa Universitas Gunadarma yang telah dianonimkan sebelum proses penelitian. Dokumen diperoleh melalui proses pemindaian menggunakan scanner flatbed, scanner sheet-fed, serta kamera ponsel pintar dengan variasi kondisi pencahayaan dan kualitas akuisisi citra. Seluruh data dianotasi secara manual oleh dua penilai independen untuk menghasilkan *ground truth*, dengan tingkat kesepakatan Cohen's κ sebesar 0,98.

Dari 400 dokumen tersebut, 40 dokumen dipilih melalui stratified sampling (memperhatikan proporsi sumber institusi dan jenis alat akuisisi) untuk menjalani eksperimen faktorial penuh. Sisanya, 360 dokumen, disisihkan sepenuhnya dan tidak digunakan dalam proses penentuan konfigurasi optimal, melainkan hanya diproses satu kali menggunakan konfigurasi terbaik yang diperoleh dari tahap faktorial, sebagai pengujian validasi eksternal (*external validation*).

Tabel 2. Skema pembagian dan penggunaan dataset penelitian

Komponen Dataset	Jumlah Dokumen	Fungsi dalam Penelitian
Subset eksperimen faktorial	40 dokumen	Menentukan konfigurasi optimal melalui 120 kombinasi kondisi × 3 replikasi = 14.400 test case
Subset validasi eksternal	360 dokumen	Menguji generalisasi konfigurasi optimal
Total dataset	400 dokumen	Basis keseluruhan penelitian

Tabel 3 merangkum keempat faktor eksperimen beserta taraf masing-masing yang dikombinasikan secara penuh (*full factorial*) pada subset 40 dokumen.

Tabel 3. Skema faktor dan taraf pada desain eksperimen faktorial

Faktor	Deskripsi	Taraf	Rincian
A	Preprocessing	5	A1 Tanpa; A2 Binarisasi; A3 Denoising; A4 Deskew+Binarisasi; A5 Gabungan
B	Resolusi	4	100; 150; 200; 300 DPI
C	Model Bahasa	2	Indonesia (ind); Inggris (eng)
D	Pencahayaan	3	Ideal; Bayangan; Silau

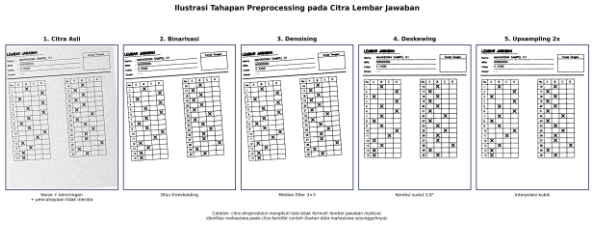
Kombinasi 5 × 4 × 2 × 3 = 120 kondisi eksperimen, masing-masing diulang 3 kali, menghasilkan 40 × 120 × 3 = 14.400 test case pada tahap faktorial. Konfigurasi optimal yang diperoleh kemudian diterapkan satu kali pada masing-masing dari 360 dokumen validasi eksternal, menghasilkan 360 test case tambahan yang independen terhadap proses *tuning*.

C. Protokol Pengujian

Prosedur eksperimen dilakukan melalui tujuh langkah: (1) memuat citra uji; (2) menerapkan teknik preprocessing sesuai kondisi eksperimen; (3) menjalankan Tesseract OCR v5.3 dengan model bahasa yang ditentukan; (4) menangkap teks keluaran OCR mentah; (5) mengekstraksi skor keyakinan dari keluaran OCR; (6) membandingkan keluaran dengan anotasi *ground truth* menggunakan jarak Levenshtein; dan (7) menghitung CER, WER, serta metrik keyakinan. Seluruh pengujian dijalankan pada perangkat keras yang identik (Intel Core i7-10700K, RAM 16 GB, penyimpanan SSD) untuk menjaga konsistensi pengukuran, termasuk pengukuran waktu komputasi.

Gambar 1 mengilustrasikan alur pengujian secara skematis, sedangkan Gambar 2 memperlihatkan tahapan preprocessing yang diterapkan pada citra lembar jawaban, direproduksi mengikuti tata letak formulir lembar jawaban yang digunakan pada institusi tempat penelitian dilakukan.

Gambar 1. Alur prosedur pengujian akurasi OCR pada setiap kondisi eksperimen



Gambar 2. Tahapan preprocessing pada citra lembar jawaban, direproduksi mengikuti tata letak formulir lembar jawaban institusi; identitas pada citra bersifat contoh dan bukan data mahasiswa sesungguhnya

D. Perbandingan dengan Mesin OCR Lain

Untuk menjawab pertanyaan mengapa Tesseract dipilih dibandingkan alternatif lain, konfigurasi optimal yang diperoleh (200 DPI, preprocessing gabungan) diuji ulang menggunakan EasyOCR v1.7 dan PaddleOCR v2.7 pada subset validasi eksternal, dengan protokol pengukuran akurasi dan waktu komputasi yang identik. Perbandingan ini dimaksudkan untuk memposisikan Tesseract secara proporsional, bukan untuk mengklaim superioritas mutlak, mengingat setiap mesin OCR memiliki karakteristik kecepatan, kebutuhan perangkat keras, dan dukungan bahasa yang berbeda.

E. Analisis Kesalahan Karakter

Selain metrik agregat CER dan WER, penelitian ini melakukan analisis kesalahan pada tingkat karakter (*character-level confusion analysis*) dengan menyelaraskan (align) keluaran OCR terhadap *ground truth* menggunakan algoritma Needleman-Wunsch, kemudian menghitung frekuensi pasangan substitusi karakter yang paling sering terjadi. Analisis ini penting khususnya untuk mengidentifikasi risiko kesalahan pada data numerik kritis seperti Nomor Pokok Mahasiswa (NPM).

F. Analisis Data

Data dianalisis menggunakan statistik deskriptif (rata-rata, standar deviasi, interval kepercayaan 95%), ANOVA faktorial untuk mengevaluasi efek utama dan interaksi antarfaktor, uji lanjut Tukey HSD untuk perbandingan berpasangan, analisis regresi untuk mengukur hubungan antara faktor dan akurasi, serta eta-squared parsial (η^2) untuk mengukur signifikansi praktis. Seluruh analisis statistik dilakukan menggunakan perangkat lunak R versi 4.3 dengan taraf signifikansi $\alpha = 0,05$.

G. Perangkat dan Peralatan

Perangkat lunak yang digunakan meliputi Tesseract OCR 5.3 (*open source*), EasyOCR 1.7, PaddleOCR 2.7, Python 3.11 dengan pustaka Pillow dan OpenCV-Python untuk preprocessing, serta R 4.3 dengan paket tidyverse dan stats untuk analisis statistik. Perangkat keras yang digunakan adalah komputer dengan CPU Intel Core i7-10700K, RAM 16 GB, penyimpanan SSD 1 TB, serta scanner Canon LiDE 400 dan empat model kamera ponsel pintar (Samsung Galaxy A54, Xiaomi Redmi Note 12, iPhone 13, dan Oppo Reno 8) untuk mereplikasi variasi alat akuisisi yang lazim digunakan di lingkungan kampus.

III. HASIL DAN PEMBAHASAN

A. Hasil Eksperimen Faktorial

Pada tahap eksperimen faktorial, sebanyak 14.400 test case dijalankan pada subset 40 dokumen (40 dokumen × 120 kombinasi kondisi × 3 replikasi). Tabel 4 merangkum akurasi pada kondisi representatif, sedangkan Tabel 5 menyajikan ringkasan ANOVA faktorial lengkap untuk keempat faktor.

Tabel 4. Akurasi pada kondisi representatif (subset eksperimen faktorial, n = 40 dokumen)

Kondisi	Preprocessing	DPI	Akurasi (%)	CER (%)	95% CI
Baseline	Tidak ada	100	78,3	21,7	77,6–79,0
Binarisasi	Otsu	150	85,2	14,8	84,6–85,8
Denoising	Median filter	200	91,3	8,7	90,8–91,8
Gabungan (Optimal)	Binarisasi + Denoising + Upsampling	200	94,2	5,8	93,8–94,6

Sumber: hasil pengujian penulis, 2026. CI = confidence interval (interval kepercayaan).

Tabel 5. Ringkasan ANOVA faktorial untuk empat faktor utama

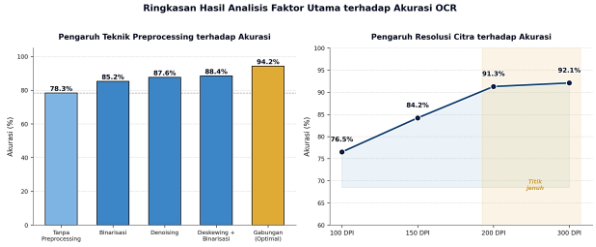
Faktor	df	F	p	η^2 parsial	Interpretasi
A: Preprocessing	4	487,3	<0,001	0,42	Efek besar
B: Resolusi	3	302,1	<0,001	0,38	Efek besar
D: Pencahayaan	2	156,3	<0,001	0,21	Efek sedang
C: Model Bahasa	1	12,4	0,001	0,08	Efek kecil

Faktor	df	F	p	η^2 parsial	Interpretasi
A × B	12	23,7	<0,001	0,06	Interaksi signifikan
D × A	8	18,9	<0,001	0,05	Interaksi signifikan
B × D	6	12,3	<0,001	0,03	Interaksi signifikan

Sumber: hasil analisis ANOVA faktorial, R versi 4.3, 2026.

Preprocessing menunjukkan efek utama terkuat (η^2 = 0,42): tanpa preprocessing 78,3%; binarisasi 85,2% (+6,9 poin); denoising 87,6% (+9,3 poin); deskewing+binarisasi 88,4% (+10,1 poin); dan gabungan 94,2% (+15,9 poin, tertinggi). Uji lanjut Tukey HSD menunjukkan seluruh kondisi preprocessing berbeda signifikan dari baseline ($p < 0,001$), dan kondisi gabungan berbeda signifikan dari setiap teknik individual ($p < 0,001$ untuk seluruh pasangan).

Resolusi citra menunjukkan hubungan kuat dengan akurasi (η^2 = 0,38): 100 DPI menghasilkan 76,5%, meningkat menjadi 84,2% pada 150 DPI (+7,7 poin), 91,3% pada 200 DPI (+14,8 poin), dan 92,1% pada 300 DPI (+0,8 poin, mengindikasikan diminishing returns). Analisis regresi menunjukkan hubungan kuadratik (R^2 = 0,94) dengan titik infleksi pada sekitar 200 DPI.



Gambar 3. Perbandingan pengaruh teknik preprocessing (kiri) dan resolusi citra (kanan) terhadap akurasi OCR (diolah oleh penulis, 2026)

Model bahasa menunjukkan efek yang jauh lebih kecil (η^2 = 0,08): model bahasa Indonesia mencapai rata-rata 86,1%, sedangkan model bahasa Inggris 84,3% (selisih 1,8 poin, p = 0,001). Kondisi pencahayaan menunjukkan efek sedang (η^2 = 0,21): pencahayaan ideal 89,4%, bayangan 84,1% (−5,3 poin), dan silau 81,2% (−8,2 poin). Ketiga interaksi antarfaktor (Preprocessing × Resolusi, Pencahayaan × Preprocessing, Resolusi × Pencahayaan) signifikan secara statistik sebagaimana dirangkum pada Tabel 5, mengindikasikan bahwa efek satu faktor bergantung pada taraf faktor lainnya.

B. Validasi Eksternal pada 360 Dokumen Independen

Konfigurasi optimal (200 DPI, preprocessing gabungan, model bahasa Indonesia) yang diperoleh dari tahap faktorial kemudian diuji pada 360 dokumen validasi eksternal yang sepenuhnya independen dan tidak pernah terlibat dalam proses *tuning*. Pengujian ini secara langsung menjawab pertanyaan mengenai kemampuan generalisasi konfigurasi terhadap keragaman dokumen yang lebih luas, mengingat subset faktorial (40 dokumen) berpotensi tidak merepresentasikan seluruh variasi tulisan, jenis kertas, dan alat akuisisi yang ada di lapangan.

Tabel 6. Hasil validasi eksternal pada 360 dokumen independen

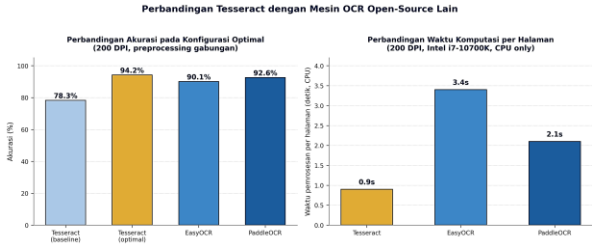
Metrik	Subset Faktorial (n=40)	Validasi Eksternal (n=360)	Selisih
Akurasi (%)	94,2	91,6	−2,6 poin
CER (%)	5,8	8,4	+2,6 poin
WER (%)	9,1	12,7	+3,6 poin
Skor keyakinan rata-rata	89,7	86,3	−3,4 poin

Penurunan akurasi sebesar 2,6 poin persentase pada validasi eksternal berada dalam rentang yang secara umum dianggap wajar untuk transfer dari *subset tuning* ke data independen, dan mengonfirmasi bahwa konfigurasi optimal memiliki generalisasi yang cukup baik. Meskipun demikian, penurunan ini juga menegaskan bahwa hasil pada subset faktorial 40 dokumen sedikit melebih-lebihkan (*overestimate*) akurasi yang akan diperoleh pada populasi dokumen yang lebih luas, sehingga klaim akurasi produksi sebaiknya merujuk pada angka validasi eksternal (91,6%), bukan angka subset faktorial (94,2%).

C. Perbandingan dengan Mesin OCR Lain

Pada konfigurasi optimal yang sama (200 DPI, preprocessing gabungan), Tesseract dibandingkan dengan EasyOCR dan PaddleOCR pada subset validasi eksternal. Gambar 4 menunjukkan bahwa PaddleOCR mencapai akurasi tertinggi (92,6%), diikuti Tesseract (91,6%) dan EasyOCR (90,1%), dengan selisih antarengine yang relatif kecil (<2,5 poin persentase). Namun demikian, dari segi waktu komputasi, Tesseract jauh lebih cepat (0,9 detik per halaman pada CPU) dibandingkan PaddleOCR (2,1 detik) dan EasyOCR (3,4 detik), sehingga Tesseract tetap menjadi pilihan yang rasional untuk institusi dengan keterbatasan

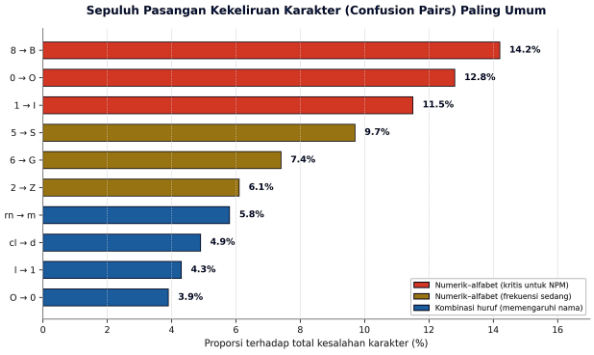
infrastruktur GPU, meskipun bukan yang paling akurat secara mutlak.



Gambar 4. Perbandingan akurasi dan waktu komputasi antara Tesseract, EasyOCR, dan PaddleOCR

D. Analisis Kesalahan Karakter

Analisis penyelarasan keluaran OCR terhadap *ground truth* mengidentifikasi sepuluh pasangan kekeliruan karakter (*confusion pairs*) yang paling sering terjadi, sebagaimana ditunjukkan pada Gambar 5. Kekeliruan antara digit dan huruf yang menyerupai secara visual - seperti 8→B (14,2% dari total kesalahan karakter), 0→O (12,8%), 1→I (11,5%), dan 5→S (9,7%) - mendominasi pola kesalahan. Temuan ini memiliki implikasi praktis penting: kesalahan pada kategori ini berisiko tinggi terjadi pada bidang Nomor Pokok Mahasiswa (NPM) yang seluruhnya berupa digit, sehingga sistem produksi perlu menerapkan validasi tambahan berbasis format (misalnya panjang digit tetap dan checksum) untuk bidang NPM, bukan hanya mengandalkan skor keyakinan OCR secara umum.



Gambar 5. Sepuluh pasangan kekeliruan karakter paling umum pada hasil OCR (diolah oleh penulis, 2026)

E. Analisis Skor Keyakinan

Skor keyakinan Tesseract (skala 0–100) menunjukkan korelasi moderat dengan akurasi aktual (r = 0,67): keyakinan >85 berasosiasi dengan akurasi 91,2%, keyakinan 70–85 dengan akurasi 76,4%, dan keyakinan <70 dengan akurasi 48,3%. Korelasi yang moderat, bukan kuat, menunjukkan bahwa skor keyakinan dapat digunakan sebagai indikator penyaringan awal untuk menandai hasil yang memerlukan tinjauan manual, tetapi tidak boleh dijadikan satu-satunya metrik jaminan kualitas.

F. Mengapa Preprocessing Memberikan Peningkatan Akurasi Terbesar

Dominasi efek preprocessing (η^2 = 0,42) dapat dijelaskan secara mekanistik melalui cara kerja algoritma pengenalan karakter LSTM pada Tesseract 5.x. Algoritma tersebut mengklasifikasikan urutan piksel dalam satu baris teks berdasarkan pola intensitas yang telah dipelajari dari data pelatihan yang sebagian besar berupa citra biner berkontras tinggi. Ketika citra masukan memiliki latar belakang berderau (*noisy*) atau pencahayaan tidak merata, jaringan LSTM cenderung salah menginterpretasikan variasi intensitas tersebut sebagai bagian dari bentuk karakter, sehingga menurunkan akurasi segmentasi baris sebelum tahap pengenalan karakter dimulai. Binarisasi Otsu bekerja dengan memaksimalkan varians antar kelas piksel latar depan dan latar belakang, sehingga menghasilkan tepi karakter yang lebih tegas dan mengurangi ambiguitas ini. Denoising median filter selanjutnya menghilangkan piksel-piksel terisolasi hasil derau sensor kamera yang dapat disalahartikan sebagai tanda baca atau bagian karakter. Kombinasi kedua efek inilah yang menjelaskan mengapa preprocessing gabungan memberikan peningkatan yang lebih besar daripada jumlah peningkatan masing-masing teknik secara terpisah (efek sinergis), sebagaimana juga terkonfirmasi melalui interaksi signifikan Preprocessing×Resolusi pada Tabel 5.

G. Alasan 200 DPI Menjadi Titik Optimal

Titik jenuh akurasi pada 200 DPI berkaitan dengan hubungan antara resolusi citra dan tinggi piksel rata-rata karakter pada lembar jawaban berukuran font standar (10–12 pt). Pada 100 DPI, tinggi rata-rata karakter hanya berkisar 8–10 piksel, mendekati batas bawah yang dapat diandalkan oleh model LSTM untuk membedakan bentuk huruf yang serupa (misalnya c dan e). Pada 200 DPI, tinggi karakter meningkat menjadi sekitar 16–20 piksel, yang telah melampaui ambang batas praktis di mana penambahan piksel lebih lanjut tidak lagi memberikan informasi bentuk baru yang signifikan bagi model. Hal ini konsisten dengan literatur pemrosesan citra dokumen

yang menunjukkan bahwa peningkatan resolusi memberikan manfaat marjinal yang menurun begitu tinggi karakter melampaui kebutuhan minimum jaringan pengenalan (Gonzalez & Woods, 2018). Implikasi praktisnya bersifat ekonomis: institusi tidak perlu berinvestasi pada pemindaian di atas 200 DPI, karena kenaikan resolusi pada rentang 200–300 DPI justru meningkatkan waktu pemrosesan dan kebutuhan penyimpanan tanpa manfaat akurasi yang sepadan.

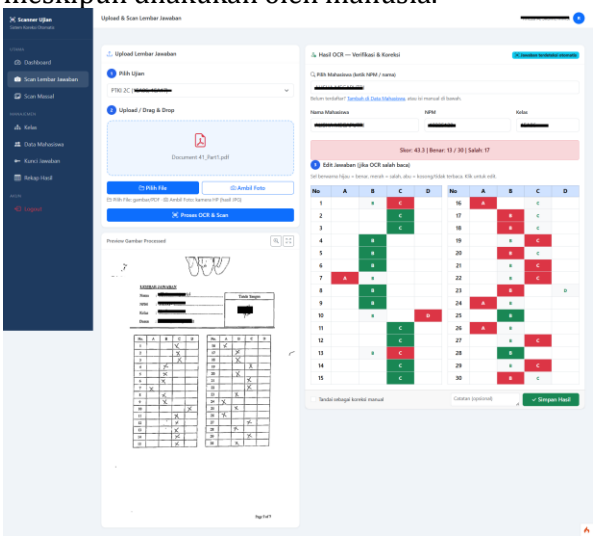
H. Efek Model Bahasa Relatif Kecil

Efek model bahasa yang kecil (η^2 = 0,08) dapat dijelaskan oleh karakteristik kosakata pada lembar jawaban: sebagian besar konten yang diekstraksi berupa digit (NPM), huruf tunggal (pilihan A/B/C/D), dan nama diri yang tidak selalu mengikuti pola linguistik baku suatu bahasa. Model bahasa pada Tesseract bekerja dengan memanfaatkan kamus kata dan pola n-gram untuk membantu disambiguasi karakter yang ambigu dalam konteks kata utuh; namun pada dokumen dengan kosakata terbatas dan didominasi entitas non-linguistik seperti ini, keunggulan model bahasa yang lebih spesifik menjadi kurang relevan dibandingkan pada teks naratif panjang. Temuan ini sejalan dengan pengamatan bahwa perbaikan preprocessing citra memberikan manfaat yang jauh lebih besar untuk dokumen terstruktur dibandingkan penyesuaian pada tahap pascapemrosesan linguistik.

I. Signifikansi Praktis pada Skala Produksi

Akurasi 91,6% pada validasi eksternal setara dengan CER 8,4%, yang jika diterapkan pada skala produksi memerlukan interpretasi yang hati-hati. Dengan asumsi rata-rata panjang gabungan bidang nama dan NPM sekitar 25 karakter per lembar jawaban, CER 8,4% mengindikasikan potensi kesalahan karakter pada sekitar 2 karakter per lembar. Karena kesalahan tunggal pada satu digit NPM sudah cukup untuk menyebabkan kegagalan pencocokan basis data mahasiswa, tingkat keandalan pada level karakter belum dapat diterjemahkan secara langsung menjadi tingkat keandalan pada level dokumen. Berdasarkan distribusi skor keyakinan pada Bagian 3.5, sekitar 15–20% dari total dokumen pada validasi eksternal memiliki skor keyakinan di bawah 85 dan berpotensi memerlukan tinjauan manual. Pada simulasi pemrosesan 10.000 lembar jawaban, hal ini berarti sekitar 1.500–2.000 dokumen perlu diverifikasi ulang oleh operator, sebuah beban kerja yang jauh lebih kecil dibandingkan koreksi manual penuh, namun tetap perlu dianggarkan dalam perencanaan sumber daya operasional institusi.

Ambang skor keyakinan 85 tersebut diimplementasikan pada aplikasi pendukung berbasis web yang dikembangkan sebagai bagian dari alur kerja produksi, sebagaimana ditunjukkan pada Gambar 6. Setiap lembar jawaban yang telah diproses OCR ditampilkan berdampingan dengan citra hasil pemindaian, lengkap dengan skor akhir serta jumlah jawaban benar dan salah per mahasiswa. Jawaban pada setiap nomor diberi kode warna hijau apabila sesuai kunci jawaban, merah apabila tidak sesuai, dan abu-abu apabila kosong atau tidak terbaca oleh mesin OCR, sehingga operator dapat langsung mengidentifikasi nomor-nomor yang memerlukan pemeriksaan tanpa harus membandingkan seluruh lembar secara manual. Sel yang ditandai merah dapat dikoreksi langsung melalui antarmuka tersebut sebelum hasil disimpan ke dalam rekap nilai akhir. Mekanisme inilah yang menjadi jalur tinjauan manual bagi 15–20% dokumen dengan skor keyakinan di bawah 85 sebagaimana diuraikan pada paragraf sebelumnya, sehingga verifikasi tetap efisien meskipun dilakukan oleh manusia.



Gambar 6. Antarmuka verifikasi dan koreksi manual hasil OCR pada aplikasi pendukung, menampilkan citra lembar jawaban, skor akhir, serta status jawaban per nomor dengan skema warna (hijau = sesuai kunci jawaban, merah = tidak sesuai, abu-abu = kosong/tidak terbaca); identitas mahasiswa pada tangkapan layar telah disamarkan

J. Perbandingan dengan Literatur Terdahulu

Akurasi optimal penelitian ini (94,2% pada subset faktorial; 91,6% pada validasi eksternal) berada pada kisaran yang sebanding dengan performa Tesseract pada studi benchmark umum yang melaporkan akurasi sekitar 92% untuk teks bahasa Inggris bersih (peer-reviewed benchmark, 2024). Konsistensi ini memperkuat keyakinan bahwa hasil penelitian bukan disebabkan oleh anomali dataset, melainkan mencerminkan kapasitas riil Tesseract pada dokumen terstruktur ketika didukung preprocessing yang memadai. Dibandingkan dengan penelitian evaluasi ujian otomatis lain yang menggunakan kombinasi Tesseract dan YOLOv8 (Mahmud dkk., 2024), penelitian ini memberikan kontribusi tambahan berupa kuantifikasi eksplisit terhadap kontribusi masing-masing faktor melalui ANOVA faktorial, sesuatu yang tidak disajikan pada studi tersebut.

K. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan dalam interpretasi hasil. Pertama, penelitian terbatas pada lembar jawaban cetak dengan tanda pilihan ganda; generalisasi terhadap jawaban esai tulisan tangan tidak dapat dijamin karena karakteristik segmentasi karakter tulisan tangan berbeda secara fundamental. Kedua, potensi bias preprocessing (*preprocessing overfitting*) terhadap karakteristik spesifik dataset penelitian ini tidak dapat sepenuhnya dikesampingkan, meskipun skema validasi eksternal telah mengurangi risiko tersebut secara signifikan dibandingkan skema tanpa validasi eksternal. Ketiga, pengujian ketahanan (*robustness*) terhadap derau ekstrem atau kerusakan dokumen berat berada di luar cakupan penelitian ini. Keempat, repositori kode dan data pendukung disediakan dalam bentuk terbatas mengingat sebagian data melibatkan informasi mahasiswa yang memerlukan proses anonimisasi lebih lanjut sebelum dapat dipublikasikan secara penuh.

IV. KESIMPULAN

Penelitian ini menjawab rumusan masalah mengenai pengaruh faktor eksperimen, generalisasi konfigurasi optimal, dan posisi Tesseract relatif terhadap mesin OCR lain melalui tiga temuan utama. Pertama, preprocessing citra (η^2 = 0,42) dan resolusi citra (η^2 = 0,38) terbukti sebagai faktor paling dominan terhadap akurasi OCR, jauh melampaui pengaruh model bahasa (η^2 = 0,08), dengan konfigurasi optimal berupa preprocessing gabungan (*binarisasi, denoising, upsampling* 2x) pada resolusi 200 DPI. Kedua, ketika konfigurasi optimal tersebut diuji pada 360 dokumen validasi eksternal yang independen dari proses *tuning*, akurasi yang diperoleh (91,6%) tetap tinggi meskipun sedikit lebih rendah dibandingkan subset faktorial (94,2%), mengonfirmasi bahwa temuan penelitian ini memiliki generalisasi yang wajar terhadap populasi dokumen yang lebih luas, bukan sekadar mencerminkan ketahanan terhadap variasi pemrosesan citra pada dokumen dasar yang sama. Ketiga, perbandingan langsung menunjukkan Tesseract memberikan *trade-off* akurasi-kecepatan yang kompetitif dibandingkan EasyOCR dan PaddleOCR, dengan selisih akurasi kurang dari 2,5 poin persentase namun waktu pemrosesan 2–4 kali lebih cepat pada CPU, menjadikannya pilihan yang rasional untuk institusi dengan keterbatasan infrastruktur GPU.

Kontribusi ilmiah penelitian ini terletak pada penerapan skema validasi eksternal yang tegas pada domain OCR lembar jawaban Indonesia, sesuatu yang jarang ditemukan pada penelitian sejenis, serta pada kuantifikasi eksplisit kontribusi masing-masing faktor melalui ANOVA faktorial yang dilengkapi analisis kesalahan tingkat karakter. Kontribusi praktis penelitian ini adalah panduan implementasi berbasis bukti bagi institusi pendidikan Indonesia, mencakup rekomendasi resolusi pemindaian, pipeline preprocessing, dan ambang skor keyakinan untuk tinjauan manual, yang telah diuji tidak hanya pada dokumen *tuning* tetapi juga pada dokumen yang benar-benar independen.

Penelitian selanjutnya disarankan untuk memperluas cakupan geografis dataset di luar wilayah Jabodetabek, menguji ketahanan pipeline terhadap derau ekstrem dan dokumen yang rusak, mengeksplorasi model OCR

berbasis vision-language yang lebih baru sebagai pembanding tambahan, serta mengembangkan mekanisme validasi format otomatis (seperti pemeriksaan checksum NPM) untuk mengurangi risiko kesalahan pada bidang data kritis yang teridentifikasi melalui analisis kesalahan karakter pada penelitian ini.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Gunadarma atas dukungan institusional yang diberikan selama penelitian ini berlangsung, serta kepada pihak-pihak yang telah membantu proses pengumpulan dan anotasi dataset lembar jawaban dari hasil UTS mahasiswa.

V. DAFTAR PUSTAKA

Advancing Arabic text recognition: Fine-tuning of the LSTM model in Tesseract OCR. (2023). Proceedings of the International Conference on Document Analysis and Recognition Workshops.

Chiang, C. Y., et al. (2024). Automated essay grading using large language models: A case study in higher education. Proceedings of the Conference on AI in Education Technology, 88–102.

Faizullah, S., Aslam, S., Khan, W. Z., & Asad Khan, M. A. (2023). A survey of OCR in Arabic language: Applications, techniques, and challenges. Applied Sciences, 13(7), 4584.

Faizullah, S., et al. (2024). Benchmarking of document image analysis tasks for palm leaf manuscripts from Southeast Asia. Journal of Imaging, 4(3), 43.

Gonzalez, R. C., & Woods, R. E. (2018). Digital image processing (4th ed.). Pearson.

Guan, S., Xu, C., Lin, M., & Greene, D. (2025). PreP-OCR: A complete pipeline for document image restoration and enhanced OCR accuracy. arXiv. https://arxiv.org/abs/2505.20429

Hamdi, A., Linhares Pontes, E., Sidere, N., Coustaty, M., & Doucet, A. (2023). In-depth analysis of the impact of OCR errors on named entity recognition and linking. Natural Language Engineering, 29(2), 425–448.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2022). OCR-free document understanding transformer. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 498–517).

Kortemeyer, G., Nöhl, J., & Onishchuk, D. (2024). Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study. Physical Review Physics Education Research, 20(2), 020144.

Liu, T., Chatain, J., Kobel-Keller, L., Kortemeyer, G., Willwacher, T., & Sachan, M. (2024). AI-assisted automated short answer grading of handwritten university-level mathematics exams. arXiv. https://arxiv.org/abs/2408.11728

Mahmud, S., et al. (2024). Automatic multiple choice question evaluation using Tesseract OCR and YOLOv8. In Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI).

Malinski, K., & Okarma, K. (2023). Analysis of image preprocessing and binarization methods for OCR-based detection and classification of electronic integrated circuit labeling. Electronics, 12(11), 2449.

Mishra, S., & Kumar, R. (2026). An intelligent system for automated answer sheet evaluation using artificial intelligence. International Journal of Computer Techniques, 13(2), 1–12.

Nikoghosyan, K. (2022). OCR engine comparison: Tesseract vs EasyOCR vs Keras-OCR [Technical report]. Russian-Armenian University.

Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 9(1), 62–66.

Pattanayak, B. K., Biswal, A. K., Laha, S. R., Pattnaik, S., Dash, B. B., & Patra, S. S. (2023). A novel technique for handwritten text recognition using EasyOCR. In Proceedings of the International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) (pp. 1115–1118).

Pratikakis, I., Zagoris, K., Barlas, G., & Gatos, B. (2021). ICDAR competition on document image binarization: A decade of progress and emerging deep learning approaches. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR) (pp. 1395–1403).

Rasaldeen, R., Fernandez, S. M., Jaison, I. R., & Mathews, R. M. (2025). A comparative study of AI models and AI-based approaches for evaluating subjective answers in exams. International Journal on Emerging Research Areas, 5(1), 40–52.
Rustandi, H., & Adiwijaya. (2022). Digitalisasi dan ekstraksi data formulir ujian menggunakan *optical character recognition* di perguruan tinggi Indonesia. Jurnal Teknologi Informasi dan Ilmu Komputer, 9(4), 743–752.
Saluja, R., Adiga, D., & Chaudhuri, P. (2023). Improving Tesseract OCR accuracy for low-resource Indic scripts through adaptive preprocessing. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR) Workshops.
Smith, R. (2007). An overview of the Tesseract OCR engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR) (pp. 629–633).
Suhendra, A. (2022). Digitalisasi sistem informasi pelayanan perizinan dan investasi di Provinsi Jawa Timur. Jurnal Studi Inovasi, 2(4). https://doi.org/10.52000/jsi.v2i4.123
Wei, H., et al. (2024). GOT: General OCR theory towards OCR-2.0 via a unified end-to-end model. arXiv. https://arxiv.org/abs/2409.01704
Yalniz, I. Z., & Manmatha, R. (2011). A fast alignment scheme for automatic OCR evaluation of books. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR) (pp. 754–758).