Evaluasi kinerja validasi KTM berbasis *Tesseract* OCR menggunakan metode *Adaptive Thresholding* dan *Levenshtein Distance*

Muhammad Faizul Ulum*, Retno Mumpuni, Afina Lina Nurlaili

Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Email: 22081010132@student.upnjatim.ac.id, retnomumpuni.if@upnjatim.ac.id, afina.lina.if@upnjatim.ac.id

Abstrak
Validasi Kartu Tanda Mahasiswa (KTM) yang dilakukan secara manual pada sistem penyewaan inventaris UKKI UPN "Veteran" Jawa Timur rentan terhadap kesalahan manusia dan risiko keamanan aset. Penelitian ini bertujuan untuk mengevaluasi kinerja sistem validasi otomatis berbasis Tesseract OCR yang mengintegrasikan metode Adaptive Gaussian Thresholding untuk pra-pemrosesan citra dan Levenshtein Distance untuk verifikasi hasil ekstrkasi teks. Penelitian menerapkan pendekatan validasi yang mengombinasikan pencocokan string dan aturan logika berbasis struktur pada Nomor Pokok Mahasiswa (NPM). Hasil pengujian menunjukkan bahwa tahapan pra-pemrosesan citra efektif meningkatkan keterbacaan teks pada kondisi pencahayaan tidak merata dan orientasi acak. Mekanisme validasi bertingkat ini terbukti efektif dalam menangani kesalahan baca (noise) pada OCR, termasuk kemampuan pemulihan data (data recovery) yang terdistorsi menggunakan logika fallback. Evaluasi kinerja sistem menggunakan Confusion Matrix terhadap 20 sampel citra menghasilkan nilai Akurasi sebesar 95%, Presisi 100%, dan Recall 93,3%. Tingkat akurasi mengindikasikan bahwa sistem mampu mencegah penerimaan dokumen tidak sah secara efektif, menjadikan solusi ini efektif untuk meningkatkan keamanan dan efisiensi verifikasi identitas pada penyewaan inventaris.

Kata Kunci: Adaptive Thresholding; Levenshtein Distance; OCR; Tesseract; Validasi Identitas.

Performance evaluation of Tesseract OCR-based KTM validation using Adaptive Thresholding and Levenshtein Distance methods

Abstract
Manual validation of Student Identity Cards (KTM) in the UKKI UPN "Veteran" Jawa Timur inventory rental system is prone to human error and asset security risks. This study aims to evaluate the performance of a Tesseract OCR-based automated validation system that integrates the Adaptive Gaussian Thresholding method for image pre-processing and Levenshtein Distance for text extraction verification. The study applies a validation approach that combines string matching and structure-based logic rules on Student Identification Numbers (NPM). The test results show that the image pre-processing stage effectively improves text readability under uneven lighting conditions and random orientations. This multi-level validation mechanism is proven effective in handling reading errors (noise) in OCR, including the ability to recover distorted data using fallback logic. System performance evaluation using Confusion Matrix on 20 image samples resulted in an Accuracy value of 95%, Precision 100%, and Recall 93.3%. The accuracy level indicates that the system is able to effectively prevent the acceptance of unauthorized documents, making this solution effective for improving the security and efficiency of identity verification in inventory rentals.

Keywords: *Adaptive Thresholding; Levenshtein Distance; OCR; Tesseract; Identity Validation.*

1. Pendahuluan
Pengelolaan aset merupakan elemen fundamental dalam menjaga keberlangsungan operasional dan stabilitas finansial sebuah institusi. Namun, manajemen inventaris yang masih mengandalkan metode konvensional seringkali menjadi penghambat utama produktivitas. Ketergantungan pada pencatatan manual atau perangkat lunak spreadsheet rentan memicu kurangnya konsistensi data akibat kesalahan input manusia (human error) serta menghambat distribusi informasi antar divisi (Wibisono et al., 2016). Hal ini diperkuat oleh studi dari Erameh dan Odoh (2021), yang menemukan bahwa pencatatan data inventaris secara manual ke dalam spreadsheet memiliki banyak kekurangan teknis dan mempersulit pemantauan stok secara efektif, sehingga tidak lagi relevan untuk diterapkan di era modern yang menuntut akurasi tinggi.

Selain aspek efisiensi, metode manual juga membuka celah fatal terhadap risiko keamanan dan kehilangan aset. Institusi pendidikan yang belum mengadopsi sistem berbasis teknologi menghadapi risiko keamanan signifikan karena minimnya validasi data yang terintegrasi (Nasution et al., 2023). Risiko ini dikonfirmasi oleh penelitian Kusumawardhani dkk. (2025) pada laboratorium universitas, yang mengungkapkan bahwa pengelolaan aset secara manual mengakibatkan kendala serius berupa pencatatan yang rawan kesalahan dan tingginya potensi kehilangan alat akibat tidak adanya pelacakan riwayat penggunaan yang akurat. Tanpa sistem pengawasan yang ketat, aset bernilai tinggi menjadi rentan terhadap penyalahgunaan atau pencurian yang sulit dilacak.

Permasalahan keamanan aset ini juga teridentifikasi pada Unit Kegiatan Kerohanian Islam (UKKI) UPN "Veteran" Jawa Timur. Saat ini, validasi identitas peminjam inventaris masih dilakukan dengan memeriksa fisik Kartu Tanda Mahasiswa (KTM) secara manual. Metode ini memiliki kelemahan fatal, yakni tidak adanya mekanisme verifikasi otomatis untuk memastikan keabsahan data mahasiswa atau status keaktifannya. Celah ini meningkatkan risiko manipulasi identitas peminjam yang dapat berujung pada hilangnya inventaris organisasi tanpa jejak administrasi yang jelas. Oleh karena itu, diperlukan sistem validasi identitas digital yang mampu mengekstraksi dan memverifikasi data KTM secara otomatis.

Sebagai solusi atas permasalahan verifikasi manual tersebut, penerapan teknologi Optical Character Recognition (OCR) menjadi pendekatan yang paling relevan. Hamad dan Kaya (2016) mendefinisikan OCR sebagai teknologi yang memungkinkan konversi otomatis dari dokumen berbasis citra menjadi format teks digital yang dapat disunting dan diolah lebih lanjut oleh komputer. Dalam implementasinya, salah satu mesin OCR open-source yang paling banyak diadopsi karena efisiensi dan akurasinya adalah Tesseract. Saoji dkk. (2021) menjelaskan bahwa Tesseract mampu mengenali karakter ASCII dari dokumen elektronik maupun tulisan tangan dengan presisi tinggi melalui mekanisme segmentasi dan pencocokan pola.

Pemanfaatan teknologi untuk ekstraksi data identitas menjadi solusi krusial dalam mengatasi kendala verifikasi konvensional. Proses verifikasi identitas yang mengandalkan input data manual tidak hanya memakan waktu, tetapi juga sangat rentan terhadap kesalahan manusia (human error), sehingga diperlukan pendekatan Document Understanding untuk mengotomatisasi ekstraksi data secara presisi (Carta et al., 2024). Dalam konteks dokumen spesifik Indonesia, implementasi Optical Character Recognition (OCR) terbukti efektif sebagai solusi ekstraksi. Hal ini didukung oleh penelitian Rusli dkk. (2021), yang mendemonstrasikan bahwa penggunaan pustaka Pytesseract (Tesseract OCR) mampu mendigitalkan informasi tekstual dari citra Kartu Tanda Penduduk (KTP) secara otomatis untuk kebutuhan pertukaran informasi digital.

Namun, kinerja Tesseract sangat bergantung pada kualitas citra input. Meskipun Tesseract memiliki performa tinggi, akurasinya dapat menurun drastis pada skenario pengambilan gambar di lingkungan alami yang tidak terkontrol (Zacharias et al., 2020). Tantangan ini dipertegas oleh Akinbade dkk. (2020), yang menyoroti bahwa kompleksitas latar belakang pada citra seringkali menghambat proses ekstraksi informasi, menyebabkan mesin gagal memisahkan karakter dari noise di sekitarnya. Kegagalan segmentasi semacam ini menuntut adanya tahapan pra-pemrosesan yang handal. Oleh karena itu, penerapan teknik perbaikan citra (image enhancement) berbasis algoritma Adaptive Thresholding menjadi syarat mutlak untuk mengisolasi teks dari gangguan visual dan memastikan data yang diperoleh benar-benar valid.

Kendala kualitas citra yang memicu kesalahan baca mesin memerlukan penanganan khusus melalui validasi pasca-pemrosesan (post-processing). Studi yang dilakukan oleh Hládek dkk. (2017) menunjukkan bahwa pendekatan berbasis metrik jarak string (string distance) sangat efektif untuk mengoreksi pola misspelling yang kerap dihasilkan oleh OCR pada dokumen teks. Relevansi algoritma ini dalam sistem pengenalan karakter juga divalidasi oleh Qhitfir dkk. (2025), yang menerapkan metode perhitungan jarak Levenshtein sebagai instrumen vital untuk menganalisis penyimpangan karakter dan memverifikasi akurasi hasil ekstraksi teks pada objek visual. Berkaca pada keberhasilan implementasi tersebut, penelitian ini mengusulkan strategi validasi hibrida: memperbaiki kualitas input citra melalui Adaptive Gaussian Thresholding, lalu memvalidasi luaran teksnya menggunakan Levenshtein Distance guna memastikan keaslian data pada Kartu Tanda Mahasiswa (KTM).

2. Metode

Penelitian ini menerapkan pendekatan eksperimental untuk mengevaluasi kinerja sistem validasi identitas otomatis. Tahapan penelitian dirancang secara sistematis melalui tiga fase utama: pra-pemrosesan citra untuk perbaikan kualitas input, ekstraksi teks menggunakan Tesseract OCR, dan validasi data menggunakan algoritma pencocokan string (string matching) terhadap basis data internal mahasiswa aktif.

2.1. Dataset dan Pra-pemrosesan Citra

Dataset yang digunakan pada penelitian ini berupa citra Kartu Tanda Mahasiswa (KTM) UPN "Veteran" Jawa Timur sebagai data positif, serta beberapa jenis kartu identitas lain sebagai data uji negatif (outliers). Seluruh citra diambil menggunakan kamera ponsel dengan kondisi lingkungan yang tidak terkontrol, sehingga memiliki variasi pencahayaan, orientasi, dan latar belakang.

Untuk mengatasi permasalahan tersebut, tahapan pra-pemrosesan citra menjadi aspek krusial. Koreksi kemiringan dan konversi ruang warna merupakan prasyarat fundamental dalam meningkatkan akurasi sistem OCR offline (Dey et al., 2022). Mengacu pada pendekatan tersebut, sistem menerapkan deteksi kontur dan transformasi perspektif (smart cropping) untuk menstandarisasi orientasi kartu, diikuti dengan konversi citra ke format grayscale.

Selain itu, dilakukan operasi inversi warna (bitwise NOT) pada area header kartu guna mengubah teks terang berlatar gelap menjadi format standar teks gelap berlatar terang agar lebih kompatibel dengan mesin Tesseract. Untuk mengatasi degradasi kualitas akibat pencahayaan tidak merata, diterapkan metode Adaptive Gaussian Thresholding. Dalam penelitian yang dilakukan oleh Sofwan dkk. (2021) menunjukkan bahwa optimasi teknik thresholding sangat efektif untuk mereduksi gangguan visual (noise) seperti efek kabur (blur) dan variasi intensitas cahaya. Pendekatan ini diperkuat oleh studi Ingle dan Kaur (2017), yang membuktikan bahwa binarisasi adaptif yang memanfaatkan kontras lokal jauh lebih tangguh (robust) dalam memisahkan goresan teks dari latar belakang yang kompleks dibandingkan metode global thresholding biasa. Tahapan ini menghasilkan citra biner dengan kontras tinggi yang optimal untuk proses ekstraksi teks.

2.2. Ekstraksi dan Validasi Hibrida

Proses ekstraksi teks dilakukan menggunakan mesin Tesseract OCR dengan dukungan bahasa Indonesia dan Inggris. Hasil ekstraksi teks mentah kemudian divalidasi melalui dua lapisan validasi yang bersifat hibrida.

Validasi pertama berfokus pada keaslian dokumen (Document Authenticity). Sistem memeriksa atribut header institusi ("UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" JAWA TIMUR") menggunakan algoritma Levenshtein Distance Ratio. Qhitfir, Sujaini, dan Sholva (2025) menyatakan bahwa metode ini krusial untuk memverifikasi akurasi teks dengan mengukur jarak penyuntingan minimum antar karakter. Dokumen dinyatakan valid jika rasio kemiripan (R) ≥ 0.75.

Validasi kedua adalah verifikasi entitas (Entity Verification) terhadap basis data mahasiswa aktif. Untuk mengantisipasi kesalahan pembacaan OCR, pencarian data NPM dan Nama dilakukan menggunakan pendekatan Fuzzy Search berbasis Levenshtein Distance. Pendekatan ini selaras dengan temuan Hládek dkk. (2017), yang menunjukkan bahwa metrik jarak string efektif untuk mengoreksi kesalahan ejaan otomatis pada luaran OCR.

Sistem menerapkan strategi verifikasi bertingkat (cascading verification), dengan memprioritaskan pencarian eksak, kemudian beralih ke pencarian samar (fuzzy search) apabila diperlukan. Data dianggap valid apabila skor kemiripan berada di atas ambang batas yang ditentukan ($R \geq 0.75$).

2.3. Skenario Pengujian

Untuk mengukur kinerja sistem secara kuantitatif, penelitian ini menggunakan metode evaluasi berbasis Confusion Matrix. Vakili dkk. (2020) mendefinisikan confusion matrix sebagai instrumen fundamental untuk memvisualisasikan performa algoritma klasifikasi dengan membandingkan hasil prediksi sistem terhadap nilai kebenaran (ground truth). Dalam konteks penelitian ini, dokumen

diklasifikasikan ke dalam empat kategori: True Positive (TP) jika KTM asli dikenali valid, False Positive (FP) jika kartu bukan KTM dikenali valid, True Negative (TN) jika kartu bukan KTM ditolak, dan False Negative (FN) jika KTM asli justru ditolak.

Berdasarkan parameter tersebut, kinerja sistem diukur menggunakan tiga metrik utama yang lazim diterapkan dalam analisis performa OCR seperti yang dilakukan oleh Prakisya dkk. (2024), yaitu:

a. Akurasi (Accuracy): Mengukur rasio prediksi benar secara keseluruhan.

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

b. Presisi (Precision): Mengukur tingkat ketepatan sistem saat menyatakan sebuah dokumen "Valid".

$$Precision = \frac{TP}{TP + FP}$$

c. Recall (Sensitivity): Mengukur kemampuan sistem dalam menemukan kembali seluruh dokumen valid yang ada dalam dataset.

$$Recall = \frac{TP}{TP + FN}$$

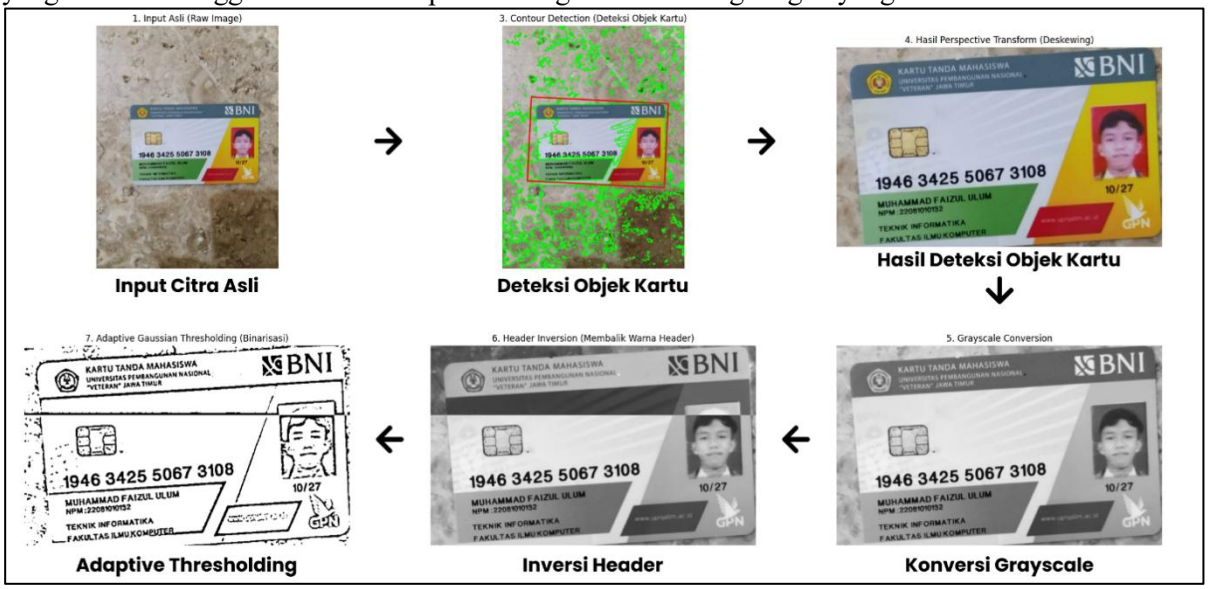
3. Hasil dan Pembahasan

Bagian ini menyajikan hasil implementasi serta analisis kinerja sistem validasi Kartu Tanda Mahasiswa (KTM) berbasis Tesseract OCR yang diusulkan. Pengujian difokuskan pada evaluasi efektivitas tahapan pra-pemrosesan citra, akurasi ekstraksi teks, serta keandalan mekanisme validasi hibrida yang mengombinasikan pencocokan header institusi dan verifikasi entitas mahasiswa.

Evaluasi dilakukan melalui beberapa skenario pengujian untuk mengamati kemampuan sistem dalam mengenali KTM UPN "Veteran" Jawa Timur secara valid serta menolak dokumen non-KTM. Hasil pengujian dianalisis secara kualitatif melalui visualisasi citra dan contoh keluaran OCR, serta secara kuantitatif menggunakan metrik evaluasi berbasis confusion matrix.

3.1. Hasil Pra-pemrosesan Citra

Tahapan pra-pemrosesan citra bertujuan untuk meningkatkan kualitas visual teks pada Kartu Tanda Mahasiswa (KTM) sebelum dilakukan proses ekstraksi menggunakan Optical Character Recognition (OCR). Gambar 1 menunjukkan alur pra-pemrosesan yang diterapkan pada citra input yang diambil menggunakan kamera ponsel dengan kondisi lingkungan yang tidak terkontrol.



Gambar 1. Alur pra-pemrosesan citra KTM

Berdasarkan hasil visualisasi, tahapan deteksi objek kartu dan koreksi perspektif berhasil menstandarisasi orientasi kartu sehingga area teks dapat diproses secara konsisten. Konversi citra ke grayscale diikuti dengan inversi warna pada area header dilakukan untuk menyesuaikan format teks menjadi lebih kontras terhadap latar belakang.

Penerapan adaptive Gaussian thresholding menghasilkan citra biner dengan pemisahan teks dan latar belakang yang lebih jelas dibandingkan citra grayscale, terutama pada kondisi pencahayaan tidak merata. Berdasarkan observasi visual, tahapan pra-pemrosesan dihentikan pada tahap thresholding adaptif karena telah memberikan tingkat keterbacaan karakter yang optimal untuk proses OCR selanjutnya.

3.2. Kinerja Ekstraksi Teks OCR

Bagian ini membahas kinerja sistem dalam mengekstraksi teks dari citra Kartu Tanda Mahasiswa (KTM) menggunakan Tesseract OCR. Evaluasi difokuskan pada kualitas hasil ekstraksi teks sebelum dilakukan proses validasi identitas, sehingga analisis pada tahap ini murni merepresentasikan performa OCR terhadap citra hasil pra-pemrosesan yang telah diusulkan. Aspek yang dianalisis meliputi hasil ekstraksi teks mentah (raw OCR output), proses pembersihan teks (text cleaning), serta nilai confidence score yang dihasilkan oleh mesin OCR.

3.2.1. Hasil Ekstraksi Teks Mentah

Hasil ekstraksi teks mentah menunjukkan bahwa metode Adaptive Gaussian Thresholding berhasil memisahkan karakter teks dari latar belakang secara optimal. Hampir seluruh karakter alfanumerik pada atribut vital (Nama, NPM) dapat dikenali. Namun, tingginya sensitivitas mesin Tesseract menyebabkan elemen non-teks seperti garis batas kartu (card borders) dan logo universitas turut terdeteksi sebagai karakter.

Berdasarkan log pengujian, gangguan yang sering muncul meliputi simbol vertikal (seperti "|", "1", atau "l") yang berasal dari deteksi garis tepi, serta simbol acak (seperti "©" atau "®") yang berasal dari interpretasi logo pada header. Selain itu, variasi jarak antar-kata pada kartu fisik seringkali diterjemahkan sebagai spasi berlebih (whitespace). Fenomena ini menegaskan perlunya tahapan pasca-pemrosesan (text cleaning) untuk memisahkan informasi relevan dari noise struktural tersebut.

3.2.2. Pembersihan Teks

Tahapan pembersihan teks (text cleaning) bertujuan untuk memperbaiki hasil ekstraksi teks mentah agar lebih terstruktur dan siap digunakan pada proses analisis selanjutnya. Pada tahap ini, sistem belum melakukan validasi terhadap basis data mahasiswa, melainkan hanya berfokus pada pengurangan noise dan normalisasi format teks hasil OCR.

Proses pembersihan teks mencakup penghapusan simbol non-relevan, normalisasi spasi berlebih, standarisasi huruf kapital, serta pemisahan baris teks berdasarkan pola atribut KTM seperti nama, NPM, program studi, dan fakultas. Pendekatan ini dirancang untuk mengurangi dampak kesalahan baca OCR tanpa melakukan koreksi berbasis referensi eksternal.

Tabel 1. Perbandingan Hasil OCR Sebelum dan Sesudah Pembersihan Teks

Teks Mentah Hasil OCR	Teks Hasil Pembersihan
NASIONAL » a BN I	NASIONAL a BN I
'— RIZKY FEBRIAN PUTRA	RIZKY FEBRIAN PUTRA
| NPM :23011010127	NPM : 23011010127
NPM :22081010039 °	NPM : 22081010039

Perbandingan hasil ekstraksi teks sebelum dan sesudah pembersihan ditunjukkan pada Tabel 1, yang memperlihatkan bagaimana simbol non-relevan dan kesalahan format berhasil direduksi tanpa mengubah makna informasi utama.

Hasil pembersihan teks menunjukkan peningkatan keterbacaan dan konsistensi struktur data dibandingkan dengan hasil teks mentah OCR. Atribut utama menjadi lebih mudah diidentifikasi secara visual, meskipun beberapa distorsi karakter minor masih ditemukan akibat kualitas citra yang beragam.

3.3. Evaluasi Validasi Identitas Menggunakan Levenshtein Distance

Pada bagian ini membahas evaluasi kinerja mekanisme validasi identitas pada sistem yang diusulkan. Evaluasi difokuskan pada kemampuan sistem dalam memastikan keaslian dokumen serta keabsahan data identitas mahasiswa berdasarkan hasil ekstraksi OCR. Validasi dilakukan menggunakan pendekatan string similarity berbasis Levenshtein Distance yang dikombinasikan dengan aturan logika (rule-based logic)

3.3.1. Mekanisme Validasi Berbasis Levenshtein Distance

Levenshtein Distance merupakan metode pengukuran jarak antar dua string berdasarkan jumlah operasi minimum yang diperlukan untuk mengubah satu string menjadi string lain, yang meliputi operasi penyisipan (insertion), penghapusan (deletion), dan substitusi (substitution). Dalam penelitian ini, Levenshtein Distance digunakan untuk menangani kesalahan minor pada hasil OCR seperti karakter terpotong, simbol tambahan, atau kesalahan ejaan (typo).

Nilai jarak yang dihasilkan kemudian dikonversi menjadi skor kemiripan (similarity ratio) untuk menentukan tingkat kecocokan antara teks hasil OCR dengan data referensi. Pendekatan ini memungkinkan sistem melakukan validasi secara toleran terhadap noise tanpa bergantung pada pencocokan teks secara eksak.

3.3.2. Validasi Keaslian Dokumen (Header Institusi)

Tahap validasi pertama bertujuan untuk memastikan bahwa dokumen yang diproses merupakan Kartu Tanda Mahasiswa (KTM) milik Universitas Pembangunan Nasional "Veteran" Jawa Timur. Validasi dilakukan dengan mencocokkan teks header institusi hasil OCR terhadap string referensi resmi menggunakan Levenshtein Distance Ratio.

Dokumen dinyatakan valid apabila skor kemiripan berada di atas ambang batas yang telah ditentukan, yaitu sebesar $R \geq 0.75$. Berdasarkan hasil pengujian, seluruh sampel KTM UPN menghasilkan skor kemiripan yang tinggi meskipun terdapat distorsi karakter minor seperti simbol tambahan atau kesalahan pemisahan kata. Hal ini menunjukkan bahwa pendekatan fuzzy matching berbasis Levenshtein Distance efektif dalam mengidentifikasi keaslian dokumen pada kondisi hasil OCR yang tidak sempurna.

Tabel 2. Hasil Validasi Header Institusi KTM Menggunakan Levenshtein Distance

Sampel	Baris Ke-	Teks Header Hasil OCR	Skor (R)	Status Baris
Data 1	1	UNIVERSITAS PEMBANGUNAN NASIONAL **:**	**0.98**	Valid
	2	"VETERAN" JAWA TIMUR	**1.00**	Valid
Data 2	1	UNIVERSITAS PEMBANGUNAN NASIONAL	**1.00**	Valid
	2	"VETERAN" JAWA TIMUR **ai TN**	**0.85**	Valid
Data 3	1	UNIVERSITAS PEMBANGUNAN NASIONAL **a BN I**	**0.94**	Valid
	2	"VETERAN JAWA TIMUR **Peat**	**0.79**	Valid
Data 4	1	UNIVERSITAS PEMBANGUNAN NASIONAL **Ds**	**0.97**	Valid
	2	"VETERAN" JAWA TIMUR	**1.00**	Valid
Data 5	1	LINIVERSITAS AIRLANGGA	**< 0.75**	Tidak Valid
	2	WWW.unair.ac.id	**< 0.75**	Tidak Valid

Berdasarkan hasil pada Tabel 1, dapat dilihat bahwa teks header institusi pada seluruh sampel KTM menghasilkan skor kemiripan di atas ambang batas yang ditentukan. Meskipun pada beberapa

kasus terjadi distorsi karakter tambahan seperti simbol atau suku kata acak akibat noise OCR, algoritma Levenshtein Distance tetap mampu mempertahankan nilai kemiripan yang tinggi. Validasi header dinyatakan berhasil apabila kedua baris utama header memenuhi kriteria kemiripan minimum, sehingga meminimalkan kemungkinan dokumen non-KTM lolos ke tahap validasi lanjutan.

3.3.3. Validasi Entitas Mahasiswa

Setelah dokumen dinyatakan valid berdasarkan header institusi, sistem melanjutkan ke tahap validasi entitas mahasiswa yang mencakup atribut Nama, Nomor Pokok Mahasiswa (NPM), Program Studi, dan Fakultas. Proses validasi ini menerapkan pendekatan hibrida yang mengombinasikan pencocokan teks berbasis Levenshtein Distance dan aturan logika berbasis struktur data.

Validasi NPM diprioritaskan karena memiliki pola numerik tetap dan diverifikasi menggunakan aturan validasi temporal untuk memastikan tahun angkatan masih berada dalam rentang mahasiswa aktif. Sementara itu, atribut berbasis teks seperti Nama, Program Studi, dan Fakultas divalidasi menggunakan fuzzy matching berbasis Levenshtein Distance Similarity Ratio dengan membandingkan hasil ekstraksi OCR terhadap data referensi.

Data entitas dinyatakan valid apabila skor kemiripan berada di atas ambang batas penerimaan ($R \geq 0.75$). Apabila pencocokan teks tidak memenuhi ambang batas tersebut, sistem menerapkan mekanisme fallback berbasis struktur kode NPM untuk memulihkan informasi Program Studi dan Fakultas.

Tabel 3. Hasil Validasi Entitas Mahasiswa Menggunakan Levenshtein Distance dan Aturan Logika

Sampel	Entitas	Teks Hasil OCR	Data Database	Skor (*R*)	Status baris
Data 1	Nama	MUHAMMAD FAIZUL ULUM	MUHAMMAD FAIZUL ULUM	**1.00**	Valid
	NPM	22081010132	22081010132	**1.00**	Valid
	Prodi	TEKNIK INFORMATIKA	TEKNIK INFORMATIKA	**1.00**	Valid
	Fakultas	SAK ULTAS JLMU.	FAKULTAS ILMU KOMPUTER	**< 0.75**	Fallback kode NPM
Data 2	Nama	AHMAD TSALITS HILMI	AHMAD TSALITS HILMI	**1.00**	Valid
	NPM	22025010056	22025010056	**1.00**	Valid
	Prodi	AGROTEKNOLOGI	AGROTEKNOLOGI	**1.00**	Valid
	Fakultas	FAKULTAS PERTANIAN	FAKULTAS PERTANIAN	**1.00**	Valid
Data 3	Nama	MUHAMMAD DIAZ SYAHMI OKTAV	MUHAMMAD DIAZ SYAHMI OKTAVIAN	**0.87**	Valid
	NPM	22081010039	22081010039	**1.00**	Valid
	Prodi	TEKNIK INFORMATIKA	TEKNIK INFORMATIKA	**1.00**	Valid
	Fakultas	FAKULTAS ILMU KOMPUTER	FAKULTAS ILMU KOMPUTER	**1.00**	Valid
Data 4	Nama	MUHAMMAD HIDAYAT NURWAHID	MUHAMMAD HIDAYAT NURWAHID	**1.00**	Valid
	NPM	22081010132	22081010132	**1.00**	Valid
	Prodi	TEKNIK INFORMATIKA	TEKNIK INFORMATIKA	**1.00**	Valid
	Fakultas	FAKULTAS ILMU KOMPUTER	FAKULTAS ILMU KOMPUTER	**1.00**	Valid
Data 5	Nama	RIZKY FEBRIAN PUTRA	RIZKY FEBRIAN PUTRA	**1.00**	Valid
	NPM	23011010127	23011010127	**1.00**	Valid
	Prodi	EKONOMI PEMBANGUNAN	EKONOMI PEMBANGUNAN	**1.00**	Valid
	Fakultas	FAKULTAS EKONOMI DAN BISNIS	FAKULTAS EKONOMI DAN BISNIS	**1.00**	Valid

Tabel 3 menyajikan hasil pengujian validasi entitas mahasiswa pada lima sampel KTM. Hasil menunjukkan bahwa pendekatan Levenshtein Distance mampu mengakomodasi kesalahan OCR seperti penghilangan karakter dan distorsi minor, sebagaimana terlihat pada kasus pemotongan sebagian nama mahasiswa pada Data 3, yang tetap menghasilkan skor kemiripan tinggi (R = 0.87) dan dapat dikenali sebagai entitas yang valid. pada Data 1. Meskipun atribut Fakultas mengalami kerusakan teks parah ("SAK ULTAS JLMU."), sistem berhasil memulihkan data menjadi "FAKULTAS ILMU KOMPUTER" melalui mekanisme fallback yang memetakan kode unik fakultas dari digit NPM. Hal ini membuktikan bahwa integrasi validasi teks dan logika struktur data memberikan fleksibilitas tinggi terhadap variasi kualitas citra input.

Pendekatan ini membuktikan bahwa integrasi Levenshtein Distance dalam proses validasi entitas tidak hanya efektif dalam menangani kesalahan ejaan, tetapi juga memberikan fleksibilitas tinggi pada kondisi citra dengan kualitas yang bervariasi.

3.4. Evaluasi Kinerja Sistem Berbasis Confusion Matrix

Tahap akhir evaluasi bertujuan untuk mengukur kinerja keseluruhan sistem dalam mengklasifikasikan dokumen secara otomatis berdasarkan hasil ekstraksi dan validasi yang telah dilakukan pada tahapan sebelumnya. Evaluasi dilakukan menggunakan pendekatan Confusion Matrix untuk membandingkan hasil prediksi sistem dengan label kebenaran (ground truth) yang ditentukan melalui verifikasi manual terhadap jenis dokumen dan kesesuaian identitas.

Pengujian dilakukan terhadap 20 sampel citra yang terdiri dari 15 citra Kartu Tanda Mahasiswa (KTM) Universitas Pembangunan Nasional "Veteran" Jawa Timur sebagai kelas positif, serta 5 citra kartu identitas lain sebagai kelas negatif. Seluruh sampel diuji pada kondisi pencahayaan normal untuk merepresentasikan skenario penggunaan sistem pada lingkungan nyata.

Berdasarkan hasil pengujian validasi yang telah dibahas pada sub-bab sebelumnya, distribusi hasil klasifikasi dipetakan ke dalam Confusion Matrix sebagaimana disajikan pada Tabel 4.

Tabel 4. Confusion Matrix Hasil Klasifikasi Dokumen

	Prediksi Sistem: Valid (KTM)	**Prediksi Sistem: Tidak Valid (Non-KTM)**
Aktual: Valid (KTM Asli)	**14** (TP)	**1** (FN)
Aktual: Tidak Valid (Kartu Lain)	**0** (FP)	**5** (TN)

Nilai performa sistem dihitung berdasarkan rumus evaluasi yang merujuk pada penelitian yang dilakukan oleh Prakisya dkk. (2024):

a. Akurasi (Accuracy)

$$Accuracy = \frac{14 + 5}{14 + 5 + 0 + 1} \times 100\%$$

$$Accuracy = \frac{19}{20} \times 100\% = \mathbf{95}\%$$

Sistem memiliki tingkat kebenaran keseluruhan sebesar 95% dalam membedakan antara KTM valid dan dokumen tidak valid.

b. Presisi (Precision)

$$Precision = \frac{14}{14 + 0} \times 100\% = \mathbf{100}\%$$

Nilai presisi sempurna (100%) mengindikasikan bahwa tidak ada dokumen palsu atau salah klasifikasi yang lolos sebagai dokumen valid. Hal ini sangat krusial dalam konteks sistem peminjaman inventaris untuk mencegah penyalahgunaan aset organisasi.

c. Recall (Sensitivity)

$$Recall = \frac{14}{14+1} \times 100\% = \mathbf{93.3}\%$$

Nilai *Recall* sebesar 93.3% menunjukkan bahwa sistem cukup sensitif dalam mendeteksi dokumen valid, meskipun masih terdapat sedikit kegagalan pada input dengan kualitas fisik yang sangat rendah.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa kombinasi metode Adaptive Gaussian Thresholding pada tahap pra-pemrosesan dan algoritma Levenshtein Distance pada tahap validasi mampu mempertahankan performa klasifikasi yang konsisten pada berbagai kondisi citra. Nilai Presisi yang mencapai 100% menandakan bahwa mekanisme validasi bertingkat efektif dalam meminimalkan kesalahan penerimaan dokumen tidak sah. Sementara itu, nilai Akurasi sebesar 95% dan Recall sebesar 93,3% menunjukkan bahwa sistem memiliki tingkat sensitivitas dan stabilitas yang baik dalam mengenali dokumen KTM pada skenario pengujian yang dilakukan.

4. Kesimpulan

Penelitian ini mengevaluasi kinerja sistem validasi Kartu Tanda Mahasiswa (KTM) berbasis Tesseract OCR dengan penerapan Adaptive Gaussian Thresholding pada tahap pra-pemrosesan serta Levenshtein Distance pada tahap validasi teks. Hasil pengujian menunjukkan bahwa kombinasi metode tersebut mampu mengekstraksi dan memvalidasi informasi identitas mahasiswa secara andal meskipun terdapat gangguan kualitas citra dan kesalahan baca OCR.

Berdasarkan evaluasi menggunakan Confusion Matrix terhadap 20 sampel citra, sistem mencapai akurasi sebesar 95%, presisi 100%, dan recall 93,3%. Nilai presisi yang tinggi menunjukkan bahwa sistem berhasil mencegah kesalahan penerimaan dokumen non-KTM, sementara mekanisme validasi bertingkat dan fallback berbasis struktur NPM meningkatkan keandalan sistem pada kondisi hasil OCR yang tidak sempurna. Dengan demikian, sistem yang diusulkan dinilai layak untuk diterapkan sebagai solusi otomatis pada proses verifikasi KTM di lingkungan operasional.

5. Ucapan terimakasih

Penulis mengucapkan terima kasih kepada Universitas Pembangunan Nasional "Veteran" Jawa Timur atas dukungan fasilitas dan lingkungan akademik yang mendukung terlaksananya penelitian ini. Ucapan terima kasih juga disampaikan kepada Unit Kegiatan Kerohanian Islam (UKKI) UPN "Veteran" Jawa Timur yang telah memberikan konteks permasalahan serta mendukung proses pengujian sistem validasi identitas yang dikembangkan.

Daftar Pustaka

Akinbade, D., Ogunde, A. O., Odim, M. O., & Oguntunde, B. O. (2020). An adaptive thresholding algorithm-based optical character recognition system for information extraction in complex images. *Journal of Computer Science*, *16*(6), 784–801. https://doi.org/10.3844/JCSSP.2020.784.801

Carta, S., Giuliani, A., Piano, L., & Tiddia, S. G. (2024). An End-to-End OCR-Free Solution For Identity Document Information Extraction. *Procedia Computer Science*, *246*(C), 453–462. https://doi.org/10.1016/j.procs.2024.09.425

Dey, R., Balabantaray, R. C., Mohanty, S., Singh, D., Karuppiah, M., & Samanta, D. (2022). Approach for Preprocessing in offline Optical Character Recognition (OCR). *2022 International Conference on Interdisciplinary Research in Technology and Management, IRTM 2022 - Proceedings*. https://doi.org/10.1109/IRTM54583.2022.9791698

Erameh, K. B., & Odoh, B. I. (2021). Design and Implementation of a Web-Based Inventory Control System Using a Small Medium Enterprise (SME) as a Case Study. *NIPES - Journal of Science and Technology Research*, *3*(3), 211–219. https://doi.org/10.37933/nipes/3.3.2021.21

Hamad, K. A., & Kaya, M. (2016). *Applied Mathematics , Electronics and Computers A Detailed*

Analysis of Optical Character Recognition Technology. https://www.researchgate.net/profile/Karez_Hamad/publication/311851325_A_Detailed_Analysis_of_Optical_Character_Recognition_Technology/links/5862191908ae8fce490767f6/A-Detailed-Analysis-of-Optical-Character-Recognition-Technology.pdf

Hládek, D., Staš, J., Ondáš, S., Juhár, J., & Kovács, L. (2017). Learning string distance with smoothing for OCR spelling correction. *Multimedia Tools and Applications*, *76*(22), 24549–24567. https://doi.org/10.1007/s11042-016-4185-5

Ingle, P. D., & Kaur, P. (2017). Adaptive thresholding to robust image binarization for degraded document images. *Proceedings - 1st International Conference on Intelligent Systems and Information Management, ICISIM 2017*, *2017-Janua*, 189–193. https://doi.org/10.1109/ICISIM.2017.8122172

Kusumawardhani, W., Purwanto, A., Fariqi, M., & Afnan G, L. (2025). Perancangan Sistem Peminjaman Barang Inventaris Berbasis Website untuk Meningkatkan Keamanan Aset Inventaris ITS. *Blantika: Multidisciplinary Journal*, *3*(3), 259–269. https://doi.org/10.57096/blantika.v3i3.300

Nasution, A. B., Aulia, H., Audiansyah, W., & Raihan, M. S. (2023). Implementasi Keamanan Aset Sekolah Angkasa Berbasis Website. *Jurnal Sains Dan Teknologi (JSIT)*, *3*(1), 68–73. https://doi.org/10.47233/jsit.v3i1.495

Prakisya, N. P. T., Kusmanto, B. T., & Hatta, P. (2024). Comparative Analysis of Google Vision OCR with Tesseract on Newspaper Text Recognition. *Media of Computer Science*, *1*(1), 31–46. https://doi.org/10.69616/mcs.v1i1.178

Qhitfir, M., Sujani, H., & Sholva, Y. (2025). Analisis Perbandingan Tesseract Ocr dan Easyocr Untuk Pengenalan Karakter Dengan Yolo Sebagai Alat Bantu Dalam Pendeteksian Plat Nomor Kendaraan. *Scientica: Jurnal Ilmiah Sains Dan Teknologi*, *3*(5), 333–340.

Rusli, F. M., Adhiguna, K. A., & Irawan, H. (2021). Indonesian ID Card Extractor Using Optical Character Recognition and Natural Language Post-Processing. *2021 9th International Conference on Information and Communication Technology, ICoICT 2021*, 621–626. https://doi.org/10.1109/ICoICT52021.2021.9527510

Saoji, S., Arora, A., Singh, R., Mangal, A., & Eqbal, A. (2021). Text Recogination and Detection From Images Using Pytesseract. *Article in Journal of Interdisiplinary Cycle Research*, *XIII*(Vii), 1674–1679. https://www.researchgate.net/publication/353679800

Sofwan, A., Sumardi, A. Y., Santoso, I., Adi Soetrisno, Y. A., Arfan, M., & Handoyo, E. (2021). Optimization of OCR in Detecting Research Proposal and Lecturer Community Service Documents using Thresholding Method. *2021 8th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE 2021*, 175–179. https://doi.org/10.1109/ICITACEE53184.2021.9617487

Vakili, M., Ghamsari, M., & Rezaei, M. (2020). *Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification*. http://arxiv.org/abs/2001.09636

Wibisono, R. S., Sofianti, T. D., & Awibowo, S. (2016). Development of A Web-Based Information System for Material Inventory Control: The Case of An Automotive Company. *CommIT (Communication and Information Technology) Journal*, *10*(2), 71. https://doi.org/10.21512/commit.v10i2.1579

Zacharias, E., Teuchler, M., & Bernier, B. (2020). *Image Processing Based Scene-Text Detection and Recognition with Tesseract*. 1–6. http://arxiv.org/abs/2004.08079