Layanan Perpustakaan Uhamka

Firman - EKSTRAKSI INFORMASI DAN PENYERAGAMAN DATA SERTIFIKAT PELATIHAN MENGGUNAKAN PENDEKATAN COM...

29/01/2026

Document Details

Submission ID
trn:oid:::3618:127187521

Submission Date
Jan 30, 2026, 3:03 PM GMT+7

Download Date
Jan 30, 2026, 3:41 PM GMT+7

File Name
Template JATi_Sitasi.docx

File Size
233.8 KB

11 Pages

6,439 Words

45,122 Characters

4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Exclusions

- 6 Excluded Matches

Top Sources

3% Internet sources
2% Publications
3% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Top Sources

3% Internet sources
2% Publications
3% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	
Universitas Muhammadiyah Purwokerto on 2025-07-09		<1%
2	Student papers	
Universitas Negeri Surabaya on 2024-07-08		<1%
3	Student papers	
Universitas Negeri Manado on 2025-06-15		<1%
4	Internet	
docplayer.info		<1%
5	Internet	
ejurnal.stmik-budidarma.ac.id		<1%
6	Internet	
jurnal.umb.ac.id		<1%
7	Internet	
jurnal.kdi.or.id		<1%
8	Student papers	
Griffth University on 2025-10-13		<1%
9	Publication	
Latifatunnisa, Wahdah Refia Rafianti. "Efektivitas Penggunaan Model Group Inve...		<1%
10	Publication	
Sulis Sutiono, Rian Farta Wijaya, Muhammad Syahputra Novelan. "Analisis Metod...		<1%
11	Internet	
www.scilit.net		<1%

12 Publication

Putu Anggi Suryantari, Faisal Muttaqin, Ani Dijah Rahajoe. "PERAN BIG DATA DAL... <1%

13 Internet

ejournal.umkla.ac.id <1%

14 Internet

journal.arteii.or.id <1%

15 Internet

journal.stie-yppi.ac.id <1%

16 Internet

repository.iainpurwokerto.ac.id <1%

17 Internet

sistemasi.ftik.unisi.ac.id <1%

Harap mengisi tabel ini, Tabel ini digunakan untuk keperluan komunikasi administrasi saja, saat publish akan dihapus oleh team editor.	
Nama Kontak	**Irfan Ricky Afandi**
Nomor WA	**082111533130**
Instansi	**Komisi Pemilihan Umum Republik Indonesia**

EKSTRAKSI INFORMASI DAN PENYERAGAMAN DATA SERTIFIKAT PELATIHAN MENGGUNAKAN PENDEKATAN COMPUTER VISION

Irfan Ricky Afandi 1*, Ratna Yuniarti 2, Ilham 3, Sewin Fathurrohman 4, Firman Noor Hasan 5, Isa Faqihuddin Hanif 6
[1, 2, 3] Komisi Pemilihan Umum Republik Indonesia
Jl. Cideng Timur No.54 DKI Jakarta, Indonesia
[4] Program Studi Sistem Informasi, Universitas Gunadarma
Jl. Margonda Raya Pondok Cina Depok, Indonesia
[5] Program Studi Teknik Informatika, Universitas Muhammadiyah Prof. Dr. Hamka
[6] Program Studi Sistem dan Teknologi Informasi, Universitas Muhammadiyah Prof. Dr. Hamka
Jl. Tanah Merdeka No.6 Jakarta, Indonesia
richky14@gmail.com

ABSTRAK

Pengelolaan sertifikat pelatihan merupakan bagian penting dalam pengembangan kompetensi sumber daya manusia, khususnya di lingkungan instansi pemerintah. Namun, sertifikat pelatihan umumnya disimpan dalam format dokumen tidak terstruktur dengan variasi tata letak dan penulisan yang tinggi, sehingga menyulitkan proses konsolidasi dan analisis data secara sistematis. Penelitian ini bertujuan untuk mengembangkan metode ekstraksi informasi dan penyeragaman data sertifikat pelatihan menggunakan pendekatan *computer vision* pada dokumen heterogen. Metode yang diusulkan berupa pipeline pemrosesan dokumen end-to-end yang mengintegrasikan identifikasi jenis dokumen, strategi ekstraksi teks adaptif, *Optical Character Recognition*, serta ekstraksi informasi berbasis template dan aturan, dilengkapi dengan mekanisme penilaian tingkat kepercayaan. Metode ini diuji pada 4.785 dokumen sertifikat pelatihan dari berbagai instansi pemerintah dengan 15 variasi format template. Hasil penelitian menunjukkan bahwa sekitar 87% dokumen berhasil diproses secara otomatis dengan tingkat kegagalan OCR yang rendah pada dokumen hasil pemindaian, serta menghasilkan data sertifikat yang terstruktur dan seragam. Temuan ini membuktikan bahwa pendekatan yang diusulkan efektif dalam mendukung pengelolaan data pelatihan dan pengambilan keputusan pengembangan kompetensi sumber daya manusia di sektor publik.

Kata kunci : *ekstraksi informasi, sertifikat pelatihan, computer vision, OCR, dokumen heterogen*

1. PENDAHULUAN

Peningkatan tuntutan terhadap penerapan *evidence-based human resource management* telah mendorong organisasi untuk mengelola data pengembangan sumber daya manusia secara lebih sistematis dan akuntabel [1]. Program pelatihan dan pengembangan profesional merupakan instrumen strategis dalam meningkatkan kompetensi pegawai, baik di sektor privat maupun sektor publik [2]. Dalam konteks tersebut, sertifikat pelatihan berfungsi sebagai bukti formal atas pencapaian kompetensi individu dan menjadi dasar penting dalam pemetaan kompetensi, pengembangan karier, serta evaluasi kinerja kelembagaan [2]. Oleh karena itu, ketersediaan data sertifikat pelatihan yang akurat, terstruktur, dan mudah dianalisis menjadi kebutuhan mendasar bagi pengambilan keputusan organisasi yang berbasis data [3].

Seiring dengan perkembangan teknologi informasi, banyak institusi telah beralih ke penyimpanan sertifikat pelatihan dalam bentuk digital, baik berupa dokumen hasil pemindaian maupun berkas *Portable Document Format* (PDF). Digitalisasi arsip ini memang mempermudah proses penyimpanan dan distribusi dokumen. Namun demikian, informasi yang terkandung di dalam sertifikat umumnya masih tersimpan dalam format tidak terstruktur, di mana elemen penting seperti nama peserta, nomor sertifikat, judul pelatihan, tanggal pelaksanaan, dan durasi pelatihan direpresentasikan sebagai teks di dalam citra dokumen. Kondisi ini menyebabkan proses konsolidasi data sertifikat masih sangat bergantung pada pemeriksaan dan pencatatan manual.

Berbagai studi terdahulu menunjukkan bahwa pemrosesan dokumen secara manual tidak hanya membutuhkan waktu dan sumber daya yang besar, tetapi juga rentan terhadap inkonsistensi serta kesalahan transkripsi, sehingga berdampak pada rendahnya kualitas data yang dihasilkan [4]. Dalam lingkungan organisasi yang memiliki volume

dokumen besar dan tersebar di berbagai unit kerja, permasalahan ini menjadi semakin kompleks dan sulit untuk diskalakan. Akibatnya, potensi pemanfaatan arsip sertifikat digital sebagai sumber data strategis untuk analisis dan perumusan kebijakan sumber daya manusia belum dapat dioptimalkan secara maksimal.

Dari sudut pandang manajerial, khususnya di lingkungan pemerintahan, keterbatasan data sertifikat yang terstruktur menimbulkan tantangan signifikan dalam proses *monitoring* dan evaluasi riwayat pelatihan pegawai. Pimpinan organisasi membutuhkan informasi yang teragregasi dan tervalidasi untuk menilai tingkat partisipasi pelatihan, mengidentifikasi kesenjangan kompetensi, serta merancang program pengembangan kapasitas yang lebih tepat sasaran [5]. Ketika data sertifikat masih berada dalam bentuk tidak terstruktur, proses verifikasi menjadi tidak efisien dan sulit mendukung kebutuhan analisis berbasis bukti, terutama dalam konteks akuntabilitas dan perencanaan sumber daya manusia sektor publik.

Dalam beberapa dekade terakhir, teknik *computer vision* dan *document image analysis* telah banyak digunakan untuk mengatasi permasalahan pengolahan dokumen digital [6]. *Optical Character Recognition* (OCR) merupakan pendekatan fundamental yang bertujuan mengonversi teks yang terkandung dalam citra menjadi representasi yang dapat dibaca oleh mesin [7]. Sistem OCR klasik umumnya melibatkan tahapan praproses citra, pengenalan karakter, serta pascaproses untuk meningkatkan akurasi pengenalan [8]. Penerapan OCR telah menunjukkan hasil yang menjanjikan pada berbagai jenis dokumen, seperti faktur, dokumen identitas, dan formulir administrasi [9].

Meskipun demikian, sebagian besar penelitian terkait OCR berfokus pada dokumen dengan tata letak yang relatif seragam atau memiliki struktur yang baku. Sertifikat pelatihan memiliki karakteristik yang berbeda, yaitu variasi tata letak antar instansi penerbit, perbedaan terminologi, serta kualitas visual dokumen yang tidak selalu konsisten akibat kondisi pemindaian. Selain itu, banyak penelitian menitikberatkan pada akurasi pengenalan karakter, tanpa membahas secara mendalam tahapan lanjutan untuk mengubah hasil OCR mentah menjadi data terstruktur yang memiliki makna semantik. Literatur analisis dokumen menegaskan bahwa ekstraksi informasi yang efektif tidak hanya bergantung pada akurasi OCR, tetapi juga memerlukan strategi pascaproses yang memanfaatkan pengetahuan domain, pola tata letak, dan konteks dokumen [10].

Kesenjangan inilah yang menunjukkan perlunya pendekatan terintegrasi yang tidak hanya mengandalkan OCR, tetapi juga mengombinasikannya dengan teknik ekstraksi informasi yang sadar terhadap konteks domain. Metode berbasis aturan (*rule-based information extraction*) masih relevan digunakan dalam pemrosesan dokumen spesifik domain, terutama ketika pola konten dan struktur dokumen dapat diidentifikasi meskipun bervariasi antar template [11]. Pendekatan semacam ini dinilai sesuai untuk lingkungan organisasi, termasuk institusi pemerintahan, di mana variasi format sertifikat cukup tinggi namun tetap memiliki elemen informasi inti yang serupa.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan sebuah *end-to-end pipeline* untuk ekstraksi informasi otomatis dari sertifikat pelatihan berbasis OCR dan *computer vision*. Pendekatan yang diusulkan mencakup identifikasi jenis dokumen, praproses citra, proses OCR, serta ekstraksi informasi berbasis aturan dan template untuk menghasilkan data sertifikat yang terstruktur. Pipeline ini dirancang agar mampu menangani dokumen PDF digital maupun hasil pemindaian, serta mengakomodasi berbagai template sertifikat yang diterbitkan oleh institusi yang berbeda. Selain itu, mekanisme penilaian tingkat kepercayaan (*confidence scoring*) diterapkan untuk mengevaluasi kelengkapan hasil ekstraksi dan mendukung keputusan apakah dokumen dapat diproses secara otomatis atau memerlukan peninjauan manual.

Tujuan penelitian ini adalah: (1) melakukan ekstraksi otomatis terhadap elemen informasi utama pada sertifikat pelatihan menggunakan kombinasi OCR dan teknik berbasis aturan; (2) mengevaluasi kinerja *pipeline* yang diusulkan pada kumpulan sertifikat pelatihan dengan format yang heterogen dari lingkungan organisasi nyata; serta (3) menganalisis kontribusi tahapan praproses dan strategi ekstraksi berbasis template terhadap keandalan pembentukan data terstruktur.

Kontribusi penelitian ini dapat dilihat dari dua perspektif. Dari sisi ilmiah, penelitian ini memperluas kajian *document image analysis* dengan memfokuskan pada sertifikat pelatihan sebagai domain dokumen yang relatif belum banyak diteliti dan memiliki tingkat heterogenitas tinggi. Penelitian ini juga menunjukkan efektivitas integrasi OCR, ekstraksi informasi berbasis aturan, dan penilaian tingkat kepercayaan dalam satu *pipeline* terpadu. Dari sisi praktis, pendekatan yang diusulkan memberikan solusi yang aplikatif bagi organisasi, khususnya instansi pemerintah, dalam mendigitalisasi dan menstrukturkan data sertifikat pelatihan secara efisien. Kemampuan ini mendukung proses verifikasi yang lebih cepat, pemetaan kompetensi yang lebih akurat, serta perencanaan pengembangan sumber daya manusia yang berbasis data dan akuntabel.

2. TINJAUAN PUSTAKA

2.1. Digitalisasi Dokumen dan Pengelolaan Data Tidak Terstruktur

Digitalisasi dokumen merupakan bagian penting dari transformasi organisasi menuju pengelolaan informasi yang lebih efisien dan berbasis data. Dalam konteks manajemen dokumen, digitalisasi tidak hanya mencakup proses konversi dokumen fisik ke bentuk digital, tetapi juga mencakup kemampuan sistem

untuk mengelola, menelusuri, dan menganalisis informasi yang terkandung di dalam dokumen tersebut [12]. Namun demikian, sebagian besar dokumen digital organisasi, termasuk sertifikat pelatihan, masih disimpan dalam bentuk data tidak terstruktur, di mana informasi penting tidak tersedia dalam format yang langsung dapat diproses oleh sistem informasi.

Penelitian di bidang *document management systems* menunjukkan bahwa dokumen tidak terstruktur menimbulkan tantangan signifikan dalam proses integrasi data dan pengambilan keputusan berbasis bukti [13]. Dokumen seperti sertifikat pelatihan mengandung informasi yang bernilai strategis, tetapi representasinya sebagai citra atau teks bebas membatasi pemanfaatannya untuk analisis lanjutan. Kondisi ini menjadi semakin kompleks dalam lingkungan organisasi besar dan institusi pemerintahan, di mana volume dokumen tinggi dan format dokumen bervariasi antar unit atau lembaga penerbit.

Permasalahan tersebut menegaskan perlunya pendekatan pemrosesan dokumen yang mampu menjembatani kesenjangan antara arsip digital dan kebutuhan data terstruktur. Untuk tujuan tersebut, bidang *document image analysis* dan *computer vision* menyediakan kerangka konseptual dan teknis yang relevan untuk mengekstraksi informasi bermakna dari dokumen berbasis citra.

2.2. Document Image Analysis dalam Konteks Computer Vision

Document image analysis merupakan subbidang dari *computer vision* yang berfokus pada pemahaman dan interpretasi dokumen dalam bentuk citra. Penelitian awal di bidang ini menekankan pentingnya analisis tata letak, segmentasi teks, dan pengenalan karakter sebagai komponen dasar sistem pemahaman dokumen [14]. Pendekatan ini menekankan bahwa pemrosesan dokumen tidak dapat bergantung pada satu tahapan tunggal, melainkan memerlukan integrasi berbagai proses visual dan struktural untuk menghasilkan informasi yang dapat digunakan.

Seiring berkembangnya penelitian, perhatian terhadap variasi struktur dokumen, degradasi kualitas citra, serta artefak hasil pemindaian semakin meningkat. Penelitian yang dilakukan oleh [15] menegaskan bahwa koleksi dokumen dunia nyata umumnya bersifat heterogen dan tidak mengikuti struktur yang seragam, sehingga menuntut strategi analisis yang fleksibel dan adaptif. Temuan ini relevan dalam konteks dokumen institusional, termasuk sertifikat pelatihan, yang sering kali memiliki variasi tata letak dan gaya visual meskipun memuat elemen informasi yang serupa.

Dengan demikian, *document image analysis* menyediakan landasan teoritis bagi pengembangan sistem yang mampu menangani variasi format dokumen. Namun, untuk mengonversi teks dalam dokumen citra menjadi data yang dapat dianalisis, diperlukan komponen teknis utama berupa *Optical Character Recognition*.

2.3. Optical Character Recognition dalam Pemrosesan Dokumen

Optical Character Recognition (OCR) merupakan teknik fundamental dalam pemrosesan dokumen citra yang bertujuan mengonversi teks visual menjadi representasi teks yang dapat dibaca oleh mesin. Sistem OCR klasik umumnya mengikuti alur pemrosesan yang terdiri atas praproses citra, pengenalan karakter, dan pascaproses untuk meningkatkan kualitas hasil pengenalan [16]. Tahapan praproses memiliki peran penting dalam meningkatkan keterbacaan teks dan mengurangi kesalahan pengenalan.

Perkembangan teknologi OCR telah memungkinkan penerapannya pada berbagai jenis dokumen administratif. Penelitian yang dilakukan oleh [17] menunjukkan bahwa mesin OCR berbasis *open-source* mampu mencapai kinerja yang kompetitif apabila dikombinasikan dengan konfigurasi dan praproses yang tepat. Meskipun demikian, sejumlah penelitian juga menunjukkan bahwa akurasi OCR sangat dipengaruhi oleh kualitas dokumen, variasi jenis huruf, serta kompleksitas tata letak [18]. Dokumen dengan format tidak seragam dan kualitas pemindaian yang beragam, seperti sertifikat pelatihan, sering kali menghasilkan keluaran OCR yang mengandung kesalahan dan inkonsistensi.

Temuan-temuan tersebut menunjukkan bahwa OCR sebaiknya dipandang sebagai tahapan awal dalam pemrosesan dokumen, bukan sebagai solusi akhir. Keluaran OCR berupa teks mentah masih memerlukan pemrosesan lanjutan agar dapat diubah menjadi informasi terstruktur yang memiliki makna semantik dan dapat digunakan dalam sistem pendukung keputusan.

2.4. Praproses Citra untuk Peningkatan Kinerja OCR

Praproses citra merupakan tahapan penting dalam *pipeline* OCR yang bertujuan meningkatkan kualitas visual dokumen sebelum dilakukan pengenalan karakter. Teknik praproses yang umum digunakan meliputi konversi ke skala abu-abu, pengurangan derau, *thresholding*, koreksi kemiringan (*skew correction*), serta peningkatan kontras. Penelitian yang dilakukan oleh [19] menunjukkan bahwa penerapan praproses yang tepat dapat meningkatkan akurasi pengenalan karakter secara signifikan, terutama pada dokumen dengan kualitas rendah.

Dalam konteks dokumen hasil pemindaian, praproses menjadi semakin krusial karena adanya variasi resolusi, pencahayaan, dan orientasi dokumen. Penelitian yang dilakukan oleh [20] menekankan bahwa strategi praproses sebaiknya disesuaikan dengan karakteristik dokumen, bukan diterapkan secara seragam. Pendekatan adaptif ini relevan untuk dokumen sertifikat pelatihan yang dapat berasal dari berkas PDF digital maupun hasil pemindaian fisik.

Oleh karena itu, praproses citra tidak dapat diperlakukan sebagai tahapan opsional, melainkan sebagai komponen integral dalam *pipeline* ekstraksi informasi. Praproses yang dirancang dengan baik akan meningkatkan kualitas keluaran OCR dan mendukung tahapan ekstraksi informasi selanjutnya.

2.5. Ekstraksi Informasi dari Keluaran OCR

Berbeda dengan OCR yang berfokus pada pengenalan karakter dan kata, ekstraksi informasi bertujuan mengidentifikasi dan menyusun entitas relevan dari teks menjadi struktur data yang bermakna. Penelitian yang dilakukan oleh [21] mengklasifikasikan pendekatan ekstraksi informasi ke dalam metode berbasis aturan dan metode statistik. Pendekatan berbasis aturan memanfaatkan pengetahuan domain, pola linguistik, serta template dokumen untuk mengenali entitas tertentu, seperti nama, tanggal, dan nomor identifikasi.

Sejumlah penelitian menunjukkan bahwa pendekatan berbasis aturan memiliki tingkat presisi yang tinggi ketika diterapkan pada dokumen dengan pola konten yang relatif dapat diprediksi [22]. Dalam lingkungan institusional, meskipun format dokumen dapat bervariasi antar lembaga, elemen informasi inti dalam sertifikat pelatihan umumnya bersifat konsisten. Kondisi ini menjadikan pendekatan berbasis aturan dan template sebagai solusi yang praktis dan dapat dikendalikan, khususnya dalam konteks organisasi dan pemerintahan.

Selain itu, Penelitian yang dilakukan oleh [23] menekankan pentingnya pascaproses OCR untuk menghubungkan teks mentah dengan konteks semantik dokumen. Integrasi antara OCR dan ekstraksi informasi berbasis domain memungkinkan sistem untuk menghasilkan data terstruktur yang lebih akurat dan dapat digunakan secara langsung dalam sistem informasi organisasi.

2.6. Penilaian Kualitas dan Tingkat Kepercayaan Ekstraksi Data

Dalam pemrosesan dokumen otomatis, kualitas hasil ekstraksi menjadi aspek penting yang memengaruhi keandalan sistem. Penilaian tingkat kepercayaan (*confidence scoring*) digunakan untuk mengukur sejauh mana hasil ekstraksi dapat dipercaya dan digunakan tanpa intervensi manusia. Berdasarkan penelitian yang dilakukan oleh [24] menunjukkan bahwa pengukuran tingkat kepercayaan pada hasil pengenalan teks dapat membantu dalam mengidentifikasi bagian dokumen yang memerlukan verifikasi manual.

Dalam konteks organisasi, khususnya sektor publik, mekanisme penilaian kepercayaan memiliki peran strategis dalam mendukung akuntabilitas dan transparansi proses administrasi. Berdasarkan penelitian yang dilakukan oleh [25] menegaskan bahwa pendekatan *human-in-the-loop* yang mengombinasikan otomatisasi dan validasi manusia dapat meningkatkan keandalan sistem berbasis data. Oleh karena itu, integrasi *confidence scoring* dalam *pipeline* ekstraksi informasi menjadi penting untuk mendukung pengambilan keputusan yang dapat dipertanggungjawabkan.

2.7. Sintesis Pustaka dan Posisi Penelitian

Berdasarkan tinjauan pustaka yang telah dibahas, dapat disimpulkan bahwa meskipun OCR dan *document image analysis* telah banyak diteliti, sebagian besar penelitian masih berfokus pada dokumen dengan struktur yang relatif standar. Transformasi keluaran OCR menjadi data terstruktur yang bermakna, khususnya pada dokumen sertifikat pelatihan dengan format heterogen, masih belum banyak dieksplorasi secara komprehensif. Selain itu, sedikit penelitian yang mengevaluasi *pipeline* pemrosesan dokumen secara *end-to-end* dalam konteks organisasi nyata.

Penelitian ini memposisikan diri untuk mengisi celah tersebut dengan mengusulkan pendekatan terintegrasi yang mengombinasikan OCR berbasis *computer vision*, praproses adaptif, ekstraksi informasi berbasis aturan dan template, serta mekanisme penilaian tingkat kepercayaan. Dengan fokus pada sertifikat pelatihan sebagai domain dokumen yang memiliki relevansi tinggi dalam manajemen sumber daya manusia, khususnya di lingkungan pemerintahan, penelitian ini diharapkan dapat memberikan kontribusi teoretis dan praktis yang signifikan. Berdasarkan posisi tersebut, bagian selanjutnya akan menguraikan metodologi penelitian yang digunakan untuk merancang dan mengevaluasi pipeline ekstraksi informasi yang diusulkan.

3. METODE PENELITIAN

3.1. Desain Penelitian

Penelitian ini menggunakan desain penelitian terapan dengan pendekatan eksperimental, yang bertujuan untuk merancang, mengimplementasikan, dan mengevaluasi sebuah metode pemrosesan dokumen otomatis dalam konteks organisasi nyata. Pendekatan ini dipilih karena penelitian tidak hanya berfokus pada pengujian konsep teoritis, tetapi juga pada pengembangan solusi yang dapat diterapkan secara langsung untuk mendukung kebutuhan operasional pengelolaan data sertifikat pelatihan. Desain penelitian terapan semacam ini umum digunakan dalam pengembangan sistem informasi dan analisis dokumen, terutama ketika tujuan utama penelitian adalah menghasilkan artefak fungsional yang dapat dievaluasi secara empiris.

Objek penelitian adalah dokumen sertifikat pelatihan pegawai yang digunakan sebagai bukti pengembangan kompetensi dalam organisasi, khususnya di lingkungan sektor publik. Penelitian ini memfokuskan diri pada permasalahan transformasi dokumen digital tidak terstruktur menjadi data terstruktur yang dapat dimanfaatkan untuk analisis dan pengambilan keputusan. Untuk mencapai tujuan tersebut, penelitian dirancang dalam bentuk *pipeline* pemrosesan *end-to-end*, yang mengintegrasikan teknik

computer vision, *Optical Character Recognition* (OCR), serta ekstraksi informasi berbasis aturan dan template.

3.2. Alur dan Tahapan Penelitian

Secara umum, alur penelitian disusun secara sistematis untuk memastikan keterulangan (*reproducibility*) dan keterlacakan (*traceability*) setiap tahapan. Penelitian dimulai dari pengumpulan data sertifikat pelatihan, dilanjutkan dengan identifikasi jenis dokumen, ekstraksi teks, ekstraksi informasi terstruktur, evaluasi kualitas hasil ekstraksi, dan diakhiri dengan pembentukan dataset terstruktur yang siap digunakan untuk analisis lanjutan.

Urutan tahapan ini dirancang berdasarkan prinsip pemrosesan dokumen yang menyatakan bahwa keberhasilan ekstraksi informasi sangat bergantung pada keterpaduan antarproses, mulai dari kualitas input hingga mekanisme validasi keluaran. Dengan demikian, setiap tahapan dalam penelitian ini saling bergantung dan membentuk satu kesatuan metodologis yang utuh.

3.3. Akuisisi Data dan Karakteristik Dataset

Data yang digunakan dalam penelitian ini berupa dokumen sertifikat pelatihan yang dikumpulkan dari beberapa institusi sektor publik. Dokumen-dokumen tersebut merupakan sertifikat autentik yang digunakan secara resmi untuk mendokumentasikan partisipasi pelatihan dan pengembangan kompetensi pegawai. Seluruh dokumen tersedia dalam format *Portable Document Format* (PDF) dan mencerminkan kondisi operasional nyata, termasuk variasi tata letak, penggunaan bahasa, serta kualitas visual dokumen.

Berdasarkan karakteristik formatnya, dataset dibedakan menjadi dua kategori utama. Kategori pertama adalah dokumen PDF digital yang memiliki lapisan teks tertanam (*embedded text layer*), sehingga memungkinkan ekstraksi teks secara langsung tanpa proses pengenalan karakter berbasis citra. Kategori kedua adalah dokumen PDF hasil pemindaian, di mana seluruh informasi tekstual direpresentasikan dalam bentuk citra raster. Pembedaan ini menjadi aspek penting dalam metodologi karena menentukan strategi ekstraksi teks yang digunakan pada tahapan berikutnya.

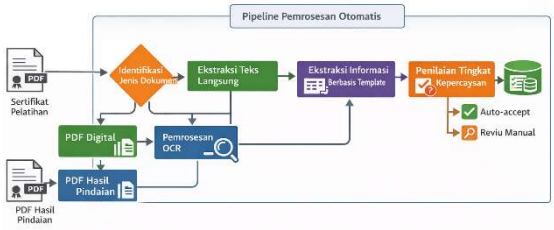
Selain perbedaan format, dataset juga menunjukkan tingkat heterogenitas yang tinggi antar institusi penerbit. Sertifikat pelatihan memiliki struktur, susunan elemen, dan terminologi yang berbeda, meskipun memuat atribut informasi inti yang relatif serupa. Kondisi ini memberikan dasar yang realistis untuk mengevaluasi kemampuan metode yang diusulkan dalam menangani variasi dokumen dunia nyata. Untuk menjaga aspek etika penelitian, seluruh dokumen dianonimkan dengan menghapus atau menyamarkan informasi identitas pribadi sebelum diproses lebih lanjut.

3.4. Arsitektur Umum Sistem Pemrosesan Dokumen

Metodologi yang diusulkan dalam penelitian ini diwujudkan dalam bentuk arsitektur sistem pemrosesan dokumen *end-to-end* yang adaptif. Arsitektur ini dirancang untuk mengakomodasi variasi karakteristik dokumen sertifikat pelatihan, baik dari sisi format, tata letak, maupun kualitas visual, dengan tetap menjaga keandalan dan efisiensi proses ekstraksi informasi. Pendekatan arsitektural ini memungkinkan integrasi berbagai teknik pemrosesan dokumen dalam satu alur yang terstruktur dan dapat direplikasi.

Secara konseptual, arsitektur sistem terdiri atas beberapa modul utama yang saling terhubung, meliputi akuisisi dokumen, identifikasi jenis dokumen, ekstraksi teks, ekstraksi informasi terstruktur, penilaian tingkat kepercayaan, dan pembentukan data keluaran. Integrasi modul-modul tersebut merepresentasikan penerapan praktis konsep *document image analysis*, OCR, dan ekstraksi informasi berbasis aturan sebagaimana dibahas dalam tinjauan pustaka.

Gambar 1 menunjukkan arsitektur sistem pemrosesan dokumen *end-to-end* yang digunakan dalam penelitian ini untuk mengekstraksi informasi terstruktur dari sertifikat pelatihan. Proses dimulai dengan masukan berupa dokumen sertifikat dalam format PDF, baik berupa dokumen digital maupun dokumen hasil pemindaian.



Gambar 1. Arsitektur sistem end-to-end untuk ekstraksi informasi otomatis dari sertifikat pelatihan

Berdasarkan Gambar 1 diatas menunjukkan arsitektur sistem pemrosesan dokumen *end-to-end* yang digunakan dalam penelitian ini untuk mengekstraksi informasi terstruktur dari sertifikat pelatihan. Proses dimulai dengan masukan berupa dokumen sertifikat dalam format PDF, baik berupa dokumen digital maupun dokumen hasil pemindaian.

Pada tahap awal, sistem melakukan identifikasi jenis dokumen untuk membedakan antara PDF digital dan PDF hasil pemindaian. Dokumen PDF digital diproses melalui modul ekstraksi teks langsung, sedangkan dokumen hasil pemindaian diproses melalui modul pengenalan karakter optik (*Optical Character Recognition*). Pembedaan jalur pemrosesan ini bertujuan untuk menjaga kualitas teks yang diekstraksi dan meminimalkan kesalahan pengenalan karakter.

Teks hasil ekstraksi kemudian diteruskan ke modul ekstraksi informasi berbasis template, di mana aturan dan pola domain digunakan untuk

mengidentifikasi atribut penting sertifikat, seperti nama peserta, judul pelatihan, nomor sertifikat, dan tanggal pelaksanaan. Selanjutnya, modul penilaian tingkat kepercayaan mengevaluasi kelengkapan hasil ekstraksi untuk menentukan apakah dokumen dapat diterima secara otomatis atau memerlukan peninjauan manual.

Keluaran akhir dari sistem berupa data sertifikat terstruktur yang siap digunakan untuk analisis lanjutan dan mendukung pengambilan keputusan dalam pengelolaan sumber daya manusia.

3.5. Strategi Ekstraksi Teks

Ekstraksi teks dilakukan menggunakan strategi adaptif yang disesuaikan dengan jenis dokumen yang telah diidentifikasi sebelumnya. Untuk dokumen PDF digital yang memiliki lapisan teks tertanam, proses ekstraksi dilakukan secara langsung dengan mengambil konten teks yang tersedia. Pendekatan ini dipilih untuk mempertahankan keakuratan karakter dan menghindari kesalahan pengenalan yang tidak perlu, sebagaimana disarankan dalam praktik pemrosesan dokumen digital.

Sebaliknya, untuk dokumen PDF hasil pemindaian, setiap halaman dokumen terlebih dahulu dikonversi menjadi representasi citra dengan resolusi tertentu guna memastikan keterbacaan teks. Tahapan praproses citra kemudian diterapkan, meliputi konversi ke skala abu-abu dan koreksi orientasi jika diperlukan. Praproses ini bertujuan untuk meningkatkan kualitas visual teks sebelum dilakukan OCR, sejalan dengan temuan bahwa praproses citra berperan signifikan dalam meningkatkan kinerja OCR [26]. Setelah praproses, OCR diterapkan untuk menghasilkan keluaran teks mentah yang selanjutnya diproses pada tahap ekstraksi informasi.

3.6. Ekstraksi Informasi Berbasis Template dan Aturan

Keluaran teks hasil ekstraksi selanjutnya diproses menggunakan pendekatan ekstraksi informasi berbasis template dan aturan (*rule-based information extraction*). Setiap dokumen sertifikat diasosiasikan dengan template yang merepresentasikan pola struktur dan konten dokumen dari institusi penerbit tertentu. Template ini memuat pengetahuan domain terkait susunan elemen, kata kunci utama, serta petunjuk kontekstual yang umum ditemukan pada sertifikat pelatihan.

Aturan ekstraksi dirancang menggunakan kombinasi *regular expressions*, pencarian kata kunci, dan heuristik kontekstual untuk mengidentifikasi atribut informasi penting, seperti nama peserta, judul pelatihan, nomor sertifikat, tanggal pelaksanaan, durasi pelatihan, dan institusi penyelenggara. Pendekatan berbasis aturan dipilih karena memberikan tingkat keterkendalian dan interpretabilitas yang tinggi, yang sangat dibutuhkan dalam pemrosesan dokumen institusional dan sektor publik [27]. Selain itu, penggunaan beberapa varian template memungkinkan sistem menangani perubahan format dokumen yang terjadi secara bertahap dalam praktik organisasi.

3.7. Penilaian Tingkat Kepercayaan dan Pengendalian Kualitas

Untuk memastikan keandalan hasil ekstraksi, penelitian ini menerapkan mekanisme penilaian tingkat kepercayaan (*confidence scoring*) pada tingkat dokumen. Skor kepercayaan dihitung berdasarkan rasio jumlah atribut yang berhasil diekstraksi terhadap jumlah atribut yang diharapkan dalam sebuah sertifikat. Pendekatan ini memberikan indikasi kuantitatif mengenai kelengkapan hasil ekstraksi.

Berdasarkan nilai skor kepercayaan, setiap dokumen diklasifikasikan ke dalam kategori hasil pemrosesan tertentu. Dokumen dengan skor tinggi dianggap berhasil diproses secara otomatis, sedangkan dokumen dengan skor menengah ditandai untuk dilakukan peninjauan manual. Dokumen dengan skor rendah diidentifikasi sebagai dokumen bermasalah yang memerlukan pemeriksaan atau pemrosesan ulang. Pendekatan ini sejalan dengan konsep *human-in-the-loop* yang menekankan keseimbangan antara otomatisasi dan validasi manusia untuk meningkatkan keandalan sistem berbasis data [28].

3.8. Strategi Evaluasi Metodologi

Evaluasi metodologi difokuskan pada kemampuan sistem dalam mengekstraksi informasi terstruktur dari dokumen sertifikat pelatihan yang heterogen. Kinerja sistem dianalisis berdasarkan tingkat kelengkapan ekstraksi setiap atribut dan skor kepercayaan pada tingkat dokumen. Penelitian ini tidak hanya menekankan akurasi pengenalan karakter secara terpisah, tetapi lebih menitikberatkan pada keandalan operasional sistem dalam menghasilkan data yang siap digunakan.

Data terstruktur yang dihasilkan selanjutnya diuji melalui integrasi dengan repositori data terpusat untuk melakukan pemeriksaan konsistensi dan deteksi duplikasi. Strategi evaluasi ini dirancang untuk mencerminkan kondisi penggunaan nyata di lingkungan organisasi, sehingga hasil penelitian dapat memberikan gambaran yang realistis mengenai manfaat dan keterbatasan metode yang diusulkan dalam mendukung pengambilan keputusan manajemen sumber daya manusia.

3.9. Ringkasan Metodologi Penelitian

Secara keseluruhan, penelitian ini mengusulkan metodologi pemrosesan dokumen terintegrasi yang mengombinasikan teknik *computer vision*, OCR, praproses adaptif, serta ekstraksi informasi berbasis aturan dan template. Metodologi ini dirancang untuk menangani heterogenitas dokumen sertifikat pelatihan dan menghasilkan data terstruktur yang dapat dimanfaatkan secara langsung dalam konteks organisasi. Dengan mengintegrasikan mekanisme penilaian tingkat kepercayaan, pendekatan yang diusulkan tidak hanya berorientasi pada otomatisasi, tetapi juga mendukung kebutuhan akuntabilitas dan

pengendalian kualitas, khususnya dalam lingkungan sektor publik.

17

4. HASIL DAN PEMBAHASAN

4.1. Penelitian Karakteristik Dataset dan Kondisi Awal Data Sertifikat

Penelitian ini menggunakan dataset yang terdiri atas 4.785 dokumen sertifikat pelatihan yang dikumpulkan pada periode September 2025 hingga Januari 2026. Dataset berasal dari berbagai instansi pemerintah dan lembaga pelatihan nasional dengan karakteristik format dan tata letak yang beragam. Kondisi ini mencerminkan situasi operasional nyata pengelolaan sertifikat pelatihan di lingkungan sektor publik.

Tabel 1. Karakteristik Dataset Penelitian

Karakteristik	Deskripsi
Jumlah dokumen	4.785 sertifikat
Periode data	Sep 2025 – Jan 2026
Jenis instansi	LAN, KPK, Kemenkeu, KKP, dll
Jenis dokumen	PDF digital dan PDF hasil pemindaian
Jumlah template	15 template
Variasi format per instansi	1–5 format

Hasil observasi awal menunjukkan bahwa meskipun dokumen sertifikat tersedia secara digital, informasi yang terkandung di dalamnya belum tersaji secara seragam akibat variasi penulisan manual dan perbedaan format antar instansi.

4.2. Ketidaksesuaian Data dan Status Ekstraksi Informasi Sertifikat

Sebelum implementasi sistem pemrosesan dokumen otomatis, seluruh berkas sertifikat pelatihan yang dikumpulkan direkap secara manual ke dalam basis data organisasi. Meskipun sebagian besar dokumen merupakan sertifikat pelatihan yang sah, hasil observasi awal menunjukkan bahwa data yang diinput tidak selalu sesuai dengan informasi yang tercantum pada sertifikat. Selain kesalahan input pada atribut utama sertifikat, ditemukan pula sejumlah kecil dokumen yang diunggah oleh pengguna namun **tidak termasuk sertifikat pelatihan**, seperti surat undangan atau dokumen kegiatan lain.

Untuk memperoleh gambaran menyeluruh mengenai kondisi awal data, dilakukan analisis terhadap kesesuaian dan ketidaksesuaian informasi sertifikat berdasarkan atribut utama, yaitu nama pelatihan, nama penyelenggara, dan nilai jam pelatihan (JP). Setiap dokumen diklasifikasikan ke dalam satu kategori berdasarkan pola kesalahan yang paling dominan. Distribusi kondisi awal data sertifikat ditunjukkan pada Tabel 2.

Tabel 2. Kondisi Kesesuaian Data pada Sertifikat Pelatihan (Observasi Awal)

Kondisi Data Sertifikat	Jumlah Dokumen	Persentase
Data sudah sesuai (nama pelatihan, penyelenggara, dan JP sesuai)	524	11%
Ketidaksesuaian tunggal (*single-field mismatch*)		
Nama pelatihan	720	15%
Nama penyelenggara	400	8%
Nilai jam pelatihan (JP)	304	6%
Ketidaksesuaian ganda (*double-field mismatch*)		
Nama pelatihan + penyelenggara	1.245	26,00%
Nama pelatihan + JP	528	11,00%
Penyelenggara + JP	337	7,00%
Ketidaksesuaian kompleks (*triple-field mismatch*)		
Nama pelatihan + penyelenggara + JP	600	12,50%
Dokumen non-sertifikat (salah unggah)	127	2,70%
Total	**4.785**	**100%**

Tabel 2 menunjukkan bahwa hanya 11% dokumen yang memiliki data sepenuhnya sesuai antara sertifikat dan hasil input manual. Sebagian besar dokumen mengalami ketidaksesuaian data, dengan pola kesalahan ganda dan kompleks sebagai kategori yang paling dominan, khususnya kombinasi kesalahan pada nama pelatihan dan nama penyelenggara. Kondisi ini mencerminkan praktik input manual yang rentan terhadap inkonsistensi terminologi dan penyederhanaan informasi. Selain itu, ditemukan pula sejumlah kecil dokumen non-sertifikat (2,7%) yang diunggah oleh pengguna, yang sejak awal tidak memenuhi kriteria sebagai riwayat pelatihan.

Setelah dilakukan pemrosesan menggunakan *pipeline* ekstraksi informasi berbasis OCR, penyeragaman data, dan penilaian tingkat kepercayaan, setiap dokumen diklasifikasikan ke dalam tiga status pemrosesan. Distribusi status pemrosesan dokumen setelah implementasi sistem ditunjukkan pada Tabel 3.

Tabel 3. Status Ekstraksi Informasi Sertifikat Setelah Implementasi Sistem

Status	Jumlah Dokumen	Persentase
Berhasil diproses otomatis	4.163	87,00%
Memerlukan validasi manual	383	8,00%
Gagal diproses (fail)	239	5,00%
Total	**4.785**	**100%**

Berdasarkan Tabel 3, dokumen yang berhasil diproses otomatis mencakup sertifikat dengan data

awal yang sesuai serta sertifikat dengan ketidaksesuaian tunggal dan sebagian ketidaksesuaian ganda yang dapat diperbaiki secara andal oleh sistem. Dokumen yang memerlukan validasi manual umumnya berasal dari kategori ketidaksesuaian ganda dan kompleks, di mana lebih dari satu atribut utama tidak sesuai sehingga sistem tidak dapat menentukan koreksi yang tepat secara otomatis. Sementara itu, dokumen yang dikategorikan sebagai gagal diproses sebagian besar merupakan dokumen non-sertifikat yang salah diunggah oleh pengguna, serta sebagian kecil dokumen sertifikat dengan kualitas visual yang sangat rendah sehingga tidak dapat diproses secara andal.

Dalam penelitian ini, penulis melakukan verifikasi terhadap hasil pemrosesan sistem melalui pemeriksaan sampel dokumen pada setiap kategori status, untuk memastikan bahwa mekanisme ekstraksi, penyeragaman, dan klasifikasi status telah berjalan sesuai dengan rancangan metodologi dan dapat dipertanggungjawabkan secara administratif.

4.3. Analisis Kinerja Berdasarkan Jenis Dokumen

Analisis kinerja sistem selanjutnya dilakukan dengan membedakan jenis dokumen berdasarkan format dan karakteristik visualnya, yaitu dokumen PDF digital dan dokumen PDF hasil pemindaian. Pembedaan ini penting karena, sebagaimana dibahas pada Bagian 4.2, kualitas dan jenis dokumen berpengaruh terhadap keberhasilan ekstraksi informasi serta kebutuhan validasi lanjutan. Dari total dataset, sebanyak 2.910 dokumen merupakan PDF digital dan 1.875 dokumen merupakan PDF hasil pemindaian. Perbandingan kinerja ekstraksi berdasarkan jenis dokumen disajikan pada Tabel 4.

Tabel 4. Kinerja Ekstraksi Berdasarkan Jenis Dokumen

Jenis Dokumen	Jumlah	Berhasil	Manual	Gagal
PDF digital	2.910	91%	7%	2%
PDF hasil pemindaian	1.875	82%	15%	**3%**

Berdasarkan Tabel 4, dokumen PDF digital menunjukkan tingkat keberhasilan ekstraksi yang lebih tinggi dibandingkan dokumen PDF hasil pemindaian. Hal ini disebabkan oleh ketersediaan lapisan teks tertanam pada PDF digital, yang memungkinkan ekstraksi teks dilakukan secara langsung tanpa melalui proses pengenalan karakter berbasis citra. Kondisi tersebut membuat sistem lebih andal dalam membaca dan menyeragamkan informasi pada dokumen PDF digital, bahkan ketika terdapat ketidaksesuaian data akibat kesalahan input manual sebagaimana dijelaskan pada Bagian 4.2.

Sebaliknya, dokumen PDF hasil pemindaian menunjukkan persentase dokumen yang memerlukan validasi manual lebih tinggi. Kondisi ini umumnya berkaitan dengan kombinasi ketidaksesuaian data dan keterbatasan kualitas visual dokumen, seperti resolusi rendah atau artefak pemindaian, yang meningkatkan ambiguitas hasil ekstraksi. Meskipun demikian, tingkat dokumen yang gagal diproses pada kategori PDF hasil pemindaian relatif rendah, yaitu sekitar 3%. Temuan ini menunjukkan bahwa strategi praproses citra dan penerapan OCR yang digunakan mampu menjaga tingkat kegagalan pada level yang dapat diterima dalam konteks operasional.

Pada dokumen PDF digital, sebagian kecil dokumen yang gagal diproses umumnya bukan disebabkan oleh kesalahan ekstraksi, melainkan oleh keberadaan dokumen non-sertifikat atau berkas yang tidak merepresentasikan riwayat pelatihan, sebagaimana telah diidentifikasi pada analisis kondisi awal data. Dengan demikian, hasil analisis ini menegaskan bahwa kegagalan pemrosesan tidak hanya dipengaruhi oleh aspek teknis OCR, tetapi juga oleh kesesuaian semantik dokumen yang diunggah oleh pengguna.

4.4. Penyeragaman Data dan Kualitas Output Terstruktur

Salah satu hasil utama penelitian ini adalah kemampuan sistem dalam menyeragamkan representasi data sertifikat pelatihan yang sebelumnya tidak konsisten akibat kesalahan input manual dan variasi penulisan antar dokumen. Penyeragaman difokuskan pada atribut utama sertifikat, yaitu nama pelatihan, nama penyelenggara, dan nilai jam pelatihan (JP), yang berdasarkan analisis pada Bagian 4.2 merupakan sumber utama ketidaksesuaian data.

Penyeragaman dilakukan melalui pemetaan istilah, normalisasi penamaan, serta validasi nilai numerik JP berdasarkan informasi yang tercantum pada sertifikat. Tabel 5 memberikan contoh perbandingan data sertifikat sebelum dan sesudah diproses oleh sistem.

Tabel 5. Contoh Penyeragaman Data Sertifikat Sebelum dan Sesudah Pemrosesan

Atribut	Sebelum	Setelah
Nama Pelatihan	AI Fluency By Microsoft (7 JP)	AI Fluency By Microsoft
Penyelenggara	LAN RI / Lembaga Administrasi Negara	Lembaga Administrasi Negara
Nama Peserta	IRFAN RICKY AFANDI / Irfan Ricky Afandi	Irfan Ricky Afandi
Jam Pelatihan (JP)	8	7

Contoh pada Tabel 5 menunjukkan bahwa sistem tidak hanya menghilangkan variasi penulisan, tetapi juga mampu mengoreksi ketidaksesuaian nilai JP berdasarkan informasi pada sertifikat. Hasil penyeragaman ini menghasilkan data sertifikat yang konsisten dan terstruktur, sehingga dapat digunakan

secara langsung untuk analisis lanjutan tanpa memerlukan proses pembersihan manual tambahan.

4.5. Evaluasi Tingkat Kepercayaan dan Pengendalian Kualitas

Untuk menjamin keandalan hasil ekstraksi, sistem menerapkan mekanisme penilaian tingkat kepercayaan (*confidence scoring*) pada tingkat dokumen. Skor kepercayaan dihitung berdasarkan kelengkapan dan konsistensi atribut utama sertifikat yang berhasil diekstraksi dan diseragamkan, khususnya nama pelatihan, nama penyelenggara, dan nilai jam pelatihan (JP), dimana ditampilkan pada Tabel 6. Skor ini digunakan sebagai dasar pengambilan keputusan apakah hasil ekstraksi dapat diterima secara otomatis atau memerlukan peninjauan manual.

Tabel 6. Interpretasi Skor Kepercayaan Ekstraksi

Rentang Skor Kepercayaan	Kategori	Tindakan
≥ 0,90	Tinggi	Diterima otomatis
0,70 – 0,89	Sedang	Validasi manual
< 0,70	Rendah	Ditandai sebagai gagal

Berdasarkan Tabel 6 diatas, dokumen dengan skor kepercayaan tinggi umumnya berasal dari sertifikat dengan data awal yang sesuai atau hanya mengalami ketidaksesuaian tunggal yang dapat diperbaiki secara otomatis. Dokumen dengan skor kepercayaan sedang umumnya memiliki ketidaksesuaian ganda atau kompleks, sehingga memerlukan validasi manual untuk menghindari kesalahan interpretasi. Sementara itu, dokumen dengan skor kepercayaan rendah mencakup dokumen non-sertifikat atau dokumen dengan kualitas informasi yang tidak memadai untuk diproses secara andal.

Penerapan mekanisme ini memungkinkan sistem untuk menjaga keseimbangan antara otomatisasi dan validasi manual (*human-in-the-loop*), sekaligus mendukung akuntabilitas dan kualitas data dalam pengelolaan riwayat pelatihan di lingkungan sektor publik.

4.6. Pembahasan Umum dan Implikasi Praktis

Hasil penelitian menunjukkan bahwa pendekatan pemrosesan dokumen terintegrasi yang diusulkan mampu menangani heterogenitas sertifikat pelatihan secara efektif, tidak hanya dari sisi format dokumen, tetapi juga dari sisi kualitas dan konsistensi data yang diinput secara manual. Tingkat keberhasilan pemrosesan otomatis yang tinggi menunjukkan bahwa sistem mampu membaca, mengekstraksi, dan menyeragamkan data sertifikat meskipun terdapat berbagai bentuk ketidaksesuaian, baik kesalahan tunggal maupun kombinasi kesalahan pada atribut utama sertifikat.

Temuan penting dari penelitian ini adalah bahwa kegagalan pemrosesan dokumen tidak semata-mata disebabkan oleh keterbatasan teknis OCR, melainkan juga oleh kesesuaian semantik dokumen yang diunggah oleh pengguna. Keberadaan dokumen non-sertifikat dalam dataset menegaskan pentingnya mekanisme identifikasi dan validasi dokumen sejak tahap awal pemrosesan. Dengan demikian, sistem tidak hanya berfungsi sebagai alat ekstraksi teks, tetapi juga sebagai mekanisme penyaring (*filtering*) dokumen yang relevan sebagai riwayat pelatihan.

Dari sisi operasional, integrasi mekanisme penilaian tingkat kepercayaan dan validasi manual (*human-in-the-loop*) memungkinkan organisasi untuk mengelola risiko kesalahan data secara terkontrol. Dokumen dengan ketidaksesuaian ganda dan kompleks dapat ditandai untuk peninjauan lanjutan, sementara dokumen dengan data yang telah sesuai atau berhasil diperbaiki secara otomatis dapat langsung digunakan. Pendekatan ini mendukung keseimbangan antara efisiensi otomatisasi dan akuntabilitas administrasi.

Implikasi praktis dari penelitian ini terlihat pada kemampuan instansi pemerintah untuk melakukan rekapitulasi pelatihan, analisis jam pelatihan, serta pemetaan kompetensi pegawai secara lebih efisien dan andal. Dengan tersedianya data sertifikat yang terstruktur dan tervalidasi, proses pengambilan keputusan terkait pengembangan sumber daya manusia dapat dilakukan secara lebih berbasis data dan dapat dipertanggungjawabkan.

5. KESIMPULAN DAN SARAN

sumber daya manusia di instansi pemerintah.

Penelitian ini menyimpulkan bahwa metode pemrosesan dokumen end-to-end berbasis *computer vision*, *Optical Character Recognition*, serta ekstraksi informasi berbasis template dan aturan mampu mengatasi permasalahan tidak teraturan dan ketidaksesuaian data sertifikat pelatihan dalam lingkungan instansi pemerintah. Berdasarkan pengujian terhadap 4.785 dokumen dengan variasi format, kualitas visual, dan tingkat kesesuaian data yang beragam, sistem menunjukkan tingkat keberhasilan pemrosesan otomatis yang dominan, dengan sebagian dokumen memerlukan validasi manual akibat kombinasi kesalahan input pada nama pelatihan, nama penyelenggara, dan nilai jam pelatihan, serta sebagian kecil lainnya diklasifikasikan sebagai dokumen non-sertifikat. Integrasi strategi ekstraksi teks adaptif, penyeragaman data, dan mekanisme penilaian tingkat kepercayaan memungkinkan sistem tidak hanya menghasilkan data sertifikat yang terstruktur dan konsisten, tetapi juga mendukung proses verifikasi riwayat pelatihan secara terkontrol dan akuntabel. Dari sisi praktis, pendekatan ini berpotensi meningkatkan efisiensi administrasi pelatihan, memperkuat kualitas data riwayat pelatihan, serta mendukung pengambilan keputusan pengembangan sumber daya manusia berbasis data di

sektor publik, sementara penelitian selanjutnya dapat diarahkan pada integrasi pendekatan pembelajaran mesin untuk meningkatkan fleksibilitas ekstraksi dan memperluas cakupan penerapan pada dokumen administratif lainnya.

DAFTAR PUSTAKA

[1] F. Muchtar, Z. Zamrudi, and S. Shaddiq, "Transformasi Evaluasi Kinerja Karyawan Berbasis Electronic Human Resource Management (E-HRM) dalam Meningkatkan Efektivitas Manajemen Sumber Daya Manusia pada Era Industry 4.0," *Jejak Digit. J. Ilm. Multidisiplin*, vol. 1, no. 6, pp. 4599–4611, 2025, doi: https://doi.org/10.63822/cc69bn29.

[2] I. R. Hawiyanti and M. Natsir, "Strategi Pengembangan Kompetensi Pegawai Negeri Sipil Di Lingkungan Pemerintah Kota Probolinggo," *J. Inov. Sekt. Publik*, vol. 4, no. 1, p. 3, 2024, doi: https://doi.org/10.38156/jisp.v4i1.222.

[3] R. A. Lubis and M. I. P. Nasution, "Penerapan Upaya Pengolahan Kualitas Data Untuk Meningkatkan Informasi yang Konsisten," *Socius J. Penelit. Ilmu-Ilmu Sos.*, vol. 2, no. 12, p. 207, 2025, doi: https://doi.org/10.5281/zenodo.15641883.

[4] A. N. Qolby and Y. Taufik, "Analisis Implementasi Arsip Digital Dokumen Dinas Penanaman Modal dan Pelayanan Terpadu Satu Pintu Kota Bandung," *Apl. Adm. Media Anal. Masal. Adm.*, vol. 28, no. 1, p. 60, 2025, doi: https://doi.org/10.30649/aamama.v28i1.278.

[5] M. N. R. Tulung, I. Pangkey, and A. R. Dilapanga, "Asesmen Kebutuhan Program Pelatihan Berbasis Kompetensi Di UPTD Balai Pelatihan Tenaga Kerja Kelas A Provinsi Sulawesi Utara," *J. Innov. Res. Knowl.*, vol. 4, no. 7, p. 4880, 2024, doi: https://doi.org/10.53625/jirk.v4i7.9157.

[6] C. Cahyaningtyas, Mira, C. Gudiato, M. Sari, and N. P, *Computer Vision untuk Pemula: Deteksi dan Analisis Ekspresi Wajah dengan CNN*, 1st ed. Kabupaten Ponorogo: Uwais Inspirasi Indonesia, 2025.

[7] Y. Darmi, M. F. Sepriansyah, Y. Darnita, and P. Pahrizal, "Penerapan Metode Optical Character Recognition (OCR) Untuk Mengidentifikasi Teks Pada Identitas Dokumen Surat Izin Mengemudi (SIM)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 4, p. 5992, 2025, doi: https://doi.org/10.36040/jati.v9i4.13987.

[8] Y. Reswan, R. Raffles, A. Wijaya, and Y. Apridiansyah, "Penerapan Algoritma OCR Untuk Ekstraksi Informasi Dari Citra Kartu Tanda Mahasiswa (KTM)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 5, p. 11005, 2024, doi: https://doi.org/10.36040/jati.v8i5.11006.

[9] G. R. Lesmana and A. Januantoro, "Pengembangan Sistem Informasi Manajemen Kearsipan Berbasis OCR Untuk Melakukan Proses Pencarian Data," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 2, p. 2203, 2025, doi: https://doi.org/10.36040/jati.v9i2.13000.

[10] A. Purdianto, *Transformasi Administrasi Publik Melalui Digitalisasi Dan Kecerdasan Buatan*, 1st ed. Indramayu: PT. Adab Indonesia, 2025.

[11] Fedhira and C. Prianto, "Systematic Literature Review: Analysis of AI Implementation for Document Verification," *J. Ilm. Sist. Inf.*, vol. 4, no. 3, p. 418, 2025, doi: https://doi.org/10.51903/kjjwk708.

[12] Y. Yusman, R. R. Putra, and I. Sinaga, *Transformasi Administrasi Di Era Digital*, 1st ed. Payakumbuh: PT. Serasi Media Teknologi, 2024.

[13] D. Prastyo, D. Irawan, and I. H. Mursyidin, "Sistem Informasi Terpusat untuk Manajemen Dokumen, Penelitian, dan Pengabdian kepada Masyarakat," *Bit-Tech*, vol. 7, no. 3, p. 759, 2025, doi: https://doi.org/10.32877/bt.v7i3.2182.

[14] K. Banu, D. Andreas, W. Anggoro, and A. Setiawan, "OCR: Masa Depan Pengenalan Karakter Optik dan Dampaknya pada Kehidupan Modern," *J. Teknol. Inf.*, vol. 9, no. 2, p. 149, 2023, doi: https://doi.org/10.52643/jti.v9i2.3798.

[15] W. Juniarrochman and R. Selamet, "Information Retrieval Dan Database Systems: Systematic Literature Review Tentang Analisis Komparatif Dan Strategi Integrasi," *NARATIF J. Nas. Ris. Apl. dan Tek. Inform.*, vol. 7, no. 2, p. 112, 2025, doi: https://doi.org/10.53580/naratif.v7i2.338.

[16] Wahyuddin and A. Hasim, "Aplikasi Ekstraksi Data Kartu Vaksin Berbasis Web Menggunakan Metode Ocr," *J. Sintaks Log.*, vol. 3, no. 2, p. 1, 2023, doi: https://doi.org/10.31850/jsilog.v3i2.2525.

[17] M. R. Toha and A. Triayudi, "Penerapan Membaca Tulisan di dalam Gambar Menggunakan Metode OCR Berbasis Website pada e-KTP," *J. Sains dan Teknol.*, vol. 11, no. 1, p. 176, 2022, doi: https://doi.org/10.23887/jstundiksha.v11i1.42279.

[18] R. Surya, "Peningkatan Akurasi OCR dalam Pemrosesan Formulir Keuangan melalui Fine-Tuning Transformer dan Strategi Pra-pemrosesan Data," *J. Inov. Inform.*, vol. 7, no. 2, p. 9, 2025, doi: https://doi.org/10.51170/jii.v7i2.85.

[19] Holila, A. R. Pratama, S. A. P. Lestari, and J. Indra, "Introduction National Identification Number And Name On Id Card Using OCR (Optical Character Recognition) Method," *Jutif*, vol. 5, no. 4, p. 1192, 2024, doi: https://doi.org/10.52436/1.jutif.2024.5.4.2242 .

[20] A. T. P. D. Akhsa, M. I. Burhan, and A. Munandar, "Integrasi OCR dan TF-IDF untuk Metadata Otomatis pada Pencarian Dokumen Digital," *J. FASILKOM (teknologi Inf. dan ILmu KOMputer)*, vol. 15, no. 2, p. 305, 2025, doi: https://doi.org/10.37859/jf.v15i2.9918.

[21] W. P. Putri, I. Munadhif, M. K. Hasin, N. Rinanto, and R. Y. Adhitya, "Implementasi OCR Dan Cnn Untuk Klasifikasi Penempatan Obat Berdasarkan Kelas Terapi Di Apotek," *J. Elkolind*, vol. 11, no. 2, p. 422, 2024, doi: https://doi.org/10.33795/elkolind.v11i2.5341.

[22] M. J. Averroes, D. E. Ratnawati, and R. S. Perdana, "Implementasi OCR Menggunakan Asprise dan Metode LSTM untuk Pengkategorian Item pada Setruk Belanja," *JUSTSI*, vol. 6, no. 2, p. 125, 2025, doi: https://doi.org/10.25126/kz7b7488.

[23] M. R. Reyvansyah, "Penerapan Metode Optical Character Recognition (OCR) Untuk Mengambil Data Arsip," *J. Tek. ELEKTRO DAN Komput. TRIAC*, vol. 10, no. 2, p. 44, 2023, doi: https://doi.org/10.21107/triac.v10i2.20809.

[24] L. P. D. Satriani and I. N. T. A. Putra, "Perancangan Dan Pemodelan Sistem Kepatuhan Cerdas UMKM Berbasis Generative AI-OCR Dengan Pendekatan Design Thinking Dan UCD," *J. Inform. Dan Tek. Elektro Terap.*, vol. 13, no. 1, p. 1986, 2025, doi: https://doi.org/10.23960/jitet.v13i3S1.8160.

[25] A. S. Firin and Sriani, "Analisis Algoritma Canny Edge Detection dengan Tesseract OCR untuk Mendeteksi Pelat Nomor Kendaraan Bermotor," *J. Ilm. Mat. Sains dan Teknol.*, vol. 13, no. 3, p. 382, 2025, doi: https://doi.org/10.37905/euler.v13i3.34868.

[26] S. Maesaroh and I. Alannurrochman, "Sistem Digitalisasi Data Pasien Klinik Bhakti Medika Dengan Image Processing Ocr (Optical Character Recognition) Menggunakan Tesseract," *J. Ilm. Sains dan Teknol. /*, vol. 9, no. 2, p. 317, 2025, doi: https://doi.org/10.5555/a8pg8a43.

[27] A. R. Azhari and M. I. P. Nasution, "Integrasi Data Ekonomi Antar Lembaga Untuk Meningkatkan Efisiensi Kebijakan Publik," *J. Ilm. Ekon. Dan Manaj.*, vol. 3, no. 6, p. 122, 2025, doi: https://doi.org/10.61722/jiem.v3i6.4996.

[28] M. H. Ilhamsyah, *Safety First 4.0: Membangun Budaya K3 di Era Digitalisasi Industri*, 1st ed. Makasar: PT. Nas Media Indonesia, 2025.