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] Komisi Pemilihan Umum Republik Indonesia
Jl. Cideng Timur No.54 DKI Jakarta, Indonesia
[3] Program Studi Ilmu Komunikasi, Universitas Esa Unggul Kampus Jakarta Barat
Jl. Arjuna Utara No.9 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* mendorong organisasi untuk mengelola data pengembangan sumber daya manusia secara lebih sistematis dan akuntabel [1]. Program pelatihan dan pengembangan profesional berperan strategis dalam peningkatan kompetensi pegawai, baik di sektor publik maupun privat, sehingga sertifikat pelatihan menjadi bukti formal pencapaian kompetensi serta dasar penting dalam pemetaan kompetensi, pengembangan karier, dan evaluasi kinerja kelembagaan [2]. Oleh karena itu, ketersediaan data sertifikat pelatihan yang akurat, terstruktur, dan mudah dianalisis menjadi kebutuhan mendasar bagi pengambilan keputusan organisasi berbasis data [3].

Seiring dengan digitalisasi arsip, sertifikat pelatihan umumnya disimpan dalam bentuk dokumen hasil pemindaian atau berkas *Portable Document Format* (PDF). Meskipun mempermudah penyimpanan dan distribusi, informasi penting di dalam sertifikat tersebut masih tersaji dalam format tidak terstruktur, sehingga proses konsolidasi data sering kali dilakukan secara manual. Berbagai penelitian menunjukkan bahwa pemrosesan manual dokumen memerlukan waktu dan sumber daya yang besar serta rentan terhadap kesalahan dan inkonsistensi data, terutama pada organisasi dengan volume dokumen yang tinggi dan tersebar lintas unit kerja [4].

Dalam konteks manajerial, khususnya di lingkungan pemerintahan, keterbatasan data sertifikat yang terstruktur menimbulkan tantangan signifikan dalam proses monitoring dan evaluasi riwayat pelatihan pegawai. Pimpinan membutuhkan informasi teragregasi dan tervalidasi untuk menilai partisipasi pelatihan, mengidentifikasi kesenjangan kompetensi, serta merancang program pengembangan sumber daya manusia yang lebih tepat sasaran [5]. Permasalahan utama yang dihadapi adalah belum tersedianya mekanisme otomatis yang andal untuk mengekstraksi dan menstrukturkan informasi utama dari sertifikat pelatihan dengan format yang heterogen, sehingga pemanfaatan arsip sertifikat digital sebagai sumber data strategis belum dapat dioptimalkan secara maksimal.

Berbagai pendekatan berbasis *document image analysis* dan *Optical Character Recognition* (OCR) telah digunakan untuk mengekstraksi teks dari dokumen digital. Namun, sebagian besar penelitian masih berfokus pada dokumen dengan tata letak relatif seragam atau menitikberatkan pada akurasi pengenalan karakter semata, sementara sertifikat

pelatihan memiliki variasi format, terminologi, dan kualitas visual yang tinggi. Literatur menunjukkan bahwa ekstraksi informasi yang andal tidak hanya bergantung pada OCR, tetapi juga memerlukan tahapan pascaproses yang mempertimbangkan konteks domain dan pola dokumen untuk menghasilkan data yang bermakna secara semantik [6].

Sebagai upaya penyelesaian permasalahan tersebut, penelitian ini bertujuan untuk mengembangkan dan mengevaluasi sebuah *end-to-end pipeline* ekstraksi informasi otomatis dari sertifikat pelatihan berbasis OCR yang dikombinasikan dengan strategi ekstraksi informasi berbasis aturan. Pendekatan ini dirancang untuk menghasilkan data sertifikat yang terstruktur dari berbagai format dan template dokumen, sehingga dapat mendukung proses verifikasi, pemetaan kompetensi, serta pengambilan keputusan pengembangan sumber daya manusia secara lebih efisien dan akuntabel.

2. TINJAUAN PUSTAKA

2.1. Digitalisasi Dokumen dan Pengelolaan Data Tidak Terstruktur

Digitalisasi dokumen merupakan bagian dari upaya organisasi dalam meningkatkan efisiensi pengelolaan informasi dan mendukung pengambilan keputusan berbasis data. Dalam konteks pengelolaan dokumen, digitalisasi tidak hanya mencakup konversi dokumen fisik ke bentuk digital, tetapi juga kemampuan sistem untuk memanfaatkan informasi yang terkandung di dalam dokumen tersebut secara optimal [7]. Namun, pada banyak organisasi, termasuk institusi pemerintahan, dokumen digital seperti sertifikat pelatihan masih tersimpan dalam bentuk data tidak terstruktur, sehingga informasinya belum dapat langsung diolah oleh sistem informasi.

Penelitian di bidang *document management systems* menunjukkan bahwa keberadaan data tidak terstruktur menjadi kendala dalam integrasi data dan analisis lanjutan, khususnya ketika volume dokumen tinggi dan format dokumen bervariasi antar unit atau lembaga penerbit [8]. Kondisi ini menegaskan perlunya pendekatan pemrosesan dokumen yang mampu mentransformasikan arsip digital menjadi data terstruktur yang bernilai guna bagi organisasi.

2.2. Document Image Analysis dalam Konteks Computer Vision

Document image analysis merupakan pendekatan dalam *computer vision* yang berfokus pada pemahaman struktur dan konten dokumen berbasis citra. Pendekatan ini menekankan pentingnya analisis tata letak dan struktur visual dokumen sebagai dasar untuk mengekstraksi informasi yang relevan [9]. Dalam praktiknya, dokumen dunia nyata sering kali memiliki variasi tata letak, kualitas pemindaian, dan desain visual yang tinggi.

Penelitian oleh [10] menunjukkan bahwa dokumen institusional umumnya bersifat heterogen dan tidak mengikuti struktur yang seragam. Temuan ini relevan dengan karakteristik sertifikat pelatihan yang diterbitkan oleh berbagai instansi dengan format berbeda, meskipun memuat elemen informasi inti yang serupa. Oleh karena itu, *document image analysis* menyediakan landasan konseptual yang penting bagi pengembangan sistem pemrosesan sertifikat pelatihan berbasis *computer vision*.

2.3. Optical Character Recognition dalam Pemrosesan Dokumen

Optical Character Recognition (OCR) berperan sebagai komponen utama dalam mengonversi teks visual pada dokumen citra menjadi teks yang dapat dibaca oleh mesin. Kinerja OCR dipengaruhi oleh kualitas dokumen, jenis huruf, serta kompleksitas tata letak [11]. Meskipun teknologi OCR telah banyak diterapkan pada dokumen administratif, dokumen dengan format tidak seragam sering kali menghasilkan keluaran OCR yang mengandung kesalahan dan inkonsistensi.

Penelitian oleh [12] menegaskan bahwa hasil OCR perlu diperlakukan sebagai teks mentah yang masih memerlukan pemrosesan lanjutan. Dengan demikian, OCR diposisikan sebagai tahapan awal dalam pipeline pemrosesan dokumen, bukan sebagai solusi akhir untuk menghasilkan data terstruktur yang siap digunakan.

2.4. Praproses Citra untuk Peningkatan Kinerja OCR

Praproses citra merupakan tahapan pendukung dalam *pipeline* OCR yang bertujuan meningkatkan kualitas visual dokumen sebelum proses pengenalan karakter dilakukan. Penelitian menunjukkan bahwa strategi praproses yang disesuaikan dengan karakteristik dokumen dapat meningkatkan keterbacaan teks dan mengurangi kesalahan pengenalan karakter [13].

Dalam konteks sertifikat pelatihan, praproses citra menjadi relevan karena dokumen dapat berasal dari berkas PDF digital maupun hasil pemindaian fisik dengan kualitas visual yang beragam. Penelitian oleh [14] menekankan bahwa pendekatan praproses adaptif lebih efektif dibandingkan penerapan teknik praproses yang seragam. Oleh karena itu, praproses citra berperan sebagai faktor pendukung penting dalam meningkatkan kualitas keluaran OCR dan tahapan ekstraksi informasi selanjutnya.

2.5. Ekstraksi Informasi dari Keluaran OCR

Ekstraksi informasi bertujuan mengidentifikasi dan menyusun elemen data penting dari teks mentah hasil OCR ke dalam struktur data yang bermakna. Pendekatan berbasis aturan memanfaatkan pengetahuan domain, pola konten, dan karakteristik dokumen untuk mengenali entitas informasi tertentu secara presisi [15]. Pendekatan ini dinilai efektif pada dokumen dengan elemen informasi inti yang relatif konsisten meskipun format visualnya bervariasi [16].

Dalam konteks sertifikat pelatihan, elemen seperti nama peserta, judul pelatihan, tanggal pelaksanaan, dan nomor sertifikat umumnya memiliki pola tertentu yang dapat dimanfaatkan dalam proses ekstraksi. Penelitian yang dilakukan oleh [17] menegaskan bahwa integrasi antara OCR dan pascaproses berbasis konteks domain memungkinkan transformasi teks mentah menjadi data terstruktur yang lebih akurat dan siap digunakan dalam sistem informasi organisasi.

2.6. Penilaian Kualitas dan Tingkat Kepercayaan Ekstraksi Data

Dalam pemrosesan dokumen otomatis, penilaian kualitas hasil ekstraksi menjadi aspek penting untuk memastikan keandalan sistem. Mekanisme penilaian tingkat kepercayaan (*confidence scoring*) digunakan untuk mengukur sejauh mana hasil ekstraksi dapat digunakan tanpa intervensi manual [18]. Pendekatan ini membantu mengidentifikasi bagian dokumen yang memerlukan verifikasi lanjutan.

Dalam lingkungan organisasi, khususnya sektor publik, penilaian tingkat kepercayaan mendukung prinsip akuntabilitas dan transparansi pengelolaan data Berdasarkan penelitian yang dilakukan oleh [19] menunjukkan bahwa integrasi antara otomatisasi dan validasi manusia dapat meningkatkan keandalan sistem berbasis data. Oleh karena itu, *confidence scoring* menjadi komponen penting dalam pipeline ekstraksi informasi sertifikat pelatihan.

2.7. Sintesis Pustaka dan Posisi Penelitian

Berdasarkan tinjauan pustaka yang telah dibahas, dapat disimpulkan bahwa penelitian terkait OCR dan *document image analysis* masih banyak berfokus pada dokumen dengan struktur relatif standar atau pada peningkatan akurasi pengenalan karakter semata. Transformasi keluaran OCR menjadi data terstruktur yang bermakna pada dokumen sertifikat pelatihan dengan format heterogen, khususnya melalui evaluasi *pipeline* pemrosesan dokumen secara *end-to-end*, masih relatif terbatas.

Penelitian ini memposisikan diri untuk mengisi celah tersebut dengan mengusulkan pendekatan terintegrasi yang mengombinasikan OCR berbasis *computer vision*, praproses citra adaptif, ekstraksi informasi berbasis aturan dan template, serta penilaian tingkat kepercayaan dalam satu *pipeline* terpadu. Dengan fokus pada sertifikat pelatihan sebagai domain dokumen yang relevan dalam manajemen sumber daya manusia, penelitian ini diharapkan dapat memberikan kontribusi teoretis dan praktis yang signifikan.

3. METODE PENELITIAN

3.1. Jenis dan Desain Penelitian

Penelitian ini merupakan penelitian terapan dengan pendekatan eksperimental yang bertujuan merancang, mengimplementasikan, dan mengevaluasi metode ekstraksi informasi otomatis dari sertifikat pelatihan. Pendekatan ini dipilih karena penelitian difokuskan pada pengembangan solusi teknis yang dapat diterapkan secara langsung untuk mengubah dokumen sertifikat pelatihan tidak terstruktur menjadi data terstruktur yang siap digunakan dalam konteks organisasi.

Objek penelitian berupa dokumen sertifikat pelatihan pegawai yang digunakan dalam proses pengembangan kompetensi, khususnya di lingkungan sektor publik. Metodologi penelitian dirancang dalam bentuk pipeline pemrosesan dokumen end-to-end yang mengintegrasikan teknik computer vision, Optical Character Recognition (OCR), dan ekstraksi informasi berbasis aturan, sebagaimana telah dibahas pada tinjauan pustaka.

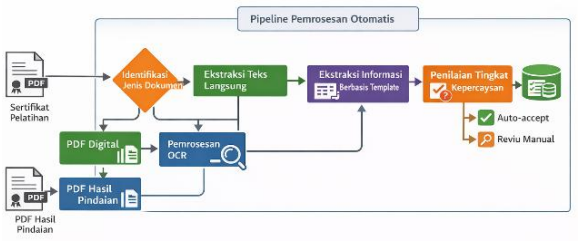
3.2. Data Penelitian dan Karakteristik Dokumen

Data penelitian terdiri atas kumpulan dokumen sertifikat pelatihan autentik yang diperoleh dari beberapa institusi sektor publik. Seluruh dokumen tersedia dalam format *Portable Document Format* (PDF) dan mencerminkan kondisi operasional nyata, termasuk variasi tata letak, terminologi, serta kualitas visual.

Berdasarkan karakteristiknya, dokumen diklasifikasikan menjadi dua jenis, yaitu PDF digital yang memiliki lapisan teks tertanam dan PDF hasil pemindaian yang merepresentasikan teks dalam bentuk citra. Pembedaan ini menjadi dasar dalam menentukan strategi ekstraksi teks pada tahapan selanjutnya. Untuk menjaga aspek etika penelitian, seluruh dokumen dianonimkan dengan menyamarkan informasi identitas pribadi sebelum diproses.

3.3. Dataset Arsitektur Sistem dan Alur Pemrosesan

Metodologi penelitian diwujudkan dalam bentuk arsitektur sistem pemrosesan dokumen end-to-end yang dirancang untuk menangani heterogenitas sertifikat pelatihan. Arsitektur ini terdiri atas beberapa modul utama, meliputi identifikasi jenis dokumen, ekstraksi teks, ekstraksi informasi terstruktur, penilaian tingkat kepercayaan, dan pembentukan data keluaran.



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

Gambar 1 menunjukkan arsitektur sistem pemrosesan dokumen yang digunakan dalam penelitian ini. Proses dimulai dengan masukan berupa dokumen sertifikat dalam format PDF. Sistem terlebih dahulu melakukan identifikasi jenis dokumen untuk membedakan antara PDF digital dan PDF hasil

pemindaian. Dokumen PDF digital diproses melalui ekstraksi teks langsung, sedangkan dokumen hasil pemindaian diproses melalui OCR. Keluaran teks dari kedua jalur tersebut kemudian diproses lebih lanjut untuk menghasilkan data sertifikat terstruktur.

3.4. Strategi Ekstraksi Teks

Ekstraksi teks dilakukan secara adaptif berdasarkan jenis dokumen. Pada dokumen PDF digital, teks diekstraksi langsung dari lapisan teks yang tersedia untuk menjaga keakuratan karakter. Sementara itu, pada dokumen PDF hasil pemindaian, setiap halaman dikonversi menjadi citra dan melalui tahapan praproses citra untuk meningkatkan keterbacaan teks sebelum dilakukan OCR.

Praproses citra mencakup penyesuaian orientasi dan peningkatan kualitas visual teks, sebagaimana disarankan dalam penelitian sebelumnya yang menunjukkan bahwa praproses berperan penting dalam meningkatkan kinerja OCR [20]. Keluaran OCR berupa teks mentah selanjutnya menjadi masukan bagi proses ekstraksi informasi.

3.5. Ekstraksi Informasi dan Penyeragaman Data

Ekstraksi informasi dilakukan menggunakan pendekatan berbasis template dan aturan yang dirancang sesuai dengan karakteristik sertifikat pelatihan. Setiap template merepresentasikan pola struktur dan konten dokumen dari institusi penerbit tertentu. Aturan ekstraksi memanfaatkan kombinasi pencarian kata kunci, pola teks, dan konteks posisi untuk mengidentifikasi atribut informasi utama, 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 penting dalam pemrosesan dokumen institusional dan sektor publik [21]. Hasil ekstraksi kemudian diseragamkan ke dalam skema data terstruktur yang konsisten untuk seluruh dokumen.

3.6. Penilaian Tingkat Kepercayaan dan Evaluasi Hasil

Penelitian ini menerapkan mekanisme penilaian tingkat kepercayaan pada tingkat dokumen. Skor kepercayaan dihitung berdasarkan tingkat kelengkapan atribut informasi yang berhasil diekstraksi. Berdasarkan skor tersebut, dokumen diklasifikasikan ke dalam kategori hasil yang dapat diterima secara otomatis atau memerlukan peninjauan manual.

Pendekatan ini mendukung prinsip *human-in-the-loop* dengan mengombinasikan otomatisasi dan validasi manusia untuk meningkatkan kualitas data yang dihasilkan [22]. Evaluasi metodologi difokuskan pada kemampuan sistem dalam menghasilkan data sertifikat terstruktur yang konsisten dan siap digunakan dalam lingkungan organisasi nyata.

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 dari Tabel 1 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

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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