Adaptif Learning Berbasis Human-in-the-Loop untuk Ekstraksi Data PDF Template

PROPOSAL TESIS

Disusun Oleh :

Moh Syaiful Rahman 247064518006

UNIVERSITAS NASIONAL

Jl. Sawo Manila No.61, RT.14/RW.7, Pejaten Barat, Ps. Minggu,

Kota Jakarta Selatan,

Daerah Khusus Jakarta 12520

DAFTAR ISI

Daftar Tabel

Daftar Gambar

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Format PDF (Portable Document Format) telah menjadi standar de facto dalam pertukaran dan penyimpanan dokumen digital di berbagai sektor, termasuk pemerintahan, bisnis, dan pendidikan. Keunggulan utama format ini terletak pada kemampuannya mempertahankan tampilan dan struktur dokumen secara konsisten lintas platform, menjadikannya pilihan utama untuk distribusi informasi digital (International Organization for Standardization, 2008).

Namun, meskipun ideal untuk keterbacaan manusia, format PDF menimbulkan tantangan signifikan dalam pemrosesan otomatis karena struktur internalnya yang kompleks. Hal ini khususnya berlaku untuk ekstraksi informasi dari PDF template, dimana struktur dokumen mengikuti template tertentu namun konten dapat bervariasi (Abiteboul, 1997). Tantangan ini umumnya ditemukan pada berbagai jenis dokumen seperti formulir administratif, dokumen perizinan, faktur, dan invoice (Dengel & Klein, 2002; Schuster et al., 2013). Proses ekstraksi data dari jenis dokumen ini umumnya masih dilakukan secara manual atau semi-otomatis, yang dapat berdampak pada efisiensi operasional dan akurasi data (Xu et al., 2020).

Pendekatan ekstraksi data tradisional menghadapi keterbatasan fundamental. Pendekatan berbasis aturan (rule-based) efektif untuk struktur konsisten namun rentan terhadap perubahan minor pada layout, sehingga memerlukan rekonfigurasi manual yang memakan waktu (Klein et al., 2004). Di sisi lain, pendekatan berbasis machine learning murni (terutama deep learning) memerlukan dataset berlabel yang sangat besar dan menghadapi tantangan dalam adaptabilitas real-time serta kebutuhan sumber daya komputasi yang tinggi (Palm et al., 2017).

Untuk mengatasi keterbatasan tersebut, paradigma Human-in-the-Loop (HITL) telah berkembang sebagai solusi yang memungkinkan integrasi expertise manusia ke dalam sistem pembelajaran mesin (Mosqueira-Rey et al., 2023). HITL memungkinkan sistem untuk belajar secara progresif melalui interaksi dengan

pengguna, menggabungkan kekuatan komputasi mesin dengan pengetahuan domain dan intuisi manusia.

Namun, berdasarkan analisis terhadap penelitian state-of-the-art (2022-2025), implementasi HITL yang efektif dalam konteks ekstraksi data PDF template masih menghadapi kesenjangan penelitian yang signifikan. Pertama, model state-of-the-art seperti Large Language Models (LLM) menunjukkan kebutuhan yang tinggi akan validasi manusia (Schroeder et al., 2025), namun tidak dirancang untuk pembelajaran adaptif incremental yang efisien karena biaya retraining yang sangat tinggi. Kedua, di sisi lain spektrum, sistem transparan seperti rule-based murni terbukti unggul dalam kepercayaan pengguna (Schleith et al., 2022) namun pada dasarnya tetap kaku dan tidak adaptif.

Kesenjangan ini menciptakan kebutuhan akan arsitektur yang menjembatani kedua ekstrem tersebut, terutama dalam skenario "data scarcity" (Gebauer et al., 2023). Secara spesifik: (1) belum ada framework yang sistematis mengintegrasikan rule-based (untuk transparansi) dengan machine learning yang efisien (seperti CRF) dalam konteks HITL adaptif; (2) mekanisme feedback yang efisien untuk mengkonversi koreksi pengguna menjadi pengetahuan sistem masih belum terkarakterisasi dengan baik untuk arsitektur hybrid ; dan (3) strategi pembelajaran adaptif real-time tanpa retraining ekstensif masih menjadi tantangan terbuka.

Penelitian ini mengusulkan sistem pembelajaran adaptif berbasis Human-in-the-Loop (HITL) untuk ekstraksi data PDF template yang secara komprehensif mengatasi kesenjangan penelitian tersebut. Sistem yang dikembangkan mengintegrasikan pendekatan rule-based sebagai baseline dengan Conditional Random Fields (CRF) sebagai model pembelajaran mesin yang efisien sumber daya dan dapat beradaptasi berdasarkan feedback pengguna dalam framework HITL yang unified.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana mengintegrasikan domain expertise pengguna ke dalam sistem ekstraksi data PDF template melalui mekanisme Human-in-the-Loop yang efisien?
2. Bagaimana merancang mekanisme pembelajaran adaptif yang dapat memanfaatkan feedback pengguna untuk meningkatkan akurasi ekstraksi secara berkelanjutan?
3. Bagaimana mengoptimalkan pola interaksi pengguna dalam sistem HITL untuk meminimalkan beban kerja sambil memaksimalkan efektivitas pembelajaran sistem?
4. Bagaimana mengevaluasi efektivitas sistem pembelajaran adaptif berbasis HITL dalam meningkatkan akurasi ekstraksi data dari dokumen PDF template?

1.3 Tujuan

Penelitian ini bertujuan untuk:

1. Mengembangkan sistem Human-in-the-Loop yang memungkinkan integrasi efektif domain expertise pengguna ke dalam proses ekstraksi data PDF template.
2. Merancang dan mengimplementasikan mekanisme pembelajaran adaptif yang dapat memanfaatkan feedback pengguna untuk meningkatkan akurasi ekstraksi secara berkelanjutan.
3. Mengembangkan workflow interaktif yang mengoptimalkan pola keterlibatan pengguna untuk meminimalkan beban kerja sambil memaksimalkan efektivitas pembelajaran sistem.
4. Mengevaluasi efektivitas sistem pembelajaran adaptif berbasis HITL dalam meningkatkan akurasi ekstraksi data seiring dengan bertambahnya interaksi pengguna.
5. Menganalisis kontribusi mekanisme HITL dan pembelajaran adaptif terhadap peningkatan performa sistem ekstraksi data secara keseluruhan.

1.4 Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Manfaat Teoritis:
    a. Kontribusi Metodologis HITL: Mengembangkan framework teoritis untuk integrasi sistematis rule-based dan machine learning dalam konteks Human-in-the-Loop, memperkaya body of knowledge dalam domain interactive machine learning untuk document processing.
    b. Teori Pembelajaran Adaptif: Memperkaya pemahaman tentang mekanisme pembelajaran adaptif berbasis feedback pengguna, khususnya dalam konteks document processing dengan variasi format, memberikan insights untuk adaptive systems design.
    c. Framework Evaluasi HITL: Mengembangkan kerangka kerja evaluasi yang komprehensif untuk mengukur efektivitas sistem HITL dari multiple dimensions: technical performance, user experience, dan learning efficiency.
2. Manfaat Praktis:
    a. Solusi Adaptif Real-world: Menghasilkan sistem yang dapat langsung diimplementasikan untuk ekstraksi data PDF template dengan kemampuan adaptasi real-time tanpa memerlukan deep technical expertise.
    b. Efisiensi Operasional: Mengurangi secara signifikan waktu dan biaya yang diperlukan untuk konfigurasi, deployment, dan maintenance sistem ekstraksi data dalam production environments.
    c. Skalabilitas Organisasi: Menyediakan solusi yang dapat diadopsi oleh organisasi dengan berbagai skala, dari SMEs hingga enterprise, untuk otomatisasi document processing.

1.5 Batasan Penelitian

Untuk memastikan fokus dan kedalaman penelitian yang optimal, batasan berikut ditetapkan:

1.5.1 Batasan Teknis

1. **Scope Dokumen:** Penelitian berfokus pada PDF template dengan teks yang dapat dibaca secara digital (machine-readable text), tidak mencakup handwritten documents atau scanned documents dengan kualitas OCR yang rendah.
2. **Model Pembelajaran Mesin:** Menggunakan Conditional Random Fields (CRF) sebagai baseline model yang dapat beradaptasi dalam framework HITL, dengan fokus pada sequence labeling untuk ekstraksi informasi terstruktur.
3. **Jenis Template:** Membatasi pada semi-structured PDF templates seperti formulir, invoice, dan dokumen administratif dengan layout yang relatif konsisten namun memiliki variasi konten.

1.5.2 Batasan Metodologi

1. **Fokus Penelitian:** Menekankan pada kontribusi mekanisme HITL dan pembelajaran adaptif sebagai inti inovasi, bukan pada pengembangan algoritma machine learning yang baru (novel).
2. **Interaksi Pengguna:** Sistem dirancang untuk pengguna non-teknis dengan fokus pada mekanisme feedback yang intuitif dan efisien, tidak memerlukan expertise dalam machine learning atau programming

1.5.3 Batasan Evaluasi

1. **Metrik Evaluasi:** Evaluasi difokuskan pada efektivitas pembelajaran, pengalaman pengguna, dan adaptabilitas sistem, dengan mengukur peningkatan performa seiring waktu (improvement over time).
2. **Dataset:** Menggunakan simulated PDF templates yang merepresentasikan dokumen administratif dan bisnis real-world, dengan fokus pada demonstrasi kapabilitas adaptif.

1.6 Metodologi Penelitian

Penelitian ini menggunakan pendekatan design science research (DSR) yang berfokus pada pengembangan dan evaluasi artefak teknologi untuk memecahkan

masalah spesifik (Hevner et al., 2004). DSR dipilih karena sesuai dengan nature penelitian yang bertujuan mengembangkan solusi praktis untuk masalah real-world.

1.6.1 Framework DSR yang Diadopsi

1. **Problem Identification & Motivation:** Mengidentifikasi kesenjangan dalam ekstraksi data PDF template dan kebutuhan akan sistem HITL yang dapat belajar secara adaptif melalui systematic literature review dan analysis of existing solutions.
2. **Objectives Definition:** Menentukan tujuan spesifik sistem pembelajaran adaptif berbasis HITL dengan measurable success criteria dan clear value proposition.
3. **Design & Development:** Merancang dan mengimplementasikan sistem dengan arsitektur modular yang mencakup:
    a. HITL interaction mechanism untuk domain expertise integration
    b. Adaptive learning engine berbasis user feedback
    c. Intelligent workflow yang mengoptimalkan user engagement
    d. CRF-based model untuk incremental learning
4. **Demonstration:** Memvalidasi feasibility sistem melalui proof-of-concept implementation dengan real-world PDF templates dari berbagai domain.
5. **Evaluation:** Menilai efektivitas sistem menggunakan multi-dimensional evaluation framework:
    a. Learning effectiveness metrics (accuracy improvement over time)
    b. User experience metrics (task completion time, cognitive load)
    c. System efficiency metrics (response time, resource utilization)
    d. Practical applicability assessment
6. **Communication:** Mendokumentasikan findings, contributions, dan lessons learned untuk academic community dan practitioners.

1.6.2 Metodologi Evaluasi

Evaluasi dilakukan menggunakan mixed-methods approach yang menggabungkan quantitative metrics dengan qualitative insights untuk memberikan comprehensive assessment terhadap sistem yang dikembangkan.

1.7 Sistematika Pembahasan

Proposal tesis ini disusun dengan sistematika sebagai berikut:

BAB I PENDAHULUAN: menyajikan foundation penelitian dengan latar belakang yang menekankan kesenjangan dalam ekstraksi data PDF template dan positioning HITL sebagai solusi, rumusan masalah yang fokus pada pembelajaran adaptif, tujuan penelitian yang terukur, batasan penelitian yang jelas, manfaat teoritis dan praktis, serta metodologi Design Science Research yang diadopsi.

BAB II TINJAUAN PUSTAKA: memberikan comprehensive review terhadap state-of-the-art dalam Human-in-the-Loop systems, adaptive learning mechanisms, PDF data extraction techniques, dan analisis mendalam terhadap research gaps yang mendasari kontribusi penelitian ini. Bab ini juga menyajikan positioning penelitian dalam konteks existing body of knowledge.

BAB III METODOLOGI PENELITIAN: mendetailkan implementasi Design Science Research framework, arsitektur sistem HITL yang dikembangkan, design rationale untuk setiap komponen sistem, implementation details dari adaptive learning mechanism, user interaction workflow design, serta comprehensive evaluation methodology yang mencakup quantitative dan qualitative measures.

BAB IV HASIL DAN PEMBAHASAN: mempresentasikan hasil implementasi sistem dengan detailed analysis, evaluation results dari multiple perspectives (accuracy, user experience, system efficiency), comparative analysis dengan existing approaches, discussion tentang findings dan implications, serta critical assessment terhadap limitations dan trade-offs.

BAB V KESIMPULAN DAN SARAN: merangkum key findings dan contributions, theoretical dan practical implications dari penelitian, lessons learned dari implementation dan evaluation, serta recommendations untuk future research directions dalam domain HITL systems dan adaptive learning untuk document processing.

BAB 2 TINJAUAN PUSTAKA

2.1 Dokumen PDF Template

2.1.1 Pengertian dan Karakteristik Dokumen PDF

Dokumen PDF (Portable Document Format) merupakan format dokumen yang dikembangkan oleh Adobe Systems pada tahun 1993 untuk menyajikan dokumen secara konsisten terlepas dari aplikasi, perangkat keras, atau sistem operasi yang digunakan untuk melihatnya (International Organization for Standardization, 2008). Format PDF telah menjadi standar ISO 32000 dan banyak digunakan dalam berbagai sektor karena kemampuannya mempertahankan tampilan dan struktur dokumen.

Beberapa karakteristik utama dokumen PDF yang membuatnya populer meliputi:

1. Portabilitas: Dokumen PDF dapat dibuka dan ditampilkan secara konsisten di berbagai platform dan perangkat.
2. Preservasi Format: PDF mempertahankan semua elemen dokumen, termasuk font, gambar, dan layout, terlepas dari aplikasi yang digunakan untuk membukanya.
3. Keamanan: PDF mendukung berbagai fitur keamanan, seperti enkripsi dan pembatasan akses.
4. Kompresi: Format PDF menggunakan teknik kompresi untuk mengurangi ukuran file tanpa mengorbankan kualitas.
5. Dukungan untuk Konten Interaktif: PDF dapat menyertakan elemen interaktif seperti hyperlink, form fields, dan JavaScript.

2.1.2 Klasifikasi Dokumen Digital

Berdasarkan strukturnya, dokumen digital dapat diklasifikasikan menjadi tiga kategori (Abiteboul, 1997):

1. Dokumen Terstruktur: Memiliki struktur yang jelas dan konsisten, seperti database atau dokumen XML dengan skema yang terdefinisi

dengan baik. Dokumen terstruktur memiliki organisasi data yang eksplisit dan dapat dengan mudah diproses secara otomatis.

2. Dokumen Semi-Terstruktur: Memiliki struktur yang dapat diidentifikasi namun dengan variasi dan fleksibilitas tertentu. Form PDF termasuk dalam kategori ini, di mana terdapat elemen statis (label, instruksi) dan elemen dinamis (data yang diisi). Dokumen semi-terstruktur menggabungkan aspek terstruktur dan tidak terstruktur, dengan beberapa bagian mengikuti format yang konsisten sementara bagian lain memiliki variasi.
3. Dokumen Tidak Terstruktur: Tidak memiliki struktur formal yang dapat diidentifikasi secara langsung, seperti teks bebas atau catatan. Dokumen tidak terstruktur memerlukan teknik pemrosesan bahasa alami atau analisis semantik untuk mengekstrak informasi yang bermakna.

2.1.3 Karakteristik PDF Template

PDF template memiliki karakteristik khusus yang membuatnya menantang untuk diekstraksi secara otomatis namun cocok untuk pendekatan Human-in-the-Loop:

1. Layout Konsisten dengan Variasi: Memiliki layout yang relatif konsisten namun dengan variasi pada posisi dan format data. Variasi ini dapat disebabkan oleh perbedaan versi formulir, perbedaan cara pengisian, atau perbedaan dalam proses digitalisasi.
2. Kombinasi Elemen Statis dan Dinamis: Mengandung kombinasi teks statis (label, instruksi, judul) dan data variabel (informasi yang diisi oleh pengguna). Elemen statis biasanya konsisten antar dokumen, sementara elemen dinamis bervariasi.
3. Struktur Kompleks: Seringkali memiliki struktur tabel, kotak centang, atau elemen interaktif yang memerlukan pendekatan khusus untuk ekstraksi.
4. Variasi Format Data: Dapat berisi data dalam berbagai format (teks, tanggal, angka, dll.) yang memerlukan normalisasi dan validasi.

5. Dependensi Kontekstual: Interpretasi data sering bergantung pada konteks, seperti label yang mendahului atau mengikuti nilai.
6. Noise dan Artefak: Dokumen yang dipindai atau dikonversi dari format lain mungkin mengandung noise, artefak, atau distorsi yang mempengaruhi kualitas ekstraksi.

Tantangan-tantangan ini membuat ekstraksi data dari form PDF template memerlukan pendekatan yang lebih canggih dibandingkan dengan dokumen terstruktur, menggabungkan teknik berbasis aturan, analisis layout, dan machine learning.

2.2 Teknik Ekstraksi Data dari Dokumen PDF

Ekstraksi data dari dokumen PDF telah menjadi bidang penelitian yang aktif dengan berbagai pendekatan yang dikembangkan. Berikut adalah pembahasan mendalam tentang teknik-teknik utama:

2.2.1 Pendekatan Berbasis Aturan

Pendekatan berbasis aturan mengandalkan seperangkat aturan yang ditentukan secara manual untuk mengidentifikasi dan mengekstrak data dari dokumen. Teknik ini umumnya menggunakan ekspresi reguler (regex), pencocokan pola, dan aturan posisional untuk menemukan informasi yang diinginkan (Dengel & Klein, 2002).

Komponen Utama Pendekatan Berbasis Aturan:

1. Ekspresi Reguler (Regex): Pola teks formal yang digunakan untuk mencocokkan dan mengekstrak informasi berdasarkan struktur sintaksis. Misalnya, pola untuk mengekstrak NIK (Nomor Induk Kependudukan) Indonesia yang terdiri dari 16 digit: \b\d{16}\b.
2. Aturan Posisional: Menggunakan koordinat atau posisi relatif untuk mengidentifikasi lokasi data dalam dokumen. Misalnya, mengekstrak data dari kolom tertentu dalam tabel berdasarkan koordinat x dan y.
3. Pencocokan Kontekstual: Menggunakan konteks sekitar (seperti label atau header) untuk mengidentifikasi data. Misalnya, mencari teks yang muncul setelah label "Nama:".

4. Validasi dan Normalisasi: Aturan untuk memvalidasi dan menormalkan data yang diekstraksi, seperti format tanggal, angka, atau teks.

Kelebihan pendekatan berbasis aturan meliputi:

1. Presisi Tinggi: Untuk dokumen dengan format yang konsisten, pendekatan ini dapat mencapai akurasi yang sangat tinggi.
2. Transparansi: Aturan dapat dengan mudah dipahami dan dimodifikasi oleh manusia.
3. Tidak Memerlukan Data Pelatihan: Dapat diimplementasikan tanpa data pelatihan yang besar.
4. Kinerja Deterministik: Hasil ekstraksi konsisten dan dapat diprediksi.

Kelemahan pendekatan berbasis aturan meliputi:

1. Kurang Fleksibel: Sulit beradaptasi dengan variasi layout dan format.
2. Pemeliharaan Tinggi: Memerlukan pembaruan manual ketika format dokumen berubah.
3. Skalabilitas Terbatas: Membuat aturan untuk setiap jenis dokumen memerlukan upaya yang signifikan.
4. Kesulitan dengan Ambiguitas: Sulit menangani kasus di mana interpretasi data bergantung pada konteks yang kompleks.

2.2.2 Pendekatan Berbasis Template

Pendekatan berbasis template menggunakan dokumen template sebagai referensi untuk mengidentifikasi lokasi data dalam dokumen target. Sistem mengidentifikasi elemen statis dalam template dan menggunakan penanda variabel seperti {nama_lengkap} atau <<nama_lengkap>> untuk menunjukkan lokasi data yang akan diekstraksi (Dengel & Klein, 2002). Penanda ini bertindak sebagai titik referensi bagi sistem untuk mengenali lokasi data yang ingin diambil.

Komponen Utama Pendekatan Berbasis Template:

1. Template Dokumen: Representasi dokumen kosong atau generik yang menandai lokasi bidang data.

2. Penanda Variabel: Simbol atau notasi khusus yang menunjukkan lokasi data yang akan diekstraksi, seperti {nama_lengkap} atau <<nama_lengkap>>.
3. Pemetaan Bidang: Hubungan antara penanda variabel dalam template dan bidang data yang akan diekstraksi.
4. Algoritma Penyelarasan: Metode untuk menyelaraskan dokumen target dengan template untuk mengidentifikasi lokasi data.

Kelebihan Pendekatan Berbasis Template:

1. Efisiensi: Untuk dokumen dengan format yang konsisten, pendekatan ini dapat mengekstrak data dengan cepat dan akurat.
2. Kemudahan Konfigurasi: Template dapat dikonfigurasi tanpa pengetahuan pemrograman yang mendalam.
3. Adaptabilitas Moderat: Dapat menangani variasi kecil dalam layout dokumen melalui teknik penyelarasan.
4. Integrasi dengan Sistem Lain: Mudah diintegrasikan dengan sistem manajemen dokumen yang ada.

Keterbatasan Pendekatan Berbasis Template:

1. Sensitif terhadap Perubahan Layout: Perubahan signifikan dalam layout dokumen dapat mengurangi akurasi.
2. Memerlukan Template untuk Setiap Jenis Dokumen: Setiap jenis dokumen memerlukan template terpisah.
3. Kesulitan dengan Variasi Besar: Sulit menangani dokumen dengan variasi format yang signifikan.
4. Keterbatasan dalam Ekstraksi Kontekstual: Kurang efektif untuk data yang interpretasinya bergantung pada konteks kompleks.

2.2.3 Pendekatan Berbasis Machine Learning

Pendekatan machine learning menggunakan algoritma pembelajaran mesin untuk mengidentifikasi dan mengekstrak data dari dokumen PDF. Pendekatan ini dapat belajar dari data pelatihan untuk mengenali pola dan struktur dokumen secara otomatis.

Jenis-jenis Pendekatan Machine Learning:

1. Supervised Learning: Menggunakan data berlabel untuk melatih model ekstraksi.
2. Unsupervised Learning: Mengidentifikasi pola dalam dokumen tanpa data berlabel.
3. Deep Learning: Menggunakan neural networks untuk ekstraksi yang lebih kompleks.
4. Sequence Labeling: Menggunakan model seperti Conditional Random Fields (CRF) atau LSTM untuk labeling sekuensial. CRF khususnya efektif untuk ekstraksi data PDF template karena kemampuannya memodelkan dependensi antar label dan mengintegrasikan berbagai jenis fitur kontekstual.

Kelebihan pendekatan machine learning:

1. Adaptabilitas: Dapat beradaptasi dengan variasi format dokumen.
2. Skalabilitas: Dapat menangani berbagai jenis dokumen dengan satu model.
3. Pembelajaran Otomatis: Dapat belajar dari data tanpa aturan manual.

Kelemahan pendekatan machine learning:

1. Kebutuhan Data: Memerlukan dataset pelatihan yang besar dan berkualitas.
2. Kompleksitas: Lebih kompleks untuk diimplementasikan dan dipelihara.
3. Interpretabilitas: Sulit untuk memahami bagaimana model membuat keputusan.

2.2.4 Conditional Random Fields

CRF memodelkan distribusi probabilitas bersyarat p(y|x) dari urutan label y diberikan urutan observasi x. Berbeda dengan model generatif seperti Hidden Markov Models (HMM), CRF adalah model diskriminatif yang langsung

memodelkan probabilitas bersyarat tanpa perlu memodelkan distribusi gabungan p(x,y).

Formulasi matematika dasar CRF adalah:

$$p(y|x) = \frac{1}{Z(x)} exp\left(\sum_{j=1}^{m} \lambda_j F_j(y, x)\right)$$

di mana:

- $Z(x)$ adalah faktor normalisasi
- λ_j adalah parameter bobot
- $F_j(y, x)$ adalah fitur global yang merupakan jumlah dari fitur lokal $f_j(y_i, y_{i-1}, x, i)$ di semua posisi i.

CRF sangat cocok untuk ekstraksi data dokumen karena beberapa alasan:

1. Pemodelan Konteks: CRF dapat memodelkan dependensi antara label yang berdekatan, memungkinkan ekstraksi yang mempertimbangkan konteks.
2. Integrasi Fitur Beragam: CRF dapat mengintegrasikan berbagai jenis fitur (teks, layout, visual) dalam satu model.
3. Penanganan Sekuens: CRF secara alami menangani data sekuensial, yang sesuai dengan struktur banyak dokumen.
4. Kinerja yang Baik dengan Dataset Kecil: Dibandingkan dengan deep learning, CRF dapat memberikan hasil yang baik bahkan dengan dataset pelatihan yang relatif kecil.

Fitur yang umum digunakan dalam CRF untuk ekstraksi dokumen meliputi:

1. Fitur Teks: Kata, n-gram, pola teks, fitur leksikal (huruf besar/kecil, angka, tanda baca).
2. Fitur Layout: Posisi (koordinat x, y), jarak dari elemen lain, alignment, indentasi.
3. Fitur Visual: Font, ukuran, gaya (tebal, miring), warna, garis, kotak.

4. Fitur Kontekstual: Kata-kata di sekitar, label tetangga, posisi dalam dokumen.
5. Fitur Domain-Specific: Pola khusus domain (misalnya format NIK, nomor telepon, kode pos).

Pelatihan CRF melibatkan estimasi parameter λ yang memaksimalkan log-likelihood dari data pelatihan, sering dengan regularisasi L1 atau L2 untuk mencegah overfitting. Algoritma optimasi seperti L-BFGS biasanya digunakan untuk pelatihan.

Inferensi dalam CRF melibatkan pencarian urutan label yang paling mungkin untuk observasi yang diberikan, biasanya menggunakan algoritma Viterbi.

Kelebihan CRF untuk Ekstraksi Data Dokumen:

1. Akurasi Tinggi: Mempertimbangkan konteks dan dependensi antar label.
2. Fleksibilitas Fitur: Dapat mengintegrasikan berbagai jenis fitur.
3. Interpretabilitas: Lebih interpretable dibandingkan deep learning models.
4. Efisiensi Komputasi: Lebih efisien dibandingkan beberapa model deep learning.

Keterbatasan CRF:

1. Kompleksitas Fitur Engineering: Memerlukan desain fitur yang cermat.
2. Skalabilitas: Dapat menjadi komputasional mahal untuk dataset besar dengan banyak fitur.
3. Keterbatasan dalam Menangkap Dependensi Jarak Jauh: Kurang efektif untuk dependensi jarak jauh dibandingkan dengan model seperti LSTM atau Transformers.

2.2.5 Pendekatan Hybrid

Pendekatan hybrid menggabungkan kekuatan dari berbagai teknik ekstraksi untuk mencapai performa yang optimal. Dalam konteks ekstraksi data PDF

template, pendekatan hybrid biasanya mengintegrasikan rule-based extraction dengan machine learning approaches.

Karakteristik Pendekatan Hybrid:

1. Multi-Strategy Integration: Menggunakan berbagai strategi ekstraksi secara bersamaan atau berurutan.
2. Adaptive Strategy Selection: Memilih strategi terbaik berdasarkan karakteristik dokumen atau confidence level.
3. Fallback Mechanisms: Menggunakan strategi alternatif ketika strategi utama gagal.
4. Confidence-based Switching: Beralih antar strategi berdasarkan tingkat kepercayaan hasil.

Kelebihan pendekatan hybrid:

1. Robustness: Lebih robust terhadap variasi dokumen dan edge cases.
2. Optimal Performance: Dapat mencapai performa terbaik dengan menggabungkan kekuatan berbagai pendekatan.
3. Flexibility: Dapat disesuaikan dengan kebutuhan spesifik domain atau jenis dokumen.

Kelemahan pendekatan hybrid:

1. Complexity: Lebih kompleks untuk didesain dan diimplementasikan.
2. Resource Requirements: Memerlukan lebih banyak sumber daya komputasi.

Maintenance Overhead: Memerlukan pemeliharaan baik rule-based maupun ML components.

2.3 Human-in-the-Loop Adaptive Learning

2.3.1 Konsep Dasar Pembelajaran Adaptif

Pembelajaran adaptif (adaptive learning) dalam konteks ekstraksi data mengacu pada kemampuan sistem untuk menyesuaikan dan meningkatkan performanya berdasarkan pengalaman dan umpan balik yang diterima (Settles, 2012). Sistem adaptif dapat belajar dari kesalahan, menyesuaikan strategi ekstraksi,

dan meningkatkan akurasi seiring waktu tanpa memerlukan reprogramming atau retraining dari awal.

Karakteristik Utama Pembelajaran Adaptif:

1. Incremental Learning: Kemampuan untuk belajar secara bertahap dari data baru tanpa melupakan pengetahuan sebelumnya.
2. Online Learning: Sistem dapat memperbarui model secara real-time saat menerima data atau feedback baru.
3. Self-Improvement: Sistem secara otomatis mengidentifikasi area yang perlu diperbaiki dan menyesuaikan strategi accordingly.
4. Context Awareness: Kemampuan untuk memahami dan beradaptasi dengan konteks yang berbeda atau perubahan domain.
5. Performance Monitoring: Sistem secara kontinyu memantau performanya dan melakukan adjustment ketika diperlukan.

Jenis-jenis Pembelajaran Adaptif dalam Information Extraction:

1. Active Learning: Sistem secara aktif memilih data yang paling informatif untuk dilabeli, meminimalkan effort annotation.
2. Transfer Learning: Memanfaatkan pengetahuan dari domain atau task yang sudah dipelajari untuk domain baru.
3. Multi-task Learning: Belajar multiple related tasks secara bersamaan untuk meningkatkan generalization.
4. Meta-Learning: Belajar bagaimana cara belajar, memungkinkan adaptasi cepat pada task baru dengan data minimal.
5. Continual Learning: Belajar sequence of tasks secara berurutan tanpa catastrophic forgetting.

Dalam domain pemrosesan dokumen, pembelajaran adaptif telah diterapkan dalam berbagai konteks:

1. Document Classification: Sistem yang dapat beradaptasi dengan kategori dokumen baru atau perubahan distribusi data.
2. Named Entity Recognition: Model yang dapat mengenali entitas baru atau beradaptasi dengan domain spesifik.

3. Information Extraction: Sistem yang dapat menyesuaikan extraction rules atau patterns berdasarkan feedback.
4. Layout Analysis: Algoritma yang dapat beradaptasi dengan variasi layout dokumen yang tidak ditemui dalam training data.

Evaluation Metrics untuk Adaptive Systems:

1. Evaluasi sistem adaptive learning memerlukan metrik khusus yang berbeda dari traditional machine learning:
2. Learning Curve Analysis: Mengukur peningkatan performa sistem seiring bertambahnya data atau interaksi.
3. Adaptation Speed: Mengukur seberapa cepat sistem dapat beradaptasi dengan perubahan atau data baru.
4. Stability Metrics: Mengevaluasi apakah sistem tetap stabil dan tidak mengalami performance degradation.
5. Forgetting Metrics: Mengukur seberapa banyak pengetahuan lama yang dilupakan ketika belajar hal baru.
6. Resource Efficiency: Mengukur computational cost dan memory usage selama proses adaptasi.

Tantangan dalam Pembelajaran Adaptif:

1. Catastrophic Forgetting: Kecenderungan untuk melupakan pengetahuan lama ketika belajar informasi baru.
2. Concept Drift: Perubahan dalam distribusi data atau target concept seiring waktu.
3. Limited Feedback: Keterbatasan dalam mendapatkan feedback yang berkualitas dan konsisten.
4. Scalability: Mengelola pembelajaran dari multiple sources atau users dengan preferensi berbeda.
5. Evaluation Complexity: Kesulitan dalam mengevaluasi sistem yang terus berubah dan beradaptasi.

Berdasarkan karakteristik dan tantangan pembelajaran adaptif di atas, pendekatan Human-in-the-Loop (HITL) muncul sebagai solusi yang menjanjikan

untuk mengatasi keterbatasan sistem adaptif murni dengan mengintegrasikan expertise manusia dalam loop pembelajaran.

2.3.2 Human-in-the-Loop Adaptive Learning

Human-in-the-Loop (HITL) adaptive learning adalah pendekatan di mana manusia berpartisipasi aktif dalam proses pembelajaran mesin, memberikan umpan balik, validasi, atau koreksi untuk meningkatkan kinerja sistem secara berkelanjutan. Pendekatan ini sangat relevan untuk ekstraksi data dokumen, di mana pengetahuan domain dan interpretasi manusia sering diperlukan.

HITL learning merupakan evolusi dari interactive machine learning yang pertama kali diperkenalkan oleh (Fails & Olsen, 2003), di mana sistem pembelajaran dapat beradaptasi secara real-time berdasarkan interaksi pengguna. Dalam konteks PDF template extraction, HITL memungkinkan sistem untuk menangani ambiguitas dan variasi format yang sulit diatasi oleh sistem otomatis.

Komponen Utama HITL Adaptive Learning:

1. User Feedback Integration: Sistem dapat menerima dan mengintegrasikan umpan balik pengguna untuk memperbaiki hasil ekstraksi.
2. Confidence-based Interaction: Sistem meminta bantuan pengguna hanya ketika confidence level rendah atau menghadapi situasi yang belum pernah ditemui.
3. Incremental Learning: Sistem dapat belajar secara bertahap dari setiap interaksi tanpa memerlukan pelatihan ulang dari awal.
4. Intelligent Sampling: Sistem secara aktif memilih contoh yang paling informatif untuk mendapatkan umpan balik pengguna.
5. Model Adaptation: Parameter model disesuaikan berdasarkan umpan balik untuk meningkatkan performa pada kasus serupa di masa depan.

Keunggulan HITL Adaptive Learning:

1. Domain Expertise Integration: Memanfaatkan pengetahuan domain pengguna yang sulit dikodekan dalam aturan.

2. Real-time Adaptation: Dapat beradaptasi secara real-time tanpa memerlukan pelatihan offline yang ekstensif.
3. Handling Edge Cases: Efektif menangani kasus-kasus khusus atau anomali yang jarang terjadi.
4. Continuous Improvement: Performa sistem terus meningkat seiring bertambahnya interaksi dengan pengguna.
5. Reduced Annotation Cost: Meminimalkan kebutuhan data berlabel dengan memanfaatkan feedback yang targeted.

Tantangan HITL Adaptive Learning:

1. User Burden: Meminimalkan beban kerja pengguna sambil memaksimalkan nilai pembelajaran.
2. Feedback Quality: Memastikan kualitas umpan balik pengguna dan menangani inkonsistensi.
3. Scalability: Mengelola pembelajaran dari multiple users dengan preferensi yang berbeda.
4. Overfitting Prevention: Mencegah overfitting pada umpan balik pengguna yang terbatas.

2.4 Penelitian Terkait

Penelitian yang relevan dengan pembelajaran adaptif berbasis HITL untuk ekstraksi data PDF template dapat dikategorikan dalam beberapa gelombang:

2.4.1 Fondasi HITL dan Pendekatan Hybrid Awall

Penelitian fondasi dalam interactive machine learning (Holzinger, 2016; Stumpf et al., 2009) dan desain interaksi Manusia-AI (Amershi et al., 2019) menekankan pentingnya umpan balik pengguna dan transparansi. Secara bersamaan, sistem hybrid awal seperti smartFIX (Dengel & Klein, 2002) dan Intellix (Schuster et al., 2013) menunjukkan kelayakan menggabungkan analisis layout dengan ML untuk dokumen bisnis. Penelitian-penelitian ini menetapkan kebutuhan akan sistem yang dapat belajar dari pengguna, meskipun mekanisme pembelajarannya seringkali terbatas.

1.

2.4.2 Tren State-of-the-Art: Model Besar dan Validasi HITL

Dalam beberapa tahun terakhir, fokus penelitian bergeser ke model deep learning yang sangat besar. Pendekatan seperti Few-Shot Learning (FSL) untuk document-level (Popovic & Färber, 2022) dan Large Language Models (LLM) (Schroeder et al., 2025) telah menunjukkan kemampuan ekstraksi yang kuat.

Namun, tren ini memunculkan tantangan baru:

1. **Kebutuhan Sumber Daya (Resource):** Pendekatan ini (misalnya, LLM seperti Gemini atau FSL SOTA) memiliki kebutuhan komputasi yang sangat tinggi (Popovic & Färber, 2022; Schroeder et al., 2025)
2. **Keterbatasan Peran HITL:** Ironisnya, model-model ini masih "sangat membutuhkan" validasi dari manusia (HITL) (Schroeder et al., 2025). Namun, karena biaya retraining yang mahal, peran HITL terbatas pada validasi hasil, bukan pembelajaran adaptif incremental pada model itu sendiri.

2.4.3 Tren Alternatif: HITL untuk Efisiensi dan Kelangkaan Data (Data Scarcity)

Sebagai tandingan dari tren model besar, penelitian lain (Gebauer et al., 2023; Schleith et al., 2022) berfokus pada peran HITL dalam skenario data scarcity atau untuk meningkatkan transparansi dan efisiensi. (Schleith et al., 2022) misalnya, menemukan bahwa sistem rule-based yang dibantu HITL dapat mengungguli sistem black-box dalam hal kepercayaan pengguna dan waktu pengerjaan end-to-end.

Penelitian-penelitian ini menunjukkan adanya kesenjangan antara (a) model besar yang boros sumber daya dan tidak adaptif-incremental, dan (b) sistem transparan yang efisien namun kaku. Penelitian ini (proposal Anda) diposisikan untuk mengisi kesenjangan tersebut.

2.5 Analisis Perbandingan Pendekatan

2.5.1 Perbandingan Sistematis dengan Penelitian Terdahulu

Tabel berikut menyajikan perbandingan sistematis antara penelitian ini dengan penelitian terdahulu.

Penelitian	Pendekatan	Adaptive Learning	Resource Req.	User Feedback	Hybrid Strategy	Evaluation
Holzinger (2016)	Interactive ML	Ya (domain kesehatan)	Sedang	Expert feedback	Tidak	Domain-specific
Schuster et al. (2013)	Template + ML	Terbatas	Sedang	Manual correction	Ya	Commercial eval
Katti et al. (2018)	Chargrid (2D CNN)	Tidak	Tinggi	Tidak	Tidak	F1-score
Palm et al. (2017)	CloudScan (RNN)	Tidak	Tinggi	Tidak	Tidak	Invoice accuracy
Schleith et al. (2022)	Rule-based + HITL	Terbatas (User membuat aturan)	Rendah	Input pembuatan aturan	Tidak	Waktu pengerjaan, Kepercayaan
Gebauer et al. (2023)	ML-based + HTL	Ya (sugesti ML)	Sedang (untuk data minim)	Keputusan anotasi	Tidak	Performa (data-mini)
Popovic et al. (2022)	Few-Shot (FSL)	Ya (Adaptasi domain)	Sangat Tinggi	Few-shot example	Tidak	F1-score (document-level)
Schroeder et al. (2025)	LLM+HITL	Tidak (HITL untuk validasi)	Sangat Tinggi	Validasi/Koreksi	Tidak	Konsistensi (vs Manusia)
Penelitian Ini	Hybrid (Rule-based+CRF) +HITL	Ya (Incremental)	Rendah-Sedang	Minimal Effort	Ya	Precision, Recall, F1, Learning Curves.

Tabel -1 Penelitian Terdahulu

Analisis Perbandingan:

Tabel 2.1 menyoroti evolusi dan kesenjangan yang jelas. Penelitian awal (misalnya Schuster et al., 2013) telah mengidentifikasi perlunya strategi hybrid. Namun, penelitian state-of-the-art (2017-2025) terpolarisasi menjadi dua kubu:

1. **Kubu Model Besar (Resource-Intensive):** Pendekatan seperti Katti et al., (2018), Palm et al., (2017) serta (Schroeder et al., (2025), berfokus pada pencapaian akurasi tinggi dengan model yang sangat kompleks (CNN, RNN, FSL, LLM). Konsekuensinya adalah kebutuhan sumber daya yang "Tinggi" atau "Sangat Tinggi". Yang paling signifikan, model-model ini tidak memiliki kemampuan "Adaptive Learning" secara incremental dari feedback pengguna; feedback hanya digunakan untuk validasi.
2. **Kubu Efisiensi (Resource-Efficient):** Pendekatan seperti Schleith et al., 2022) dan (Gebauer et al., 2023) secara eksplisit menargetkan efisiensi, transparansi, atau skenario data scarcity. Mereka membuktikan nilai HITL. Namun, mereka tidak mengintegrasikannya ke dalam arsitektur hybrid yang dapat belajar secara statistik.

2.5.2 Perbandingan Komprehensif Pendekatan Ekstraksi Data

Tabel berikut memberikan perbandingan komprehensif dari teknik ekstraksi yang mendasari penelitian ini.

Aspek	Rule-Based	Template-Based	CRF (Baseline)	Hybrid (Penelitian Ini)
Akurasi	Sedang-Tinggi	Tinggi	Tinggi	Tinggi (Adaptif)
HITL Compatibility	Rendah	Rendah	Sedang	Tinggi
Real-time Adaptation	Tidak ada	Tidak ada	Terbatas	Baik
User Burden	Tinggi	Rendah	Sedang	Rendah
Feedback Integration	Manual	Manual	Sulit	Semi-otomatis
Learning Efficiency	Tidak ada	Tidak ada	Rendah	Sedang-Tinggi

Data Training Required	Tidak perlu	Template Only	Sedang	Minimal
Interpretability	Sangat Tinggi	Tinggi	Tinggi	Tinggi
Computational Cost	Rendah	Rendah	Sedang	Sedang
Deployment Complexity	Rendah	Rendah	Sedang	Sedang
Adaptabilitas	Rendah	Rendah	Sedang	Tinggi
Sequential Modeling	Tidak ada	Tidak ada	Sangat Baik	Sangat Baik
Use Case Optimal	Format tetap	Consistent docs	Sequence tasks	Adaptive PDF templates

Tabel 2-2 Pendekatan Ekstraksi Data

2.5.3 Positioning Penelitian dalam Spektrum Pendekatan

Penelitian ini diposisikan secara unik untuk mengisi kesenjangan praktis di antara kedua kubu yang teridentifikasi di Tabel 2.1. Seperti yang ditunjukkan pada Tabel 2.2 dan baris terakhir Tabel 2.1, penelitian ini mengadopsi pendekatan hybrid namun dengan fokus modern pada:

1. Adaptabilitas Tinggi: Menggunakan HITL untuk pembelajaran incremental (yang tidak dimiliki LLM seperti pada Schroeder et al., 2025).
2. Efisiensi Sumber Daya: Menggunakan CRF (bukan RNN/LLM) untuk menjaga kebutuhan sumber daya "Rendah-Sedang".
3. Hybrid Strategy: Menggabungkan Rule-based (untuk transparansi, seperti Schleith et al., 2022) dan CRF (untuk adaptasi ML, seperti Gebauer et al., 2023) dalam satu framework terpadu.

Penelitian ini tidak bertujuan mengalahkan LLM dalam benchmark akurasi murni, melainkan mengusulkan arsitektur yang unggul dalam **keseimbangan antara akurasi, efisiensi sumber daya, dan kemampuan adaptasi real-time**.

2.6 Research Gap

2.6.1 Rangkuman Tinjauan Pustaka

Berdasarkan tinjauan pustaka yang telah dilakukan, beberapa poin penting dapat disimpulkan:

1. Dokumen PDF Template memiliki karakteristik semi-terstruktur yang menantang untuk ekstraksi otomatis.
2. Teknik Ekstraksi Data telah berkembang dari berbasis aturan yang kaku, hingga machine learning yang lebih fleksibel.
3. Conditional Random Fields (CRF) menunjukkan keunggulan dalam sequence labeling untuk dokumen.
4. Human-in-the-Loop Adaptive Learning muncul sebagai paradigma yang menjanjikan untuk mengatasi keterbatasan sistem otomatis murni .

2.6.2 Identifikasi Research Gap

Meskipun tinjauan pustaka menunjukkan evolusi teknik ekstraksi, analisis terhadap penelitian state-of-the-art (termasuk periode 2022-2025) mengidentifikasi beberapa kesenjangan kritis yang baru:

1. Kesenjangan Kebutuhan Sumber Daya (Resource Gap): Terdapat tren penelitian yang kuat ke arah penggunaan model yang sangat besar, seperti Few-Shot Learning (FSL) berbasis document-level (misalnya, Popovic & Färber, 2022) dan Large Language Models (LLM) (misalnya, Schroeder et al., 2025). Meskipun kuat, pendekatan ini memiliki kebutuhan sumber daya komputasi dan data yang sangat tinggi. Hal ini menciptakan kesenjangan praktis untuk adopsi di banyak organisasi.
2. Peran HITL yang Terbatas pada Model Besar: Ironisnya, bahkan model LLM terbesar pun terbukti masih sangat memerlukan Human-in-the-Loop (Schroeder et al., 2025). Namun, karena kompleksitas dan biaya retraining model-model ini, peran HITL seringkali terbatas hanya

sebagai mekanisme validasi atau koreksi, bukan sebagai pemicu pembelajaran adaptif incremental pada model itu sendiri.

3. Keterbatasan Sistem Transparan (Rule-Based): Di sisi lain spektrum, penelitian terbaru (misalnya, Schleith et al., 2022) mengkonfirmasi bahwa sistem HITL yang transparan (seperti rule-based) unggul dalam hal kepercayaan pengguna (user trust) dan efisiensi end-to-end. Namun, sistem ini pada dasarnya tetap kaku dan tidak memiliki kemampuan pembelajaran adaptif otomatis.
4. Minimnya Arsitektur Hybrid Adaptif yang Efisien: Kesenjangan utama terletak di antara dua ekstrem tersebut. Masih sangat minim penelitian yang berfokus pada arsitektur hybrid yang efisien sumber daya, yang secara cerdas menggabungkan transparansi rule-based dengan kemampuan sequence modeling (seperti CRF), dan—yang terpenting—menggunakan HITL sebagai pendorong utama pembelajaran adaptif incremental yang resource-efficient.

2.6.3 Positioning Penelitian

Penelitian ini memposisikan diri secara strategis untuk mengisi kesenjangan yang teridentifikasi tersebut. Alih-alih bersaing dalam paradigma resource-intensive (LLM/FSL), penelitian ini mengusulkan sebuah **alternatif arsitektur hybrid yang praktis dan adaptif**

1. Menjembatani Paradigma: Mengintegrasikan CRF sebagai sequence model yang resource-efficient dengan pendekatan rule-based sebagai baseline yang transparan.
2. HITL sebagai Pemicu Pembelajaran: Memanfaatkan paradigma HITL secara penuh, dalam sistem ini, feedback pengguna secara aktif digunakan untuk pembelajaran adaptif incremental model CRF.
3. Fokus pada Efisiensi: Menargetkan solusi yang meminimalkan beban kerja pengguna dan memiliki kebutuhan sumber daya rendah-sedang, sehingga praktis untuk implementasi di dunia nyata, terutama dalam skenario "data scarcity"

Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan sistem ekstraksi data yang lebih adaptif, efisien, dan user-friendly.

BAB 3 METODOLOGI PENELITIAN

3.1 Desain Penelitian

Penelitian ini menggunakan pendekatan design science research (DSR) yang berfokus pada pengembangan dan evaluasi artefak teknologi untuk memecahkan masalah spesifik (Hevner et al., 2004). Dalam konteks ini, artefak yang dikembangkan adalah sistem ekstraksi data adaptif dari form PDF semi-terstruktur dengan mekanisme pembelajaran berbasis masukan pengguna.

3.1.1 Kerangka Design Science Research

Metodologi DSR yang digunakan dalam penelitian ini mengadopsi kerangka kerja yang diusulkan oleh (Peffers et al., 2007), yang terdiri dari enam langkah utama:

1. Identifikasi Masalah dan Motivasi: Pada tahap ini, penelitian mendefinisikan masalah spesifik dalam ekstraksi data dari form PDF PDF template dan menjelaskan pentingnya solusi. Masalah utama yang diidentifikasi adalah:
    a. Kesulitan dalam mengekstrak data dari form PDF dengan format yang bervariasi
    b. Kebutuhan akan sistem yang dapat beradaptasi dengan berbagai jenis dokumen
    c. Keterbatasan pendekatan berbasis aturan statis dan kebutuhan data pelatihan yang besar untuk pendekatan machine learning murni
2. Definisi Tujuan Solusi: Berdasarkan masalah yang diidentifikasi, penelitian menetapkan tujuan spesifik untuk artefak yang dikembangkan:
    a. Mengembangkan sistem pembelajaran adaptif berbasis HITL yang dapat beradaptasi dengan berbagai jenis PDF template
    b. Mengintegrasikan pendekatan berbasis aturan dan machine learning (CRF) dalam sistem HITL yang koheren
    c. Merancang mekanisme interaksi HITL yang efektif untuk mengoptimalkan pembelajaran adaptif dari expertise pengguna

    d. Mencapai peningkatan akurasi yang signifikan melalui feedback loop HITL dan pembelajaran berkelanjutan
3. Perancangan dan Pengembangan: Tahap ini melibatkan perancangan dan implementasi artefak, termasuk:
    a. Arsitektur sistem pembelajaran adaptif berbasis HITL
    b. Komponen analisis template dan ekstraksi berbasis aturan
    c. Model pembelajaran adaptif berbasis Conditional Random Fields (CRF)
    d. Mekanisme interaksi HITL dan antarmuka validasi pengguna
    e. Integrasi komponen-komponen dalam sistem HITL yang koheren
4. Demonstrasi: Pada tahap ini, penelitian menunjukkan penggunaan artefak untuk menyelesaikan satu atau lebih contoh masalah. Demonstrasi meliputi:
    a. Ekstraksi data dari berbagai jenis PDF template dengan format yang berbeda menggunakan sistem HITL
    b. Simulasi proses interaksi HITL dan pembelajaran adaptif
    c. Visualisasi peningkatan akurasi melalui feedback loop HITL
5. Evaluasi: Tahap ini melibatkan pengamatan dan pengukuran seberapa baik artefak mendukung solusi untuk masalah. Evaluasi meliputi:
    a. Metrik Akurasi Ekstraksi Data: - Precision, Recall, dan F1-score untuk field-level extraction - Character-level accuracy untuk extracted values - Template-level success rate (percentage of completely correct extractions) - Error rate analysis berdasarkan field types (text, numeric, date)
    b. Metrik Pembelajaran Adaptif: - Learning curve analysis: akurasi improvement per iteration - Convergence rate: jumlah feedback iterations untuk mencapai stable performance - Retention rate: kemampuan sistem mempertahankan learned knowledge - Adaptation speed: time to adapt to new template variations
    c. Metrik Kemampuan Adaptasi: - Cross-template generalization: performance pada unseen templates - Variation tolerance:

accuracy degradation dengan template modifications - Robustness metrics: performance consistency across different document qualities - Scalability assessment: performance dengan increasing template complexity

   d. Metrik User Experience dan Efisiensi: - User burden metrics: average correction time per document - Feedback quality: correlation antara user corrections dan system improvement - Interaction efficiency: ratio of successful learning per user intervention - System responsiveness: processing time untuk feedback integration - User satisfaction scores melalui usability testing

6. Komunikasi: Tahap terakhir melibatkan komunikasi masalah, solusi, dan efektivitasnya kepada peneliti dan praktisi yang relevan. Komunikasi meliputi:
   a. Dokumentasi hasil penelitian dalam bentuk tesis
   b. Publikasi ilmiah tentang pendekatan dan hasil yang diperoleh
   c. Diseminasi kode sumber dan dataset untuk penelitian lebih lanjut

3.1.2 Pendekatan Iteratif

Penelitian ini mengadopsi pendekatan iteratif dalam pengembangan dan evaluasi artefak, dengan beberapa siklus iterasi:

1. Iterasi 1: Analisis Template dan Ekstraksi Berbasis Aturan
   a. Pengembangan komponen analisis template PDF
   b. Implementasi ekstraksi berbasis aturan dengan pattern matching
   c. Evaluasi baseline akurasi ekstraksi
2. Iterasi 2: Mekanisme Interaksi HITL dan Antarmuka Pengguna
   a. Pengembangan antarmuka validasi dan koreksi berbasis HITL
   b. Implementasi mekanisme pengumpulan feedback pengguna
   c. Evaluasi efektivitas interaksi HITL
3. Iterasi 3: Model Pembelajaran Adaptif CRF
   a. Implementasi model Conditional Random Fields (CRF)

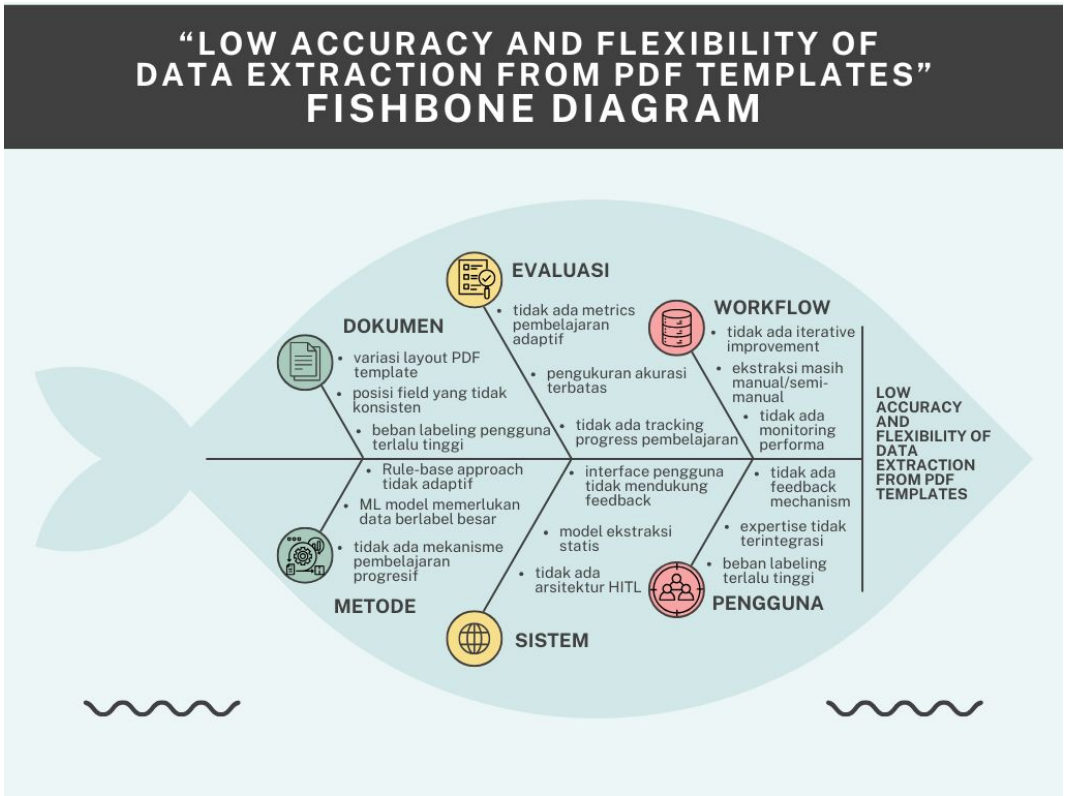
b. Integrasi dengan feedback loop HITL
c. Evaluasi peningkatan akurasi melalui pembelajaran adaptif

4. Iterasi 4: Integrasi dan Optimasi Sistem HITL
    a. Integrasi semua komponen dalam sistem HITL yang koheren
    b. Optimasi performa dan efisiensi interaksi HITL
    c. Evaluasi komprehensif efektivitas sistem pembelajaran adaptif

Pendekatan iteratif ini memungkinkan penyempurnaan bertahap dari artefak berdasarkan hasil evaluasi pada setiap iterasi.

3.2 Identifikasi Permasalahan

Untuk merumuskan strategi perancangan sistem yang tepat, dilakukan analisis akar masalah menggunakan pendekatan Fishbone (Ishikawa & Ishikawa, 1987). Tujuannya adalah mengidentifikasi faktor-faktor penyebab utama dari kegagalan proses ekstraksi data dari dokumen PDF semi-terstruktur, seperti ditunjukkan pada Gambar 3.1.



Gambar 3-1 Diagram Fishbone Akar Masalah Ekstraksi Data PDF

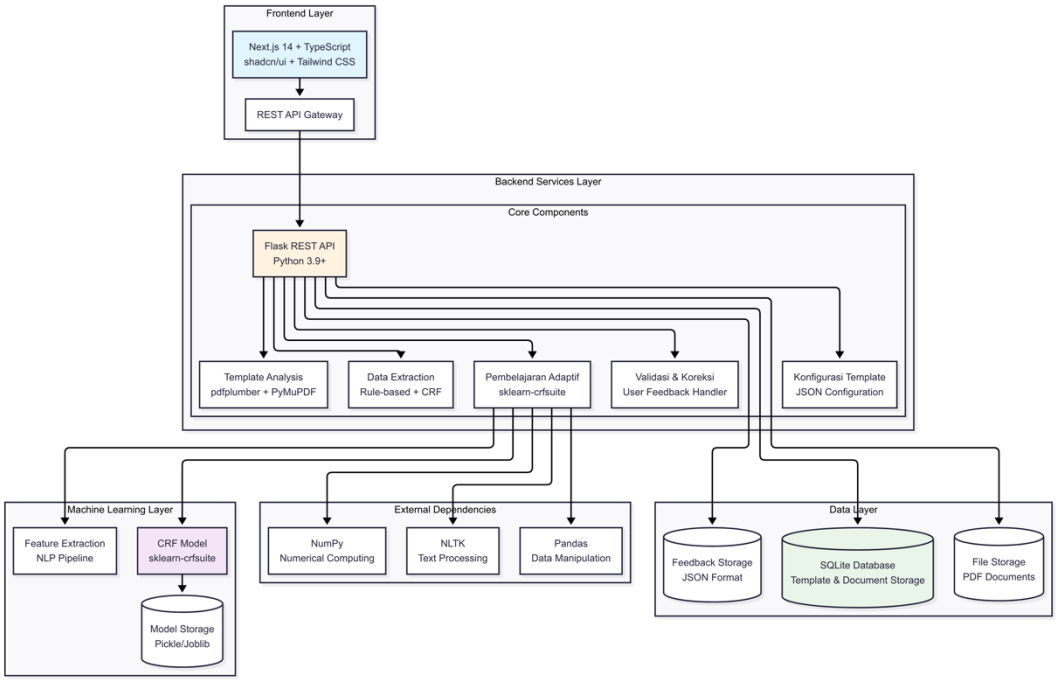
Dari diagram tersebut, dapat disimpulkan bahwa kegagalan ekstraksi data dari PDF template disebabkan oleh enam faktor utama: (1) keterbatasan metode ekstraksi tradisional yang tidak adaptif, (2) minimnya integrasi expertise pengguna dalam proses pembelajaran, (3) variasi struktur dan layout PDF template yang kompleks, (4) ketiadaan arsitektur sistem yang mendukung Human-in-the-Loop, (5) proses ekstraksi yang statis tanpa mekanisme perbaikan berkelanjutan, dan (6) kurangnya framework evaluasi untuk mengukur efektivitas pembelajaran adaptif. Oleh karena itu, penelitian ini mengusulkan sistem pembelajaran adaptif berbasis Human-in-the-Loop (HITL) yang mengintegrasikan pendekatan rule-based dengan Conditional Random Fields (CRF) sebagai baseline, serta memanfaatkan feedback pengguna untuk mendukung pembelajaran adaptif secara progresif.

3.3 Arsitektur Sistem

Sistem ekstraksi data adaptif yang dikembangkan terdiri dari beberapa komponen utama yang saling terintegrasi dalam arsitektur modular. Arsitektur ini dirancang untuk mendukung alur kerja ekstraksi data adaptif, dari analisis template hingga pembelajaran dari umpan balik pengguna.

3.3.1 Arsitektur Keseluruhan

Arsitektur sistem secara keseluruhan ditunjukkan pada Gambar 3.2, yang mengilustrasikan komponen utama dan interaksi antar komponen.



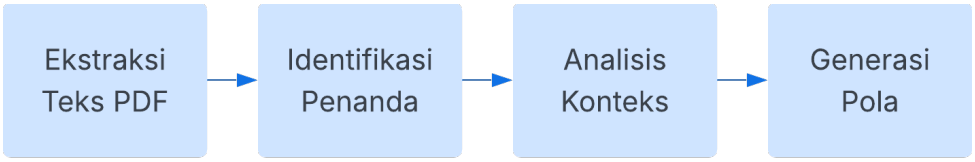
Gambar 3-2 Arsitektur Keseluruhan Sistem Ekstraksi Data Adaptif

Alur kerja utama dalam sistem meliputi:

1. Analisis Template: Menganalisis dokumen template untuk mengidentifikasi bidang dan mengusulkan pola ekstraksi awal.
2. Konfigurasi Template: Menyimpan dan mengelola konfigurasi template, termasuk bidang, pola, dan metadata.
3. Ekstraksi Data: Mengekstrak data dari dokumen PDF menggunakan pola atau model machine learning.
4. Validasi & Koreksi: Memungkinkan pengguna memvalidasi dan mengoreksi hasil ekstraksi.
5. Pembelajaran Adaptif: Melatih dan memperbarui model berdasarkan umpan balik pengguna.

3.3.2 Komponen Analisis Template

Komponen analisis template bertanggung jawab untuk menganalisis struktur dokumen template dan mengidentifikasi bidang-bidang yang perlu diekstraksi. Arsitektur internal komponen ini ditunjukkan pada Gambar 3.3.



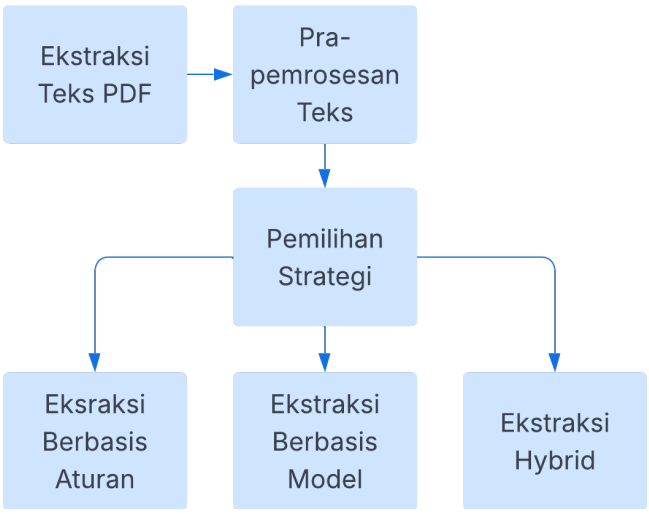
Gambar 3-3 Arsitektur Komponen Analisis Template

Subkomponen utama meliputi:

1. Ekstraksi Teks PDF: Mengekstrak teks dari dokumen PDF template menggunakan pustaka pdfplumber untuk mendapatkan informasi posisi koordinat yang akurat.
2. Identifikasi Penanda: Mengidentifikasi penanda variabel dalam teks template, seperti {nama} atau ${tanggal_lahir}.
3. Analisis Konteks: Menganalisis teks sebelum dan sesudah penanda untuk memahami konteks bidang.
4. Generasi Pola: Menghasilkan pola ekstraksi awal berdasarkan analisis konteks dan karakteristik bidang.

3.3.3 Komponen Ekstraksi Data

Komponen ekstraksi data bertugas untuk mengekstrak informasi dari dokumen PDF berdasarkan konfigurasi template. Arsitektur internal komponen ini ditunjukkan pada Gambar 3.4.



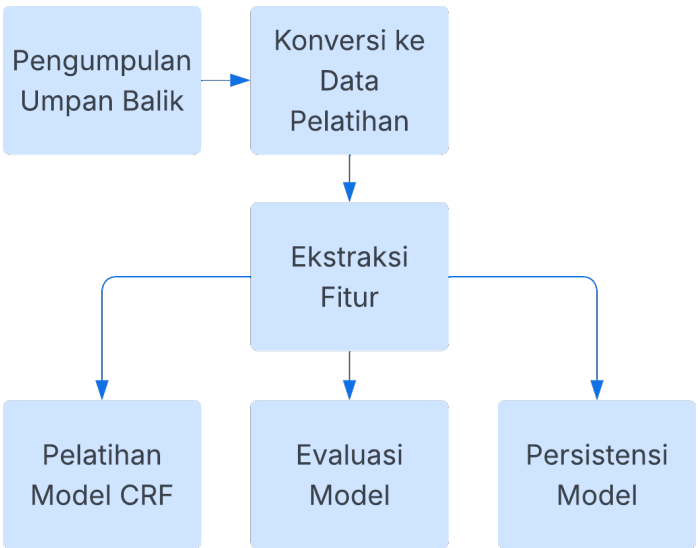
Gambar 3-4 Arsitektur Komponen Ekstraksi Data

Subkomponen utama meliputi:

1. Ekstraksi Teks PDF: Mengekstrak teks dari dokumen PDF yang akan diekstraksi datanya.
2. Pra-pemrosesan Teks: Melakukan normalisasi dan pembersihan teks untuk memudahkan ekstraksi.
3. Pemilihan Strategi: Memilih strategi ekstraksi yang sesuai berdasarkan ketersediaan model dan karakteristik dokumen.
4. Ekstraksi Berbasis Aturan: Mengekstrak data menggunakan ekspresi reguler dan aturan posisional.
5. Ekstraksi Berbasis Model: Mengekstrak data menggunakan model Conditional Random Fields yang telah dilatih.
6. Ekstraksi Hybrid: Menggabungkan hasil dari ekstraksi berbasis aturan dan model untuk meningkatkan akurasi.

3.3.4 Komponen Pembelajaran Adaptif

Komponen pembelajaran adaptif bertanggung jawab untuk melatih dan memperbarui model machine learning berdasarkan umpan balik pengguna. Arsitektur internal komponen ini ditunjukkan pada Gambar 3.5.



Gambar 3-5 Arsitektur Komponen Pembelajaran Adaptif

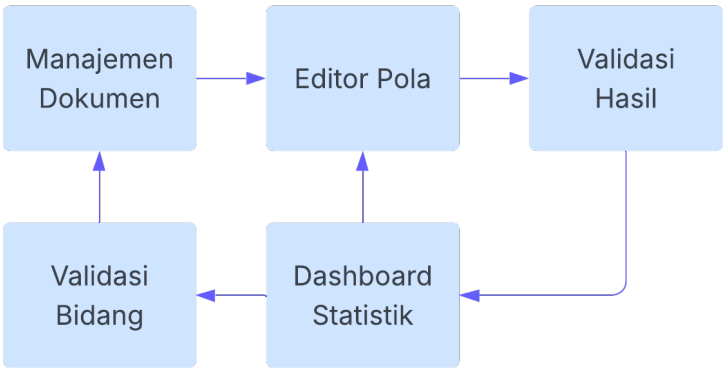
Subkomponen utama meliputi:

1. Pengumpulan Umpan Balik: Mengumpulkan dan menyimpan koreksi yang diberikan oleh pengguna.

2. Konversi ke Data Pelatihan: Mengubah umpan balik menjadi format yang sesuai untuk pelatihan model.
3. Ekstraksi Fitur: Mengekstrak fitur yang relevan dari teks untuk digunakan dalam model CRF.
4. Pelatihan Model CRF: Melatih atau memperbarui model Conditional Random Fields dengan data umpan balik.
5. Evaluasi Model: Mengevaluasi kinerja model pada data pengujian untuk memantau peningkatan.
6. Persistensi Model: Menyimpan dan memuat model yang telah dilatih untuk penggunaan di masa mendatang.

3.3.5 Antarmuka Pengguna

Antarmuka pengguna menyediakan cara bagi pengguna untuk berinteraksi dengan sistem, termasuk mengunggah dokumen, melihat dan mengedit pola ekstraksi, memvalidasi hasil, dan memvisualisasikan bidang yang diekstraksi. Arsitektur antarmuka pengguna ditunjukkan pada Gambar 3.6.



Gambar 3-6 Arsitektur Antarmuka Pengguna

Komponen utama antarmuka pengguna meliputi:

1. Manajemen Dokumen: Antarmuka untuk mengunggah, melihat, dan mengelola dokumen template dan dokumen yang akan diekstraksi.
2. Editor Pola: Antarmuka untuk melihat dan mengedit pola ekstraksi untuk setiap bidang.
3. Validasi Hasil: Antarmuka untuk memvalidasi dan mengoreksi hasil ekstraksi.

4. Visualisasi Bidang: Komponen untuk memvisualisasikan bidang yang diekstraksi pada dokumen PDF.
5. Dashboard Statistik: Tampilan statistik tentang akurasi ekstraksi dan peningkatan performa seiring waktu.

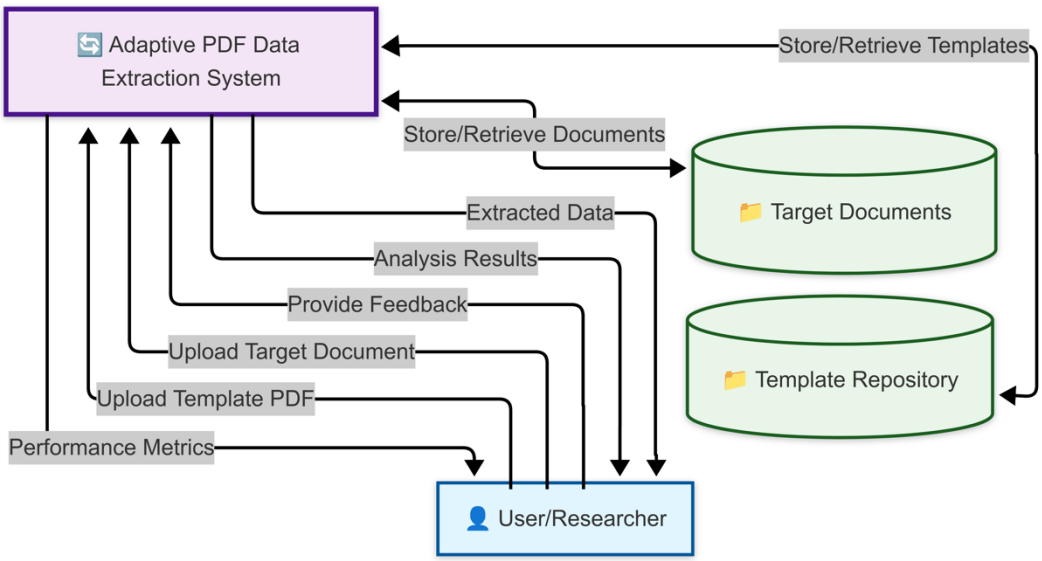
Antarmuka pengguna diimplementasikan sebagai aplikasi web modern menggunakan Next.js 14 dengan TypeScript sebagai frontend dan Flask sebagai REST API backend. UI components menggunakan shadcn/ui dengan Tailwind CSS untuk styling, dengan fokus pada pengalaman pengguna yang intuitif, responsif, dan type-safe.

3.4 Data Flow Diagram (DFD) dan Model Data

Untuk memberikan pemahaman yang lebih komprehensif tentang alur data dan struktur data dalam sistem, bagian ini menyajikan Data Flow Diagram (DFD) dan model data yang mendukung arsitektur sistem ekstraksi data adaptif.

3.4.1 Data Flow Diagram

DFD Level 0 menggambarkan sistem ekstraksi data PDF adaptif sebagai satu proses utama dengan entitas eksternal yang berinteraksi dengannya:

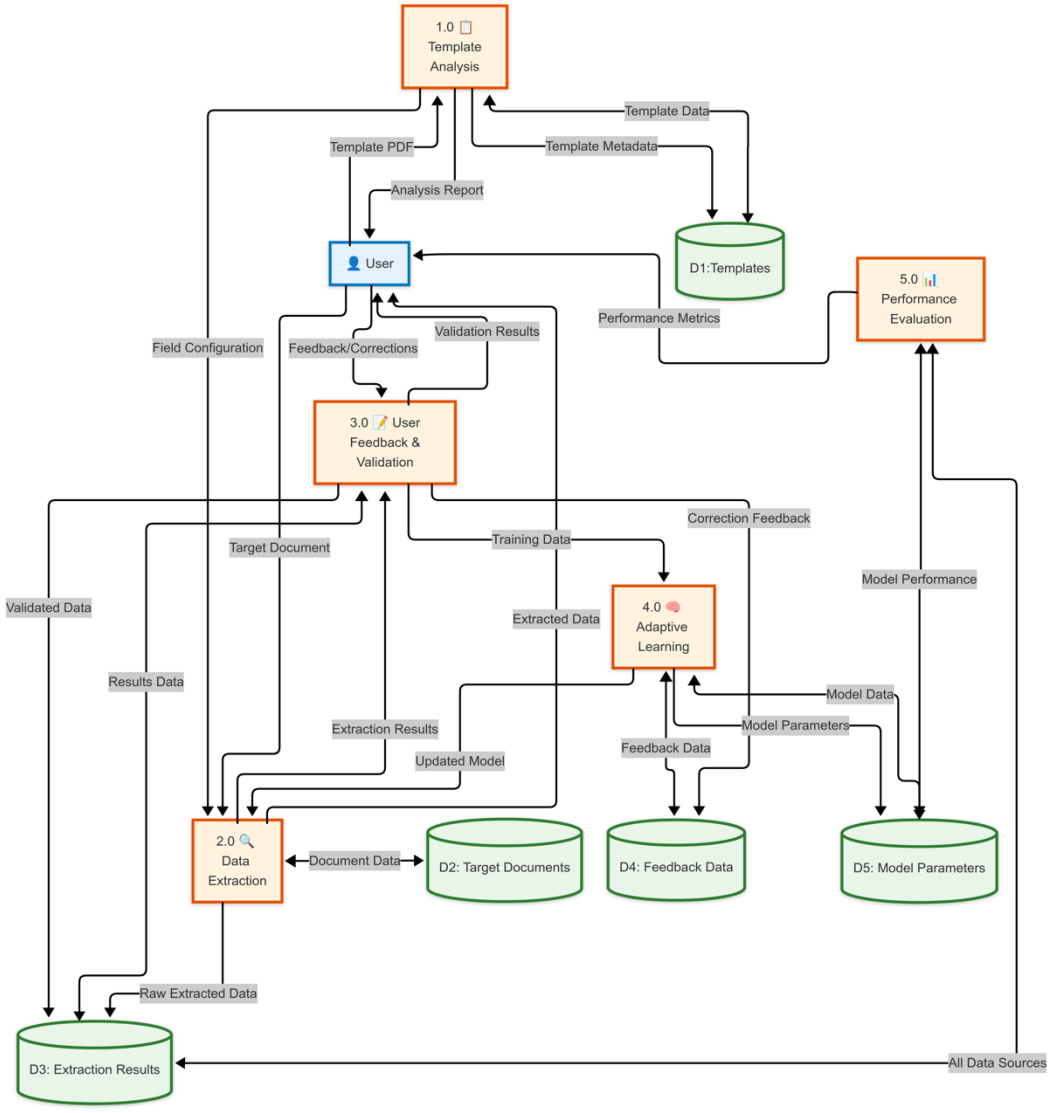


Gambar 3-7 Data Flow Diagram Level 0 - Context Diagram

Context diagram menunjukkan interaksi sistem dengan entitas eksternal:

- User/Researcher: Pengguna yang mengunggah template, dokumen target, dan memberikan feedback
- Template Repository: Penyimpanan template PDF dan konfigurasi analisis
- Target Documents: Penyimpanan dokumen target yang akan diekstraksi datanya

DFD Level 1 mendekomposisi sistem menjadi lima proses utama yang sesuai dengan arsitektur modular:



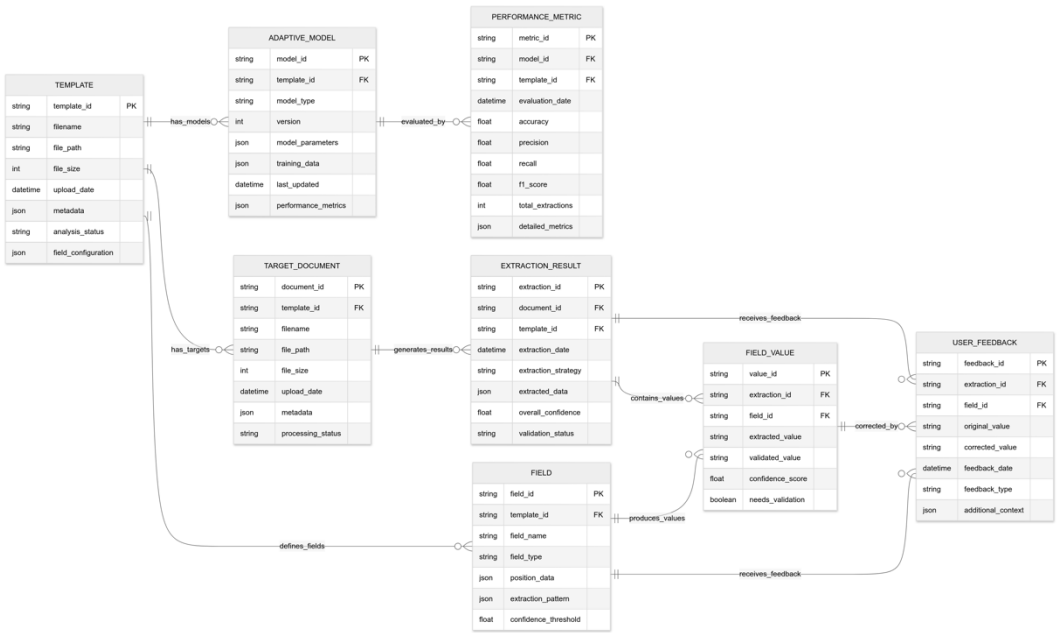
Gambar 3-8 Data Flow Diagram Level 1 - Process Decomposition

DFD Level 1 menunjukkan detail proses utama sistem:

1. Template Analysis (1.0): Menganalisis struktur template PDF dan mengidentifikasi field
2. Data Extraction (2.0): Mengekstrak data dari dokumen target menggunakan konfigurasi template
3. User Feedback & Validation (3.0): Memvalidasi hasil ekstraksi dan mengumpulkan feedback pengguna
4. Adaptive Learning (4.0): Melatih dan memperbarui model berdasarkan feedback
5. Performance Evaluation (5.0): Mengevaluasi dan melaporkan performa sistem.

3.4.2 Model Data Sistem

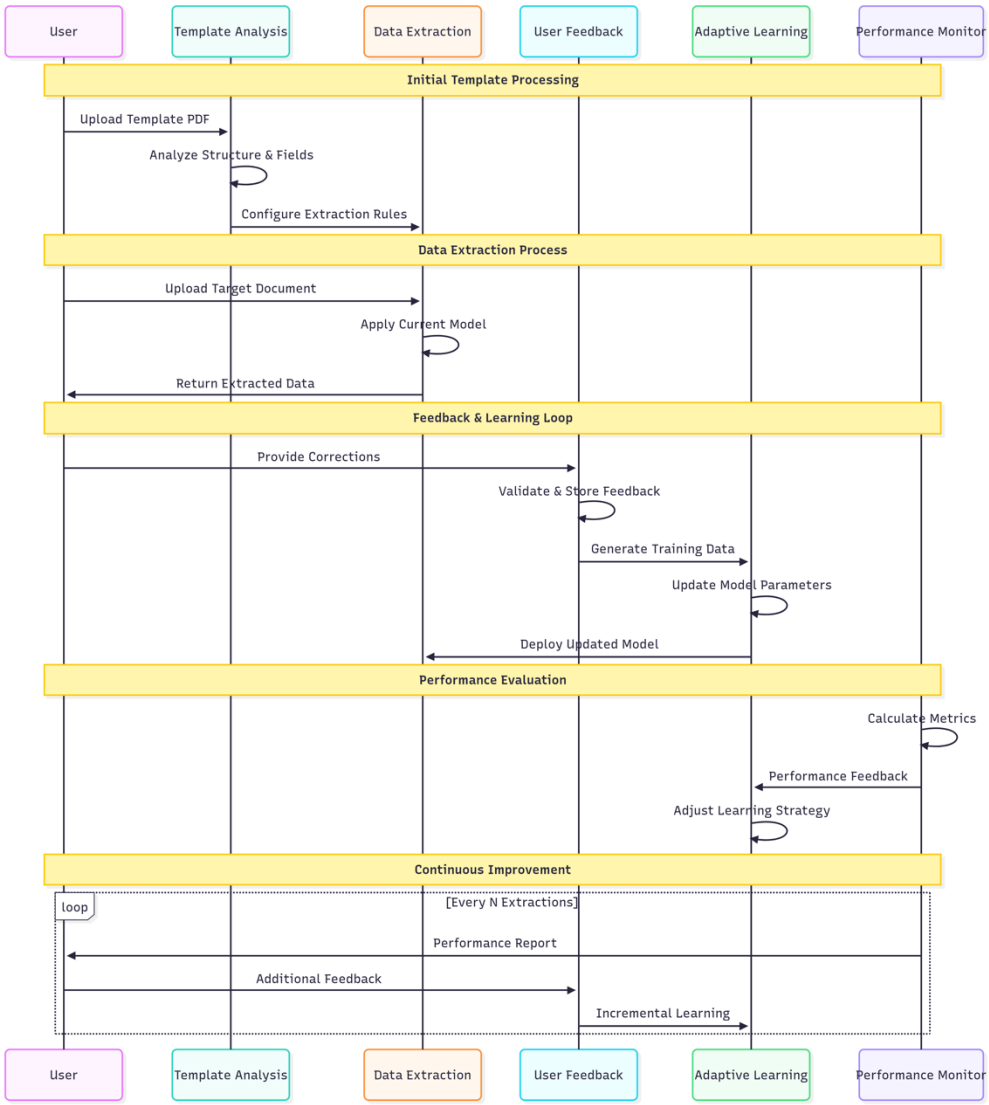
Model data sistem menggambarkan entitas utama dan hubungan antar entitas dalam sistem ekstraksi data adaptif:



Gambar 3-9 Entity Relationship Diagram - Model Data Sistem

3.4.3 Alur Pembelajaran Adaptif

Untuk menjelaskan mekanisme pembelajaran adaptif secara detail, berikut adalah sequence diagram yang menggambarkan alur feedback loop dan proses update model:



Gambar 3-10 Sequence Diagram - Alur Pembelajaran Adaptif

3.4.4 Kamus Data

Tabel berikut menjelaskan elemen data kunci dan relevansinya terhadap tujuan penelitian:

Entitas	Atribut	Tipe	Deskripsi	Relevansi Penelitian
Template	field_configuration	JSON	Field yang terdeteksi dan pola ekstraksi	Inti dari pembelajaran adaptif

Field	extraction_pattern	JSON	Pola berbasis aturan + ML	Implementasi pendekatan hybrid
ExtractionResult	extraction_strategy	String	Rule/CRF/Hybrid	Perbandingan strategi penelitian
UserFeedback	corrected_value	String	Koreksi pengguna	Data pelatihan untuk adaptasi
AdaptiveModel	model_parameters	JSON	Parameter model ML	Mekanisme penyimpanan pembelajaran
PerformanceMetric	Accuracy / precision / recall	Float	Metrik evaluasi	Pengukuran efektivitas penelitian

Tabel 3-1 Kamus elemen data kunci

3.4.5 Integrasi dengan Metodologi Penelitian

DFD dan model data yang disajikan memiliki relevansi langsung dengan metodologi Design Science Research (DSR) yang digunakan:

1. DFD Level 0: Menunjukkan sistem adaptif sebagai artefak utama dengan feedback loop yang menjadi kontribusi penelitian
2. DFD Level 1: Mendetailkan lima komponen utama yang sesuai dengan arsitektur modular yang dirancang
3. Model Data: Struktur data yang mendukung pendekatan hybrid dan pembelajaran adaptif
4. Sequence Diagram: Proses iterative improvement yang menjadi core contribution penelitian

Diagram-diagram ini berfokus pada aspek document processing methodology dan adaptive learning mechanism, bukan pada traditional business process atau database design, sehingga sesuai dengan nature penelitian DSR dalam bidang ekstraksi data dokumen.

3.5 Desain Rinci Komponen dan Spesifikasi Teknis

Desain sistem ekstraksi data adaptif dirancang dengan menggunakan arsitektur modular dan teknologi yang mendukung pembelajaran adaptif. Bagian ini menjelaskan spesifikasi desain komponen-komponen utama sistem dan interaksi antar komponen secara konseptual.

3.5.1 Teknologi dan Pustaka

Sistem dirancang menggunakan teknologi dan pustaka berikut:

Bahasa Pemrograman: Python 3.9+ dipilih karena kekayaan pustaka untuk pemrosesan dokumen dan machine learning, serta kemudahan pengembangan dan debugging.

1. Pustaka Ekstraksi PDF:
    a. pdfplumber untuk ekstraksi teks dengan informasi posisi koordinat yang akurat
    b. Pillow (PIL) untuk manipulasi gambar dan visualisasi field highlighting
    c. pdf2image untuk konversi PDF ke format gambar (PNG/JPEG)
2. Framework Backend:
    a. Flask sebagai web framework untuk REST API backend
    b. Flask-CORS untuk menangani Cross-Origin Resource Sharing
    c. Werkzeug untuk file upload handling dan security
3. Framework Frontend:
    a. Next.js 14 sebagai React framework dengan App Route
    b. TypeScript untuk type safety dan better development experience
    c. Tailwind CSS untuk utility-first CSS styling
    d. shadcn/ui untuk komponen UI yang konsisten dan modern
4. Pustaka Machine Learning:
    a. sklearn-crfsuite untuk implementasi Conditional Random Fields
    b. scikit-learn untuk evaluasi model dan pemrosesan data
    c. numpy untuk operasi numerik dan array processing
5. Pustaka Data Processing:
    a. pandas untuk manipulasi dan analisis data terstruktur

    b. regex (re) untuk pattern matching dan text processing
    c. json untuk serialisasi data dan konfigurasi sistem
6. Development Tools:
    a. pytest untuk unit testing dan integration testing
    b. black untuk code formatting dan consistency
    c. flake8 untuk code linting dan quality assurance
7. Deployment & Infrastructure:
    a. File-based storage untuk prototype dan development
    b. JSON-based configuration untuk template dan model parameters
    c. RESTful API architecture untuk modular system integration

Pemilihan teknologi ini didasarkan pada kebutuhan sistem untuk:

- Mengekstrak teks dari dokumen PDF dengan akurasi tinggi
- Melakukan pemrosesan teks dan pattern recognition
- Melatih dan mengevaluasi model machine learning adaptif
- Menyediakan antarmuka pengguna yang modern dan responsif
- Mendukung workflow penelitian yang iteratif dan eksperimental
- Memfasilitasi deployment dan maintenance yang mudah

3.5.2 Spesifikasi Teknis dan Persyaratan Sistem

Persyaratan Hardware Minimum:

1. CPU: Intel Core i5 atau AMD Ryzen 5 (minimum 4 cores)
2. RAM: 8 GB (recommended 16 GB untuk processing dokumen besar)
3. Storage: 10 GB free space untuk sistem dan temporary files
4. GPU: Tidak diperlukan (CPU-based processing)

Persyaratan Software:

1. Operating System: Windows 10+, macOS 10.15+, atau Linux Ubuntu 18.04+
2. Python: Version 3.9 atau lebih tinggi
3. Node.js: Version 18+ untuk frontend development

4. Browser: Chrome 90+, Firefox 88+, Safari 14+ untuk web interface

Performance Benchmarks:

1. Template Analysis: < 5 detik untuk dokumen PDF hingga 10 halaman
2. Data Extraction: < 3 detik per dokumen untuk rule-based, < 10 detik untuk CRF-based
3. Model Training: < 30 detik untuk incremental learning dengan 20 feedback samples
4. API Response Time: < 2 detik untuk operasi CRUD standard
5. Concurrent Users: Mendukung hingga 10 concurrent users untuk prototype

Scalability Limits:

1. Document Size: Maksimal 50 MB per PDF file
2. Template Complexity: Hingga 50 fields per template
3. Feedback Volume: Optimal performance dengan < 1000 feedback entries per field
4. Storage: File-based storage suitable untuk < 10,000 documents

3.5.3 Desain Komponen Analisis Template

Komponen analisis template bertanggung jawab untuk menganalisis struktur dokumen template dan mengidentifikasi bidang-bidang yang perlu diekstraksi. Desain metodologi komponen ini meliputi:

Metodologi Analisis Template:

1. Ekstraksi Struktural: Pendekatan untuk mengekstrak struktur dokumen PDF dengan mempertahankan informasi layout dan positioning
2. Identifikasi Penanda: Metodologi pattern recognition untuk mengidentifikasi marker variabel dengan berbagai format konvensi
3. Analisis Kontekstual: Pendekatan untuk menganalisis konteks di sekitar penanda untuk memahami semantik bidang
4. Generasi Konfigurasi: Framework untuk menghasilkan konfigurasi ekstraksi berdasarkan analisis struktural dan kontekstual

Keluaran Metodologi:

1. Framework untuk deteksi field dengan metadata positioning
2. Metodologi konfigurasi ekstraksi per field type
3. Strategi generasi pola ekstraksi berbasis aturan

3.5.4 Desain Komponen Ekstraksi Data

Komponen ekstraksi data dirancang dengan metodologi multi-strategi yang dapat digunakan secara individual atau kombinasi:

Metodologi Ekstraksi:

1. Pendekatan Berbasis Aturan (Rule-Based)
   - Metodologi pattern matching dengan positioning constraints
   - Cocok untuk dokumen dengan struktur konsisten
   - Memberikan predictability dan interpretability tinggi
   - Framework confidence scoring berbasis pattern matching
2. Pendekatan Berbasis Model (CRF)
   - Metodologi sequence labeling dengan probabilistic framework
   - Kemampuan adaptasi terhadap variasi format dokumen
   - Memerlukan strategi training data preparation
   - Framework confidence scoring berbasis probabilistic inference
3. Pendekatan Hybrid
   - Metodologi kombinasi multi-approach dengan intelligent selection
   - Strategy untuk memilih hasil optimal berdasarkan confidence metrics
   - Framework robustness dengan fallback mechanisms
   - Pendekatan ensemble untuk handling edge cases

Framework Proses Ekstraksi:

- Metodologi tokenisasi dan preprocessing
- Strategy feature extraction untuk model-based approaches
- Framework aplikasi multi-strategy extraction
- Metodologi confidence scoring dan result selection

- Pendekatan validasi dan normalisasi output

3.5.5 Desain Komponen Pembelajaran Aktif

Komponen pembelajaran adaptif dirancang dengan metodologi Human-in-the-Loop (HITL) untuk pembelajaran berkelanjutan melalui interaksi pengguna.

Framework Metodologi HITL:

1. Metodologi Pengumpulan Feedback:
    a. Framework interface design untuk user correction workflows
    b. Strategi structured feedback collection dengan quality metadata
    c. Pendekatan assessment untuk feedback reliability dan consistency
2. Framework Konversi Training Data:
    a. Metodologi transformasi user corrections ke format pembelajaran
    b. Strategi feature extraction dari document context
    c. Pendekatan integration dengan historical training data
3. Metodologi Pembelajaran Inkremental:
    a. Framework incremental learning untuk model adaptation
    b. Strategi kombinasi historical dan new training data
    c. Pendekatan performance monitoring dan model validation
4. Framework Strategi Adaptif:
    a. Metodologi intelligent triggering berdasarkan feedback metrics
    b. Pendekatan performance degradation detection
    c. Framework adaptive learning parameter adjustment

Kriteria Motodologis Update Model:

- Framework threshold determination untuk feedback quantity
- Metodologi quality scoring untuk feedback reliability
- Strategi performance degradation indicators
- Pendekatan time-based update scheduling

Keluaran Framework:

- Metodologi model improvement dengan accuracy enhancement
- Framework performance metrics dan learning curve analysis
- Strategi feedback quality assessment

3.5.6 Desain Antarmuka Pengguna

Antarmuka pengguna dirancang menggunakan teknologi web modern untuk memberikan pengalaman yang intuitif dan responsif dalam proses HITL.

Arsitektur Frontend:

1. Technology Stack:
   - Next.js dengan TypeScript untuk type safety
   - shadcn/ui untuk consistent UI components
   - Tailwind CSS untuk responsive styling
2. Component Architecture:
   - Layout components untuk navigation dan structure
   - Page components untuk major workflows
   - Feature components untuk specialized functionality
   - Reusable UI components untuk consistency

Key User Interface Components:

1. Template Analysis Interface:
   - File upload dengan drag-and-drop support
   - Visual field highlighting pada PDF template
   - Configuration options untuk analysis strategies
2. Data Extraction Interface:
   - Target document upload dan processing
   - Real-time extraction progress monitoring
   - Results display dengan confidence indicators
3. Validation dan Feedback Interface:
   - Side-by-side comparison untuk original vs extracted values
   - Inline editing untuk user corrections
   - Batch validation untuk multiple fields

- Feedback quality indicators

4. Performance Monitoring Dashboard:
    - Learning curve visualization
    - Accuracy metrics over time
    - System performance indicators

Prinsip Desain User Experience:

- Metodologi intuitive workflow design dengan clear navigation patterns
- Framework immediate feedback untuk user action responsiveness
- Pendekatan progressive disclosure untuk complex feature management
- Strategi accessibility compliance untuk inclusive design principles

Framework Komponen Interface:

1. Metodologi Document Management:
    a. Framework upload dan management workflow untuk PDF documents
    b. Strategi document organization dan retrieval
    c. Pendekatan document preview dan metadata handling
2. Framework Pattern Editor:
    a. Metodologi pattern visualization dan editing interface
    b. Strategi syntax highlighting dan validation
    c. Pendekatan pattern testing dan optimization workflow
3. Framework Validation Interface:
    a. Metodologi result presentation dan correction workflow
    b. Strategi side-by-side comparison untuk validation efficiency
    c. Pendekatan feedback collection dan quality assurance
4. Framework Field Visualization:
    a. Metodologi overlay visualization pada PDF documents
    b. Strategi color coding dan field differentiation
    c. Pendekatan interactive navigation dan field management

5. Framework Performance Dashboard:
    a. Metodologi metrics visualization dan trend analysis
    b. Strategi learning curve presentation dan interpretation
    c. Pendekatan comparative analysis untuk strategy evaluation

3.6 Pengumpulan dan Pengolahan Data

Pengumpulan dan pengolahan data dalam penelitian ini melibatkan beberapa jenis data yang berbeda, mulai dari dokumen PDF template, dokumen target, hingga feedback pengguna untuk pembelajaran adaptif.

Penelitian ini menggunakan berbagai jenis data untuk pengembangan dan evaluasi sistem ekstraksi data adaptif. Bagian ini menjelaskan metode pengumpulan dan pengolahan data yang digunakan dalam penelitian.

3.6.1 Jenis Data

Penelitian ini menggunakan beberapa jenis data, antara lain:

1. Dokumen Template: Dokumen PDF yang berisi penanda variabel yang menunjukkan bidang yang perlu diekstraksi. Dokumen template ini menjadi dasar untuk analisis struktur dan pembuatan pola ekstraksi awal.
2. Dokumen Terisi: Dokumen PDF yang telah diisi dengan data aktual, yang akan diekstrak oleh sistem. Dokumen ini digunakan untuk evaluasi akurasi ekstraksi dan sebagai sumber data pelatihan setelah divalidasi.
3. Data Umpan Balik: Data koreksi yang diberikan oleh pengguna terhadap hasil ekstraksi. Data ini mencakup nilai yang diekstrak oleh sistem dan nilai yang dikoreksi oleh pengguna.
4. Data Pelatihan: Data yang digunakan untuk melatih model machine learning, yang dihasilkan dari konversi data umpan balik.
5. Data Pengujian: Data yang digunakan untuk mengevaluasi kinerja model, yang dipisahkan dari data pelatihan untuk memastikan evaluasi yang objektif.

3.6.2 Sumber Data

Data yang digunakan dalam penelitian ini berasal dari beberapa sumber:

1. Template Dokumen Simulasi: Dokumen PDF template yang dibuat khusus untuk penelitian ini, menyerupai format dokumen administratif seperti formulir pendaftaran, surat keterangan, dan dokumen identitas. Template ini dirancang untuk representatif terhadap kasus penggunaan nyata tanpa menggunakan data sensitif.
2. Dokumen Terisi Simulasi: Dokumen yang dihasilkan dari template dengan menggunakan data dummy yang bervariasi untuk menguji kemampuan adaptasi sistem terhadap variasi konten dan format.
3. Umpan Balik Pengguna: Data koreksi yang diberikan oleh pengguna selama penggunaan sistem. Data ini merupakan sumber utama untuk pembelajaran adaptif.
4. Umpan Balik Simulasi: Data koreksi yang dihasilkan secara otomatis untuk simulasi proses pembelajaran dalam jumlah besar tanpa memerlukan interaksi pengguna yang ekstensif.

3.6.3 Metodologi Pengumpulan Data

Pengumpulan data dilakukan melalui beberapa metodologi yang sistematis:

1. Metodologi Pengumpulan Dokumen: Framework untuk mengumpulkan dokumen template dan dokumen terisi melalui strategi sampling yang representatif untuk penelitian.
2. Metodologi Pengumpulan Feedback: Framework pengumpulan umpan balik melalui desain interface validasi yang memungkinkan pengguna memberikan koreksi terstruktur pada hasil ekstraksi.
3. Metodologi Simulasi Feedback: Pendekatan untuk mempercepat pengumpulan data pelatihan melalui mekanisme simulasi yang menghasilkan data koreksi berdasarkan variasi yang dikonfigurasi secara sistematis.

3.6.4 Pra-pemrosesan Data

Sebelum digunakan untuk pelatihan model atau evaluasi, data melalui beberapa tahap pra-pemrosesan yang sistematis:

1. Framework Normalisasi Teks:
    a. Metodologi Standardisasi: Pendekatan sistematis untuk konsistensi encoding dan format
    b. Strategi Kapitalisasi: Framework normalisasi berdasarkan tipologi field
    c. Pendekatan Whitespace: Metodologi standardisasi spasi dan formatting
    d. Framework Karakter Khusus: Strategi normalisasi diacritics dan special characters
    e. Metodologi Number Formatting: Pendekatan standardisasi format numerik dan temporal
2. Framework Tokenisasi:
    a. Metodologi Word-level: Strategi tokenisasi untuk teks bahasa Indonesia
    b. Pendekatan Subword: Framework handling compound words dan abbreviations
    c. Strategi Punctuation: Metodologi pemisahan dan preservasi punctuation
    d. Framework Boundary Detection: Pendekatan identifikasi multi-word entities
3. Framework Pelabelan BIO:
    a. Metodologi Sequence Labeling: Pendekatan Beginning-Inside-Outside tagging
    b. Strategi Entity Recognition: Framework untuk named entity identification
    c. Pendekatan Boundary Detection: Metodologi untuk entity boundary determination
4. Framework Ekstraksi Fitur:
    a. Metodologi Lexical: Strategi ekstraksi fitur berbasis kata

b. Pendekatan Orthographic: Framework fitur berbasis pola tipografi
c. Strategi Contextual: Metodologi fitur berbasis konteks
d. Framework Semantic: Pendekatan fitur berbasis semantic
e. Metodologi Layout: Strategi fitur berbasis struktur PDF

5. Framework Pembagian Data:
    a. Metodologi Data Splitting: Strategi pembagian training/validation/test
    b. Pendekatan Stratified Sampling: Framework mempertahankan distribusi
    c. Strategi Temporal Split: Metodologi pembagian berbasis waktu

3.6.5 Penyimpanan dan Pengelolaan Data

Data yang dikumpulkan dan diolah dikelola melalui framework sistematis:

1. Metodologi Format Penyimpanan: Strategi penyimpanan terstruktur untuk persistensi dan pemrosesan data penelitian.
2. Framework Struktur Data: Pendekatan struktural untuk metadata feedback yang mencakup informasi kontekstual dan temporal.
3. Strategi Pengelompokan Data: Metodologi kategorisasi data berdasarkan tipologi bidang untuk optimasi pembelajaran.
4. Framework Versioning: Pendekatan sistematis untuk pelacakan evolusi data dan model dalam konteks penelitian longitudinal.

3.6.6 Pertimbangan Etis Penelitian

Dalam konteks penelitian yang melibatkan data dokumen, beberapa pertimbangan etis perlu diperhatikan:

1. Anonimisasi Data: Metodologi untuk memastikan data sensitif dalam dokumen template tidak mengandung informasi pribadi yang dapat diidentifikasi.
2. Consent Framework: Pendekatan untuk memperoleh persetujuan penggunaan data dalam konteks penelitian akademis.

3. Data Minimization: Strategi untuk menggunakan hanya data yang diperlukan untuk tujuan penelitian tanpa mengumpulkan informasi berlebihan.
4. Research Ethics Compliance: Framework untuk memastikan penelitian mematuhi standar etika penelitian akademis yang berlaku.

3.7 Metode Evaluasi

Untuk mengevaluasi efektivitas sistem ekstraksi data adaptif, penelitian ini menggunakan berbagai metrik dan metode evaluasi. Bagian ini menjelaskan metrik, metode, dan skenario pengujian yang digunakan dalam evaluasi.

3.7.1 Metrik Evaluasi

Beberapa metrik evaluasi yang digunakan dalam penelitian ini meliputi:

1. Presisi: Proporsi bidang yang diekstrak dengan benar dari semua bidang yang diekstrak.

$$Presisi = TP/(TP + FP)$$

di mana:

- TP (True Positive): Jumlah bidang yang diekstrak dengan benar
- FP (False Positive): Jumlah bidang yang diekstrak secara salah

2. Recall: Proporsi bidang yang diekstrak dengan benar dari semua bidang yang seharusnya diekstrak.

$$Recall = TP/(TP + FN)$$

di mana:

- FN (False Negative): Jumlah bidang yang seharusnya diekstrak tetapi tidak diekstrak

3. F1-Score: Rata-rata harmonik dari presisi dan recall, yang memberikan ukuran keseimbangan antara keduanya.

$$F1 = 2 * (Presisi * Recall)/(Presisi + Recall)$$

4. Akurasi per Bidang: Presisi, recall, dan F1-score untuk setiap bidang individual.
5. Akurasi Keseluruhan: Proporsi semua bidang yang diekstrak dengan benar dari semua bidang.

$$Akurasi = (TP + TN)/(TP + TN + FP + FN)$$

di mana:

- TN (True Negative): Jumlah bidang yang dengan benar tidak diekstrak

6. Waktu Ekstraksi: Waktu yang dibutuhkan untuk mengekstrak data dari dokumen.
7. Kurva Pembelajaran: Peningkatan akurasi seiring bertambahnya data umpan balik.

3.7.2 Framework Metodologi Evaluasi

Evaluasi sistem akan menggunakan beberapa metode validasi:

1. Cross-Validation Framework: Metodologi k-fold cross-validation untuk evaluasi kinerja model dengan data terbatas. Framework ini menggunakan strategi pembagian data sistematis untuk validasi robustness.
2. Hold-Out Validation Strategy: Framework pembagian data menjadi training dan testing sets untuk evaluasi objektif. Metodologi ini menggunakan strategi sampling yang representatif.
3. Learning Curve Analysis Framework: Metodologi analisis peningkatan kinerja model seiring bertambahnya data pelatihan. Framework ini mengukur convergence rate dan learning efficiency.
4. Error Analysis Methodology: Framework analisis sistematis untuk kategorisasi dan identifikasi pola kesalahan sistem. Metodologi ini menggunakan pendekatan kualitatif dan kuantitatif.
5. Ablation Study Framework: Metodologi evaluasi kontribusi komponen sistem terhadap kinerja keseluruhan. Framework ini menggunakan controlled experimentation approach.

3.7.3 Skenario Pengujian

Untuk evaluasi komprehensif, penelitian ini menggunakan beberapa skenario pengujian yang berfokus pada efektivitas sistem HITL:

1. Baseline Ekstraksi Berbasis Aturan:

a. Objective: Evaluasi kinerja ekstraksi berbasis aturan tanpa interaksi HITL
b. Test Data: 100 dokumen dengan 5 template types berbeda
c. Metrics: Precision, Recall, F1-score per field type
d. Expected Results: Baseline performance untuk comparison

2. Pembelajaran Adaptif HITL:
    a. Objective: Evaluasi peningkatan kinerja melalui feedback loop HITL
    b. Methodology: Incremental learning dengan 5, 10, 15, 20 feedback samples
    c. Measurement: Learning curve analysis dan convergence rate
    d. Success Criteria: ≥15% improvement dalam F1-score setelah 20 iterations
3. Variasi PDF Template:
    a. Template Categories:
        i. Form-based documents (formulir pendaftaran)
        ii. Letter-based documents (surat keterangan)
        iii. Table-based documents (laporan data)
        iv. Mixed-layout documents (dokumen kompleks)
    b. Evaluation: Cross-template generalization ability
    c. Metrics: Accuracy drop ketika testing pada unseen templates
4. Simulasi Interaksi HITL Realistis:
    a. User Simulation: Realistic error patterns dan correction behaviors
    b. Feedback Quality: Varying levels of user expertise (novice, expert)
    c. Temporal Dynamics: Learning retention over time
    d. Scalability: Performance dengan increasing document volume
5. Comparative Analysis:
    a. Rule-based vs CRF vs Hybrid: Performance comparison
    b. Static vs Adaptive: Learning capability assessment
    c. Manual vs Automated: Efficiency comparison

d. Single-user vs Multi-user: Collaborative learning evaluation

3.7.4 Framework Implementasi Evaluasi

Evaluasi sistem dilakukan melalui framework metodologi sistematis:

1. Framework Persiapan Data: Metodologi penyiapan ground truth data dan reference standards untuk evaluasi komprehensif.
2. Framework Eksekusi: Strategi systematic execution dengan controlled conditions untuk konsistensi hasil.
3. Framework Perhitungan Metrik: Metodologi standardized metric calculation dengan statistical validation.
4. Framework Analisis: Pendekatan systematic analysis untuk identifikasi patterns dan insights.
5. Framework Visualisasi: Metodologi data presentation untuk interpretasi dan communication hasil.
6. Framework Dokumentasi: Strategi comprehensive documentation untuk reproducibility dan future reference.

3.8 Rencana Eksperimen

Untuk menguji hipotesis penelitian dan mengevaluasi kinerja sistem secara komprehensif, penelitian ini menggunakan rencana eksperimen yang terstruktur. Bagian ini menjelaskan rencana eksperimen yang digunakan dalam penelitian.

3.8.1 Tujuan Eksperimen

Eksperimen dalam penelitian ini bertujuan untuk:

1. Mengevaluasi akurasi ekstraksi data dengan pendekatan berbasis aturan sebagai baseline
2. Mengukur peningkatan akurasi dengan pembelajaran adaptif berbasis umpan balik pengguna
3. Menilai kemampuan adaptasi sistem terhadap variasi dalam format dokumen
4. Mengevaluasi efisiensi pembelajaran dengan jumlah data umpan balik yang terbatas

5. Membandingkan kinerja berbagai strategi ekstraksi (berbasis aturan, berbasis model, hybrid)

3.8.2 Desain Eksperimen

Desain eksperimen meliputi beberapa tahap dan skenario:

1. Eksperimen Baseline: Evaluasi kinerja ekstraksi berbasis aturan tanpa pembelajaran adaptif pada set dokumen standar.
2. Eksperimen Pembelajaran Inkremental:
   - Mulai dengan model tanpa pelatihan
   - Tambahkan data umpan balik secara bertahap (5, 10, 15, 20 contoh per bidang)
   - Evaluasi kinerja pada setiap tahap
   - Analisis kurva pembelajaran
3. Eksperimen Variasi Dokumen:
   - Evaluasi kinerja pada dokumen dengan format yang berbeda
   - Analisis pengaruh variasi format terhadap akurasi ekstraksi
   - Evaluasi kemampuan adaptasi sistem terhadap variasi
4. Eksperimen Strategi Ekstraksi:
   - Bandingkan kinerja ekstraksi berbasis aturan, berbasis model, dan hybrid
   - Analisis kekuatan dan kelemahan setiap strategi
   - Identifikasi skenario optimal untuk setiap strategi

3.8.3 Framework Metodologi Eksperimen

Setiap eksperimen dilaksanakan dengan framework metodologi sistematis:

1. Framework Persiapan Data: Metodologi penyiapan ground truth dan reference standards untuk eksperimen terkontrol
2. Framework Konfigurasi: Strategi konfigurasi sistematis sesuai dengan scenario experimental design
3. Framework Eksekusi: Pendekatan systematic execution dengan controlled conditions untuk konsistensi

4. Framework Pengumpulan Data: Metodologi structured data collection dengan standardized metrics
5. Framework Analisis: Strategi systematic analysis dengan statistical validation
6. Framework Dokumentasi: Pendekatan comprehensive documentation untuk reproducibility

3.8.4 Analisis Hasil Eksperimen

Hasil eksperimen dianalisis dengan metode statistik dan komparatif berikut:

1. Framework Analisis Metrik:
    a. Metodologi Per-Field Analysis: Framework evaluasi accuracy, precision, recall per field type
    b. Strategi Overall Performance: Pendekatan macro dan micro-averaged metrics assessment
    c. Framework Confidence Analysis: Metodologi correlation analysis untuk confidence validation
    d. Pendekatan Error Pattern: Framework systematic error categorization dan analysis
2. Framework Analisis Komparatif:
    a. Metodologi Statistical Testing: Framework statistical significance testing untuk strategy comparison
    b. Strategi Effect Size: Pendekatan practical significance measurement
    c. Framework Confidence Intervals: Metodologi uncertainty quantification untuk metrics
    d. Pendekatan Multi-group Analysis: Framework comparative analysis across template types
3. Framework Pembelajaran Adaptif:
    a. Metodologi Learning Curve: Framework curve modeling untuk learning progression analysis
    b. Strategi Convergence: Pendekatan improvement rate dan plateau detection

c. Framework Sample Efficiency: Metodologi feedback effectiveness measurement
d. Pendekatan Retention: Framework performance sustainability analysis

4. Framework Analisis Efisiensi:
    a. Metodologi Complexity Analysis: Framework computational efficiency assessment
    b. Strategi Resource Usage: Pendekatan memory dan processing requirement analysis
    c. Framework Scalability: Metodologi performance scaling assessment
    d. Pendekatan Response Time: Framework latency analysis untuk practical usage
5. Framework Kualitas Feedback:
    a. Metodologi Impact Assessment: Framework quantitative feedback effectiveness measurement
    b. Strategi Consistency Analysis: Pendekatan inter-rater reliability assessment
    c. Framework Feedback Type: Metodologi comparative feedback effectiveness analysis
    d. Pendekatan Active Learning: Framework informative sample identification
6. Framework Visualisasi:
    a. Metodologi Dashboard: Framework interactive visualization untuk key metrics
    b. Strategi Learning Curves: Pendekatan progress tracking visualization
    c. Framework Error Visualization: Metodologi confusion matrix dan error pattern display
    d. Pendekatan Performance Curves: Framework classification performance visualization

3.8.5 Framework Expected Outcomes

Framework Primary Success Criteria:

1. Metodologi Accuracy Improvement: Framework measurement untuk significant F1-score enhancement melalui adaptive learning
2. Strategi Learning Efficiency: Pendekatan assessment untuk rapid improvement dengan minimal feedback samples
3. Framework User Acceptance: Metodologi usability assessment dengan validated measurement scales
4. Pendekatan System Responsiveness: Framework performance measurement untuk practical usage requirements

Framework Secondary Success Criteria:

1. Metodologi Generalization: Framework assessment untuk cross-template performance consistency
2. Strategi Retention: Pendekatan long-term performance sustainability measurement
3. Framework Scalability: Metodologi performance scaling assessment dengan increasing complexity

Framework Risk Mitigation:

1. Metodologi User Adoption: Framework interface design dan training strategy untuk user engagement
2. Strategi Training Data: Pendekatan synthetic data generation untuk bootstrapping pembelajaran
3. Framework Performance: Metodologi optimization dan caching strategy untuk system efficiency
4. Pendekatan Technical Robustness: Framework error handling dan fallback mechanism design

3.8.6 Framework Expected Outcomes

3.9 Ringkasan Metodologi

Bab ini telah menjelaskan metodologi penelitian yang komprehensif untuk pengembangan sistem ekstraksi data PDF adaptif berbasis Human-in-the-Loop. Metodologi ini mencakup:

Framework Penelitian:

1. Design Science Research methodology untuk systematic development
2. Iterative approach dengan continuous evaluation dan improvement
3. Integration antara technical development dan user-centered design

Framework Arsitektur Sistem:

1. Modular design dengan clear separation of concerns
2. Hybrid extraction approach combining rule-based dan machine learning
3. Adaptive learning mechanism melalui user feedback integration
4. Scalable architecture untuk real-world deployment

Framework Evaluasi dan Validasi:

1. Multiple evaluation metrics untuk comprehensive assessment
2. User studies untuk real-world validation
3. Statistical analysis untuk rigorous result interpretation
4. Framework untuk validity, reliability, dan generalizability

Metodologi ini menyediakan foundation yang solid untuk pengembangan dan evaluasi sistem yang dapat memberikan kontribusi signifikan dalam domain ekstraksi data dokumen adaptif.

Daftar Pustaka

Abiteboul, S. (1997). Querying semi-structured data. In S. Abiteboul, *Lecture Notes in Computer Science* (pp. 1–18). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-62222-5_33

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Dengel, A. R., & Klein, B. (2002). smartFIX: A Requirements-Driven System for Document Analysis and Understanding. In A. R. Dengel & B. Klein, *Lecture Notes in Computer Science* (pp. 433–444). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45869-7_47

Fails, J. A., & Olsen, D. R. (2003). Interactive machine learning. *Proceedings of the 8th International Conference on Intelligent User Interfaces*, 39–45. https://doi.org/10.1145/604045.604056

Gebauer, M., Maschhur, F., Leschke, N., Grünewald, E., & Pallas, F. (2023). A 'Human-in-the-Loop' approach for Information Extraction from Privacy Policies under Data Scarcity. *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 76–83. https://doi.org/10.1109/EuroSPW59978.2023.00014

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004a). *Design Science in Information Systems Research*.

Hevner, March, Park, & Ram. (2004b). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75. https://doi.org/10.2307/25148625

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131. https://doi.org/10.1007/s40708-016-0042-6

International Organization for Standardization. (2008). *Document management—Portable document format—Part 1: PDF 1.7* (No. ISO 32000-1:2008). ISO. https://www.iso.org/standard/51502.html

Ishikawa, K., & Ishikawa, K. (1987). *What is total quality control? The Japanese way* (6. print). Prentice-Hall.

Katti, A. R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., & Faddoul, J. B. (2018). *Chargrid: Towards Understanding 2D Documents* (No. arXiv:1809.08799). arXiv. https://doi.org/10.48550/arXiv.1809.08799

Klein, B., Dengel, A., & Fordan, A. (2004). smartFIX: An Adaptive System for Document Analysis and Understanding. In B. Klein, A. R. Dengel, & A. Fordan, *Lecture Notes in Computer Science* (pp. 166–186). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24642-8_11

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of

the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Palm, R. B., Winther, O., & Laws, F. (2017). CloudScan—A Configuration-Free Invoice Analysis System Using Recurrent Neural Networks. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 406–413. https://doi.org/10.1109/icdar.2017.74

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Popovic, N., & Färber, M. (2022). Few-Shot Document-Level Relation Extraction. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5733–5746. https://doi.org/10.18653/v1/2022.naacl-main.421

Schleith, J., Hoffmann, H., Norkute, M., & Cechmanek, B. (2022). *Human in the loop information extraction increases efficiency and trust*. https://doi.org/10.18420/MUC2022-MCI-WS12-249

Schroeder, N. L., Jaldi, C. D., & Zhang, S. (2025). *Large Language Models with Human-In-The-Loop Validation for Systematic Review Data Extraction* (No. arXiv:2501.11840). arXiv. https://doi.org/10.48550/arXiv.2501.11840

Schuster, D., Muthmann, K., Esser, D., Schill, A., Berger, M., Weidling, C., Aliyev, K., & Hofmeier, A. (2013). Intellix—End-User Trained Information

Extraction for Document Archiving. *2013 12th International Conference on Document Analysis and Recognition*, 101–105. https://doi.org/10.1109/icdar.2013.28

Settles, B. (2012). *Active Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-031-01560-1

Stumpf, S., Rajaram, V., Li, L., Wong, W.-K., Burnett, M., Dietterich, T., Sullivan, E., & Herlocker, J. (2009). Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, *67*(8), 639–662. https://doi.org/10.1016/j.ijhcs.2009.03.004

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). *LayoutLM: Pre-training of Text and Layout for Document Image Understanding*. 1192–1200. https://doi.org/10.1145/3394486.3403172