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 (*Document Management. Portable Document Format PDF 2.0*, 2020). Meski demikian, PDF dirancang untuk keterbacaan manusia, bukan untuk pemrosesan mesin, ekstraksi data terstruktur dari dokumen-dokumen ini tetap menjadi tantangan yang persisten. Survei terkini tentang ekstraksi informasi dokumen (Cui et al., 2021) menyoroti bahwa dokumen berbasis template, di mana struktur konsisten namun konten sangat bervariasi, menciptakan kesenjangan signifikan antara format yang dapat dibaca manusia dan data yang dapat diproses mesin.

Pendekatan ekstraksi tradisional menghadapi keterbatasan fundamental. Pendekatan berbasis aturan efektif untuk struktur yang konsisten namun rentan terhadap perubahan tata letak minor, memerlukan rekonfigurasi manual yang memakan waktu. Sebaliknya, pendekatan pembelajaran mesin murni, termasuk model pre-trained terkini untuk pemahaman dokumen (Hong et al., 2022; Huang et al., 2022), sering memerlukan data berlabel serta sumber daya komputasi yang signifikan; pada dokumen semi-terstruktur berbasis template, pendekatan berbasis pembelajaran juga menghadapi tantangan praktis dalam pengembangan dan penerapan (Cui et al., 2021). Contoh penerapannya dapat ditemukan pada dokumen bisnis seperti invoice (Hamdi et al., 2021), dokumen legal seperti kontrak/Non-Disclosure Agreement (Hendrycks et al., 2021; Stanisławek et al., 2021), serta dokumen hasil pemindaian yang memerlukan pemahaman struktur dokumen (Mathew et al., 2020).

Untuk mengatasi keterbatasan ini, paradigma *Human-in-the-Loop* (HITL) telah muncul sebagai pendekatan yang menjanjikan yang memungkinkan integrasi keahlian manusia ke dalam sistem pembelajaran mesin (Mosqueira-Rey et al., 2023). HITL memungkinkan sistem untuk belajar secara progresif melalui interaksi

pengguna, menggabungkan kekuatan komputasi mesin dengan pengetahuan domain dan intuisi manusia. Penelitian terkini menekankan pentingnya kolaborasi manusia-AI yang efektif (Bansal et al., 2021) dan strategi pembelajaran aktif (Ren et al., 2022) dalam membangun sistem yang dapat beradaptasi secara efisien dengan beban pengguna yang minimal.

Penelitian ini mengusulkan arsitektur hibrid yang mengintegrasikan strategi berbasis aturan dan Conditional Random Fields (CRF) dalam kerangka kerja HITL untuk mencapai keseimbangan yang lebih baik antara akurasi ekstraksi, efisiensi sumber daya komputasi, dan kemampuan adaptasi real-time. Sistem yang dikembangkan memungkinkan pembelajaran inkremental dari umpan balik pengguna tanpa memerlukan dataset berlabel yang besar atau infrastruktur komputasi yang mahal.

1.2 Rumusan Masalah

Meskipun paradigma HITL menjanjikan, analisis terhadap penelitian *state-of-the-art* mengungkapkan bahwa implementasi HITL yang efektif dalam konteks ekstraksi data PDF template masih menghadapi kesenjangan penelitian. Pertama, model *state-of-the-art* seperti *Large Language Models* (LLM) menunjukkan kebutuhan tinggi untuk validasi manusia (Schroeder et al., 2025), namun tidak dirancang untuk pembelajaran adaptif inkremental yang efisien karena biaya pelatihan ulang yang sangat tinggi. Kedua, di sisi lain spektrum, sistem transparan seperti berbasis aturan murni telah terbukti superior dalam kepercayaan pengguna (Schleith et al., 2022) namun secara fundamental kaku dan non-adaptif.

Kesenjangan ini menciptakan peluang untuk arsitektur yang menjembatani kedua ekstrem, terutama dalam skenario "kelangkaan data" (Gebauer et al., 2023). Secara spesifik:

1. Penelitian terbatas pada kerangka kerja sistematis yang mengintegrasikan berbasis aturan (untuk transparansi) dengan pembelajaran mesin yang efisien (seperti CRF) dalam konteks HITL adaptif, khususnya dalam skenario kelangkaan data dan keterbatasan sumber daya.

2. Mekanisme umpan balik yang efisien untuk mengonversi koreksi pengguna menjadi pengetahuan sistem belum terkarakterisasi dengan baik untuk arsitektur hibrid dengan dual learning mechanism.
3. Strategi pembelajaran adaptif real-time tanpa pelatihan ulang ekstensif tetap menjadi tantangan terbuka.

Berdasarkan kesenjangan tersebut, penelitian ini merumuskan pertanyaan penelitian sebagai berikut:

1. Bagaimana merancang arsitektur hibrid yang mengintegrasikan strategi berbasis aturan dan CRF dalam kerangka kerja HITL untuk ekstraksi data PDF template?
2. Bagaimana mengimplementasikan mekanisme umpan balik yang efisien untuk mengonversi koreksi pengguna menjadi peningkatan sistem secara inkremental?
3. Bagaimana mencapai pembelajaran adaptif real-time dengan data pelatihan minimal dan sumber daya komputasi terbatas?

1.3 Tujuan

Penelitian ini bertujuan untuk:

1. Mengembangkan arsitektur hibrid yang mengintegrasikan strategi ekstraksi berbasis aturan dan CRF dengan mekanisme pemilihan berbasis confidence untuk ekstraksi data PDF template.
2. Merancang dan mengimplementasikan loop umpan balik HITL yang efisien untuk pembelajaran inkremental melalui pembelajaran pola (rule-based) dan pelatihan ulang CRF inkremental.
3. Mengevaluasi efektivitas sistem dalam hal akurasi ekstraksi, efisiensi pembelajaran, dan efisiensi sumber daya pada berbagai jenis template dokumen.

1.4 Batasan Penelitian

Penelitian ini memiliki batasan sebagai berikut:

1. Jenis Dokumen: Penelitian difokuskan pada dokumen PDF berbasis template dengan struktur yang relatif konsisten, bukan dokumen dengan layout yang sepenuhnya bebas.
2. Jenis Template: Evaluasi dilakukan pada empat jenis template (Form, Table, Letter, Mixed) yang mewakili kasus penggunaan umum.
3. Sumber Daya: Sistem dirancang untuk lingkungan dengan keterbatasan sumber daya (CPU-only, tanpa GPU).
4. Skala Dataset: Evaluasi dilakukan pada 140 dokumen sintetis dengan 2.800 field, yang didesain untuk merepresentasikan skenario deployment organisasi kecil hingga menengah.

1.5 Manfaat Penelitian

1.5.1 Manfaat Teoritis

1. Memberikan kontribusi pada literatur pembelajaran adaptif dengan mengusulkan kerangka kerja sistematis untuk integrasi strategi berbasis aturan dan pembelajaran mesin dalam konteks HITL.
2. Memperkaya literatur tentang ekstraksi informasi dokumen dengan pendekatan hibrid yang menyeimbangkan transparansi, efisiensi, dan adaptabilitas.
3. Memberikan bukti empiris tentang efektivitas pembelajaran inkremental dengan data minimal dalam domain ekstraksi dokumen.

1.5.2 Manfaat Praktis

1. Menyediakan solusi praktis untuk organisasi kecil dan menengah yang memerlukan sistem ekstraksi dokumen dengan keterbatasan data dan sumber daya komputasi.
2. Mengurangi beban pengguna dalam proses ekstraksi dokumen melalui pembelajaran adaptif yang efisien dengan tingkat koreksi yang minimal.
3. Memungkinkan deployment yang lebih efisien untuk sistem ekstraksi dokumen tanpa memerlukan dataset berlabel yang besar atau infrastruktur komputasi yang mahal.

4. Memberikan sistem yang dapat beroperasi dengan waktu respons yang cepat pada perangkat keras standar tanpa memerlukan infrastruktur GPU.

1.6 Metodologi Penelitian

Penelitian ini menggunakan pendekatan Design Science Research (DSR) yang fokus pada pengembangan dan evaluasi artefak teknologi untuk memecahkan masalah spesifik (Peffers et al., 2020; Wagner et al., 2021). DSR dipilih karena sesuai dengan nature penelitian yang bertujuan mengembangkan solusi praktis untuk masalah real-world dalam ekstraksi data PDF template.

Framework DSR yang diadopsi mencakup enam tahapan: (1) identifikasi masalah dan motivasi melalui analisis kesenjangan penelitian state-of-the-art, (2) definisi objektif sistem pembelajaran adaptif berbasis HITL dengan kriteria sukses yang terukur, (3) desain dan pengembangan arsitektur hibrid dengan mekanisme pembelajaran inkremental, (4) demonstrasi feasibility melalui proof-of-concept implementation, (5) evaluasi komprehensif menggunakan metrik multi-dimensional (akurasi, efisiensi pembelajaran, efisiensi sumber daya), dan (6) komunikasi temuan untuk komunitas akademik dan praktisi.

Evaluasi dilakukan menggunakan mixed-methods approach yang menggabungkan quantitative metrics (accuracy, precision, recall, F1-score) dengan qualitative insights untuk memberikan comprehensive assessment terhadap efektivitas sistem yang dikembangkan. Detail metodologi penelitian dijelaskan secara lengkap pada BAB III.

1.7 Sistematika Penulisan

Tesis ini disusun dengan sistematika sebagai berikut:

BAB I PENDAHULUAN: Bab ini membahas latar belakang penelitian, rumusan masalah, tujuan penelitian, batasan penelitian, manfaat penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA: Bab ini membahas landasan teori dan penelitian terkait yang menjadi dasar penelitian, meliputi teknik ekstraksi data PDF, sistem Human-in-the-Loop, pembelajaran adaptif, dan penelitian terkait lainnya.

BAB III METODOLOGI PENELITIAN: Bab ini menjelaskan metodologi penelitian yang digunakan, meliputi desain penelitian, arsitektur sistem, strategi ekstraksi, mekanisme HITL, metrik evaluasi, dan prosedur eksperimen.

BAB IV HASIL DAN PEMBAHASAN: Bab ini menyajikan hasil eksperimen dan pembahasan mendalam tentang kinerja sistem, efektivitas pembelajaran adaptif, dan perbandingan dengan pendekatan lain.

BAB V KESIMPULAN DAN SARAN: Bab ini menyimpulkan hasil penelitian, menjawab pertanyaan penelitian, dan memberikan saran untuk penelitian lanjutan.