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

TESIS

Disusun Oleh :

Moh Syaiful Rahman

NPM: 247064518006

PROGRAM STUDI MAGISTER TEKNOLOGI INFORMASI

FAKUSTAS TEKNOLOGI KOMUNIKASI DAN INFORMATIKA

UNIVERSITAS NASIONAL

TAHUN 2026

HALAMAN PENGESAHAN

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

Untuk memenuhi sebagai persyaratan memperoleh Gelar Magister Komputer

Disusun oleh:

Moh. Syaiful Rahman

247064518006

Tesis ini telah diuji dan dinyatakan lulus pada

21 Februari 2026

Telah diperiksa dan disetujui oleh:

Dosen Pembimbing,	Mengetahui, Ketua Program Studi Magister Teknologi Informasi
Dr. Ir. Andriamingsih, S. Kom., MMSI	**Ir. Asrul Sani, ST., M.T., M.Kom., Ph.D.**
NIDN. 0303097902	**NIDN. 0303067003**

HALAMAN PENGESAHAN

TESIS

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

Moh Syaiful Rahman

247064518006

Dosen Pembimbing

Dr. Ir. Andrianingsih, S. Kom., MMSI

LEMBAR PERSETUJUAN TESIS

.

Tugas Akhir dengan judul :

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

Dibuat untuk melengkapi salah satu persyaratan menjadi Magister Komputer pada Program Studi Magister Teknologi Informasi, Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional. Tesis ini diujikan pada Sidang Tesis Semester Ganjil 2025-2026 pada tanggal 21 Februari Tahun 2026

Dosen Pembimbing

Dr. Ir. Andrianingsih, S. Kom., MMSI
NIDN. 0303097902

Ketua Program Studi

Ir. Asrul Sani, ST., M.T., M.Kom., Ph.D.
NIDN. 0303067003

PERNYATAAN KEASLIAN TESIS

Saya menyatakan dengan sesungguhnya bahwa Tesis dengan judul :

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

Yang dibuat untuk melengkapi salah satu persyaratan menjadi Magister Komputer pada Program Studi Magister Teknologi Informasi Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional, sebagaimana yang saya ketahui adalah bukan merupakan tiruan atau publikasi dari Tesis yang pernah diajukan atau dipakai untuk mendapatkan gelar di lingkungan Universitas Nasional maupun perguruan tinggi atau instansi lainnya, kecuali pada bagian – bagian tertentu yang menjadi sumber informasi atau acuan yang dicantumkan sebagaimana mestinya.

Jakarta, 28 Februari 2026

Moh. Syaiful Rahman
NPM: 247064518006

LEMBAR PERSETUJUAN JUDUL YANG TIDAK ATAU YANG DIREVISI

Nama : Moh. Syaiful Rahman

NPM : 247064518006

Fakultas/Akademi : Fakultas Teknologi Komunikasi dan Informatika

Program Studi : Magister Teknologi Informasi

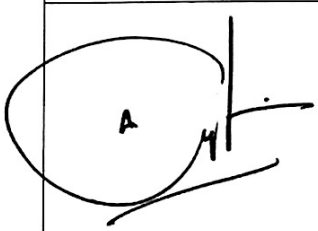
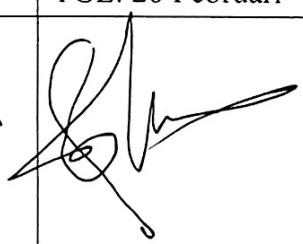
Tanggal Sidang : 21 Februari 2026

JUDUL DALAM BAHASA INDONESIA :

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

JUDUL DALAM BAHASA INGGRIS :

Adaptive Learning System Based on Human-in-the-Loop for PDF Template Data Extraction

TANDA TANGAN DAN TANGGAL		
Pembimbing 1	**Ka. Prodi**	**Mahasiswa**
TGL: 28 Februari	TGL: 28 Februari	TGL: 28 Februari
		

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT atas segala rahmat, taufik, dan hidayah-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul "Sistem Pembelajaran Adaptif Berbasis Human-in-the-Loop untuk Ekstraksi Data PDF Template" ini. Tesis ini disusun sebagai salah satu syarat untuk memperoleh gelar Magister Teknologi Informasi pada Program Studi Magister Teknologi Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional.

Penulis menyadari sepenuhnya bahwa penyelesaian tesis ini tidak terlepas dari bantuan, bimbingan, dan dukungan berbagai pihak. Untuk itu, dengan segala kerendahan hati, penulis menyampaikan penghargaan dan ucapan terima kasih yang sebesar-besarnya kepada:

1. Kedua orang tua penulis, Masrikun dan Samaniyah, kakak Jupriadi, adik Ina Fitriani, serta istri tercinta Widi Setyowati, dan buah hati tersayang Prisya Silvana dan Yovika Sklodowska, atas doa yang tidak pernah putus, dukungan moral, dan pengorbanan yang tidak ternilai selama penulis menempuh pendidikan.
2. Jajaran pimpinan Universitas Nasional, mulai dari Rektor, Direktur Program Pascasarjana, hingga Ketua Program Studi Magister Teknologi Informasi, atas segala fasilitas, kebijakan akademik, dan dukungan yang telah diberikan selama masa studi.
3. Dr. Ir. Andrianingsih, S. Kom., MMSI, selaku Dosen Pembimbing Utama, yang dengan penuh kesabaran telah meluangkan waktu untuk memberikan bimbingan, arahan, dan masukan yang sangat berarti dalam setiap tahap penyelesaian tesis ini.
4. Ir. Asrul Sani, ST., M.T., M.Kom., Ph.D. dan Prof. Agung Triayudi, S.Kom, M.Kom, Ph.D., selaku Dosen Penguji, atas kritik dan saran yang tajam dan membangun dalam penyempurnaan tesis ini.
5. Seluruh dosen pengajar Program Studi Magister Teknologi Informasi Universitas Nasional, atas ilmu pengetahuan dan pengalaman akademik yang telah diberikan selama masa perkuliahan.

6. Seluruh staf dan karyawan Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional, atas bantuan dalam kelancaran proses administrasi selama masa studi.
7. Rekan-rekan mahasiswa Program Magister Teknologi Informasi angkatan 2024, atas semangat, diskusi, dan kebersamaan yang menjadi bagian tak terlupakan dari perjalanan studi ini.
8. Semua pihak yang tidak dapat penulis sebutkan satu per satu, yang telah memberikan kontribusi nyata dalam penyelesaian tesis ini.

Penulis menyadari bahwa tesis ini masih memiliki keterbatasan dan kekurangan. Oleh karena itu, kritik dan saran yang konstruktif dari berbagai pihak sangat penulis harapkan demi penyempurnaan ke depan. Penulis berharap tesis ini dapat memberikan kontribusi yang berarti, khususnya dalam pengembangan sistem ekstraksi data berbasis kecerdasan buatan dan penerapan pendekatan Human-in-the-Loop di Indonesia.

Jakarta, 28 Februari 2026

Moh. Syaiful Rahman

2470645138006

ABSTRAK

Pada dokumen PDF berbasis template, pemetaan fields sering bergantung pada isyarat tata letak. Perubahan kecil seperti jarak antar-elemen, pergeseran posisi label, atau pergeseran baris pada tabel dapat mengubah hasil pemetaan tersebut. Oleh sebab itu, keluaran ekstraksi dapat berbeda ketika dokumen berasal dari sumber yang berbeda, walaupun jenis template-nya sama. Model pre-trained berukuran besar dapat memberikan akurasi tinggi, tetapi biasanya menuntut komputasi yang besar dan korpus berlabel yang ekstensif, sehingga penerapannya tidak selalu realistis pada lingkungan dengan sumber daya terbatas. Di sisi lain, pemrosesan berbasis aturan relatif ringan dan mudah diaudit, namun sensitif terhadap perubahan tata letak sehingga kinerjanya dapat menurun. Penelitian ini merancang alur pembelajaran adaptif berbasis pendekatan hibrid yang menggabungkan komponen berbasis aturan dengan Conditional Random Fields (CRF) untuk meningkatkan toleransi terhadap variasi tata letak pada kondisi data dan infrastruktur yang terbatas.

Sistem menjalankan beberapa strategi ekstraksi secara paralel dan memilih keluaran berdasarkan kriteria confidence, kemudian mengintegrasikan Human-in-the-Loop (HITL) untuk mendukung perbaikan inkremental. Koreksi pengguna dikonversi menjadi pembaruan pola pada komponen berbasis aturan, sedangkan model CRF dilatih ulang secara berkala dalam batch kecil untuk menangkap fields yang bergantung pada konteks dan sulit ditentukan sepenuhnya melalui aturan. Pendekatan ini tidak dimaksudkan untuk menggantikan salah satu metode, melainkan memanfaatkan aturan untuk interpretabilitas dan pembaruan cepat, serta CRF untuk generalisasi ketika konteks lokal berperan, dengan tetap menjaga kebutuhan deployment yang ringan.

Evaluasi tidak hanya berfokus pada akurasi ekstraksi, tetapi juga pada besarnya usaha koreksi dari pengguna serta efisiensi operasional. Pertimbangan ini penting karena rancangan sistem ditujukan untuk skenario dengan kapasitas anotasi yang terbatas dan ketersediaan komputasi yang tidak selalu memadai. Dievaluasi pada 140 dokumen PDF sintetis di 4 jenis template (Form, Table, Letter, Mixed),

sistem mencapai akurasi 98,61% dengan data pelatihan 25–35 dokumen dan tingkat koreksi pengguna 7,0% (196 dari 2.800 fields), menunjukkan efisiensi pembelajaran 7,55 koreksi per peningkatan poin persentase. Akurasi dihitung pada keluaran ekstraksi sebelum koreksi manual terhadap ground truth, sedangkan koreksi pengguna merepresentasikan akumulasi koreksi sepanjang proses pembelajaran adaptif. Peningkatan performa dari baseline (72,65%) ke sistem adaptif (98,61%) signifikan secara statistik (paired t-test: $t=17,89$, $p<0,001$, Cohen's $d=8,94$). Sistem beroperasi pada perangkat keras CPU-only dengan footprint memori 37–162 MB, ukuran model 1,75 MB, dan waktu pemrosesan median 45 detik per dokumen (termasuk review HITL), sementara waktu ekstraksi inti per dokumen (tanpa HITL) berdasarkan pencatatan waktu sistem berada pada rentang 0,131–0,685 detik (median 0,265 detik)

Secara keseluruhan, hasil ini mengindikasikan bahwa alur kerja hibrid berbasis HITL dapat menyeimbangkan akurasi dengan kebutuhan data dan sumber daya pada lingkungan terbatas. Namun, karena evaluasi menggunakan dokumen sintetis dan koreksi yang disimulasikan, diperlukan validasi lanjutan pada kumpulan dokumen real-world dan studi pengguna sebelum menarik kesimpulan terkait kinerja pada deployment produksi.

Kata Kunci: Conditional Random Fields, Pemrosesan Dokumen, Umpan Balik Pengguna, Skor Kepercayaan, Pelatihan Inkremental

ABSTRACT

Extracting structured fields from template-based PDF documents remains difficult in practice because layouts may vary slightly across sources, while the content is only semi-structured. Although large pre-trained models can deliver strong accuracy, they typically demand substantial compute and sizable labeled corpora, which are not always feasible in resource-constrained settings. At the other end of the spectrum, rule-based parsing is lightweight and transparent, but it can degrade quickly when a document layout shifts. This study therefore explores a hybrid alternative: an adaptive learning workflow that combines rule-based extraction with Conditional Random Fields (CRF) to better accommodate layout variation under limited data and infrastructure

The system runs multiple extraction strategies in parallel and selects outputs using confidence-based criteria, then integrates Human-in-the-Loop (HITL) review to enable incremental improvement. User corrections are converted into pattern updates for the rule-based components, while the CRF models are retrained periodically in small batches to capture context-dependent fields that are difficult to encode as rules. The goal is not to replace one approach with the other, but to use rules for interpretability and fast updates, and CRF for generalization when local context matters, while keeping the deployment footprint lightweight.

Beyond extraction accuracy, the evaluation also considers user effort and operational efficiency, because the approach targets scenarios with limited annotation capacity and constrained compute resources. The correction-driven update mechanism is intended to reduce labeling overhead by focusing user review on low-confidence fields while preserving a transparent workflow. To support reproducibility of the evaluation, templates and fields are defined consistently across document types, and the reported metrics are computed against ground truth before any user correction is applied.

Evaluated on 140 synthetic PDF documents across 4 template types (Form, Table, Letter, Mixed), the system achieves 98.61% accuracy using 25–35 training documents and a 7.0% user correction rate (196 out of 2,800 fields), corresponding

to a learning efficiency of 7.55 corrections per percentage point improvement. Accuracy is measured on pre-correction extraction output against ground truth, while user corrections represent the accumulated corrections applied throughout the adaptive learning process. The performance improvement from baseline (72.65%) to the adaptive system (98.61%) is statistically significant (paired t-test: t=17.89, p<0.001, Cohen's d=8.94). The system operates on CPU-only hardware with 37–162 MB memory footprint, 1.75 MB total model size, and median 45 seconds end-to-end processing time per document (including HITL review), while the core extraction time per document (excluding HITL) based on system-recorded timing ranges from 0.131–0.685 seconds (median 0.265 seconds).

Taken together, these results suggest that a hybrid HITL workflow can balance accuracy with data and resource requirements in constrained environments. Never-theless, the evaluation relies on synthetic documents and simulated corrections, so additional validation on real-world document collections and user studies is needed before drawing conclusions about production performance.

Keywords: Conditional Random Fields, Rule-Based Parsing, User Feedback, Confidence Scoring, Incremental Training

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL