BAB IV
PERANCANGAN

4.1 Analisis Sistem terdahulu

Sebelum merancang sistem validasi entry data berbasis integrasi OCR, NLP, dan metode CRF, dilakukan terlebih dahulu kajian terhadap beberapa sistem yang telah dikembangkan pada penelitian sebelumnya. Tujuannya adalah untuk mengidentifikasi kelemahan dan keterbatasan pendekatan terdahulu serta sebagai dasar dalam merancang solusi yang lebih komprehensif dan efektif.

Salah satu sistem yang relevan dikembangkan oleh Nurhaliza dan Lussiana (2022), di mana mereka membangun sistem pengenalan karakter otomatis menggunakan teknologi OCR untuk memproses dokumen izin distribusi alat kesehatan. Dengan memanfaatkan pustaka Tesseract OCR, sistem tersebut berhasil memperoleh tingkat akurasi sebesar 98,78% dalam proses pembacaan karakter. Meskipun begitu, sistem tersebut hanya sebatas membaca karakter dan belum dilengkapi dengan mekanisme validasi struktur data maupun pemahaman konteks, sehingga masih memiliki keterbatasan ketika dihadapkan pada dokumen dengan kualitas gambar yang rendah, adanya noise, atau penggunaan jenis huruf yang tidak umum.

Sistem lain yang dikembangkan oleh Rusli et al. (2020) menggunakan kombinasi antara OCR dan pendekatan NLP sederhana untuk mengekstrak informasi dari dokumen KTP. Sistem ini mampu melakukan segmentasi data dan normalisasi teks secara lebih baik, namun proses validasinya masih mengandalkan aturan tetap (rule-based) yang sulit beradaptasi dengan variasi struktur dokumen. Selain itu, belum terdapat kemampuan untuk memverifikasi hubungan antar entitas dalam teks, seperti urutan logis antara NIK, nama, dan alamat.

Seiring berkembangnya teknologi pemrosesan bahasa alami, pendekatan menggunakan Conditional Random Field (CRF) mulai banyak diterapkan untuk tugas pengenalan entitas (NER). Model CRF terbukti efektif dalam mengolah teks berurutan dan memberikan label berdasarkan konteks

kata sebelumnya maupun sesudahnya. Dalam beberapa penelitian, seperti yang dilakukan oleh Permana (2019) dan Husein (2022), CRF mampu meningkatkan akurasi pelabelan entitas dalam dokumen berbahasa Indonesia. Namun, integrasi metode CRF dalam sistem yang menggabungkan hasil ekstraksi OCR dan pemrosesan NLP masih jarang dikembangkan secara menyeluruh.

Dari berbagai studi tersebut, dapat disimpulkan bahwa sebagian besar sistem terdahulu masih fokus pada satu aspek teknologi, baik OCR maupun NLP, dan belum banyak yang menggabungkan seluruh elemen—OCR, NLP, dan CRF—dalam satu sistem yang utuh. Maka dari itu, penelitian ini bertujuan mengembangkan sistem validasi entry data secara otomatis dan menyeluruh dengan menggabungkan ketiga pendekatan tersebut, guna meningkatkan akurasi, mengurangi potensi kesalahan input, dan mempercepat proses validasi data dari dokumen semi-struktur seperti invoice.Berikut adalah tabel untuk proses pengujian white box testing pada implementasi algoritma CRF

4.2 Spesifikasi Kebutuhan Sistem Baru

Agar sistem validasi entry data dapat berjalan dengan optimal, maka diperlukan penjabaran kebutuhan sistem yang akan dikembangkan. Bagian ini menyajikan uraian mendetail mengenai spesifikasi teknis dan fungsional dari sistem baru yang dirancang untuk menggantikan proses manual dalam mengekstraksi informasi dari dokumen invoice. Spesifikasi ini mencakup kebutuhan proses, struktur data, pengguna, serta spesifikasi perangkat keras dan lunak yang dibutuhkan untuk menjalankan sistem secara efisien.

4.2.1 Spesifikasi Proses

Spesifikasi proses merinci alur operasional sistem yang dikembangkan, mulai dari input pengguna hingga keluaran hasil akhir. Proses dalam sistem validasi entry data berbasis OCR, NLP, dan metode CRF ini terdiri atas beberapa tahapan penting yang saling terhubung secara sistematis untuk memastikan ketepatan ekstraksi informasi dari dokumen invoice.

Berikut ini adalah penjabaran dari tahapan proses yang dilakukan:

1. Pengunggahan Dokumen Invoice

Pengguna (admin/operator) melakukan upload dokumen invoice hasil pemindaian, baik dalam format gambar (.jpg, .jpeg, .png) maupun PDF. Sistem memverifikasi jenis file agar sesuai dengan format yang didukung. Jika file berupa gambar, maka akan dilanjutkan ke proses OCR.

2. Pemrosesan OCR (Optical Character Recognition)

Setelah file berhasil diunggah, sistem akan mengekstraksi teks dari gambar menggunakan Tesseract.js. Tahapan ini mengubah konten visual dokumen menjadi data teks mentah yang dapat dianalisis lebih lanjut.

3. Pembersihan dan Analisis Teks (NLP Preprocessing)

Teks hasil OCR selanjutnya diproses melalui tahapan NLP, seperti normalisasi teks, penghapusan karakter yang tidak relevan, dan segmentasi baris. Tujuan dari tahap ini adalah menyederhanakan struktur teks agar lebih siap untuk diekstraksi menggunakan algoritma CRF.

4. Ekstraksi Data menggunakan CRF (Conditional Random Field)

Algoritma CRF digunakan untuk mengenali pola dan entitas penting dari teks, seperti: tanggal invoice, nomor invoice, deskripsi barang, jumlah, harga satuan, total, hingga nilai total keseluruhan. Algoritma ini bekerja dengan mengenali hubungan antar kata dan konteksnya berdasarkan pelatihan data sebelumnya.

5. Menampilkan dan Menyimpan Hasil Ekstraksi

Data yang berhasil diekstraksi secara otomatis akan ditampilkan dalam form auto-fill di halaman web. Pengguna dapat melakukan pengecekan ulang dan mengklik tombol "Save" untuk menyimpan data ke dalam tabel hasil validasi.

6. Tabel Ringkasan dan Pengelolaan Data

Setelah data disimpan, sistem akan menampilkan semua hasil OCR & NLP dalam tabel rekap yang dapat dikelola oleh pengguna (misalnya menghapus baris tertentu jika ada kesalahan).

4.2.2 Spesifikasi Data

Dalam merancang sistem validasi entry data menggunakan kombinasi teknologi OCR, NLP, dan algoritma Conditional Random Field (CRF), sangat penting untuk mendefinisikan spesifikasi data yang digunakan. Subbab ini menguraikan jenis, struktur, dan sumber data yang menjadi dasar pengolahan dalam sistem yang dikembangkan..

1) Jenis Data yang Digunakan

Sistem ini dirancang untuk memproses data dalam bentuk dokumen invoice hasil pemindaian (scan) yang biasanya berformat JPG, PNG, atau PDF. Dokumen tersebut mengandung informasi tidak terstruktur yang perlu diubah menjadi bentuk yang dapat dianalisis dan disimpan dalam sistem.

Data utama yang diekstraksi meliputi:

a) Tanggal Invoice
b) Nomor Invoice
c) Tanggal Terima
d) Deskripsi Barang
e) Jumlah (QTY)
f) Harga per Unit
g) Total Invoice
a) Total Amount

Setelah dilakukan pemrosesan OCR, data yang awalnya berupa gambar akan dikonversi menjadi teks mentah yang akan diolah lebih lanjut menggunakan NLP dan metode CRF...

2) Sumber dan Teknik Pengumpulan Data

Untuk membangun model CRF yang akurat, sistem membutuhkan dataset pelatihan yang telah diberi anotasi (labeled data). Dalam tugas akhir ini, pengumpulan data dilakukan dari:

a) Dataset invoice hasil scan internal
b) Dokumen invoice sample dari berbagai perusahaan
c) Data manual yang ditulis ulang dan dianotasi sesuai kebutuhan pelabelan CRF

Setiap baris data dianotasi dengan label seperti *DATE, INVOICE_NO, ITEM_DESC, QTY, PRICE, TOTAL,* dan *AMOUNT*, agar CRF dapat mengenali pola hubungan antar kata..

3) Format dan Struktur Dataset

Data pelatihan untuk CRF biasanya diformat dalam bentuk baris teks, di mana setiap kata memiliki label-nya masing-masing. Format umum yang digunakan:

Tabel 4. 1 *Format Umum*

Styrene	ITEM_DESC
Monomer	ITEM_DESC
1000	QTY
15000	PRICE
15000000	TOTAL

Setiap baris data dianotasi dengan label seperti *DATE, INVOICE_NO, ITEM_DESC, QTY, PRICE, TOTAL,* dan *AMOUNT*, agar CRF dapat mengenali pola hubungan antar kata..

4) Perhitungan dan Pelatihan algoritma CRF

Dalam metode CRF, proses pelatihan dilakukan menggunakan pendekatan probabilistik. CRF menghitung kemungkinan hubungan antar kata berdasarkan fitur yang dikandung masing-masing token, misalnya:

a) Bentuk kata (huruf besar semua, angka, gabungan huruf dan angka)
b) Posisi dalam baris
c) Kata sebelum dan sesudahnya

CRF akan mencari parameter θ (theta) yang memaksimalkan kemungkinan prediksi label Y terhadap input X. Secara matematis, CRF memaksimalkan:

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_i \sum_k \lambda_k f_k(y_i - 1, y_i, x, i)\right)$$

Di mana:

a) f_k adalah fungsi fitur
b) λ_k adalah bobot parameter
c) Z(X) adalah fungsi normalisasi untuk semua kemungkinan label

Pelatihan dilakukan menggunakan library Python seperti **sklearn-crfsuite** atau **PyCRFSuite**, lalu hasil model digunakan untuk mengekstrak data dari hasil OCR secara otomatis.

4.2.3 Spesifikasi Data

Sistem ini dirancang untuk digunakan oleh satu jenis aktor, yaitu Administrator/Operator: Bertugas mengunggah dokumen invoice, memverifikasi hasil ekstraksi data, dan menyimpan informasi ke sistem. Pengguna diharapkan memiliki pengetahuan dasar tentang pengelolaan dokumen digital serta pemahaman sederhana terhadap elemen-elemen invoice.Sistem dirancang agar mudah digunakan bahkan oleh pengguna dengan latar belakang non-teknis.

Tabel 4. 2 *Spesifikasi Data*

No	User	Penjelasan
1	Admin	Admin akan memiliki hak akses untuk mengelola data yang sudah di upload.
2	Operator	Operator akan memiliki akses yang terbatas dibanding dengan admim hanya bisa meng upload invoice dan menyimpan datanya.

No	Tampilan	Fitur	Aktor
1	Halaman Login	Menampilkan form email dan password untuk mengakses halaman dashboard.	Admin, Operator
2	Halaman *Dashboard*	Menampilkan fitur yang digunakan dalam sistem dan menampilkan tampilan selamat datang.	Admin, Operator
3	Halaman validasi	Berisi upload invoice dan merubah gambar menjadi text,.	Admin, manager

4.2.4 Spesifikasi Perangkat Keras

Agar sistem dapat berjalan lancar, dibutuhkan perangkat keras dengan spesifikasi minimum sebagai berikut:

a) Prosesor: Intel Core i3 atau setara
b) RAM: Minimal 4 GB
c) Penyimpanan: Tersedia ruang kosong minimal 500 MB
d) Display: Resolusi minimal 1366x768 pixel
e) Koneksi Internet: Dibutuhkan untuk akses pustaka eksternal (seperti Tesseract.js dan model NLP)

Spesifikasi tersebut cukup untuk mendukung aktivitas upload file, pemrosesan teks, dan tampilan antarmuka berbasis web.

4.2.5 Spesifikasi Perangkat Lunak

Komponen perangkat lunak yang dibutuhkan dalam pengembangan dan implementasi sistem ini mencakup:

Tabel 4. 3 *Spesifikasi Perangkat Lunak*

Sistem Operasi	Windows 10 / Linux Ubuntu / macOS
Web Browser	Google Chrome
Bahasa Pemrograman	HTML, CSS, JavaScript
Text Editor	Visual Studio Code atau setara
Library Tambahan	Bootstrap 5, Tesseract.js, dan CRF model yang di-load via JavaScript atau Python

4.2.6 Spesifikasi Kebutuhan Sistem Baru

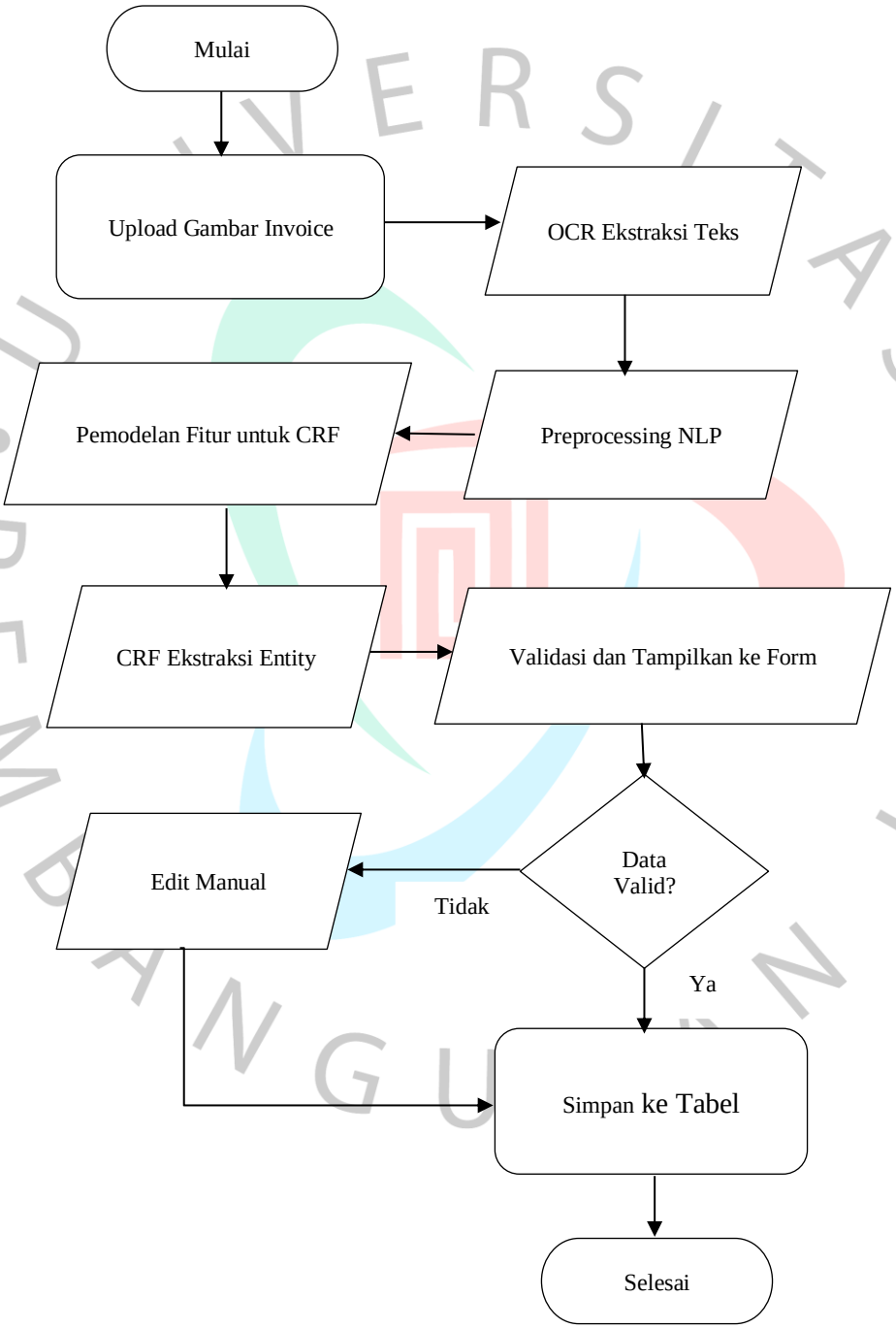
Pada tahap ini, dilakukan penyusunan rancangan sistem yang bertujuan untuk mengembangkan aplikasi validasi entry data secara otomatis. Rancangan ini menjadi pondasi awal sebelum sistem dikembangkan dan diimplementasikan secara menyeluruh. Perancangan dilakukan berdasarkan spesifikasi kebutuhan yang telah dianalisis sebelumnya, dengan fokus pada integrasi teknologi OCR (Optical Character Recognition), pemrosesan NLP (Natural Language Processing), serta penerapan algoritma Conditional Random Field (CRF) untuk akurasi ekstraksi data yang lebih tinggi. Subbab ini memuat rancangan visual serta logika dari alur sistem yang akan dibangun, mulai dari flowchart proses hingga tampilan antarmuka aplikasi..

4.2.7 Spesifikasi Kebutuhan Sistem Baru

Pada tahap ini, dilakukan penyusunan rancangan sistem yang bertujuan untuk mengembangkan aplikasi validasi entry data secara otomatis. Rancangan ini menjadi pondasi awal sebelum sistem dikembangkan dan diimplementasikan secara menyeluruh. Perancangan dilakukan berdasarkan spesifikasi kebutuhan yang telah dianalisis sebelumnya, dengan fokus pada integrasi teknologi OCR (Optical Character Recognition), pemrosesan NLP (Natural Language Processing), serta penerapan algoritma Conditional Random Field (CRF) untuk akurasi ekstraksi data yang lebih tinggi. Subbab ini memuat rancangan visual serta logika dari alur sistem yang akan dibangun, mulai dari flowchart proses hingga tampilan antarmuka aplikasi..

4.3 Flowchart Metode CRF

Flowchart menggambarkan langkah-langkah alur kerja sistem secara urut dari awal hingga akhir. Proses dimulai dari pengunggahan gambar invoice oleh pengguna, dilanjutkan dengan pengenalan teks menggunakan OCR, lalu pemrosesan teks menggunakan NLP untuk normalisasi dan tokenisasi.

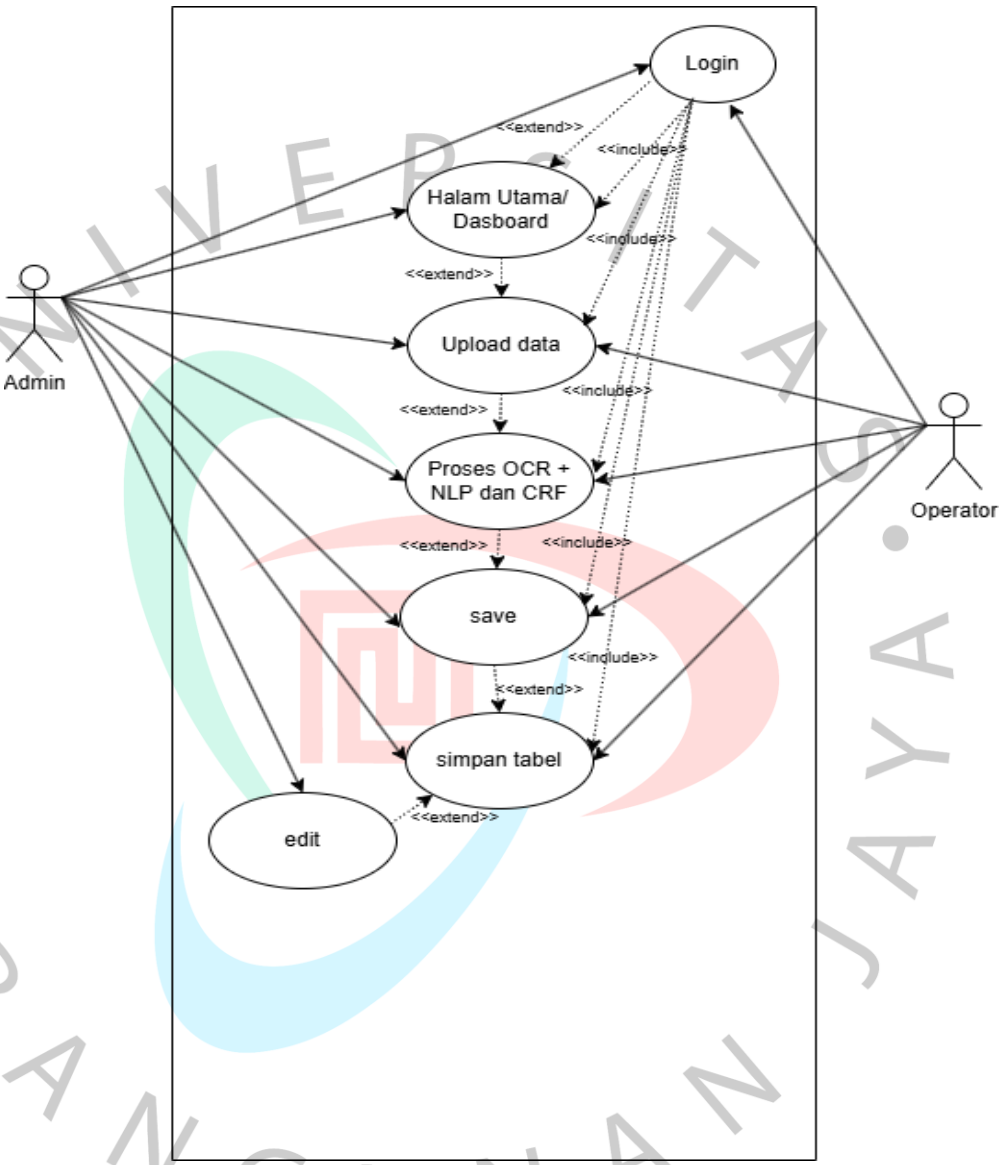


Gambar 4. 1 *Flowchart Metode CRF*

(1) Sistem diaktifkan, dan pengguna mengakses halaman validasi invoice.
(2) Pengguna mengunggah file invoice hasil pemindaian dengan format gambar (JPG, PNG) atau PDF.
(3) Sistem menjalankan proses Optical Character Recognition (OCR) untuk mengubah gambar menjadi teks digital.
(4) Hasil teks dari OCR dibersihkan dari noise, kesalahan spasi, karakter tak dikenal, dan dilakukan tokenisasi (pemecahan kata).
(5) Data teks diproses dengan teknik NLP seperti Named Entity Recognition (NER), POS Tagging, dan pattern matching untuk menyiapkan input ke CRF.
(6) Setiap token dianalisis dan dilabeli berdasarkan model CRF untuk menentukan entitas seperti: tanggal, nomor invoice, deskripsi barang, jumlah, harga, total, PPN, dan PPH.
(7) Hasil label CRF diverifikasi oleh sistem (dan/atau pengguna) untuk mengecek akurasi output.
(8) Sistem secara otomatis mengisi field isian dengan hasil ekstraksi CRF ke dalam antarmuka pengguna (form).
(9) Jika validasi selesai, pengguna dapat menyimpan data ke tabel dan menyimpan ke dalam database.
(10) Proses validasi dan ekstraksi selesai. Pengguna dapat memproses invoice berikutnya.

4.3.1 *Use Case*

Flowchart menggambarkan langkah-langkah alur kerja sistem secara urut dari awal hingga akhir. Proses dimulai dari pengunggahan gambar invoice oleh pengguna, dilanjutkan dengan pengenalan teks menggunakan OCR, lalu pemrosesan teks menggunakan NLP untuk normalisasi dan tokenisasi.



Gambar 4. 2 *Use Case* Diagram proses OCR + NPL metode CRF

Agar informasi tentang skenario *Use Case* dapat disajikan dengan jelas dan mudah dipahami, maka perlu dibuat tabel skenario. Tabel skenario harus dibuat dengan pihak yang terlibat, nama skenario, ringkasan singkat, skenario normal, dan skenario alternatif. Tabel skenario ini harus dibuat agar informasi tentang skenario *use case* jelas dan mudah dipahami.

Use Case	Halaman Login
Penjelasan	Admin dan operator dihadapkan dengan halaman login.
Skenario Utama	1. Admin dan operator masuk ke halaman login. 2. Form *login* berhasil ditampilkan oleh sistem.
Skenario Alternatif	1. User mengalami kesalahan saat memasukkan email/password. 2. Sistem menampilkan pesan error.
Kondisi Akir	User berhasil masuk kedalam aplikasi setelah melakukan validasi *login.*

Tabel 4. 4 *Login*

Use Case	Halaman Utama/*Dashboard*
Penjelasan	Admin dan operator mengakses halaman utama/*dashboard* aplikasi.
Skenario Utama	1. Aktor membuka aplikasi. 2. Sistem menampilkan halaman utama/*dashboard* dengan informasi dan *navigasi* ke fitur-fitur utama.
Skenario Alternatif	1. Aktor mengalami kesalahan saat membuka aplikasi. 2. Sistem menampilkan pesan error.
Kondisi Akir	Aktor berada di halaman utama aplikasi

Tabel 4. 5 *Menu Dasboar*

Use Case	Halaman Validasi
Penjelasan	Admin dan operator mengakses validasi aplikasi.
Skenario Utama	1. Admin dan operator memilih menu “Validasi data”. 2. Admin dan operator memilih menu “Data . 3. Admin dan operator memilih menu “Validasi Invoice sesuai PT 4. Aktor memiliki akses untuk mengubah, menambahkan dan . Save data
Skenario Alternatif	1. Aktor mengalami kesalahan saat membuka aplikasi. 2. Sistem menampilkan pesan error.
Kondisi Akir	Aktor berada di halaman validasi data

Tabel 4. 6 *Menu Validasi*

Use Case	Mengelola Upload data
Penjelasan	Admin dan operator mengakses upload data.
Skenario Utama	1. Admin dan operator memilih menu “Upload data”. 2. Aktor memiliki akses untuk upload data 3. Gambar ditampilkan di table preview,
Skenario Alternatif	1. Aktor mengalami kesalahan saat meng upload. 2. Sistem menampilkan pesan error.
Kondisi Akir	Aktor berada di halaman validasi data

Tabel 4.7 *Menu Upload Data*

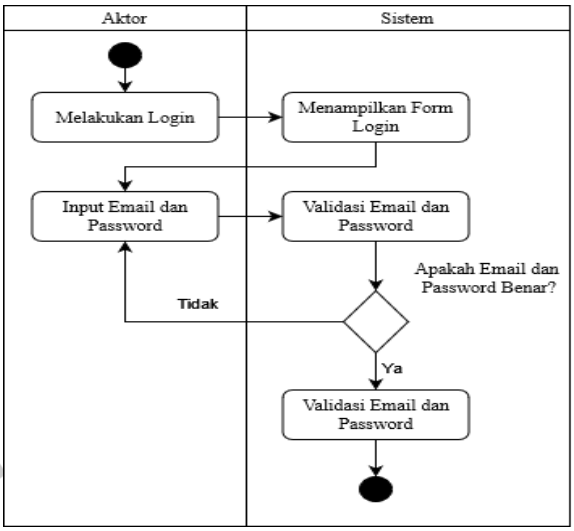
Use Case	Mengelola OCR + NLP metode CRF
Penjelasan	Admin dan operator mengakses extrak dari gambar ke text.
Skenario Utama	1. Admin dan operator memilih menu “Scan”. 2. Proses extrak berjalan 3. Hasi text akan masuk sesuai dengan tabelnya
Skenario Alternatif	1. Aktor mengalami kesalahan saat extrak. 2. Sistem menampilkan blank table.
Kondisi Akir	Aktor berada di halaman validasi data

Tabel 4. 8 *Menu Proses Validasi*

4.3.2 *Use Case*

Setelah perancangan Use Case diselesaikan, langkah berikutnya yang dilakukan oleh peneliti adalah menyusun activity diagram untuk masing-masing aktivitas yang terdapat di dalamnya. Diagram ini berfungsi untuk memvisualisasikan alur proses kerja sistem dalam setiap aktivitas yang berlangsung. Selanjutnya, peneliti menguraikan activity diagram berdasarkan peran atau pengguna yang terlibat dalam sistem, guna memberikan gambaran yang lebih rinci mengenai interaksi antar komponen dalam proses bisnis tersebut.

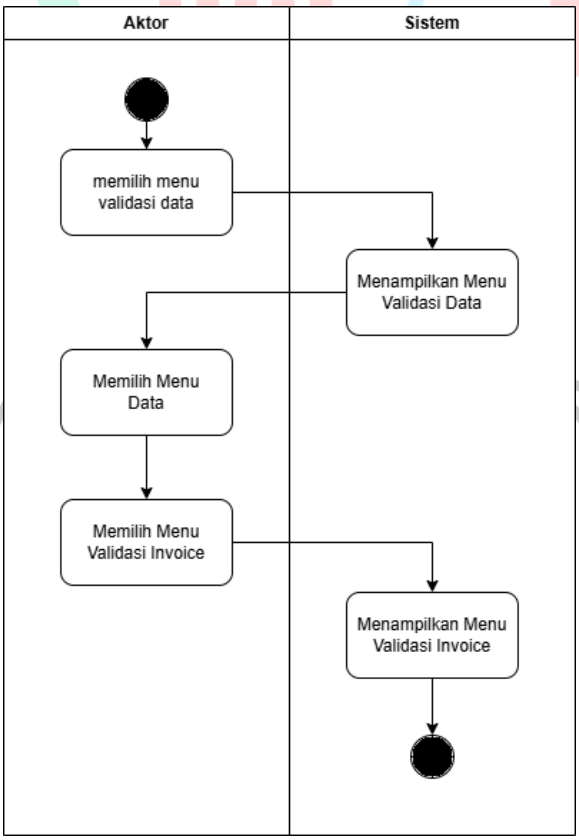
1. *Diagram Activity Login*



Gambar 4. 3 Diagram proses login

2. *Diagram Aktivity Validasi Data Menu*

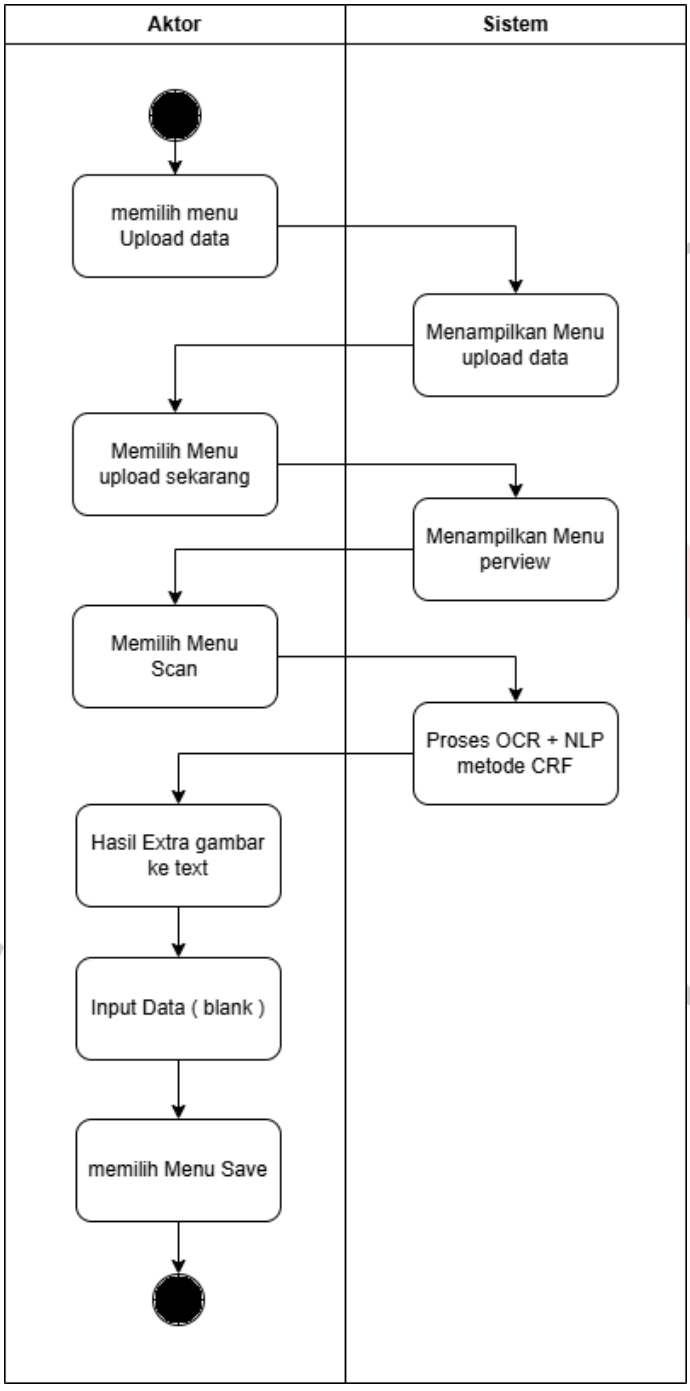
Diagram aktivity ini menjelaskan urutan proses ketika operator memilih menu mulai dashboard ke tampilan menu validasi data. Diagram aktivity dashboard operator digambarkan pada Gambar berikut



Gambar 4. 4 Diagram proses menu validasi

3. *Diagram Activity Validasi Data OCR + NLP metode CRF*

Diagram aktivity ini menunjukkan urutan prosedur yang digunakan ketika operator meng upload data. Data dapat ditambahkan operator. Aktivitas upload data ini akan mengahilakan data berupa text yang akan di simpan atau di edit sesuai table yang sudah di sediakan digambarkan pada gambar 4.16 berikut

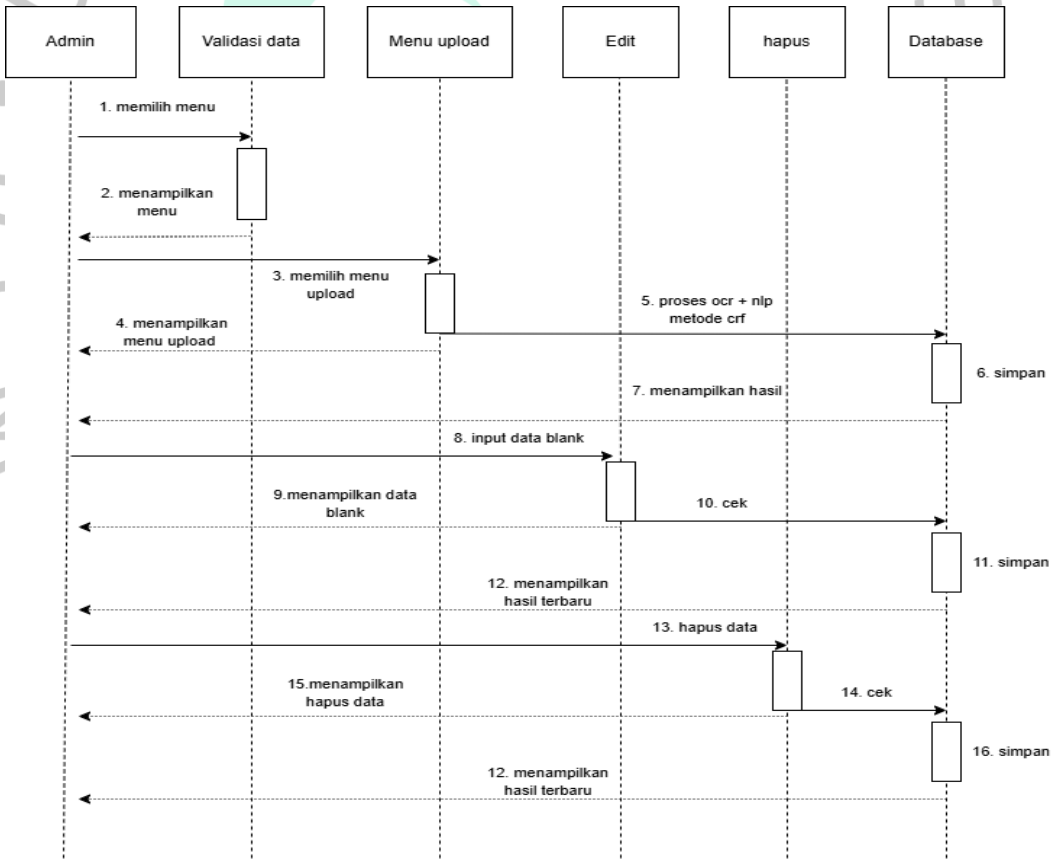


Gambar 4. 5 Diagram proses validasi invoice

4.3.3 *Sequence Diagram*

Sequence diagram merupakan salah satu jenis diagram dalam Unified Modeling Language (UML) yang berfungsi untuk menggambarkan interaksi antar objek dalam sebuah sistem berdasarkan urutan kejadian waktu. Diagram ini memperlihatkan komunikasi antar objek melalui pengiriman pesan selama proses pelaksanaan suatu skenario atau use case. Setiap objek direpresentasikan oleh garis vertikal yang dikenal sebagai lifeline, sementara pesan yang dikirimkan antar objek digambarkan melalui panah horizontal yang menghubungkan antar lifeline. Sequence diagram memudahkan pemahaman terhadap alur proses sistem, menetapkan peran serta tanggung jawab masing-masing objek, dan menyajikan dokumentasi skenario dinamis secara sistematis dan terstruktur.

1. *Sequence Diagram Validasi Data*



Gambar 4. 6 *Sequence Diagram Validasi Data*