

Implementasi Sistem Ekstraksi Data Otomatis Dari Formulir Administrasi Penduduk Menggunakan Optical Character Recognition Tesseract Di Kelurahan Berkas Kota Bengkulu

¹⁾ Ramalia Apriani, ²⁾ Khairil, ³⁾ Devi Sartika

¹ Mahasiswa, Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Dehasen Bengkulu

Alamat : Jalan pari rt 04 rw 02 kel berkas kec teluk segara

email : rmliapri18@gmail.com

^{2,3} Dosen Tetap, Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Dehasen Bengkulu

Jl. Meranti Raya No. 32 Kota Bengkulu 38228 Telp. (0736) 22027, 26957 Fax. (0736) 341139;

e-mail: khairil@unived.ac.id, devisaritika@unived.ac.id

(Received: Juni 2026, Revised: Agustus 2026, Accepiend: Oktober 2026)

Abstrak-Penelitian ini bertujuan membangun aplikasi digitalisasi dokumen administrasi penduduk di Kelurahan Berkas Kota Bengkulu menggunakan Optical Character Recognition (OCR) dan Rule-Based Text Extraction. Aplikasi dikembangkan menggunakan PHP dan MySQL untuk mengubah dokumen citra menjadi data terstruktur. Tesseract OCR digunakan sebagai metode utama dalam proses pembacaan teks, sedangkan EasyOCR digunakan sebagai metode perbandingan untuk mengevaluasi hasil pembacaan pada kondisi dokumen yang berbeda. Pengujian dilakukan menggunakan Black Box Testing terhadap 31 dokumen administrasi. Hasil pengujian menunjukkan seluruh fungsi aplikasi berjalan sesuai rancangan. Tesseract menghasilkan rata-rata 921,88 karakter dan 122,98 kata per dokumen, sedangkan EasyOCR menghasilkan rata-rata 908,26 karakter dan 121,16 kata per dokumen. Hasil perbandingan menunjukkan bahwa EasyOCR memberikan hasil pembacaan yang lebih baik pada beberapa kondisi dokumen, yaitu kertas bercak atau bernoda, kusut atau kerumuk, miring, dan buram, berdasarkan kesesuaian hasil pembacaan terhadap ground truth serta tingkat kesalahan pembacaan. Selanjutnya, Rule-Based Text Extraction berhasil mengenali 230 dari 248 field dengan tingkat keberhasilan sebesar 92,74%. Hasil tersebut menunjukkan bahwa aplikasi dapat membantu mempercepat proses digitalisasi dan pengolahan data dokumen administrasi penduduk.

Kata Kunci: Digitalisasi Dokumen, Tesseract OCR, EasyOCR, Rule-Based Text Extraction, Optical Character Recognition (OCR).

Intisari- ujuan penelitian ini adalah untuk mengembangkan aplikasi untuk mendigitalisasi dokumen administrasi penduduk di Kelurahan Berkas, Kota Bengkulu, dengan memanfaatkan Optical Character Recognition (OCR) dan Ekstraksi Teks Berbasis Aturan. Aplikasi ini dikembangkan menggunakan PHP dan MySQL untuk mengubah gambar dokumen menjadi data terstruktur. Tesseract OCR berfungsi sebagai metode utama untuk pembacaan teks, sedangkan EasyOCR digunakan sebagai metode perbandingan untuk mengevaluasi kinerja pada berbagai kondisi dokumen. Pengujian Black Box dilakukan pada 31 dokumen administratif, dan hasilnya menunjukkan bahwa semua fungsi aplikasi beroperasi sesuai dengan yang dirancang. Tesseract menghasilkan rata-rata 921,88 karakter dan 122,98 kata per dokumen, sedangkan EasyOCR menghasilkan rata-rata 908,26 karakter dan 121,16 kata per dokumen. Analisis perbandingan menunjukkan bahwa EasyOCR memberikan hasil pembacaan yang lebih baik dalam kondisi tertentu—seperti dokumen yang ternoda, kusut, miring, atau buram—berdasarkan kesesuaian hasil keluaran dengan data acuan (ground truth) serta tingkat kesalahannya. Selain itu, komponen Ekstraksi Teks Berbasis Aturan berhasil mengidentifikasi 230 dari 248 bidang,

dengan tingkat keberhasilan sebesar 92,74%. Hasil ini menunjukkan bahwa aplikasi ini dapat membantu mempercepat digitalisasi dan pemrosesan data dokumen administrasi penduduk.

Kata kunci: Digitalisasi Dokumen, Tesseract OCR, EasyOCR, Ekstraksi Teks Berbasis Aturan, Pengenalan Karakter Optik (OCR).

I. PENDAHULUAN

Pelayanan administrasi di tingkat kelurahan merupakan salah satu bentuk pelayanan publik yang paling sering bersentuhan langsung dengan masyarakat. Berbagai kebutuhan administrasi, seperti pembuatan Surat Keterangan Usaha, Surat Keterangan Tidak Mampu (SKTM), formulir permohonan KTP, maupun surat pengantar administrasi lainnya, hampir setiap hari diproses oleh petugas kelurahan. Meskipun sebagian pelayanan telah memanfaatkan komputer, proses pengelolaan data pada formulir administrasi masih banyak dilakukan secara manual. Informasi yang terdapat pada dokumen hasil pemindaian atau formulir yang telah diisi masyarakat umumnya masih harus diketik kembali ke dalam sistem sehingga membutuhkan waktu, tenaga, dan ketelitian yang tinggi.

Proses tersebut terlihat sederhana, tetapi dalam praktiknya sering menimbulkan berbagai kendala. Kesalahan pengetikan, data yang tidak konsisten, pengisian berulang, hingga lamanya proses pencarian arsip menjadi permasalahan yang masih ditemui dalam pengelolaan administrasi. Ketika jumlah dokumen terus bertambah, pekerjaan petugas menjadi semakin berat karena harus melakukan pemeriksaan satu per satu terhadap data yang tersimpan maupun dokumen fisik yang diarsipkan. Kondisi ini tentu dapat memengaruhi kecepatan pelayanan kepada masyarakat. Salah satu teknologi yang dapat dimanfaatkan untuk membantu permasalahan tersebut adalah Optical Character Recognition (OCR). Teknologi ini mampu mengenali karakter pada dokumen hasil pemindaian maupun foto, kemudian mengubahnya menjadi teks digital yang

dapat diproses oleh komputer. Dengan adanya OCR, proses pengetikan ulang data dapat dikurangi sehingga pekerjaan menjadi lebih cepat dan risiko kesalahan akibat *human error* juga dapat diminimalkan.

Namun demikian, hasil yang diperoleh dari OCR masih berupa kumpulan teks mentah. Sistem belum dapat membedakan secara otomatis mana yang merupakan Nomor Induk Kependudukan (NIK), nama penduduk, alamat, tempat dan tanggal lahir, maupun informasi lainnya. Apabila seluruh hasil OCR langsung disimpan ke dalam basis data tanpa proses identifikasi, maka informasi tersebut belum dapat dimanfaatkan secara optimal dalam proses administrasi.

Untuk mengatasi permasalahan tersebut, diperlukan mekanisme yang mampu mengelompokkan setiap hasil pembacaan OCR sesuai dengan jenis datanya. Salah satu pendekatan yang dapat digunakan adalah Rule-Based Text Extraction, yaitu teknik yang memanfaatkan aturan tertentu untuk mengenali pola informasi pada dokumen. Dengan pendekatan ini, setiap hasil ekstraksi dapat dipetakan ke dalam atribut yang sesuai sehingga menghasilkan data yang lebih terstruktur dan siap disimpan ke dalam basis data.

Beberapa penelitian sebelumnya telah membuktikan bahwa teknologi OCR memiliki tingkat akurasi yang cukup tinggi dalam proses ekstraksi teks. Penelitian Panuntun dan Sejati (2024) menunjukkan bahwa integrasi *Convolutional Neural Network* (CNN) dan OCR mampu meningkatkan akurasi verifikasi data kependudukan hingga 92%. Penelitian Nashiruddin dkk. (2025) juga menunjukkan bahwa OCR Tesseract mampu mengenali teks dengan tingkat akurasi mencapai 95%, meskipun masih diperlukan proses pasca-pemrosesan agar hasil ekstraksi menjadi lebih baik. Penelitian lain oleh Sangaji dkk. (2022) dan Reksiyono dkk. (2025) membuktikan bahwa pendekatan *rule-based* mampu meningkatkan ketepatan identifikasi data hasil OCR melalui proses klasifikasi berdasarkan aturan tertentu.

II. TINJAUAN PUSTAKA

Beberapa penelitian berikut akan menggambarkan dan memperkuat arah penelitian yang akan dilakukan. Selain itu melalui penelitian terdahulu ini juga penulis dapat melakukan pendekatan terhadap tema penelitian.

No	Peneliti dan Tahun	Metode yang Digunakan	Objek Penelitian	Hasil Penelitian	Kesimpulan Penelitian
1	Reksiyono dkk. (2025)	Analisis data kependudukan menggunakan OCR Tesseract dan Python	Identifikasi data kependudukan	Hasil penelitian menunjukkan bahwa OCR Tesseract mampu mengenali data kependudukan dengan tingkat akurasi yang tinggi.	OCR Tesseract dapat digunakan untuk mengotomatiskan proses identifikasi data kependudukan.
2	Nashiruddin dkk. (2025)	Analisis data kependudukan menggunakan OCR Tesseract dan Python	Identifikasi data kependudukan	Hasil penelitian menunjukkan bahwa OCR Tesseract mampu mengenali data kependudukan dengan tingkat akurasi yang tinggi.	OCR Tesseract dapat digunakan untuk mengotomatiskan proses identifikasi data kependudukan.
3	Panuntun dan Sejati (2024)	Analisis data kependudukan menggunakan OCR Tesseract dan Python	Identifikasi data kependudukan	Hasil penelitian menunjukkan bahwa OCR Tesseract mampu mengenali data kependudukan dengan tingkat akurasi yang tinggi.	OCR Tesseract dapat digunakan untuk mengotomatiskan proses identifikasi data kependudukan.
4	Sangaji dkk. (2022)	Analisis data kependudukan menggunakan OCR Tesseract dan Python	Identifikasi data kependudukan	Hasil penelitian menunjukkan bahwa OCR Tesseract mampu mengenali data kependudukan dengan tingkat akurasi yang tinggi.	OCR Tesseract dapat digunakan untuk mengotomatiskan proses identifikasi data kependudukan.
5	Reksiyono dkk. (2025)	Analisis data kependudukan menggunakan OCR Tesseract dan Python	Identifikasi data kependudukan	Hasil penelitian menunjukkan bahwa OCR Tesseract mampu mengenali data kependudukan dengan tingkat akurasi yang tinggi.	OCR Tesseract dapat digunakan untuk mengotomatiskan proses identifikasi data kependudukan.

Gambar 1. Penelitian Terdahulu

Administrasi Kependudukan adalah rangkaian kegiatan penataan dan penertiban dalam penerbitan dokumen dan data kependudukan. Ini mencakup pendaftaran penduduk, pencatatan sipil, pengelolaan informasi administrasi, serta pendayagunaan hasilnya untuk pelayanan publik dan pembangunan sektor (Undang-Undang Republik Indonesia Nomor 24 Tahun 2013, 2013).

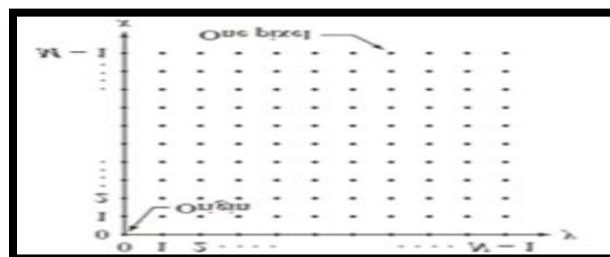
Citra atau gambar dapat didefinisikan sebagai fungsi dua dimensi $f(x,y)$, dimana x dan y adalah kordinat bidang datar; dan harga fungsi f di setiap pasangan kordinat (x,y) disebut intensitas atau level keabuan (*greylevel*) dari gambar di titik itu. Apabila nilai x,y dan f secara keseluruhan berhingga (*finite*) dan bernilai diskrit maka dapat dikatakan bahwa citra tersebut adalah citra digital (Jumadi et al., 2020).

Pengolahan citra merupakan cabang ilmu dalam Artificial Intelegence yang menggunakan objek citra dalam bentuk digital untuk penyelesaian kasusnya. Metode dalam citra dapat digunakan baik perhitungan matematis pada objek secara piksel ataupun geometris. Masing-masing objek citra memiliki nilai perbedaan yang dapat diperhitungkan secara matematis, sehingga menunjukkan ciri yang berbeda antara objek satu dengan yang lain. Penciri dari perbedaan setiap objek dapat ditentukan dari warna, tekstur, ataupun bentuk (Jumadi et al., 2021)



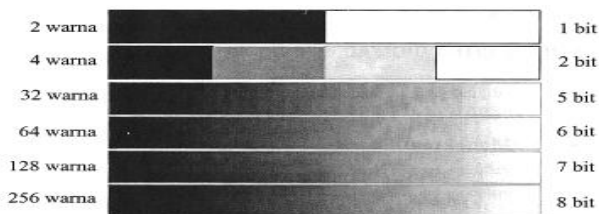
Gambar 2. Ilustrasi Citra Digital

Nilai suatu pixel memiliki nilai dalam rentang tertentu, dari nilai minimum sampai nilai maksimum. Jangkauan yang digunakan berbeda-beda tergantung dari jenis warnanya. Namun secara umum jangkauannya adalah 0 – 255. Citra dengan penggambarannya seperti ini digolongkan kedalam citra integer. Berikut adalah jenis-jenis citra berdasarkan nilai pixelnya. Citra biner adalah citra digital yang hanya memiliki dua kemungkinan nilai pixel yaitu hitam dan putih. Citra biner juga disebut sebagai citra *B&W (black and white)* atau citra monokrom. hanya dibutuhkan 1 bit untuk mewakili nilai setiap pixel dari citra biner



Gambar 3. Citra Biner

Citra *grayscale* menangani gradasi warna hitam dan putih, yang tentu saja menghasilkan warna abu-abu. Dalam hal ini intensitas berkisar antara 0 sampai dengan 255. Nilai 0 menyatakan hitam dan nilai 255 menyatakan putih. Banyaknya warna tergantung pada jumlah bit yang disediakan dimemori untuk menampung kebutuhan warna ini. Semakin besar jumlah bit warna yang disediakan di memori, semakin halus gradasi warna yang terbentuk.



Gambar. 4. Citra Perbandingan Gradasi
Warna 1 bit, 2 bit, 5 bit, 6

Setiap piksel pada citra warna memiliki warna yang merupakan kombinasi dari tiga warna dasar RGB (*Red, Green, Blue*). Setiap warna dasar menggunakan penyimpanan 8 bit = 1 byte, yang berarti setiap warna mempunyai gradasi sebanyak 255 warna. Berarti setiap piksel mempunyai kombinasi warna sebanyak $28 \cdot 28 \cdot 28 = 224 = 16$ juta warna lebih.



Gambar 5. 8 Bit true color

Tahap *preprocessing* merupakan fondasi utama dalam sistem ekstraksi data otomatis. Teknologi OCR Tesseract sangat bergantung pada kualitas visual citra input; adanya gangguan seperti derau (*noise*), kemiringan dokumen, atau kontras yang rendah dapat mengakibatkan kegagalan pengenalan karakter (Ponggohong et al., 2025).

Optical Character Recognition (OCR) adalah metode yang memungkinkan komputer untuk mengenali dan mengekstrak teks dari gambar atau dokumen digital (Darmi et al., 2025). Sedangkan menurut Zulhijar et al (2024) OCR adalah sebuah teknologi yang memungkinkan perangkat untuk membaca dan mengonversi teks yang tercetak pada dokumen fisik menjadi data digital yang dapat diproses lebih lanjut dengan sistem komputer.

Tesseract merupakan *engine* OCR sumber terbuka yang pada versi terbarunya (versi 4 dan 5) telah mengadopsi arsitektur *Neural Network* berbasis *Long Short-Term Memory* (LSTM). Implementasi LSTM dalam sistem OCR bertujuan untuk meningkatkan akurasi pengenalan teks melalui pendekatan *sequence learning*, di mana mesin tidak hanya mengenali karakter secara individual, tetapi juga mempelajari urutan dan konteks dari string yang dibaca (Sugiarta et al., 2021).

Ekstraksi teks berbasis aturan (*Rule-Based Text Extraction*) adalah sebuah metodologi dalam pemrosesan bahasa alami (*Natural Language Processing*) yang menggunakan sekumpulan instruksi logis, pola deterministik, atau aturan heuristik yang telah ditentukan sebelumnya untuk mengidentifikasi dan mengambil informasi spesifik dari teks tidak terstruktur (Colter et al., 2022).

Implementasi *Regular Expression* (Regex) dalam penelitian ini berfungsi sebagai algoritma penapis (*filtering*) yang bekerja pada tahap pasca-pemrosesan OCR. Tesseract seringkali menghasilkan *raw text* yang mengandung karakter sampah akibat garis tepi formulir atau bercak pada dokumen (Zulhijar et.al., 2024). Oleh karena itu, Regex digunakan untuk melakukan *pattern matching* yang kaku, seperti pencarian pola 16 digit angka secara berurutan untuk mengidentifikasi NIK (Sugiarta et.al., 2021).

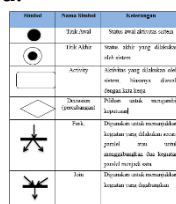
Pembersihan data (*Data Cleaning*) adalah tahap krusial pasca-OCR untuk memperbaiki kesalahan pembacaan mesin (*misrecognition*) dan menghilangkan karakter sampah (*noise*) yang muncul akibat kualitas citra yang rendah atau gangguan garis pada formulir.

Salah satu bahasa pemrograman yang mulai banyak digunakan saat ini untuk pengembangan berbagai macam aplikasi berbasis desktop, web, dan mobile adalah bahasa pemrograman python. Kepopuleran penggunaan python menempatkannya menjadi bahasa pemrograman yang mulai banyak di pelajari oleh mahasiswa di kampus yang berbasis IT, guna menyelesaikan tugas kuliah, tugas akhir maupun tugas penelitian. Python merupakan bahasa pemrograman tingkat tinggi yang yang dibuat oleh Guido Van Rossum dan dirilis pada tahun 1991. Python juga merupakan bahasa yang sangat populer belakangan ini. Selain itu python juga merupakan bahasa pemrograman yang multi fungsi contohnya python dapat digunakan untuk *Machine Learning* dan *Deep Learning*. Python dipilih sebagai penelitian karena python memiliki penulisan sintaksis yang mudah selain itu python juga memiliki *library* yang lengkap dan memiliki dukungan komunitas yang kuat karena python bersifat *open source*. Untuk menuliskan *source code* python anda dapat menggunakan IDE seperti *Visual Studio Code*, *Sublime Text*, *PyCharm* atau anda juga dapat menggunakan IDE online seperti *Jupyter notebook* dan *google colab*. (Alfarizi, et.all, 2023).

UML merupakan kepanjangan dari *Unified Modeling Language* yang memiliki arti bahasa pemodelan standar. Wazlawick mengatakan "... *language that can be used to describe things*". Yang apabila diterjemahkan adalah bahwa UML dapat digunakan untuk mendeskripsikan sesuatu. Menurut Wazlawick terdapat simbol penulisan yang membantu penulis

untuk menjelaskan mengenai hasil yang diekspektasikan. Simbol-simbol yang dimaksud disebut dengan notasi dan proses. (Limantoro & Kristiadi, 2021).

Diagram aktivitas (*Activity Diagram*) adalah tipe khusus dari diagram status yang memperlihatkan aliran dari suatu aktivitas ke aktivitas lainnya dalam suatu sistem (Limantoro & Kristiadi, 2021), *Activity diagram* digunakan untuk mendokumentasikan alur kerja pada sebuah sistem, yang dimulai dari pandangan business level hingga ke operational level. Pada dasarnya, *activity diagram* merupakan variasi dari statechart diagram. *Activity diagram* mempunyai peran seperti halnya flowchart, akan tetapi perbedaannya dengan flowchart adalah *activity diagram* bisa mendukung perilaku paralel sedangkan flowchart tidak bisa.



Gambar 6. Notasi Activity Diagram

Diagram ini memperlihatkan himpunan *use case* dan aktor-aktor (suatu jenis khusus dari kelas). Diagram ini terutama sangat penting untuk mengorganisasi dan memodelkan perilaku suatu sistem yang dibutuhkan serta diharapkan pengguna (Limantoro & Kristiadi, 2021). *Use Case diagram* yaitu model hasil analisis perancangan sistem yang bertujuan untuk mendeskripsikan kebutuhan sistem. Kebutuhan sistem tersebut akan diterapkan oleh pengguna sehingga perancangan sistem dapat tergambarkan.

Class diagram merupakan kumpulan dari beberapa class dan relasinya. Class identik dengan entity yang direpresentasikan dalam bentuk persegi dimana pada bagian atas ditulis nama class, kemudian ke bawah ditulis attribute yang terdapat pada class, kemudian ke bawah lagi ditulis metode yang ada pada class. Class Diagram adalah sebuah spesifikasi yang jika diinstansiasi akan menghasilkan sebuah objek dan merupakan inti dari pengembangan dan desain berorientasi objek. (Andhika, et.all, 2022).

Flowchart yang sering disebut dengan diagram alir adalah suatu jenis diagram yang merepresentasikan algoritma atau langkah-langkah instruksi yang berurutan dalam sistem. flowchart dibuat dengan tujuan agar melihatkan setiap proses yang harus dilalui dalam suatu sistem (Febryana et al., 2024).

III. METODOLOGI PENELITIAN

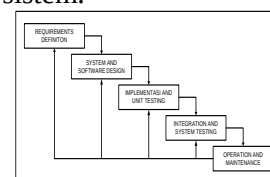
Penelitian ini menggunakan jenis penelitian Rekayasa Perangkat Lunak (Software Engineering Research) dengan pendekatan eksperimental. Penelitian difokuskan pada perancangan dan

pembangunan sistem ekstraksi data otomatis dari formulir administrasi penduduk menggunakan Optical Character Recognition (OCR) Tesseract dan Rule-Based Text Extraction, kemudian mengevaluasi kinerja sistem berdasarkan tingkat akurasi ekstraksi data yang dihasilkan.

Subjek dalam penelitian ini adalah formulir administrasi penduduk yang digunakan pada pelayanan di Kelurahan Berkas Kota Bengkulu, yaitu Surat Keterangan Usaha dan Surat Keterangan Tidak Mampu (SKTM). Dokumen yang digunakan sebagai data penelitian berupa citra digital hasil pemindaian (*scanning*) maupun hasil pengambilan gambar menggunakan kamera dengan variasi kualitas pencahayaan, resolusi, dan tingkat kemiringan (*skew*). Variasi tersebut digunakan untuk menguji kemampuan sistem dalam melakukan ekstraksi informasi secara otomatis menggunakan OCR Tesseract dan Rule-Based Text Extraction.

Waktu penelitian dimulai dari bulan November 2025 sampai dengan Mei 2026. Penelitian dilakukan di Kantor Kelurahan Berkas yang beralamat di Jalan Pari RT.01, Bengkulu, Kota Bengkulu, Bengkulu 38114.

Metode penelitian yang digunakan pada penelitian ini adalah metode *waterfall*. Dengan metode *waterfall* mampu melakukan analisa bertahap. Analisa kebutuhan digunakan untuk mengetahui dari kelemahan sistem yang lama, kemudian membuat desain dari rancangan tersebut dan dilanjutkan dengan pembuatan rancangan sistem baru yang meliputi kode-kode program. Setelah sistem baru selesai di ujikan sistem tersebut. Jika tidak ada kesalahan, maka sistem akan diimplementasikan dan pemeliharaan sistem.



Gambar 7. Diagram *Waterfall*

Mengumpulkan kebutuhan secara lengkap kemudian kemudian dianalisis dan didefinisikan kebutuhan yang harus dipenuhi oleh program yang akan dibangun. Fase ini harus dikerjakan secara lengkap untuk bisa menghasilkan desain yang lengkap. Pada tahap ini pengembang sistem diperlukan suatu komunikasi yang bertujuan untuk memahami software yang diharapkan pengguna dan batasan *software*. Informasi ini biasanya dapat diperoleh melalui wawancara, survey atau diskusi. Informasi tersebut dianalisis untuk mendapatkan data yang dibutuhkan oleh pengguna.

Desain dikerjakan setelah kebutuhan selesai dikumpulkan secara lengkap. Kebutuhan dari tahap pertama akan dipelajari dalam fase ini dan desain

sistem disiapkan. Desain Sistem membantu dalam menentukan perangkat keras dan sistem persyaratan dan juga membantu dalam mendefinisikan arsitektur sistem secara keseluruhan.

Pada tahap ini, sistem pertama kali dikembangkan di program kecil yang disebut unit, yang terintegrasi dalam tahap berikutnya. Desain program diterjemahkan ke dalam kode-kode dengan menggunakan bahasa pemrograman yang sudah ditentukan.

Penyatuan unit-unit program kemudian diuji secara keseluruhan (*system testing*). Semua unit yang dikembangkan dalam tahap implementasi diintegrasikan ke dalam sistem setelah pengujian masing-masing unit. Pasca integrasi seluruh sistem diuji untuk mengecek setiap kesalahan dan kegagalan.

Ini merupakan tahap terakhir dalam model waterfall. *Software* yang sudah jadi dijalankan serta dilakukan pemeliharaan. Pemeliharaan termasuk dalam memperbaiki kesalahan yang tidak ditemukan pada langkah sebelumnya. Perbaikan implementasi unit sistem dan peningkatan jasa sistem sebagai kebutuhan baru.

Tahapan penelitian ini akan menggambarkan *flow diagram* yang akan dilakukan mulai dari identifikasi masalah sampai dengan evaluasi dan analisa hasil penelitian dengan tujuan agar proses penelitian berjalan secara sistematis.



Gambar 8. Tahapan Penelitian

Tahap Inisiasi dan Literasi

Pada tahap awal, dilakukan Studi Literatur yang mendalam mengenai komponen utama sistem, yaitu *engine Optical Character Recognition (OCR) Tesseract*, teknik pengolahan citra menggunakan library OpenCV, dan mekanisme pencocokan pola menggunakan Regular Expression (Regex).

Tahap Pengumpulan Data

Tahapan ini difokuskan pada perolehan dataset yang representatif. Peneliti mengumpulkan sampel formulir administrasi penduduk baik dalam bentuk dokumen fisik yang dipindai (*scanning*) maupun tangkapan kamera digital. Dataset ini disusun dengan berbagai variasi kondisi fisik untuk menguji reliabilitas sistem, mencakup variasi tingkat pencahayaan, adanya *noise* (bercak), hingga kemiringan posisi dokumen.

Tahap Analisa dan Perancangan

Pada tahap ini, dilakukan analisis kebutuhan fungsional dan non-fungsional sistem. Peneliti

melakukan pemetaan entitas data yang perlu diekstraksi (seperti NIK, Nama, dan Alamat) serta merancang arsitektur perangkat lunak yang mengintegrasikan lapisan preprocessing, mesin OCR, dan algoritma *Rule-Based Text Extraction*. Rancangan ini divisualisasikan melalui diagram arsitektur dan UML untuk mempermudah implementasi kode.

Tahap Implementasi Sistem

Tahap implementasi melibatkan proses coding menggunakan bahasa pemrograman Python. Peneliti membangun modul *image preprocessing* menggunakan OpenCV (seperti *grayscale*, *binarization*, dan *deskewing*), mengintegrasikan *engine Tesseract OCR*, dan mengembangkan algoritma *Rule-Based Text Extraction*. Algoritma ini dirancang khusus untuk melakukan klasifikasi teks mentah hasil OCR menjadi data terstruktur melalui logika *keyword matching* dan *Regex pattern matching*.

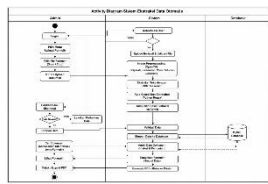
Sistem yang telah dibangun kemudian diuji menggunakan dataset yang telah disiapkan. Parameter pengujian difokuskan pada akurasi ekstraksi entitas kunci seperti NIK (16 digit angka) dan Nama Lengkap. Pada tahap ini akan dilakukan pengamatan dan evaluasi seperti:

- Jika berhasil (akurasi telah memenuhi target): akan melanjutkan ke tahap analisis hasil dan pembahasan.
- Jika tidak berhasil (akurasi belum memenuhi target): Terjadi proses *loopback* (umpan balik) ke tahap implementasi. Dimana sistem akan melakukan optimasi pada parameter *preprocessing* atau memperbaiki pola *Regex* hingga sistem mampu mencapai tingkat akurasi yang diharapkan.

Tahap Akhir

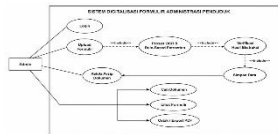
Tahapan terakhir adalah penyusunan secara menyeluruh. Dan melakukan sintesis atas hasil eksperimen, menarik kesimpulan yang menjawab rumusan masalah, serta memberikan saran untuk pengembangan penelitian selanjutnya.

Perancangan arsitektur sistem merupakan gambaran umum mengenai alur kerja sistem yang akan dibangun pada penelitian ini. Arsitektur sistem disusun untuk memperlihatkan hubungan antarproses mulai dari proses akuisisi dokumen, pengolahan citra, ekstraksi informasi, penyimpanan data, hingga penyajian kembali hasil ekstraksi kepada pengguna. Melalui perancangan ini, setiap tahapan pengolahan data dapat dilakukan secara sistematis sehingga tujuan penelitian untuk mendigitalisasi formulir administrasi penduduk dapat tercapai.



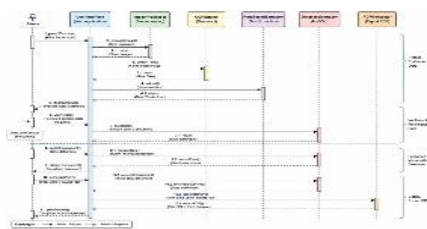
Gambar 9. Perancangan Arsitektur Sistem

Use Case Diagram digunakan untuk menggambarkan hubungan antara aktor dengan fungsi-fungsi utama yang tersedia pada sistem. Pada penelitian ini hanya terdapat satu aktor, yaitu **Admin**, yang berperan sebagai pengguna sistem dan bertanggung jawab terhadap seluruh proses digitalisasi formulir administrasi penduduk. Admin memiliki hak akses untuk mengunggah dokumen, melakukan verifikasi hasil ekstraksi, mengelola arsip dokumen, melakukan pencarian data, menampilkan kembali formulir administrasi, serta mencetak dokumen apabila diperlukan.



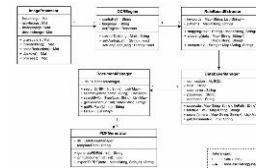
Gambar 10. Use Case Sistem

Activity Diagram digunakan untuk menggambarkan alur aktivitas sistem secara berurutan mulai dari proses unggah dokumen hingga dokumen administrasi berhasil disimpan dan dapat digunakan kembali. Diagram ini memperlihatkan interaksi antara Admin, Sistem, dan Basis Data dalam menjalankan proses digitalisasi formulir administrasi penduduk. Melalui diagram ini, setiap tahapan proses dapat dipahami secara sistematis sehingga aliran data yang terjadi pada sistem menjadi lebih jelas.



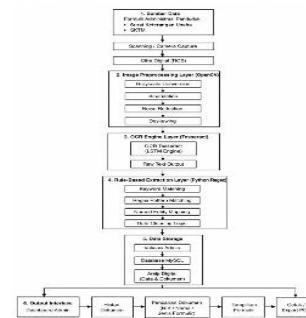
Gambar 11. Activity Diagram

Class Diagram digunakan untuk menggambarkan struktur statis sistem yang dibangun melalui hubungan antar kelas, atribut, dan operasi (*method*). Diagram ini memberikan gambaran mengenai pembagian tanggung jawab setiap kelas dalam menjalankan proses digitalisasi formulir administrasi penduduk. Dengan adanya pemisahan fungsi pada setiap kelas, sistem menjadi lebih terstruktur, mudah dikembangkan, serta memudahkan proses pemeliharaan perangkat lunak.



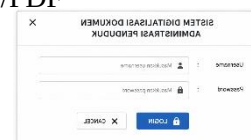
Gambar 12. Class Diagram

Sequence Diagram digunakan untuk menggambarkan urutan interaksi antar objek dalam sistem berdasarkan waktu (*time sequence*). Diagram ini menunjukkan bagaimana setiap objek saling berkomunikasi melalui pengiriman pesan (*message*) mulai dari proses unggah formulir, pengolahan citra, ekstraksi data, penyimpanan ke basis data, hingga penyajian kembali dokumen administrasi kepada pengguna. Dengan demikian, Sequence Diagram memberikan gambaran yang lebih rinci mengenai mekanisme kerja sistem dibandingkan *Use Case Diagram* maupun *Activity Diagram*.



Gambar 13. Sequence Diagram

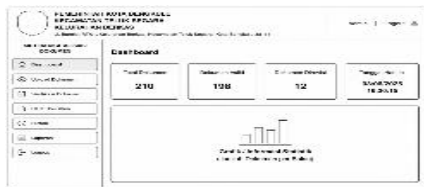
Perancangan *interface* aplikasi bertujuan memberikan gambaran mengenai tampilan sistem yang akan dibangun. Rancangan antarmuka disusun berdasarkan kebutuhan fungsional sistem yang telah dianalisis pada subbab sebelumnya sehingga setiap halaman memiliki fungsi yang mendukung proses digitalisasi formulir administrasi penduduk. Perancangan ini meliputi halaman Login, Dashboard, Upload Dokumen, Verifikasi Hasil OCR, Data Dokumen, Histori Dokumen, Detail Formulir, dan Print Preview/PDF



Gambar 14. Perancangan Login Aplikasi



Gambar 15. Halaman Dashboard Aplikasi



Gambar 16. Halaman Upload Dokumen



Gambar 17. Rancangan Halaman Verifikasi Hasil OCR



Gambar 18. Rancangan Halaman Data Dokumen



Gambar 19. Rancangan Halaman Detail Formulir



Gambar 20. Rancangan Halaman Histori Dokumen



Gambar 21. Rancangan Halaman Print Preview

Rancangan pengujian dan evaluasi dilakukan untuk mengetahui apakah sistem yang dibangun telah berfungsi sesuai dengan kebutuhan pengguna serta mampu menghasilkan proses ekstraksi data yang akurat. Tahap pengujian tidak hanya berfokus pada keberhasilan fungsi aplikasi, tetapi juga mengevaluasi kemampuan sistem dalam mengenali karakter pada dokumen administrasi penduduk menggunakan metode Optical Character Recognition (OCR) yang dikombinasikan dengan Rule-Based Text Extraction. Pengujian dilakukan mulai dari proses login, unggah dokumen, ekstraksi data menggunakan OCR, verifikasi hasil ekstraksi, penyimpanan data, pencarian dokumen, hingga pencetakan dokumen dalam format PDF. Selain itu,

penelitian ini juga mengukur tingkat akurasi hasil ekstraksi dengan membandingkan data yang dihasilkan sistem terhadap data asli (*ground truth*). Metode pengujian yang digunakan pada penelitian ini terdiri atas dua tahap, yaitu pengujian fungsional menggunakan Black Box Testing dan pengujian akurasi menggunakan Word Error Rate (WER) serta Character Error Rate (CER).

Pengujian fungsional dilakukan menggunakan metode Black Box Testing. Metode ini digunakan untuk menguji setiap fungsi pada aplikasi berdasarkan masukan (*input*) dan keluaran (*output*) yang dihasilkan tanpa memperhatikan struktur maupun kode program yang digunakan. Tujuan pengujian ini adalah memastikan bahwa seluruh fitur pada sistem dapat berjalan sesuai dengan kebutuhan pengguna. Setiap fungsi diuji menggunakan beberapa skenario yang mewakili aktivitas utama pada aplikasi, kemudian hasil yang diperoleh dibandingkan dengan hasil yang diharapkan. Apabila hasil yang diperoleh sesuai dengan hasil yang diharapkan, maka fungsi tersebut dinyatakan berhasil.

Selain melakukan pengujian terhadap fungsi aplikasi, penelitian ini juga mengevaluasi tingkat akurasi sistem dalam mengenali teks pada dokumen administrasi penduduk. Pengujian dilakukan dengan membandingkan hasil ekstraksi yang dihasilkan oleh sistem terhadap data asli (*ground truth*). Semakin kecil tingkat kesalahan yang dihasilkan, maka semakin baik kemampuan sistem dalam mengenali informasi pada dokumen.

Pengujian akurasi dilakukan menggunakan dokumen administrasi penduduk yang terdiri atas Surat Keterangan Usaha dan Surat Keterangan Tidak Mampu (SKTM) sebagai data uji. Setiap dokumen dipindai menggunakan scanner maupun kamera, kemudian diproses oleh sistem melalui tahapan image preprocessing, Optical Character Recognition (OCR), Rule-Based Text Extraction, verifikasi hasil ekstraksi, hingga penyimpanan ke dalam basis data. Selanjutnya hasil ekstraksi dibandingkan dengan data asli (*ground truth*) untuk memperoleh nilai Word Error Rate (WER) dan Character Error Rate (CER). Selain itu, seluruh fungsi aplikasi juga diuji menggunakan Black Box Testing guna memastikan bahwa sistem mampu menjalankan setiap proses sesuai dengan kebutuhan pengguna.

Output pengujian merupakan hasil yang diharapkan setelah seluruh proses pada sistem digitalisasi dokumen administrasi penduduk selesai dijalankan. Tahap ini bertujuan untuk memastikan bahwa sistem tidak hanya mampu melakukan ekstraksi data dari dokumen administrasi menggunakan Optical



Character Recognition (OCR), tetapi juga mampu mengelola hasil ekstraksi tersebut menjadi informasi yang terstruktur, tersimpan dengan baik, serta dapat digunakan kembali dalam proses pelayanan administrasi. Dari sisi fungsional, sistem diharapkan mampu menjalankan seluruh proses sesuai dengan rancangan yang telah dibuat, mulai dari proses login, unggah dokumen, image preprocessing, ekstraksi teks menggunakan OCR, Rule-Based Text Extraction, verifikasi hasil ekstraksi, penyimpanan data ke dalam basis data, pencarian dokumen, penampilan kembali formulir administrasi, hingga proses cetak dan export PDF.

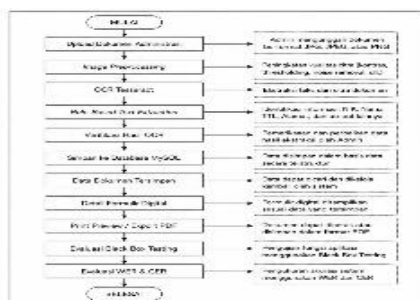
Gambar 21. Alur Output Sistem dan Evaluasi

IV. HASIL DAN PEMBAHASAN

Tahap implementasi merupakan proses penerapan hasil perancangan sistem yang telah dijelaskan pada Bab III menjadi sebuah aplikasi yang dapat digunakan untuk mendukung proses digitalisasi dokumen administrasi penduduk di Kelurahan Berkas Kota Bengkulu. Sistem dikembangkan berbasis web menggunakan bahasa pemrograman PHP dengan MySQL sebagai media penyimpanan data. Pada proses pengenalan teks, Optical Character Recognition (OCR) Tesseract digunakan sebagai metode utama sesuai dengan fokus penelitian untuk mengubah dokumen dalam bentuk citra menjadi teks digital. Selanjutnya, hasil pembacaan OCR diproses menggunakan Rule-Based Text Extraction untuk mengidentifikasi dan mengambil informasi penting dari dokumen administrasi.



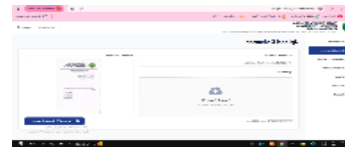
Gambar 22. Implementasi Halaman Login Aplikasi



Gambar 23. Implementasi Halaman Dashboard



Gambar 24. Implementasi Halaman Upload Dokumen



Gambar 25. Implementasi Halaman Upload Setelah Dokumen Dipilih



Gambar 26. Implementasi Halaman Verifikasi Dokumen



Gambar 27. Implementasi Halaman Data Dokumen



Gambar 28. Implementasi Halaman Detail Dokumen



Gambar 29. Implementasi Halaman Histori Dokumen



Gambar 4.9 Implementasi Halaman Laporan

Pada tahap ini dilakukan implementasi metode Optical Character Recognition (OCR) dan Rule-Based Text Extraction yang digunakan dalam penelitian untuk mendukung proses digitalisasi dokumen administrasi penduduk di Kelurahan Berkas Kota Bengkulu. Tesseract OCR digunakan sebagai metode utama sesuai dengan fokus penelitian, yaitu melakukan pengenalan teks dari dokumen administrasi dalam bentuk citra menjadi teks digital. Hasil pembacaan OCR selanjutnya diproses menggunakan Rule-Based Text Extraction untuk mengidentifikasi dan mengambil informasi penting sesuai dengan struktur formulir administrasi penduduk.

Tahap setelah dokumen berhasil diunggah adalah proses ekstraksi teks menggunakan metode *Optical Character Recognition* (OCR). Pada implementasinya, sistem menggunakan dua mesin OCR, yaitu Tesseract OCR sebagai metode utama sesuai dengan fokus penelitian dan EasyOCR sebagai metode pembanding.



Gambar 30. Implementasi Proses OCR Menggunakan EasyOCR



Gambar 31. Implementasi Proses OCR Menggunakan Tesseract OCR

Setelah proses pembacaan dokumen menggunakan Tesseract OCR dan EasyOCR selesai, hasil pembacaan dari kedua metode selanjutnya dievaluasi untuk menentukan hasil OCR yang paling sesuai digunakan dalam proses ekstraksi informasi. Tesseract OCR tetap menjadi metode utama yang menjadi fokus penelitian, sedangkan EasyOCR digunakan sebagai metode pembandingan. Berdasarkan hasil pengujian terhadap 31 dokumen, EasyOCR menunjukkan hasil pembacaan yang lebih baik dibandingkan Tesseract pada data penelitian yang digunakan.



Gambar 32. Implementasi Rule-Based Text Extraction



Gambar 33. Implementasi Hasil Pengujian OCR dan Rule-Based Text Extraction

Pengujian sistem dilakukan untuk mengetahui kemampuan aplikasi dalam membaca dan mengekstraksi informasi dari dokumen administrasi penduduk menggunakan Tesseract OCR sebagai metode utama sesuai dengan fokus penelitian, EasyOCR sebagai metode pembandingan, serta Rule-Based Text Extraction untuk mengidentifikasi informasi penting dari hasil pembacaan OCR. Pengujian dilaksanakan terhadap 31 dokumen yang terdiri atas Surat Keterangan Tidak Mampu (SKTM) dan Surat Keterangan Usaha (SKU). Setiap dokumen diproses menggunakan Tesseract OCR dan EasyOCR secara bersamaan sehingga hasil pembacaan kedua metode dapat dibandingkan pada dokumen dan kondisi pengujian yang sama.

Pengujian EasyOCR dilakukan untuk mengetahui kemampuan metode OCR dalam mengenali karakter dan kata pada dokumen administrasi penduduk. Pengujian dilakukan terhadap 31 dokumen yang terdiri atas Surat Keterangan Tidak Mampu (SKTM) dan Surat Keterangan Usaha (SKU). Dokumen yang digunakan merupakan dokumen yang sama dengan pengujian Tesseract OCR sehingga hasil kedua metode dapat dibandingkan pada kondisi pengujian yang sama. Setelah kedua metode OCR diuji, hasil pembacaan EasyOCR dan Tesseract OCR dibandingkan menggunakan parameter yang sama. Perbandingan dilakukan berdasarkan jumlah karakter, jumlah kata, Character Error Rate (CER), dan Word Error Rate (WER). Jumlah karakter dan jumlah kata digunakan sebagai informasi statistik hasil pembacaan, sedangkan CER dan WER digunakan sebagai parameter utama untuk mengetahui tingkat kesalahan hasil OCR terhadap *ground truth*.

Parameter	EasyOCR	Tesseract OCR
Karakter Ground Truth	823	823
Kata Ground Truth	120	120
Jumlah karakter	835	911
Jumlah kata	117	135
CER	10,94%	18,99%
WER	24,17%	25,80%

Gambar 34. Implementasi Perbandingan Hasil EasyOCR dan Tesseract OCR

Parameter	EasyOCR	TesseractOCR
Karakter Ground Truth	823	823
Kata Ground Truth	120	120
Jumlah karakter	835	911
Jumlah kata	117	135
CER	10,94%	18,99%
WER	24,17%	25,80%

Gambar 35. Perbandingan EasyOCR dan TesseractOCR Setelah proses Optical Character Recognition (OCR) selesai dilakukan, hasil pembacaan teks diproses menggunakan Rule-Based Text Extraction untuk mengambil informasi penting dari dokumen administrasi penduduk. Pengujian dilakukan untuk mengetahui kemampuan aturan yang diterapkan dalam mengenali dan mengambil informasi dari hasil pembacaan OCR. Hasil ekstraksi kemudian dibandingkan dengan data yang terdapat pada dokumen asli untuk menentukan keberhasilan setiap field.

Berdasarkan hasil pengujian yang telah dilakukan, Tesseract OCR sebagai metode utama kemudian EasyOCR sebagai metode pembandingan, sama-sama mampu membaca teks pada dokumen administrasi penduduk. Tesseract OCR sebagai metode utama sesuai dengan fokus penelitian menghasilkan teks yang dapat digunakan dalam proses ekstraksi data. Namun, hasil pembacaan Tesseract masih ditemukan beberapa kesalahan karakter dan kata, terutama pada dokumen yang memiliki kondisi citra kurang baik.

EasyOCR digunakan sebagai metode pembandingan pada dokumen yang sama. Hasil pengujian menunjukkan bahwa EasyOCR juga mampu mengenali teks dan pada beberapa kondisi menghasilkan pembacaan yang lebih mendekati *ground truth* dibandingkan Tesseract OCR.

Perbedaan kemampuan kedua metode terlihat pada kondisi fisik dokumen yang digunakan dalam penelitian. Dokumen pengujian tidak seluruhnya berada dalam kondisi ideal. Beberapa dokumen memiliki bercak atau noda, kertas kusut atau kerumuk, posisi kertas miring, serta kondisi kertas yang buram. Kondisi tersebut dapat menyebabkan karakter menjadi tidak jelas, bentuk karakter terganggu, dan posisi teks berubah.

V. PENUTUP

Kesimpulan

Berdasarkan hasil perancangan, implementasi, dan pengujian yang telah dilakukan pada penelitian Implementasi Sistem Ekstraksi Data Otomatis dari Formulir Administrasi Penduduk Menggunakan Optical Character Recognition Tesseract di Kelurahan Berkas Kota Bengkulu, dapat disimpulkan sebagai berikut.

1. Penelitian berhasil membangun sistem ekstraksi data otomatis dari formulir administrasi penduduk yang mampu mengubah dokumen dalam bentuk citra menjadi data digital terstruktur. Sistem menyediakan fitur unggah dokumen, preprocessing, proses Tesseract OCR, pengujian pembandingan menggunakan EasyOCR, Rule-Based Text Extraction, verifikasi hasil ekstraksi, penyimpanan data, histori digitalisasi, dan pembuatan laporan. Sistem dapat membantu mengurangi proses penginputan data secara manual dan membuat pengelolaan dokumen menjadi lebih terstruktur.
2. Tesseract OCR sebagai metode utama mampu mengenali teks pada formulir administrasi penduduk dan menghasilkan teks yang dapat digunakan sebagai masukan pada proses ekstraksi data. Berdasarkan pengujian terhadap 31 dokumen, Tesseract mampu membaca sebagian besar informasi yang terdapat pada dokumen. Namun, hasil pembacaan masih dipengaruhi oleh kondisi fisik dokumen dan kualitas citra, terutama pada dokumen yang memiliki bercak atau noda, kusut atau kerumuk, miring, dan buram.
3. Pengujian pembandingan menggunakan EasyOCR menunjukkan bahwa hasil OCR dipengaruhi oleh kondisi fisik dokumen. Pada kondisi dokumen yang diuji, EasyOCR menunjukkan hasil pembacaan yang lebih baik pada beberapa kondisi, terutama pada kertas bercak atau bernoda, kertas kusut atau kerumuk, kertas miring, dan kertas buram. Hasil tersebut ditunjukkan melalui

pembacaan yang lebih mendekati *ground truth* serta tingkat kesalahan yang lebih rendah pada kondisi tersebut. Dengan demikian, hasil pengujian menunjukkan bahwa EasyOCR lebih unggul pada kondisi dokumen tertentu yang digunakan dalam penelitian, sedangkan Tesseract tetap digunakan sebagai metode utama sesuai dengan fokus penelitian.

4. Rule-Based Text Extraction berhasil mengubah hasil pembacaan Tesseract OCR menjadi informasi yang lebih terstruktur. Ekstraksi dilakukan terhadap field administrasi yang telah ditentukan sesuai dengan jenis formulir. Dari 248 field yang dievaluasi pada 31 dokumen, sebanyak 230 field berhasil dikenali dan 18 field belum berhasil dikenali. Dengan demikian, tingkat keberhasilan ekstraksi mencapai 92,74%. Hasil tersebut menunjukkan bahwa aturan berbasis Regex mampu mengenali sebagian besar informasi penting dari hasil pembacaan OCR.
5. Kondisi fisik dokumen menjadi salah satu faktor yang memengaruhi keberhasilan proses OCR dan ekstraksi data. Bercak atau noda dapat menutupi karakter, lipatan pada kertas kusut atau kerumuk dapat mengubah bentuk karakter, kemiringan dokumen dapat memengaruhi posisi pembacaan teks, sedangkan kondisi buram dapat mengurangi ketajaman karakter. Kondisi tersebut dapat menyebabkan hasil OCR mengalami kesalahan dan selanjutnya memengaruhi proses Rule-Based Text Extraction. Oleh karena itu, hasil ekstraksi tetap memerlukan verifikasi administrator sebelum data disimpan ke dalam basis data.
6. Berdasarkan pengujian fungsional menggunakan metode Black Box Testing, seluruh fitur utama aplikasi dapat berjalan sesuai dengan fungsi yang telah dirancang. Fitur tersebut meliputi login, unggah dokumen, preprocessing, Tesseract OCR, pengujian pembandingan, Rule-Based Text Extraction, verifikasi hasil ekstraksi, pengelolaan data dokumen, histori digitalisasi, dan pembuatan laporan. Dengan demikian, sistem yang dikembangkan dapat digunakan untuk mendukung proses digitalisasi dan ekstraksi data formulir administrasi penduduk di Kelurahan Berkas Kota Bengkulu.

Saran

Berdasarkan hasil penelitian dan keterbatasan yang ditemukan selama proses implementasi dan pengujian, beberapa saran untuk pengembangan selanjutnya adalah sebagai berikut.

1. Tahap preprocessing citra perlu dikembangkan lebih lanjut untuk meningkatkan kualitas dokumen sebelum diproses menggunakan Tesseract OCR. Pengembangan dapat mencakup peningkatan kontras, pengurangan noise, thresholding, dan koreksi kemiringan dokumen.

Tahapan tersebut diharapkan dapat membantu meningkatkan pembacaan pada dokumen yang memiliki noda, kusut atau kerumuk, posisi miring, maupun kondisi buram.

2. Rule-Based Text Extraction dapat dikembangkan dengan aturan yang lebih fleksibel untuk menangani variasi format, tata letak, dan penulisan dokumen administrasi. Pengembangan terutama diperlukan pada field seperti alamat dan keperluan yang memiliki variasi penulisan lebih tinggi.
3. Sistem dapat ditambahkan validasi otomatis terhadap hasil ekstraksi, seperti pemeriksaan jumlah digit NIK, format nomor surat, nomor KK, format tanggal, serta pemeriksaan terhadap field wajib. Validasi tersebut dapat membantu administrator menemukan kesalahan hasil OCR sebelum data disimpan.
4. Pengujian selanjutnya disarankan menggunakan jumlah dokumen yang lebih banyak dan kondisi dokumen yang lebih beragam. Pengujian dapat mencakup dokumen dengan bercak atau noda, kertas kusut atau kerumuk, dokumen miring, dokumen buram, resolusi berbeda, hasil pemindaian, foto dokumen, serta variasi tata letak. Dengan demikian, kemampuan Tesseract sebagai metode utama dapat dievaluasi pada kondisi dokumen yang lebih luas.
5. Penelitian selanjutnya dapat mengembangkan mekanisme pemilihan hasil OCR berdasarkan kondisi dokumen. Hasil penelitian menunjukkan bahwa meskipun Tesseract digunakan sebagai metode utama, EasyOCR memberikan hasil yang lebih baik pada beberapa kondisi dokumen, khususnya dokumen yang memiliki noda, kusut atau kerumuk, kemiringan, dan kondisi buram. Pengembangan selanjutnya dapat memanfaatkan kondisi tersebut sebagai dasar untuk menentukan metode OCR yang menghasilkan pembacaan paling sesuai.

DAFTAR PUSTAKA

- [1] Chen, Y., Cui, T., Capstick, A., Fletcher-Loyd, N., & Barnaghi, P. (2024). Enabling Regional Explainability by Automatic and Model-agnostic Rule Extraction. *ArXiv Preprint ArXiv:2406.17885*, 1.
- [2] Colter, Z., Fayazi, M., Youbi, Z. B. El, Kamp, S., Yu, S., & Dreslinski, R. (2022). Tablext: A combined neural network and heuristic based table extractor. *Array*, 15. <https://doi.org/10.1016/j.array.2022.100220>
- [3] Darmawan, R., Nasuha, A., Zaman, L., & Armanto, H. (2021). Sistem Manajemen Kartu Nama dengan OCR dan Ekstraksi Informasi Otomatis. *Journal of Intelligent System and Computation*, 3(2), 61–72. <https://doi.org/10.52985/insyst.v3i2.194>
- [4] Darmi, Y., Sepriansyah, M. F., & Darnita, Y. (2025). Penerapan Metode Optical Character Recognition (OCR) Untuk Mengidentifikasi Teks Pada Identitas Dokumen Surat Izin Mengemudi (SIM). 9(4), 5992–5998.
- [5] Darpito, M. N., Firdausy, K., & Fadlil, A. (2025). Perbandingan Unjuk Kerja Library Optical Character Recognition (OCR) dalam Pengenalan Teks pada Dokumen Digital. *Jurnal Informatika Polinema*, 11(3), 273–282. <https://doi.org/10.33795/jip.v11i3.7025>
- [6] Febryana, S., Wulandari, I. A., & Prabowo, D. (2024). Rancang Bangun Aplikasi Pengolahan Data Inventory Barang Berbasis Desktop Pada Cv. Hamim Group. *Jurnal Mahasiswa Sistem Informasi (JMSI)*, 6(1), 24–35. <https://doi.org/10.24127/jmsi.v6i1.7553>
- Undang-Undang Republik Indonesia Nomor 24 Tahun 2013, (2013).
- [7] Jumadi, J., Akbar, A. Al, Informatika, D., & Bengkulu, U. D. (2020). Implementasi Metode Template Matching Untuk Klasifikasi Citra Anggrek Pensil Bengkulu. 1–7.
- [8] Jumadi, J., Yupianti, & Sartika, D. (2021). Pengolahan Citra Digital Untuk Identifikasi Objek Menggunakan Metode Hierarchical Agglomerative Clustering. *Jurnal Sains Dan Teknologi*, 10(2), 148–156.
- [9] M. Irfan. (2021). Implementasi Kebijakan Administrasi Kependudukan Di Kota Bima (Studi Kasus di Dinas Catatan Sipil dan Kependudukan Kota Bima). *Jurnal Ilmu Administrasi Negara*, 18(2).
- [10] Nashiruddin, M., Dhaza, F., Praditya, N., & Mufazzal, A. F. (2025). Implementasi OCR berbasis Tesseract untuk Ekstraksi data kartu mahasiswa UMKLA. 01(1), 11–14.
- [11] Noiija, H. C., Wemaf, P. A., Nurdianty, O. A., Sohila, W., Haumahu, S., Yusuf, H., Tomagola, K., Wa Rindi La Ana, & Salaiswa, T. S. (2023). Perancangan Sistem Informasi Akuntansi Untuk Analisis Siklus Pendapatan Pada Orantata Celular Menggunakan DFD Dan Flowchart. *Jurnal Bisnis Dan Manajemen (JURBISMAN)*, 1(2), 577–592. <https://doi.org/10.61930/jurbisman.v1i2.188>
- [12] Okta, M. D. U., Aulia, S., & Burhanuddin. (2021). Pengenalan Pola Berbasis OCR untuk Pengambilan Data Bursa Saham. *Jurnal Rekayasa ElektriKa*, 17(2), 100–106. <https://doi.org/10.17529/jre.v17i2.19656>
- [13] Panuntun, F. G., & Sejati, R. H. P. (2024). Sistem Otomatisasi Deteksi dan Ekstraksi Data KTP Berbasis Convolutional Neural Network dan Optical Character Recognition. 7(3), 464–471.
- [14] Ponggohong, A. R. I., Rorimpandey, G. C., &

- Maswonggo, V. V. (2025). Penerapan Optical Character Recognition Tesseract Dan Gemini Pada Sistem Buku Tamu Digital Berbasis Web. *Jointer : Journal of Informatics Engineering*, 06(02), 2025.
- [15] Reksiyano, R. D., Pane, S. F., & Awangga, R. M. (2025). *Enhancing OCR Accuracy on Indonesian ID Cards Using Dual-Pipeline Tesseract and Post-Processing*. 10(2), 119–127.
- [16] Reyvansyah, M. R., Setiawan, E., Indarti, R., Adianto, & Munadhif, I. (2023). Penerapan Metode Optical Character Recognition (OCR) Untuk Mengambil Data Arsip. *Jurnal Teknik Elektro Dan Komputer TRIAC*, 10(2), 44–50. <https://doi.org/10.21107/triac.v10i2.20809>
- [17] Sangaji, A. H., Pamungkas, Y., Risqiwati, D., Mardi, S., Nugroho, S., & Dharma, A. (2022). *Rule-based disease classification using text mining on symptoms extraction from electronic medical records in Indonesian*. 4(1), 69–80.
- [18] Sugiarta, G., Andini, D. P., & Hidayatullah, S. (2021). Ekstraksi Informasi/Data e-KTP Menggunakan Optical Character Recognition Convolutional Neural Network. *JTERA (Jurnal Teknologi Rekayasa)*, 6(1), 1. <https://doi.org/10.31544/jtera.v6.i1.2021.1-6>
- [19] Zulhijar, A., Basri, M., & Marlina. (2024). Implementasi Optical Character Recognition (OCR) untuk Otomasi Penghitungan Tagihan Listrik. *Jurnal Publikasi Manajemen Informatika*, 3(3), 162–171. <https://doi.org/10.55606/jupumi.v3i3.3752>