PEMBUATAN DATASET TERINTEGRASI JAWI-RUMI UNTUK EVALUASI KUALITAS DAN AKURASI MULTI-MODEL (OCR DAN TRANSLITERASI)

TUGAS AKHIR

Diajukan Oleh:

Rozatul Jannah

220705053

Mahasiswa Fakultas Sains dan Teknologi

Program Studi Teknologi Informasi

PROGRAM STUDI TEKNOLOGI INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI
AR-RANIRY BANDA ACEH

2026 M / 1448 H

LEMBAR PENGESAHAN

PEMBUATAN DATASET TERINTEGRASI JAWI-RUMI UNTUK EVALUASI KUALITAS DAN AKURASI MULTI-MODEL (OCR DAN TRANSLITERASI)

TUGAS AKHIR

Telah Diuji Oleh Panitia Ujian Munaqasyah Tugas Akhir
Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh dan Dinyatakan Lulus
Serta Diterima Sebagai Salah Satu Beban Studi Program Sarjana (S1)
Dalam Program Studi Teknologi Informasi

Pada Hari/Tanggal: Jumat, 14 Agustus 2026 M
1 Rabiul Awal 1448 H

Di Darussalam, Banda Aceh

Panitia Ujian Munaqasyah Tugas Akhir:

Ketua,	Sekretaris,
Bustami, M.Sc.	**Khairan AR, M.Kom.**
NIP. 198604082014031001	**NIP. 198607042014031001**
Penguji I,	Penguji II,
Dr. Hendri Ahmadian, S.Si.,M.IM.	**Mulkan Fadhli, S.T.,M.T.**
NIP. 198301042014031002	**NIP. 198811282020121006**

Mengetahui:
Dekan Fakultas Sains dan Teknologi
UIN Ar-Raniry Banda Aceh

Prof. Dr. Ir. Muhammad Dirhamsyah, M.T., IPU
NIP. 19621002198811101

LEMBAR PERSETUJUAN

PEMBUATAN DATASET TERINTEGRASI JAWI-RUMI UNTUK EVALUASI KUALITAS DAN AKURASI MULTI-MODEL (OCR DAN TRANSLITERASI)

TUGAS AKHIR

Diajukan Kepada Fakultas Sains dan Teknologi
Universitas Islam Negeri (UIN) Ar-Raniry Banda Aceh
Sebagai Salah Satu Beban Studi Memperoleh Gelar Sarjana (S1)
Dalam Ilmu Teknologi Informasi

Oleh:
Rozatul Jannah
NIM.220705053

Mahasiswa Fakultas Sains dan Teknologi
Program Studi Teknologi Informasi

Disetujui Untuk Dimunaqasyahkan Oleh:

Pembimbing I,	Pembimbing II,
Bustami, M.Sc.	**Khairan AR, M.Kom.**
NIP. 198604082014031001	**NIP. 198607042014031001**

Mengetahui,
Ketua Program Studi Teknologi Informasi

Malahayati, M.T.
NIP. 198301272015032003

LEMBAR PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan di bawah ini:

Nama	:	Rozatul Jannah
NIM	:	220705053
Program Studi	:	Teknologi Informasi
Judul	:	Pembuatan dataset terintegrasi jawi-rumi untuk evaluasi kualitas dan akurasi multi-model (ocr dan transliterasi)

Dengan ini menyatakan bahwa dalam penulisan tugas akhir ini, saya:

1. Tidak menggunakan ide orang lain tanpa mampu mengembangkan dan mempertanggungjawabkan;
2. Tidak melakukan plagiasi terhadap naskah karya orang lain;
3. Tidak menggunakan karya orang lain tanpa menyebutkan sumber asli atautanpa izin pemilik karya;
4. Tidak memanipulasi dan memalsukan data;
5. Mengerjakan sendiri karya ini dan mampu bertanggungjawab atas karya ini.

Bila dikemudian hari ada tuntutan dari pihak lain atas karya saya, dan telah melalui

pembuktian yang dapat dipertanggungjawabkan dan ternyata memang ditemukan bukti bahwa saya telah melanggar pernyataan ini, maka saya siap dikenai sanksi berdasarkan aturan yang berlaku di Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh.

Demikian pernyataan ini saya buat dengan sesungguhnya dan tanpa paksaan dari pihak manapun.

Banda Aceh, 25 Agustus 2026
Yang Menyatakan,

Rozatul Jannah

ABSTRAK

Nama	: Rozatul Jannah
NIM	: 220705053
Program Studi	: Teknologi Informasi
Judul	: Pembuatan dataset terintegrasi jawi-rumi untuk evaluasi kualitas dan akurasi multi-model (OCR dan Transliterasi)
Tanggal Sidang	: 14 Agustus 2026
Jumlah Halaman:	: 66 Halaman
Pembimbing I	: Bustami, M.Sc.
Pembimbing II	: Khairan AR, M.Kom.
Kata Kunci	: Arab-Jawi, Data sintetik dan autentik, OCR, TrOCR, K-Fold Cross Validation, CER, WER

Arab-Jawi merupakan warisan aksara Nusantara hasil adaptasi tulisan Arab yang berperan penting dalam penyebaran ilmu agama, hukum, dan sastra Melayu-Islam, namun ketersediaan dataset Arab-Jawi hingga kini masih terbatas dari segi jumlah, keragaman, dan aksesibilitas publik. Penelitian ini bertujuan membangun dataset terintegrasi Jawi-Rumi yang representatif dan terdokumentasi dengan baik, serta menguji validitas, konsistensi, dan keandalannya melalui skema 10-fold cross-validation berbasis GroupKFold. Dataset dibangun melalui dua jalur, yaitu data autentik dari kitab Arab-Jawi fisik dan digital melalui segmentasi layout, koreksi bounding box, dan anotasi manual, serta data sintetik hasil crawling artikel keagamaan yang dirender menjadi citra dengan delapan jenis augmentasi, menghasilkan data gambar-teks data autentik dan data sintetik). Pengujian reliabilitas menggunakan model TrOCR (*microsoft/trocr-base-handwritten*) menunjukkan performa stabil pada seluruh *fold*, dengan rata-rata CER dan WER pada data autentik sebesar 16,35% dan 51,70% (standar deviasi 0,94% dan 1,65%), serta pada data sintetik sebesar 0,10% dan 0,45% (standar deviasi 0,01% dan 0,05%). Perbandingan lintas arsitektur dengan Qwen2.5-VL-3B-Instruct (QLoRA) menghasilkan selisih CER dan WER tidak lebih dari 3%, membuktikan kualitas dataset bersifat objektif dan tidak bias terhadap model tertentu. Dataset Jawi-Rumi yang dibangun terbukti valid, konsisten, dan layak dijadikan fondasi bagi pengembangan sistem OCR dan transliterasi Arab-Jawi yang lebih akurat di masa mendatang.

ABSTRACT

Name	: Rozatul Jannah
Student ID	: 220705053
Study Program	: *Information Technology*
Title	: *Development of an Integrated Jawi-Rumi Dataset for Multi-Model Quality and Accuracy Evaluation (OCR and Transliteration)*
Date	: 14 August 2026
Number of Pages	: 66 Pages
Supervisor I	: Bustami, M.Sc.
Supervisor II	: Khairan AR, M.Kom.
Keywords	*: Jawi Script, Synthetic and Authentic Data, OCR, TrOCR, K-Fold Cross Validation, CER, WER*

Jawi script is a Nusantara writing heritage adapted from Arabic script that has played an important role in spreading religious knowledge, law, and Malay-Islamic literature, yet the availability of Jawi datasets remains limited in terms of quantity, diversity, and public accessibility. This study aims to build an integrated Jawi-Rumi dataset that is representative and well-documented, and to test its validity, consistency, and reliability through a 10-fold cross-validation scheme based on GroupKFold. The dataset was built through two pathways: authentic data from physical and digital Jawi manuscripts through layout segmentation, bounding box correction, and manual annotation, and synthetic data from crawled religious articles rendered into images with eight types of augmentation, producing image-text data from authentic data and synthetic data. Reliability testing using the TrOCR model (microsoft/trocr-base-handwritten) showed stable performance across all folds, with average CER and WER on authentic data of 16.35% and 51.70% (standard deviation 0.94% and 1.65%), and on synthetic data of 0.10% and 0.45% (standard deviation 0.01% and 0.05%). Cross-architecture comparison with Qwen2.5-VL-3B-Instruct (QLoRA) yielded CER and WER differences of no more than 3%, proving that dataset quality is objective and not biased toward a particular model. The resulting Jawi-Rumi dataset is proven to be valid, consistent, and suitable as a foundation for developing more accurate OCR and transliteration systems for Jawi script in the future

KATA PENGANTAR

Puji syukur senantiasa penulis panjatkan ke hadirat Allah SWT, Tuhan Yang Maha Esa, atas segala rahmat, hidayah, dan karunia-Nya sehingga penulis dapat menyelesaikan penulisan skripsi yang berjudul "Pembuatan Dataset Terintegrasi Jwi-Rumi untuk Evaluasi Kualitas dan Akurasi Multi-Model (OCR dan Transliterasi)". Salawat serta salam tak lupa senantiasa tercurahkan kepada junjungan Nabi Besar Muhammad SAW, beserta keluarga, sahabat, dan para pengikutnya.

Penyusunan skripsi ini merupakan salah satu syarat akademik untuk menyelesaikan pendidikan program sarjana (S1) pada Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, UIN Ar-Raniry Banda Aceh.

Penulis menyadari sepenuhnya bahwa penyelesaian skripsi ini tidak lepas dari dukungan, bimbingan, arahan, serta doa dari berbagai pihak sejak awal masa perkuliahan hingga tahap penyusunan akhir. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Kepada *raisul usrah* (pemimpin keluarga), Ayahanda tercinta Ismandani, sosok panutan sekaligus cinta pertama penulis, tiada kata yang mampu mewakili rasa terima kasih penulis. Terima kasih telah berjuang tanpa lelah demi keluarga dan telah mengusahakan segala hal terbaik bagi penulis serta adik-adik penulis. Meski tak banyak kata manis yang terucap, perjuangan dan kasih sayang yang telah ditunjukkan selama ini berbicara lebih banyak daripada kata-kata apa pun.
2. Kepada *madrasatul ula* (sekolah pertama), Ibunda Rita Susanti, S.Pd., Gr., pintu surga dan bidadari yang paling penulis cintai, terima kasih atas doa yang tak pernah putus, dukungan yang selalu hadir di setiap langkah, serta nasihat dan pelukan hangat yang menjadi tempat penulis pulang dan bercerita. Ibu yang menyayangi penulis dengan segenap jiwa dan raga, yang telah berjuang dengan pengorbanan dan kekuatan jauh melampaui batas kemampuannya.

3. Kepada adik-adik penulis yang paling penulis cintai dan sayangi, Zahbi, Zahlul, dan Zafran, meski tak jarang diwarnai pertengkaran kecil, kalian adalah salah satu alasan dan penyemangat penulis untuk terus maju dan menyelesaikan pendidikan ini dengan baik. Canda tawa dan kehadiran kalian selalu menjadi pelipur lelah di tengah kesibukan penulis. Terima kasih telah menjadi adik-adik yang selalu mendoakan dan mendukung penulis dengan caranya masing-masing, karena tanpa disadari, hal itu menjadi salah satu kekuatan penulis untuk terus melangkah hingga sampai di titik ini.
4. Kepada seluruh keluarga besar penulis yang tidak dapat disebutkan satu per satu, terima kasih atas doa yang tak pernah putus, kasih sayang yang senantiasa dicurahkan, serta dukungan moral maupun material yang telah diberikan, sehingga penulis dapat menyelesaikan pendidikan ini dengan baik.
5. Ibu Malahayati, M.T., selaku Ketua Program Studi Teknologi Informasi.
6. Bapak Bustami, M.Sc.,selaku Dosen Pembimbing I, yang telah membantu penulis dalam menyelesaikan penelitian dan skripsi ini. Terima kasih telah menyempatkan waktu dan pikiran untuk membimbing penulis, serta senantiasa sabar mengajarkan banyak ilmu dan pengalaman baru, meski penulis masih banyak melakukan kesalahan. Bapak Khairan AR, M.Kom., selaku Dosen Pembimbing II, yang telah meluangkan banyak waktu, tenaga, dan pikiran untuk memberikan bimbingan, masukan berharga, serta motivasi yang tak ternilai harganya kepada penulis selama proses penyusunan skripsi ini.
7. Seluruh Bapak/Ibu Dosen serta Staf Akademik Program Studi Teknologi Informasi yang telah membagikan ilmu yang sangat bermanfaat selama masa perkuliahan dan banyak membantu urusan administrasi penulis.
8. Penulis mengucapkan terima kasih kepada sahabat-sahabat dekat penulis, ketiga perempuan hebat, Adea Marlinda, Naqqa Anfasa, dan Nurul Mirdawati, yang telah setia menemani dan mendengarkan keluh kesah penulis sepanjang proses penyusunan skripsi ini, serta senantiasa meluangkan waktu dan tenaga untuk menemani penulis melewati setiap

prosesnya dan yang telah membersamai penulis sepanjang perjalanan masa perkuliahan, berbagi suka maupun duka, hingga penulis mampu sampai pada titik ini.

9. Dua rekan tim penelitian penulis yang telah bekerja sama membagi tugas dalam pengumpulan dan pengolahan data, serta seluruh rekan-rekan mahasiswa Teknologi Informasi angkatan 2022 yang telah bersama-sama berjuang, belajar, dan saling memberikan semangat.
10. Terakhir, izinkan penulis mengucapkan terima kasih kepada diri sendiri, Rozatul Jannah, yang selama ini terus berusaha meski sering merasa lelah dan ragu. Skripsi ini bukan hanya soal menyelesaikan tugas akhir, tapi juga bukti bahwa penulis mampu bertahan menghadapi banyak hal yang tidak mudah, mulai dari revisi yang berulang, waktu yang terkuras, hingga rasa lelah yang datang silih berganti. Terima kasih telah bertahan sejauh ini dan terus berjalan melewati segala tantangan yang datang. Terima kasih karena tetap berani memberanikan diri dan tidak menyerah dalam menyelesaikan penelitian dan skripsi ini. Bangga atas setiap langkah kecil yang telah diambil, atas setiap usaha yang mungkin tidak selalu terlihat orang lain. Walau terkadang harapan tidak selalu sejalan dengan kenyataan, tetaplah belajar menerima dan mensyukuri apa pun yang telah dicapai. Jangan lelah untuk terus berusaha, dan berbahagialah di mana pun berada.
11. Semua pihak yang tidak dapat penulis sebutkan satu per satu, yang secara langsung maupun tidak langsung telah memberikan bantuan, dukungan, dan motivasi selama proses penyusunan skripsi ini, serta pihak-pihak yang telah memberikan inspirasi kepada penulis untuk mendalami bidang *Artificial Intelligence* sehingga penelitian ini dapat terselesaikan dengan baik.

Penulis menyadari bahwa di dalam penulisan dan penyusunan skripsi ini masih terdapat banyak kekurangan karena keterbatasan ilmu dan pengalaman. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun dari semua pihak.

Akhir kata, penulis berharap semoga skripsi ini dapat memberikan manfaat yang nyata bagi pengembangan ilmu pengetahuan, pelestarian naskah budaya Arab-jawi, serta dapat menjadi referensi yang berguna bagi penelitian-penelitian selanjutnya di masa yang akan datang.

Banda Aceh, 12 Agustus 2026

Penulis

Rozatul Jannah

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan peradaban manusia senantiasa berjalan seiring dengan perubahan cara berkomunikasi, dan di kawasan Nusantara hal ini tercermin dalam penggunaan Arab-Jawi sebagai sistem tulisan hasil adaptasi aksara Arab yang mulai digunakan sejak abad ke-14 dan disesuaikan dengan karakter fonologis bahasa Melayu. Arab-Jawi tidak hanya berfungsi sebagai alat komunikasi, tetapi juga menjadi hal penting dalam penyebaran ilmu agama, hukum, sastra, serta nilai-nilai sosial yang membentuk identitas Melayu-Islam di Nusantara (Andriani et al., 2025).

Penyesuaian tersebut diwujudkan melalui penambahan huruf-huruf khusus seperti ڤ (pa), چ (cha), ڠ (nga), ݢ (Ga), ڽ (nya), dan ۏ (va) untuk merepresentasikan bunyi yang tidak terdapat dalam aksara Arab standar. Namun, berbeda dengan tulisan Arab, teks Arab-Jawi umumnya ditulis tanpa harakat, sehingga berpotensi menimbulkan perbedaan penafsiran, yang menuntut ketelitian dalam proses penyalinan, pendokumentasian, dan anotasi agar dataset Arab-Jawi yang dihasilkan akurat dan andal (Ramala, 2020).

Dalam beberapa tahun terakhir, pengembangan dataset Arab-Jawi masih terbatas akibat tingginya biaya digitalisasi dan anotasi. Sebagai upaya awal pelestarian dan pengolahan tulisan Jawi, dikembangkan DPJ (*Database Printed Jawi*), yaitu basis data karakter Jawi cetak untuk penelitian pengenalan karakter. Dataset ini dibentuk melalui proses pengetikan, pencetakan, dan pemindaian, serta mencakup citra karakter dan teks Jawi berupa kata dan kalimat, dengan total 1.524 karakter dari empat jenis font dan 168 citra kata serta kalimat Jawi cetak, dan menjadi salah satu basis data awal yang secara khusus dikembangkan untuk karakter Jawi cetak (Saddami et al., 2016).

Sebagai pengembangan dari penelitian sebelumnya, disusun pula dataset tulisan tangan Jawi yang dikenal sebagai DHJ (*Dataset Handwritten Jawi*). Dataset ini menyoroti variasi bentuk tulisan tangan Jawi yang lebih kompleks dibandingkan tulisan cetak. Data dikumpulkan dari sepuluh penulis dengan latar belakang yang beragam dan mencakup citra digital karakter, kata, serta kalimat Jawi, dengan total 3.810 karakter, 300 kata, dan 40 kalimat (Saddami et al., 2017).

Selain Dataset *Handwritten Jawi (DHJ)* dan *Database Printed Jawi (DPJ)*, dataset Arab-Jawi yang pernah tersedia adalah *Jawi-OCR-v1,* yang dikembangkan dari surat kabar Arab-Jawi *Warta Malaya* terbitan tahun 1930-an. Dataset ini memiliki dokumen dengan tata letak yang relatif rapi dan terstruktur. Selain itu, dataset ini hanya berasal dari satu sumber dokumen dengan gaya bahasa dan format yang kurang bervariasi, sehingga keragaman data relatif terbatas. Meskipun dataset *Jawi-OCR-v1* pernah dipublikasikan secara terbuka, saat ini dataset tersebut tidak lagi tersedia secara public. Namun demikian, dataset Arab-Jawi yang tersedia saat ini masih terbatas dari segi jumlah dan keragaman data, serta belum sepenuhnya dapat diakses secara publik, sehingga sulit diverifikasi kebenarannya dan dimanfaatkan sebagai rujukan penelitian lanjutan.

Oleh karena itu, penelitian ini bertujuan membangun dataset Arab-Jawi dalam jumlah yang memadai serta terdokumentasi dengan baik, guna mendukung pengembangan dan evaluasi model OCR dan transliterasi secara lebih optimal. Dataset dengan jumlah data yang memadai, variasi citra yang beragam, serta proses dokumentasi dan validasi yang jelas diharapkan memiliki kredibilitas tinggi dan menjadi rujukan yang andal bagi penelitian Arab-Jawi di masa mendatang.

1.2 Rumusan Masalah

Berdasarkan analisis mendalam terhadap urgensi pelestarian dan kompleksitas teknis yang ada, penelitian ini dirumuskan untuk menjawab pertanyaan-pertanyaan berikut:

1. Bagaimana proses pembuatan dataset terintegrasi Arab-Jawi yang dilakukan secara sistematis sehingga menghasilkan data yang representatif, terdokumentasi dengan baik?
2. Sejauh mana dataset Arab-Jawi yang dibangun memiliki validitas, konsistensi, dan keseimbangan data serta anotasi yang andal sehingga layak digunakan dan tidak menimbulkan bias?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah diatas, tujuan penelitian ini adalah:

1. Menguraikan dan mendokumentasi proses pembuatan dataset terintegrasi Arab-Jawi yang dilakukan secara sistematis, sehingga dihasilkan data yang representatif dan terdokumentasi dengan baik.
2. Melakukan validasi dan pengujian kredibilitas dataset Arab-Jawi untuk memastikan konsistensi data, keseimbangan isi dataset, serta keandalan anotasi agar dataset tidak bias dan layak diimplementasi di berbagai model.

1.4 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah:

1. Penelitian ini memberikan gambaran dan acuan tentang cara pembuatan, pendokumentasian, dan validasi dataset Arab-Jawi secara terstruktur dan sistematis.
2. Penelitian ini menghasilkan dataset Arab-Jawi yang rapi, terdokumentasi, dan dapat dipercaya sehingga dapat digunakan oleh peneliti, akademisi, dan lembaga dalam kegiatan pelestarian dan pengelolaan naskah

1.5 Batasan Penelitian

Untuk menjaga fokus dan kedalaman analisis, penelitian ini dibatasi oleh beberapa ruang lingkup yang telah ditetapkan secara cermat:

1. Penelitian ini dibatasi pada proses pembangunan, pendokumentasian, serta validasi kredibilitas dataset Arab-Jawi,bukan pada pengembangan atau optimasi suatu model.
2. Dataset yang dibangun bersumber dari naskah Arab-Jawi dalam bentuk cetak atau digital yang diperoleh dari koleksi pribadi dan repositori daring.
3. Penelitian ini tidak membahas pengembangan, pelatihan, atau optimasi model OCR maupun transliterasi. Metrik akurasi seperti *Character Error Rate* (CER) dan *Word Error Rate* (WER) digunakan semata-mata sebagai instrumen ukur untuk menilai konsistensi dataset yang telah dibangun bukan sebagai tolok ukur performa atau optimasi model.

4. Penelitian ini tidak menghitung akurasi hasil transliterasi (Jawi ke Rumi) maupun akurasi transkrip (hasil OCR berupa teks Jawi) secara terpisah, karena evaluasi yang dilakukan berfokus pada pengukuran konsistensi dataset melalui CER dan WER, bukan pada penilaian ketepatan hasil transliterasi atau transkrip itu sendiri.

BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Tabel 2. 1 Penelitian terkait

No.	Peneliti	Judul Penelitian	Fokus & Metode	Hasil Penelitian
1.	(Saddami et al., 2016)	A Database of Printed Jawi Character Image (DPJ)	Membangun dataset karakter Jawi cetak dari 4 jenis font (Times New Roman, Sakkal Majalla, Courier New, Segoe UI), dicetak lalu discan menjadi citra karakter, kata, dan kalimat	Dihasilkan 1.524 citra karakter, 132 citra kata, dan 36 citra kalimat Jawi cetak; merupakan dataset Jawi publik pertama, namun terbatas pada bentuk cetak dan belum mencakup variasi tulisan tangan
2.	(Saddami et al., 2017)	DHJ: A Database of Handwritten Jawi for Recognition Research	Mengumpulkan dataset tulisan tangan Jawi dari 10 penulis dengan latar belakang usia,	Diperoleh 3.810 citra karakter, 300 citra kata, dan 40 citra kalimat

			gender, dan pendidikan berbeda, mencakup karakter, kata, dan kalimat.	tulisan tangan Jawi (total 4.180 citra) melengkapi DPJ namun jumlah penulis dan sampel masih relatif kecil.
3.	(Razali et al., 2024)	Augmentation of Additional Arabic Dataset for Jawi Writing and Classification Using Deep Learning	Mengaugmentasi 6 karakter tambahan Jawi (*nya, ca, nga, pa, ga, va*) dari dataset arab yaitu HMBD, AHAWP, dan HUCD melalui manipulasi titik, serta menguji klasifikasinya menggunakan model CNN (InceptionV3 dan ResNet34).	Model ResNet34 mencapai akurasi 96% dan InceptionV3 mencapai 95% pada klasifikasi 22 bentuk karakter tambahan Jawi. Kendati demikian, performa menurun pada kelas karakter bersambung dengan kemiripan bentuk visual

				(seperti *"Nga Middle"* dan *"Pa Middle"*) serta pengujian masih terbatas pada skala karakter terisolasi (*cropped*).
4.	(Nasrudin et al., 2008)	Handwritten Cursive Jawi Character Recognition: A Survey	Merangkum penelitian pengenalan tulisan tangan Jawi berdasarkan lima tahapan umum: pra-pemrosesan, representasi huruf, pemisahan huruf/kata, ekstraksi ciri, dan pengenalan akhir model.	Jawi bersambung sulit dikenali (huruf menumpuk, berubah bentuk, gaya tulisan beragam). Metode terbatas pada ANN, akurasi 82–99,75%. Belum ada dataset publik standar.
5.	(Abdalkafor et al., 2019)	A Novel Database for Arabic Handwritten	Membangun database tulisan tangan Arab (NDAHR) yang	Dihasilkan 61.500 gambar karakter

		Recognition (NDAHR) System	lebih lengkap dari database sebelumnya, mencakup huruf, angka, simbol, tanda baca, dan diakritik sekaligus. Data ditulis 100 orang, discan, dipotong per karakter, dan dibersihkan dari noise.	dalam 123 kelas (80% latih, 20% uji). NDAHR mengatasi keterbatasan database sebelumnya misalnya HACDB tanpa titik huruf) dengan mencakup seluruh elemen tulisan Arab sekaligus, meski belum diuji dengan model OCR.

Berdasarkan penelitian-penelitian terdahulu pada tabel 2.1 di atas, terlihat bahwa penyediaan dataset Arab-jawi masih berada pada tahap pengembangan. Memulai Upaya yang telah dirintis pada dataset Jawi, baik cetak (DPJ) maupun tulisan tangan (DHJ) (Saddami et al., 2016), namun jumlah sampelnya masih terbatas. Hal ini juga sejalan dengan catatan (Nasrudin et al., 2008) bahwa dataset publik standar untuk Jawi bersambung memang belum banyak tersedia.

Kondisi ini mendorong (Razali et al., 2024) menempuh jalur lain, yaitu mengaugmentasi karakter Jawi dari dataset Arab yang sudah mapan. Langkah ini menunjukkan bahwa penyediaan data Jawi secara langsung dari sumber aslinya memang masih menjadi tantangan tersendiri. Sementara itu, dataset Arab seperti

NDAHR (Abdalkafor et al., 2019)memperlihatkan bahwa kualitas sistem pengenalan tulisan sangat dipengaruhi oleh kelengkapan dan kerapian proses penyusunan datasetnya.

Berdasarkan uraian di atas, penelitian dataset Jawi yang ada selama ini masih terbatas pada upaya penyediaan data, dengan pengujian kualitas dan keandalan yang belum banyak dilakukan, terlebih dataset Arab-Jawi sendiri hingga kini masih sangat minim tersedia. Kondisi inilah yang menjadi dasar penelitian ini, yaitu membangun dataset Arab-Jawi yang lebih lengkap sekaligus disertai pengujian reliabilitas yang lebih terukur, sehingga dapat menjadi landasan yang lebih kuat bagi pengembangan penelitian selanjutnya, khususnya dalam sistem pengenalan dan transliterasi Arab-jawi.

2.2 Landasan Teori

Terdapat beberapa teori yang menjadi dasar pendukung dalam penelitian ini, yang diuraikan sebagai berikut :

2.2.1 Arab Jawi

Arab Jawi adalah salah bentuk Bahasa yang digunakan oleh bangsa Melayu dengan memadukan seluruh karakter aksara Arab dengan menambahkan aksara lain yang tidak termasuk dalam huruf Arab seperti huruf چ، ڠ، ڤ، ک، ڽ، ز untuk melengkapi huruf dalam bahasa Arab yang tidak terdapat pada abjad Arab standar. Penyesuaian tersebut menjadikan Jawi lebih sesuai dengan struktur fonologi bahasa Melayu, sekaligus membedakannya dari tulisan Arab meskipun keduanya memiliki akar skrip yang sama (Ramala, 2020)

Tabel 2. 2 Abjad Arab Jawi

Shen	ش	**Khaa**	خ	**Alif**	ا
Saad	ص	**Dal**	د	**Baa**	ب
Daad	ض	**Zal**	ذ	**Taa**	ت
Tah	ط	**Raa**	ر	**Thaa**	ث
Zah	ظ	**Zin**	ز	**Gem**	ج
Ain	ع	**Sin**	س	**Haa**	ح
Nya	ڽ	**Non**	ن	**Gen**	غ

Ca	چ	**Waw**	و	**Faa**	ف
Nga	ڠ	**Ha**	ه	**Qaf**	ق
Pa	ڤ	**Hamzah**	ء	**Kaf**	ك
Ga	ݢ	**Yaa_Dot**	ي	**Lam**	ل
Va	ڤ	**Yaa**	ى	**Mem**	م

Arab Jawi digunakan secara luas dalam berbagai konteks, khususnya untuk penulisan teks keagamaan literatur Islam, dan juga dalam karya sejarah, sastra, serta dokumen budaya masyarakat Melayu di Nusantara. Seiring perkembangan zaman, penggunaan Arab Jawi semakin menurun karena huruf Latin lebih banyak digunakan dalam bidang pendidikan dan administrasi.

Kondisi ini menjadikan digitalisasi dokumen Arab-Jawi penting untuk melestarikan warisan bahasa dan budaya Melayu. Di sisi lain, pemrosesan aksara Arab-Jawi oleh komputer masih menghadapi berbagai tantangan, salah satunya keterbatasan data beranotasi yang dapat digunakan untuk melatih dan mengevaluasi sistem pengenalan karakter secara optimal.(Coluzzi, 2020)

2.2.2 Transliterasi

Pengubahan teks dari satu sistem aksara ke aksara lain dengan menjaga arti aslinya dikenal sebagai transliterasi. Pada riset ini, transliterasi diterapkan untuk mengubah naskah Jawi menjadi huruf Latin. Pekerjaan ini memiliki tingkat kesulitan tersendiri karena aksara Jawi kerap ditulis tanpa tanda baca vokal, memiliki beragam variasi bentuk penulisan, dan mengacu pada aturan ejaan yang tidak selalu seragam.

Istilah Rumi yang digunakan dalam penelitian ini bukanlah istilah yang dibuat sendiri oleh penulis, melainkan terminologi baku yang lazim dipakai dalam kajian linguistik dan pengolahan bahasa Melayu untuk merujuk pada sistem tulisan Latin bahasa Melayu, berbeda dengan sistem tulisan Jawi yang berbasis huruf Arab. Pembedaan Jawi-Rumi ini digunakan secara konsisten dalam literatur mengenai digitalisasi naskah Melayu, termasuk dalam penelitian-penelitian tentang pelestarian aksara Jawi.(Coluzzi, 2022), sehingga penggunaan istilah Rumi pada skripsi ini konsisten dengan ketentuan penulisan yang berlaku, bukan sekadar istilah umum romanisasi yang dapat merujuk pada bahasa apa pun.

2.2.3 Definisi Dataset

Secara umum, dataset dapat diartikan sebagai sekumpulan data yang disusun secara teratur, baik dalam bentuk terstruktur maupun semi-terstruktur, dan digunakan untuk keperluan analisis, pelatihan, pengujian, serta evaluasi dalam suatu penelitian. Dalam bidang ilmu komputer dan kecerdasan buatan, dataset berfungsi sebagai representasi dari kondisi atau fenomena dunia nyata yang ingin dipelajari oleh sistem komputasi. Dataset merupakan elemen utama yang menentukan bagaimana suatu sistem memahami pola, karakteristik, dan variasi data yang ada di dunia nyata (Gong et al., 2023)

2.2.4 Data Autentik

Data autentik adalah data yang diperoleh langsung dari sumber aslinya tanpa rekayasa, seperti dokumen historis, manuskrip lama, atau hasil observasi, sehingga mencerminkan kondisi nyata dengan variasi alami, noise, dan ketidaksempurnaan. Meskipun bernilai tinggi, data autentik memiliki keterbatasan berupa jumlah yang terbatas, distribusi data yang tidak seimbang, serta biaya dan waktu anotasi yang besar, terutama pada naskah lama seperti tulisan Arab Jawi yang sering mengalami kerusakan fisik dan hilangnya tanda diakritik. Oleh karena itu, data autentik umumnya dikombinasikan dengan pendekatan lain untuk melengkapi dataset.

2.2.5 Data Sintetik

Data sintetik merupakan data yang dihasilkan secara artifisial melalui proses komputasi, seperti rendering teks, simulasi citra, atau transformasi visual. Data ini digunakan untuk menambah jumlah data dan memperluas variasi dalam dataset, terutama ketika data autentik sulit diperoleh. Augmentasi data adalah teknik untuk memperkaya variasi dataset dengan menerapkan transformasi tertentu pada data yang sudah ada tanpa mengubah makna atau label aslinya. Dalam domain citra, augmentasi dapat berupa rotasi, perubahan skala, penyesuaian pencahayaan, dan penambahan noise (Atzori et al., 2025).

2.2.6 Data Augmentasi

Augmentasi data adalah teknik untuk menambah jumlah data pelatihan dengan cara memodifikasi data yang sudah ada, misalnya dengan rotasi, pembalikan, atau penyesuaian warna pada citra. Tujuan utamanya adalah meningkatkan keberagaman

data agar model dapat belajar lebih baik dan melakukan generalisasi dengan lebih akurat. Dalam konteks tulisan Jawi, augmentasi data efektif untuk memperbanyak variasi karakter dan meningkatkan kualitas dataset tanpa harus membuat data baru dari awal. Berbeda dengan data sintetik, augmentasi tetap mempertahankan ciri utama data autentik, namun prosesnya harus dilakukan secara hati-hati agar tidak menghasilkan data yang tidak realistis atau mengubah makna serta label sebenarnya.

2.2.7 K fold cross validation

K-fold cross validation merupakan salah satu teknik validasi statistik yang digunakan untuk menguji kestabilan dan reliabilitas performa suatu model maupun dataset dengan cara membagi data menjadi beberapa bagian *(fold)* yang berukuran relatif sama. Pada penelitian ini, digunakan skema *10-fold cross validation* (k=10), sehingga dataset dibagi menjadi 10 subset data. Proses pengujian dilakukan sebanyak 10 iterasi, di mana pada setiap iterasi satu *fold* digunakan sebagai data *testing,* sedangkan 9 fold sisanya digunakan sebagai data *training*. Proses ini diulang secara bergiliran hingga seluruh *fold* pernah berperan sebagai data testing, sehingga pada akhirnya seluruh sampel dalam dataset pernah diuji oleh model.

Penggunaan cross validation dalam penelitian ini bertujuan untuk menguji reliabilitas dataset, bukan sekadar performa model. Pembagian data satu kali (misalnya 80% *training* dan 20% *testing*) berisiko bias karena hasil evaluasi sangat bergantung pada kebetulan pembagian tersebut (Wijiyanto et al., 2024). Dengan k=10, dataset diuji pada 10 kombinasi subset berbeda, sehingga setiap fold dapat dievaluasi secara terpisah untuk mendeteksi apabila ada fold dengan *CER/WER* yang menyimpang jauh menandakan ada bagian data yang kurang baik, seperti kualitas gambar yang kurang baik, variasi augmentasi yang tidak merata atau keragaman gaya tulisan kitab jawi yang luas.

2.2.8 Metrik Evaluasi Dataset

Evaluasi kualitas dan akurasi dataset pada penelitian ini dilakukan menggunakan beberapa metrik standar yang umum digunakan di bidang OCR dan transliterasi teks, yaitu untuk menilai seberapa dekat hasil pengenalan karakter dengan label ground truth pada dataset.

2.2.8.1 Character Error Rate (CER) & Word Error Rate (WER)

Untuk menilai reliabilitas dataset, digunakan beberapa metrik yaitu *Character Error Rate (CER)* dan *Word Error Rate (WER).* CER menghitung tingkat kesalahan pada tingkat huruf, sehingga sangat peka terhadap kesalahan karakter atau tanda baca, khususnya pada tulisan Arab-Jawi yang memiliki banyak huruf mirip. Sementara itu, WER mengukur kesalahan pada tingkat kata dan menunjukkan seberapa mudah teks hasil OCR dapat dibaca.

1. Character Error Rate (CER)

   CER digunakan untuk mengukur akurasi hasil OCR pada level karakter, yaitu seberapa besar perbedaan antara teks yang dikenali sistem dengan teks referensi (ground truth). Metrik ini sangat relevan untuk tulisan Arab-Jawi mengingat karakteristiknya yang kursif serta banyaknya huruf yang memiliki bentuk dasar serupa dan hanya dibedakan oleh jumlah dan posisi titik. Beberapa contoh kesalahan yang tertangkap oleh CER antara lain:

   a. Substitusi (karakter salah dikenali sebagai karakter lain), misalnya huruf "ب" salah dikenali sebagai "ت" akibat kekeliruan posisi titik.
   b. Delesi (karakter hilang dari hasil prediksi), misalnya ligatur "لا" hanya terbaca sebagai "ل"
   c. Insersi (karakter tambahan yang tidak ada pada teks asli), misalnya huruf "س" terbaca ganda menjadi "سس"

   Perhitungan CER mengikuti persamaan berikut:

$$CER = \frac{S_c + D_c + I_c}{N_c} \quad (1)$$

   dengan Sc, Dc, dan Ic berturut-turut menyatakan jumlah substitusi (S), delesi (D), dan insersi (I) pada level karakter (c), sementara Nc menyatakan jumlah total karakter pada teks referensi.

2. *Word Error Rate (WER)*

WER mengukur kesalahan pada level kata dengan logika perhitungan yang sama *(substitusi, delesi, insersi),* namun unit yang dibandingkan adalah kata, bukan karakter.

Perhitungan WER mengikuti persamaan berikut:

$$WER = \frac{S_w + D_w + I_w}{N_w} \quad (2)$$

dengan Sw, Dw, dan Iw berturut-turut menyatakan jumlah substitusi (S), delesi (D), dan insersi (I) pada level kata (w), sementara Nw menyatakan jumlah total kata pada teks referensi.

Penggunaan CER dan WER secara bersamaan memberikan gambaran yang lebih lengkap dalam mengevaluasi kualitas dataset. CER membantu menemukan kesalahan teknis pada pengenalan karakter, sedangkan WER menunjukkan pengaruh kesalahan tersebut terhadap keterbacaan teks secara keseluruhan.

2.2.8.2 Mean (Rata-rata)

Pada pengujian *k-fold cross validation* dengan $k = 10$, proses pelatihan dan pengujian dilakukan sebanyak 10 kali dengan pembagian data yang berbeda-beda tiap putaran, sehingga dihasilkan 10 nilai CER dan 10 nilai WER. Untuk merangkum hasil tersebut menjadi satu nilai yang representatif, digunakan *Mean* (rata-rata), yaitu penjumlahan seluruh nilai CER atau WER dari 10 fold dibagi dengan jumlah fold. Nilai Mean yang diperoleh mencerminkan performa rata-rata dataset. Mean dirumuskan sebagai:

$$x = \frac{x_1 + x_2 + x_3 + \cdots + x_{10}}{10} \quad (3)$$

dengan $x_1, x_2, \ldots, x_{10}$masing-masing menyatakan nilai CER atau WER yang diperoleh dari fold ke-1 hingga fold ke-10.

2.2.8.3 Standar Deviasi

Nilai Mean saja belum cukup untuk menyimpulkan kualitas dataset, karena dua dataset bisa memiliki Mean yang sama tetapi tingkat kestabilan yang berbeda. Oleh karena itu, digunakan *Standar Deviasi* untuk melihat seberapa besar variasi nilai *CER/WER* antar-fold terhadap nilai rata-ratanya. *Standar Deviasi* inilah yang menjadi indikator utama untuk menjawab apakah dataset reliable (menghasilkan performa yang konsisten di seluruh fold) atau justru timpang (performanya berbeda jauh antara satu fold dengan fold lainnya, yang bisa menandakan ada bagian data yang kualitasnya tidak merata).

Nilai Standar Deviasi menunjukkan seberapa konsisten hasil CER/WER antar fold. Semakin kecil nilainya, semakin konsisten dan reliable dataset tersebut. Sebaliknya, semakin besar nilainya, semakin besar kemungkinan dataset tidak stabil. Standar Deviasi dihitung dengan menjumlahkan selisih kuadrat antara setiap nilai fold dengan nilai rata-rata (mean), kemudian membaginya dengan jumlah fold, lalu mengambil akar kuadratnya, seperti pada rumus berikut:

$$\sigma = \sqrt{\frac{\big(x_1 - (x)\big)^2 + \big(x_2 - (x)\big)^2 + \cdots + \big(x_{10} - (x)\big)^2}{10}} \qquad (4)$$

dengan $x_1, x_2, \ldots, x_{10}$ adalah nilai CER atau WER pada tiap fold, dan $\bar{x}$ adalah nilai *Mean* yang sudah dihitung pada rumus sebelumnya.

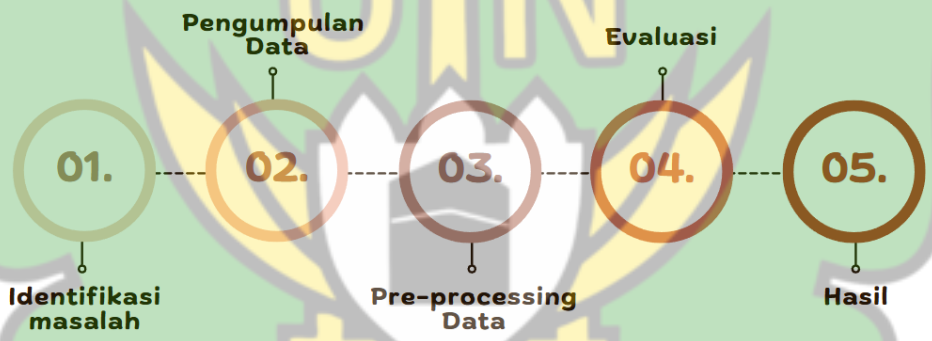
BAB III
METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian ini merupakan penelitian deskriptif-evaluatif yang bertujuan membangun dan mengevaluasi kredibilitas dataset Arab-Jawi, dengan fokus pada kelengkapan data, keberagaman representasi teks, konsistensi anotasi, serta kejelasan proses pendokumentasian dan validasi. Evaluasi ini dilakukan untuk memastikan dataset layak dijadikan rujukan dalam penelitian digitalisasi Arab-Jawi.

3.2 Tahapan Penelitian

Untuk memastikan kredibilitas dataset yang dihasilkan, penelitian ini disusun melalui tahapan-tahapan pembuatan dataset yang terstruktur dan divisualisasikan dalam diagram alur pada gambar 3.1 berikut.



Gambar 3. 1 Alur Penelitian

3.2.1 Studi Literatur dan Identifikasi masalah

Tahap awal penelitian diawali dengan penelaahan literatur mengenai pengelolaan dataset citra teks dan digitalisasi aksara non-Latin, khususnya Arab Jawi, untuk mengetahui kondisi dataset yang telah ada beserta keterbatasannya. Berdasarkan kajian tersebut, teridentifikasi sejumlah tantangan pada data Arab Jawi, seperti bentuk huruf yang saling terhubung, variasi penulisan, kondisi dokumen yang kurang baik, serta keterbatasan jumlah data. Hal ini mendasari tujuan penelitian untuk menyusun dataset Arab Jawi dengan jumlah data yang memadai, variasi citra yang beragam, serta proses dokumentasi dan validasi yang jelas, sehingga dataset yang dihasilkan kredibel dan dapat dijadikan rujukan dalam penelitian digitalisasi Arab Jawi.

3.2.2 Pengumpulan dan Persiapan Data

Dataset dalam penelitian ini berasal dari dua sumber utama yaitu hasil pemotretan langsung terhadap kitab-kitab berArab-jawi dengan kondisi fisik yang bervariasi, mulai dari yang masih terjaga baik hingga yang mengalami kerusakan atau penurunan kualitas cetak, serta cuplikan layar dari dokumen PDF yang diperoleh secara daring dari berbagai sumber di internet. Dalam penelitian ini digunakan dua jenis data, yaitu data autentik dan data sintetik, yang saling melengkapi untuk menghasilkan dataset yang lebih kuat, beragam, dan representatif. Data tersebut dibagi menjadi dua jenis:

1. Data Autentik

   Data autentik dalam penelitian ini bersifat variatif dan memiliki kualitas citra yang bervariasi, mengingat sumbernya berasal dari naskah asli berupa kitab Arab-Jawi. Kitab jawi tersebut terdiri dari kitab fisik (cetak) dan kitab digital. Kitab fisik diperoleh melalui pemotretan menggunakan kamera dengan orientasi *portrait*.

   Pada halaman dengan pencahayaan kurang optimal, pengambilan gambar dibantu aplikasi *CamScanner* untuk menyesuaikan sudut pengambilan serta meningkatkan kontras hasil foto. Adapun kitab digital lainnya didapatkan dalam bentuk PDF dari internet.

2. Data Sintetik

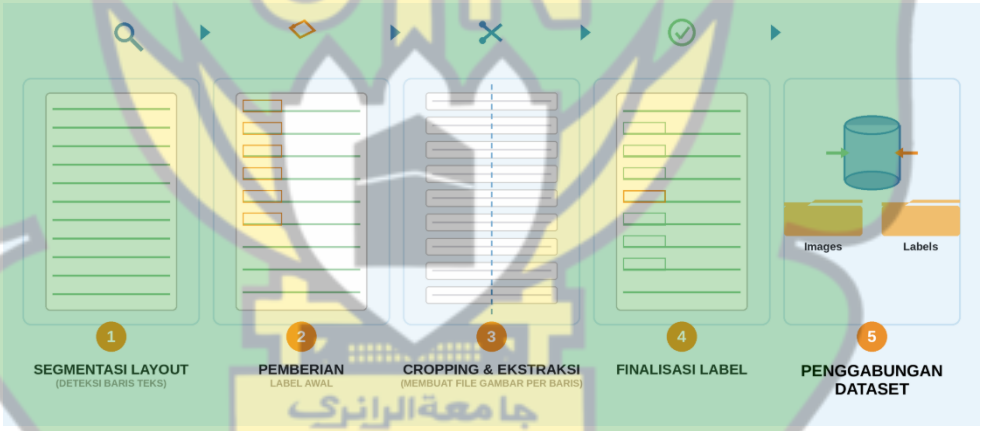
   Data sintetik adalah data buatan yang dihasilkan secara otomatis menggunakan komputer (*generated data*), bertujuan untuk menambah jumlah dan memperluas variasi data latih dalam penelitian ini. Data sintetik dalam penelitian ini dikumpulkan melalui proses *web crawling* pada situs-situs yang memuat konten islami. Berbeda dengan data autentik yang bersifat variatif dan memiliki kualitas citra yang bervariasi, data sintetik ini memiliki kualitas yang lebih bersih dan terkontrol, mengingat teks yang diperoleh dalam bentuk digital (bukan hasil pemindaian atau pemotretan).

3.2.3 *Pre-processing* Data

Pre-processing data dilakukan secara terpisah untuk data autentik dan data sintetik karena karakteristik dan sumbernya berbeda. Data autentik diperoleh dari manuskrip fisik yang digitalisasi sehingga memerlukan segmentasi dan pelabelan manual, sedangkan data sintetik dibuat secara otomatis menggunakan program. Kedua alur ini bertujuan sama, yaitu menghasilkan data yang bersih, dan siap digunakan.

3.2.3.1 Data Autentik

Persiapan data autentik dilakukan menggunakan sebuah pipeline yang dirancang secara sistematis, mencakup rangkaian langkah terstandar yang memastikan setiap sampel data terolah dan tervalidasi, sehingga dataset yang dihasilkan reliabel untuk digunakan dalam proses evaluasi. Berikut alur *pre-processing* data autentik yang di tampilkan pada gambar 3.2 berikut.



Gambar 3. 2 Alur pre-processing data autentik

1. Segmentasi layout (deteksi baris teks)

   Tujuannya menemukan dan menandai area tiap baris teks pada halaman *(bounding box)*. Modul Kraken digunakan untuk mendeteksi serta menghasilkan bounding box secara otomatis pada setiap baris teks dalam halaman dokumen. Hasil segmentasi tersebut kemudian diperiksa kembali, dan apabila ditemukan kesalahan, dilakukan penyempurnaan secara manual menggunakan VGG Image Annotator (VIA) (*VGG Image Annotator (VIA)*, n.d.) dengan keluaran anotasi dalam format JSON.

2. Pemberian label awal

   Label awal untuk sebagian besar data dihasilkan secara otomatis menggunakan model Gemini melalui Gemini API, sedangkan halaman yang gagal terbaca diproses secara manual menggunakan Document AI OCR dari Google Cloud (*Google Cloud OCR*, n.d.). Hasil transkripsi tersebut kemudian divalidasi keselarasan jumlah barisnya dengan entri anotasi JSON hasil segmentasi pada tahap sebelumnya, sebagai acuan pemotongan citra pada tahap selanjutnya.

3. *Cropping Image* (membuat file gambar per baris)

   Tujuannya adalah mengonversi setiap bounding box hasil anotasi menjadi file gambar terpisah melalui proses cropping, sehingga setiap potongan gambar yang telah ditandai dipotong dan dipisahkan menjadi satu gambar individual yang merepresentasikan satu baris teks Jawi. Setelah proses pengambilan atau pemotongan, sistem secara otomatis menghasilkan dua luaran yaitu: Citra baris teks *(cropped image)* dan Data label berformat CSV.

4. Finalisasi Label

   Proses ini melibatkan pemeriksaan silang (*cross-check*) secara manual oleh peneliti antara teks Jawi digital (label) dan potongan gambar naskah asli secara baris demi baris. Untuk mempermudah proses validasi, digunakan antarmuka berbasis web Gradio. Pemeriksaan ini bertujuan memastikan bahwa ejaan Jawi digital sepenuhnya selaras dengan tampilan visual pada gambar. Apabila ditemukan kesalahan ejaan, termasuk kekeliruan titik, perbaikan dilakukan secara langsung menggunakan keyboard Jawi. Setelah proses koreksi selesai, label dikunci dan ditetapkan sebagai ground truth.

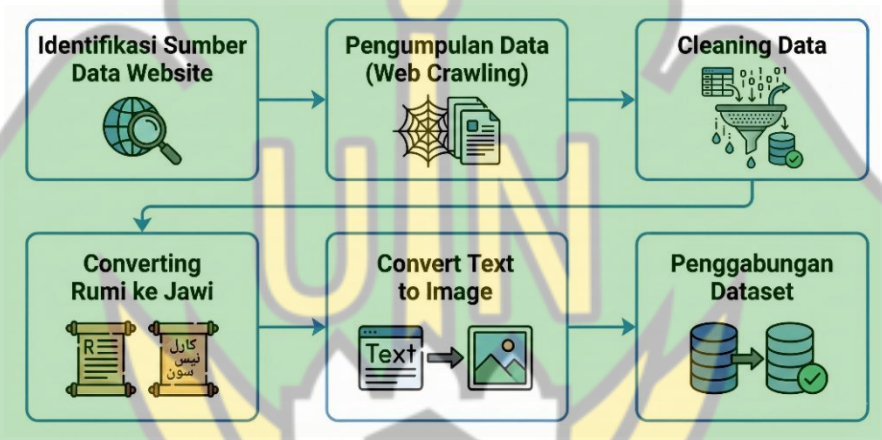
5. Penggabungan Dataset

   Penggabungan dataset dilakukan dengan menyatukan seluruh citra dan label dari setiap naskah yang telah melalui beberapa tahapan sebelumnya

ke dalam satu file CSV terstruktur, sehingga terbentuk dataset terintegrasi yang siap digunakan pada tahap pengujian selanjutnya.

3.2.3.2 Data Sintetik

Persiapan data sintetik dilakukan menggunakan sebuah pipeline yang dirancang secara sistematis, mencakup rangkaian tahapan terstandar yang memastikan setiap sampel data terolah dan tervalidasi, sehingga dataset yang dihasilkan reliabel untuk digunakan dalam proses evaluasi. Berikut alur *pre-processing* data sintetik yang ditampilkan pada Gambar 3.3 berikut.



Gambar 3. 3 Alur pre-processing data sintetik

1. Identifikasi Sumber Data

   Tahap awal penelitian dimulai dengan identifikasi sumber data digital berupa artikel-artikel keislaman dari berbagai situs web. Pemilihan sumber dilakukan menggunakan kata kunci spesifik, seperti tata cara salat, sedekah, dan materi fiqih lainnya, guna memastikan relevansi isi artikel dengan topik penelitian.

2. Pengumpulan Data *(Web Crawling)*

   Setelah sumber data ditentukan, dilakukan proses *web crawling* untuk mengambil isi artikel dari situs-situs rujukan keislaman tersebut. Teks artikel yang diperoleh kemudian dipecah menjadi satuan kalimat (*sentence splitting*) dan melalui tahap pembersihan data sebelum dikonversi ke dalam aksara Arab-Jawi.

3. *Cleaning Data*

   Setelah kalimat dipisahkan, dilakukan proses pembersihan data untuk menghapus elemen yang tidak diperlukan. Proses ini mencakup penghapusan kalimat duplikat, kalimat yang tidak lengkap atau tidak bermakna, serta baris teks yang hanya berisi angka atau simbol tidak terstruktur. Panjang kalimat juga dibatasi, misalnya minimal 3 kata dan maksimal 19 kata, agar gambar yang dihasilkan tidak terlalu panjang. Tanda baca berlebihan seperti tanda kurung, tanda seru, tanda tanya, dan simbol lainnya dibatasi jumlahnya untuk menjaga keberagaman data, sementara simbol-simbol tidak lazim dalam teks juga dihilangkan.

4. *Converting Rumi ke Jawi*

   Langkah berikutnya adalah mengonversi teks Rumi menjadi aksara Jawi. Setiap kata dicocokkan terlebih dahulu dengan kamus referensi Rumi-Jawi apabila kata tidak ditemukan dalam kamus, sistem melakukan konversi karakter perkarakter berdasarkan pemetaan huruf Rumi ke Jawi yang telah ditentukan.

5. *Convert Jawi Text to Image*

   Pada tahap ini, teks Jawi hasil konversi diubah menjadi data gambar menggunakan beberapa jenis font Arab-Melayu, di antaranya Amiri, Harmattan, Lateef, Scheherazade New, dan Noto (Naskh/Sans Arabic). Sebagian gambar dihasilkan dalam kondisi bersih, sedangkan sebagian lainnya diberi variasi tambahan seperti perubahan kecerahan, rotasi, dan gangguan visual lain, untuk memperkaya keragaman data.

6. Penggabungan Dataset

   Setelah teks Jawi dan Rumi dibersihkan dan dikonversi, langkah berikutnya adalah menggabungkan seluruh data ke dalam satu dataset. Pada tahap ini, setiap baris teks Jawi dipasangkan dengan teks Rumi yang sesuai beserta lokasi file gambarnya, sehingga dihasilkan satu file CSV yang berisi teks Jawi, teks Rumi, dan *path* gambar. Dengan cara ini, setiap pasangan teks dan gambar

tetap terkait dengan benar, sehingga terbentuk dataset terintegrasi yang siap digunakan pada tahap pengujian selanjutnya.

3.3 Evaluasi

Reliabilitas dataset diuji menggunakan *k-fold cross validation* (k=10) untuk mengetahui apakah kualitas data merata atau timpang pada bagian tertentu. Pengujian dilakukan terpisah pada dataset autentik dan sintetik, dengan prosedur yang sama untuk keduanya:

1) Dataset diacak (*shuffle*) terlebih dahulu agar data dari sumber yang sama tidak mengumpul di satu fold. Tanpa langkah ini, satu fold berisiko hanya berisi data yang mirip, sehingga hasil pengujian menjadi kurang representatif.
2) Dataset dibagi menjadi 10 fold. Pada setiap dari 10 percobaan, satu fold berperan sebagai data uji dan 9 fold sisanya sebagai data latih, sehingga setiap fold tepat satu kali menjadi data uji.
3) Setiap fold dievaluasi dengan CER dan WER, lalu hasil dari 10 fold dirangkum menjadi Mean dan Standar Deviasi. Standar deviasi rendah menunjukkan performa stabil antar fold (dataset *reliable)* jika standar deviasinya tinggi menunjukkan kualitas data timpang pada bagian tertentu.

Melalui kombinasi metrik CER, WER, Mean, dan Standar Deviasi pada kedua dataset tersebut, dapat diketahui apakah dataset autentik maupun sintetik sudah cukup andal (*reliable*) dan seimbang untuk mendukung proses pelatihan model OCR Jawi. Kedua dataset selanjutnya digabungkan dan dilatih menggunakan skema pelatihan tunggal (*single run*), lalu dievaluasi kembali pada data uji terpisah untuk dataset autentik dan sintetik (*held-out test set*) untuk hasil yang lebih objektif.

3.4 Hasil

Tahap akhir dari proses validasi ini meliputi analisis dan penyusunan hasil secara komprehensif. Hasil pengujian *k-fold cross validation* pada dataset autentik dan sintetik, beserta hasil pelatihan pada skema *single run*, disajikan dalam bentuk visualisasi data seperti grafik dan tabel perbandingan nilai CER, WER, Mean, dan Standar Deviasi. Pengujian reliabilitas terhadap dataset yang sama turut dilakukan menggunakan model lain pada penelitian terkait, sehingga kualitas dataset dapat

dinilai secara lebih menyeluruh dan tidak hanya bergantung pada hasil dari satu model saja. Keseluruhan rangkaian pengujian, mulai dari prosedur eksperimental, analisis temuan, hingga penarikan kesimpulan mengenai kualitas dataset, dicatat dan disusun secara sistematis dalam laporan tugas akhir.

3.5 Waktu dan Lokasi Penelitian

Penelitian dilaksanakan bertempat pada Laboratorium Komputasi Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, UIN Ar-Raniry Banda Aceh. Selain itu, untuk mendukung kebutuhan komputasi intensif (GPU) selama proses *fine-tuning* model, digunakan pula platform *cloud computing* seperti Google Colab Pro dan Kaggle guna akselerasi proses pelatihan. Penelitian dimulai sejak tanggal 1 Maret 2026.

3.6 Alat Bantu Penelitian

Pada penelitian ini, digunakan alat bantu berupa hardware dan software. Untuk kebutuhan pengembangan kode dan pemrosesan data ringan, digunakan satu unit komputer dengan spesifikasi sebagai berikut:

1. Processor AMD Ryzen Gen 11
2. RAM 16 GB DDR4
3. Storage SSD 512 GB

Adapun proses pelatihan model, yang membutuhkan komputasi intensif berbasis GPU, dijalankan pada lingkungan komputasi awan yaitu Google Colab Pro dan Kaggle Notebook, memanfaatkan akselerasi GPU yang disediakan kedua platform tersebut.

Untuk software, penelitian ini menggunakan bahasa pemrograman Python dengan deep learning framework PyTorch. Pemodelan dan pelatihan memanfaatkan pustaka Transformers dari Hugging Face, sedangkan evaluasi dan perhitungan metrik CER serta WER menggunakan pustaka JiWER. Proses pengolahan data memanfaatkan pustaka Pillow (PIL) untuk manipulasi citra dan Pandas untuk manipulasi data tabular, sedangkan pembagian data k-fold menggunakan Scikit-learn. Visualisasi hasil dilakukan menggunakan Matplotlib, dan penyusunan laporan tabel evaluasi menggunakan OpenPyXL untuk ekspor ke format Excel.

BAB IV
HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil penelitian dalam dua bagian utama. Pertama, dipaparkan proses pembangunan dataset terintegrasi Jawi-Rumi, mulai dari pengumpulan data hingga verifikasi, sehingga dihasilkan dataset yang representatif dan terdokumentasi secara sistematis. Kedua, disajikan hasil pengujian validitas, konsistensi, dan keseimbangan dataset menggunakan skema K-Fold Cross Validation (k=10) pada model TrOCR (trocr-base-handwritten), yang diterapkan pada data autentik dan data sintetik untuk memastikan dataset layak digunakan tanpa menimbulkan bias.

4.1 Gambaran Umum Hasil Penelitian

Penelitian ini menghasilkan dataset terintegrasi Jawi–Rumi yang dibangun dari data autentik dan data sintetik, disertai proses dokumentasi mulai dari pengumpulan hingga verifikasi data. Dataset tersebut kemudian diuji reliabilitasnya menggunakan model TrOCR (microsoft/trocr-base-handwritten) melalui skema K-Fold Cross Validation (k = 10) pada masing-masing dataset, dilanjutkan dengan satu pelatihan menggunakan data gabungan serta pembandingan terhadap model lain untuk melihat konsistensi hasil. Evaluasi menggunakan CER dan WER corpus-level, dengan mean dan standar deviasi tiap fold sebagai indikator kestabilan hasil. Bagian berikut menjelaskan proses pembuatan dan verifikasi dataset, hasil pengujian pada tiap skema, perbandingan kinerja antar sumber data, hingga perbandingan konsistensi antar-model sebagai indikator kualitas dataset secara keseluruhan.

4.2 Proses Pembuatan Dataset Terintegrasi Jawi-Rumi

Pembuatan dataset terintegrasi Jawi-Rumi dilakukan melalui dua jalur pengerjaan yang berjalan secara terpisah, yaitu penyusunan data autentik dari manuskrip kitab Jawi dan penyusunan data sintetik dari teks web yang dikonversi menjadi gambar tulisan Jawi. Masing-masing jalur dikerjakan melalui serangkaian tahap yang saling bergantung, mulai dari pengolahan sumber data, pemeriksaan , koreksi, hingga penyusunan menjadi data siap pakai,sebagaimana dipaparkan pada

pembahasan berikut. Kedua jalur pengerjaan tersebut menghasilkan dua berkas dataset yang tetap terpisah, yaitu dataset autentik dan dataset sintetik, dengan total keseluruhan 34.380 data yang siap digunakan pada tahap pengujian reliabilitas.

4.2.1 Proses Pembuatan Data Autentik

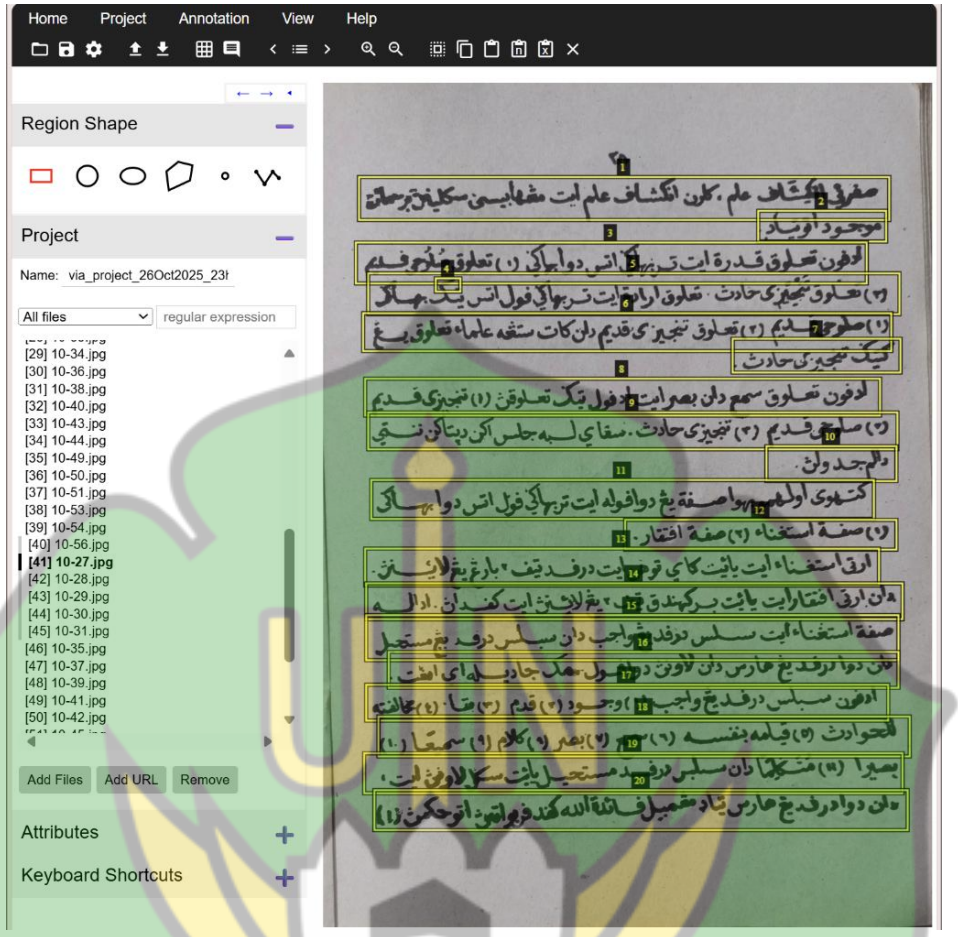
Data autentik bersumber dari halaman-halaman kitab Jawi yang telah dipindai. Pembuatannya dilaksanakan melalui beberapa tahapan berikut ini.

1. Segmentasi *Layout* (Deteksi Baris Teks)

   Seluruh halaman kitab Jawi hasil pindaian dijalankan ke modul Kraken untuk seluruh halaman dalam satu proses, guna mendeteksi baris tulisan dan menghasilkan *bounding box* (kotak penanda posisi) pada tiap barisnya secara otomatis.Hasil deteksi otomatis ini kemudian ditinjau ulang satu per satu melalui *VGG Image Annotator (VIA)* untuk memastikan hasil deteksi telah sesuai dan akurat.Tingkat keberhasilan deteksi ini bervariasi tergantung pada kondisi fisik naskah. pada halaman dengan tulisan yang rapi dan kontras tinta yang jelas, *bounding box* umumnya sudah sesuai dan hanya memerlukan sedikit penyesuaian saja.

   Sebaliknya, pada halaman dengan kualitas naskah yang kurang baik, ditemukan lebih banyak kesalahan, di antaranya kotak yang terpotong karena baris tulisan miring, kotak yang tidak muncul sama sekali pada baris dengan tinta pudar, serta kotak yang menangkap dua baris sekaligus karena jaraknya terlalu rapat. Kesalahan-kesalahan tersebut diperbaiki secara manual di VIA. kotak yang meleset digeser dan disesuaikan ulang ukurannya, baris yang tidak terdeteksi ditambahkan kotak baru, dan kotak yang menangkap dua baris dipisah menjadi dua kotak tersendiri. Proses ini dilakukan halaman demi halaman hingga setiap kotak sesuai dengan posisi baris tulisan pada gambar Jawi asli.

   Gambar dibawah ini menunjukkan proses anotasi bounding box melalui *VGG Image Annotator (VIA)* yang di tampilkan pada gambar 4.1 di bawah ini.

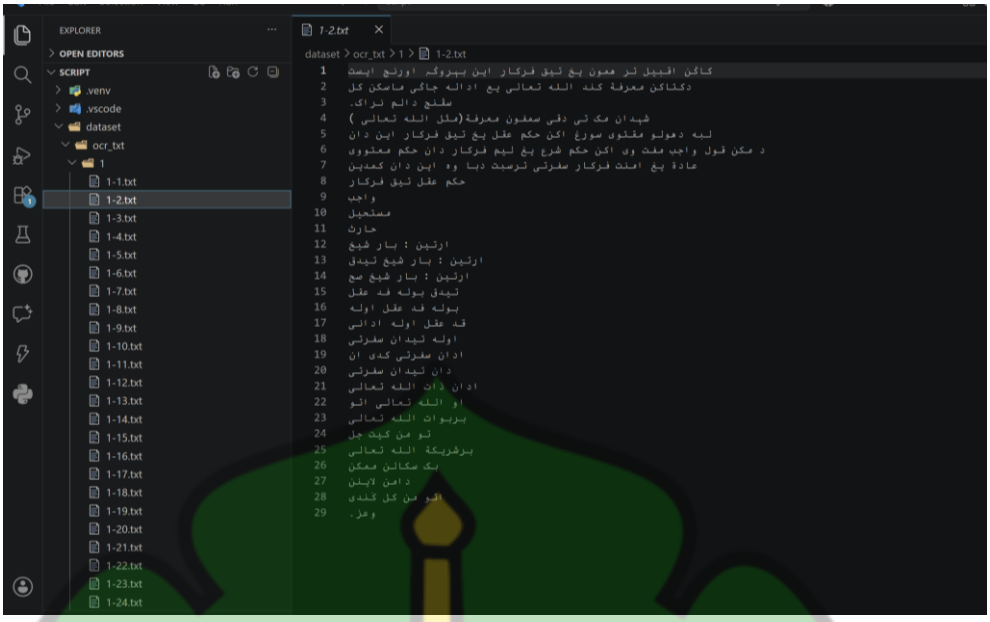


Gambar 4. 1 proses proses anotasi bounding box

Seluruh hasil koreksi kemudian diekspor menjadi satu berkas JSON berisi koordinat tiap baris, yang menjadi acuan pada tahap cropping.

2. Pemberian Label Awal

Transkripsi awal dihasilkan melalui dua tahap. Sebagian besar halaman gambar kitab jawi diekstraksi secara otomatis menggunakan model Gemini (gemini-1.5-flash) melalui Gemini API, yang membaca tulisan pada tiap gambar halaman dan mengonversinya menjadi teks digital dalam Arab-jawi, seperti yang ditunjukkan pada gambar 4.2 di bawah ini.



Gambar 4. 2 Proses ekstraksi teks dari gambar halaman kitab

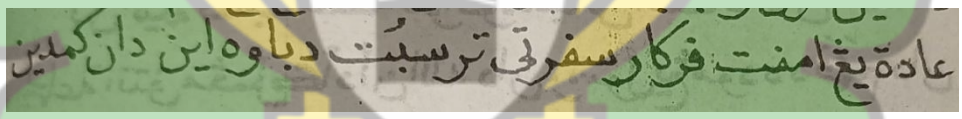
Sementara itu, halaman-halaman yang gagal terbaca melalui proses tersebut kemudian diproses secara manual menggunakan demo Document AI OCR dari Google Cloud, dengan mengunggah gambar halaman yang bersangkutan langsung ke situs tersebut. seperti yang ditunjukkan pada gambar 4.3 berikut.

Gambar 4. 3 Hasil OCR halaman kitab pada Document AI Google Cloud

Setiap halaman yang telah diekstraksi disimpan sebagai satu berkas .txt dengan format nama {Nomor Folder}-{Nomor Halaman}.txt, lalu dikelompokkan ke dalam folder sesuai nomornya. Sebagai contoh, berkas 1-1.txt berisi teks Halaman 1 dari Folder 1, dan 1-2.txt berisi teks Halaman 2 dari Folder 1. Baris-baris teks pada tiap berkas kemudian disesuaikan secara manual dengan urutan bounding box hasil segmentasi, sehingga susunan baris pada file .txt mencerminkan urutan baca sesungguhnya pada citra halaman. Label hasil OCR pada tahap ini masih berupa bacaan mentah, bukan hasil final. Oleh sebab itu, seluruh isi berkas .txt pada tahap ini masih bersifat sementara dan akan melalui proses koreksi menyeluruh pada tahap finalisasi label.

3. *Cropping image*

Proses cropping dilakukan berdasarkan koordinat bounding box tiap baris teks yang tercatat dalam file JSON, meliputi titik pojok kiri-atas serta lebar dan tinggi area baris tersebut. Berdasarkan keempat nilai ini, script Python secara otomatis menghitung area pemotongan, mengeksekusi cropping, lalu menyimpan hasilnya sebagai citra baris individual seperti yang di tampilkan pada gambar 4.4 berikut.

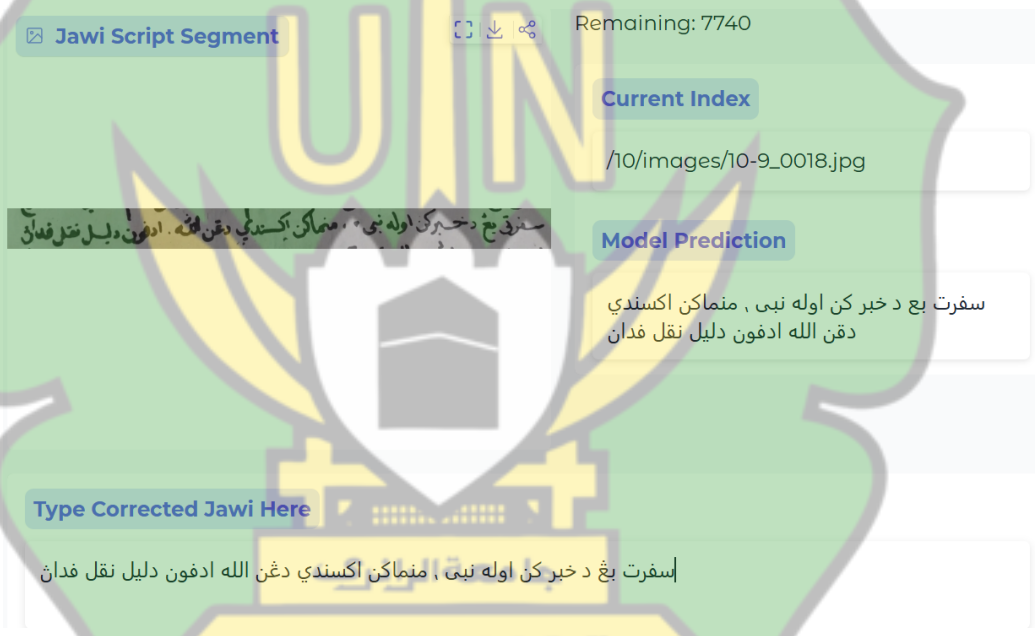


Gambar 4. 4 hasil pemotongan citra perbaris

Setiap citra hasil pemotongan diberi nama file berdasarkan nomor halaman dan urutan barisnya, sehingga posisi asal tiap citra pada kitab dapat ditelusuri kembali dengan mudah. Seluruh proses ini dilakukan secara otomatis untuk semua halaman, tanpa penyesuaian manual. Bersamaan dengan proses cropping, sistem juga menghasilkan file CSV awal yang memasangkan path citra baris (hasil cropping) dengan label teks hasil OCR untuk baris tersebut, sehingga terbentuk data mentah citra-teks. Label ini masih berupa hasil bacaan OCR awal, sehingga belum tentu akurat.

4. Finalisasi dan Koreksi Label

Proses koreksi label dilakukan melalui antarmuka web berbasis Gradio yang dirancang khusus untuk mendukung anotasi teks Jawi secara interaktif. Antarmuka ini menampilkan segmen citra baris teks di satu sisi, berdampingan dengan hasil prediksi awal dari model pada sisi lainnya, sehingga memudahkan proses pembandingan antara citra asli dan keluaran sistem. Tersedia pula kolom input untuk mengetikkan hasil transkripsi yang telah dikoreksi, apabila prediksi model masih mengandung kesalahan. Tampilan antarmuka tersebut dapat dilihat pada gambar 4.5 di bawah ini.



Gambar 4. 5 Tampilan web gradio

Gambar di atas menunjukkan antarmuka web berbasis Gradio untuk koreksi label teks Jawi, yang menampilkan citra baris teks, hasil prediksi awal model, dan kolom input untuk mengetikkan hasil koreksi. Kesalahan yang paling sering ditemukan pada hasil prediksi awal OCR berkaitan dengan karakter Arab-Jawi yang memiliki bentuk dasar serupa namun berbeda jumlah atau posisi titik, sehingga model kerap keliru mengenali satu huruf sebagai

huruf lain yang mirip. Kesalahan lain yang ditemukan meliputi huruf hilang, urutan huruf terbalik, serta kata yang sama sekali tidak terbaca.

Jenis-jenis kesalahan inilah yang menjadi fokus utama proses koreksi, untuk memastikan setiap karakter pada hasil transkripsi sesuai dengan tulisan pada gambar aslinya. Koreksi dilakukan menggunakan dua keyboard virtual Jawi secara bergantian, yaitu Lexigos Arabic dan Arab-Jawi, yang saling melengkapi jika suatu karakter tidak tersedia pada satu keyboard, annotator berpindah ke keyboard lainnya. Khusus beberapa potongan doa pada data autentik yang tidak dapat disusun lewat kedua keyboard tersebut, teksnya dicari dan dicocokkan secara manual melalui internet, kemudian disalin ke kolom koreksi.

Setiap hasil koreksi disimpan ke berkas CSV melalui tombol submit begitu satu gambar selesai dikoreksi, sistem menyimpan pasangan nama file dan teks hasil koreksi sebagai satu baris data baru, lalu otomatis melanjutkan ke gambar berikutnya hingga seluruh antrian selesai diproses. Dengan demikian, seluruh hasil koreksi label dari gambar pertama hingga terakhir terkumpul dalam satu berkas CSV yang sama.

5. Transliterasi Jawi ke Rumi dan Penyusunan Dataset Akhir

Setelah label Jawi selesai dikoreksi secara manual oleh annotator, sistem secara otomatis menghasilkan pasangan teks Rumi untuk setiap label tersebut. memanfaatkan kamus Rumi–Jawi dari basis data Firebase yang dibalik arah pemetaannya menjadi kamus Jawi-Rumi. Proses pencocokan mempertimbangkan variasi ejaan yang umum terjadi pada naskah Jawi klasik maupun hasil OCR, mengingat sejumlah huruf tambahan dalam ejaan Jawi (seperti ڤ dan چ) merupakan modifikasi dari huruf Arab dasar (ف dan ج) yang tidak selalu digunakan konsisten.

Sistem terlebih dahulu menormalkan ejaan dan mencoba beberapa variasi penulisan sebelum mencocokkan ke kamus; apabila kata tidak ditemukan persis, pencocokan dilakukan berdasarkan kerangka konsonan (skeleton matching), dan jika masih tidak ditemukan, digunakan transliterasi huruf per huruf sebagai upaya terakhir. Untuk kata dengan lebih dari satu kemungkinan pasangan Rumi, sistem memilih kandidat paling sesuai konteks

menggunakan model bahasa sederhana berbasis frekuensi kata dan pasangan kata (bigram) dari korpus Rumi lokal.

Proses ini dilengkapi mekanisme checkpoint yang menyimpan progres secara berkala, sehingga jika terhenti di tengah jalan misalnya karena gangguan koneksi proses dapat dilanjutkan tanpa mengulang dari awal.

Hasil transliterasi berbasis kamus ini kemudian diverifikasi ulang menggunakan model bahasa besar (Gemini) untuk mengoreksi kasus-kasus yang tidak tertangani dengan baik oleh pencocokan kamus, sebelum akhirnya melalui pemeriksaan dan koreksi manual sebagai tahap validasi akhir. Berkas akhir berupa satu berkas Excel yang memuat kolom jalur citra, teks Jawi terkoreksi, dan teks Rumi hasil transliterasi yang telah diverifikasi.

4.2.2 Proses Pembuatan Data Sintetik

Data sintetik bersumber dari teks-teks keislaman yang dikumpulkan dari internet dan dikonversi menjadi gambar tulisan Jawi. Pembuatannya dilaksanakan melalui beberapa tahap sebagai berikut.

1. Pengumpulan dan Pembersihan Data

   Tahap pertama dilakukan dengan mengumpulkan teks digital dari internet sebagai bahan baku data sintetik. Sebanyak 50 tautan situs keislaman dikurasi secara manual, dipilih berdasarkan kesesuaian topik dengan konteks penggunaan Arab-Jawi pada kitab-kitab keagamaan, seperti tata cara salat, sedekah, dan fiqih. Script crawling berhasil mengambil isi artikel dari 36 domain situs keislaman, menghasilkan 1.582 artikel, dengan lima kontributor terbesar yaitu islampos.com sebanyak 188 artikel, alkhoirot.com 181 artikel, voa-islam.com 178 artikel, konsultasisyariah.com 167 artikel, dan muslim.or.id 131 artikel.

   Artikel tersebut kemudian dipecah menjadi satuan kalimat (*sentence splitting*), menghasilkan 53.810 kalimat mentah**.** Kalimat-kalimat ini melewati proses pembersihan otomatis mencakup penghapusan kalimat duplikat, pembatasan panjang kalimat pada rentang 3-19 kata, penghapusan kalimat tidak bermakna serta baris berisi angka/simbol acak, dan pembatasan tanda baca berlebihan agar bentuk kalimat wajar saat dirender menjadi gambar.

Melalui proses ini, jumlah kalimat menyusut dari 53.810 menjadi 26.478 kalimat bersih**.**

2. Konversi Teks Rumi ke Arab-jawi

Kalimat-kalimat bersih hasil crawling, yang masih berbentuk teks Rumi (Latin), dikonversi ke Arab-Jawi menggunakan pendekatan hybrid dua lapis. Pada lapis pertama, setiap kata dicocokkan dengan kamus rujukan Rumi-Jawi yang disimpan di Firebase Firestore. Jika ditemukan, ejaan Jawi baku dari kamus langsung digunakan karena penulisan Jawi tidak selalu bisa diturunkan murni dari bunyi huruf.

Jika kata tidak ditemukan dalam kamus (*out-of-vocabulary*, misalnya kata serapan atau istilah baru), sistem beralih ke lapis kedua berupa algoritma konversi karakter-per-karakter berdasarkan aturan baku transliterasi Rumi-Jawi, mencakup pemetaan huruf tunggal maupun gabungan (ng, ny, sy, kh, dh), penanganan huruf vokal (i, e, o, u) yang bentuknya berubah sesuai posisi di awal/tengah/akhir kata, konversi angka Latin ke angka Arab, serta konversi tanda baca koma, tanda tanya, dan titik koma ke bentuk RTL (*Right-to-Left*, arah penulisan kanan-ke-kiri sesuai kaidah Arab-Jawi).

Proses ini dijalankan dengan mekanisme checkpoint otomatis setiap 100 baris, sehingga apabila terhenti dapat dilanjutkan dari baris terakhir tanpa mengulang dari awal. Setelah seluruh data dikonversi, dilakukan pemeriksaan sampel (*spot-check*) secara manual untuk memastikan hasil konversi sesuai kaidah ejaan Jawi yang benar.

3. Rendering Teks Jawi Menjadi citra dan Augmentasi Data

Teks Jawi hasil konversi diubah menjadi data citra, karena model OCR pada penelitian ini membutuhkan input berupa gambar untuk proses pelatihan dan evaluasi. Rendering dilakukan menggunakan Playwright dengan mesin Chromium (*headless browser,* browser tanpa tampilan visual yang tetap dapat memuat dan merender halaman web secara utuh). Setiap kalimat Jawi disusun dalam HTML/CSS dengan arah tulisan kanan-ke-kiri (RTL) dan ligatur Arab diaktifkan agar sambungan antar-huruf tampak alami, serta *line-height* diperbesar menjadi 2,2 kali ukuran font untuk mengakomodasi

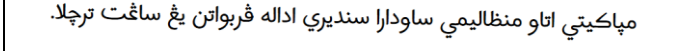
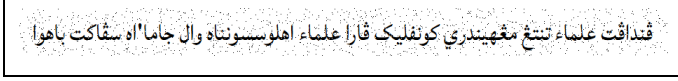
ascender/descender huruf Arab yang lebih tinggi dari huruf Latin, sehingga tidak terpotong saat gambar diambil. Setelah dirender, sistem mengukur posisi dan ukuran teks (*bounding box*), lalu meng-screenshot seluruh halaman dan meng-*crop*-nya sesuai bounding box tersebut, sehingga gambar akhir hanya berisi tulisan Jawi tanpa ruang kosong berlebih dan tanpa huruf terpotong.

Untuk memperkaya variasi gaya tulisan, digunakan tujuh font Arab-Melayu yang dipilih secara acak berbobot (*weighted random*) per kalimat, terdiri dari lima font Naskh/tradisional (Noto Naskh Arabic, Scheherazade New, Amiri, Lateef, Harmattan) dan dua font sans modern (Noto Sans Arabic, IBM Plex Sans Arabic), dengan bobot font tradisional dibuat lebih besar mengikuti karakter naskah Jawi asli yang umumnya bergaya Naskh.

Selanjutnya, untuk mensimulasikan kondisi tulisan atau hasil pemindaian yang bervariasi tanpa perlu menambah jumlah kalimat sumber, setiap gambar diberi salah satu dari delapan jenis augmentasi, Gaussian blur, perubahan kecerahan/kontras, Gaussian noise, latar belakang keabu-abuan, rotasi ±5°, morfologi (erosi/dilasi), salt-and-pepper noise, dan motion blur yang masing-masing meniru kondisi nyata seperti fokus kamera kurang tajam, pencahayaan tidak merata, derau sensor, kertas tidak bersih, dokumen miring, tinta tidak konsisten, kualitas scan rendah, dan guncangan kamera. Komposisi akhir dataset dibuat 30% citra bersih (*clean*) dan 70% citra teraugmentasi dengan bobot proporsional antar jenis.

Seluruh proses rendering dan augmentasi ini dilengkapi mekanisme *resume* yang mengecek gambar mana yang belum berhasil dibuat, sehingga proses berskala besar dapat dilanjutkan tanpa mengulang dari awal bila terhenti. Berikut ditampilkan beberapa contoh hasil render teks Jawi menjadi citra beserta variasi augmentasinya pada Gambar 4.1 berikut.

Tabel 4. 1 Contoh Hasil Augmentasi pada Citra Tulisan Jawi

Gambar augmentasi	Jenis
	rotation
	salt_pepper

ڤنتيڠ باکي ستياڤ مسلم اونتوق منجاک ليسن دان توليسنڽ،	brightness
هوسنودزون (برڤيکير ڤوسيتيف) ترهادڤ سسام مسلم داڤت ممبنتو	gaussian_noise

4. Finalisasi Dataset

Pada tahap akhir proses rendering, setiap gambar yang dihasilkan dipasangkan dengan path berkasnya, label teks Jawi, teks Rumi, dan jenis augmentasi yang diterapkan, kemudian disimpan ke dalam satu file Excel . Sebagai pemeriksaan akhir, sebagian sampel diperiksa secara acak untuk memastikan pasangan gambar dan label tidak tertukar. Melalui rangkaian proses ini, diperoleh 18.535 pasangan data gambar-teks sintetik yang siap digunakan.

4.3 Hasil Pembuatan Dataset

Dataset dibangun melalui dua jalur pembuatan berbeda, data autentik bersumber dari kitab Jawi (manuskrip fisik), dan data sintetik dibuat dari teks digital hasil crawling yang direnderer menjadi citra menggunakan font komputer. Dari kedua jalur tersebut, diperoleh total 34.380 data yang dinyatakan layak dan siap digunakan dalam penelitian ini.

Kedua dataset disimpan dalam format Excel dengan struktur kolom nama file/path gambar, teks Jawi, dan teks Rumi. Pada data autentik, struktur kolom terdiri atas image_path, jawi_text, dan rumi_text. Pada data sintetik, struktur kolom sama namun ditambahkan satu kolom aug_type untuk menandai jenis augmentasi yang diterapkan pada tiap gambar .Perbedaan ini muncul karena data sintetik melalui tahap augmentasi, sedangkan data autentik tidak.

4.4 Distribusi Dataset Final

Setelah proses pengumpulan, pembuatan, dan verifikasi data yang telah dijelaskan sebelumnya, diperoleh dataset final yang terdiri dari dua jenis, yaitu dataset autentik dan dataset sintetik. Bagian ini memaparkan statistik deskriptif dan distribusi dari kedua dataset tersebut, meliputi jumlah data, variasi teks, dan distribusi

augmentasi, sebagai gambaran karakteristik dataset sebelum digunakan pada tahap pengujian reliabilitas berikutnya.

4.4.1 Deskripsi Dataset Autentik

Dataset autentik berasal dari hasil cropping baris teks pada kitab-kitab Arab Jawi fisik maupun digital. Panjang teks dalam dataset bervariasi dalam jumlah karakter maupun jumlah kata perbaris. Variasi ini menunjukkan keragaman struktur baris pada manuskrip Jawi, sehingga dataset autentik yang terbentuk cukup representatif untuk mendukung tujuan penelitian dalam membangun dataset.

4.4.2 Deskripsi Dataset Sintetik

Dataset sintetik dihasilkan melalui proses crawling artikel keagamaan Islam, yang kemudian dikonversi menjadi gambar teks Arab Jawi dengan berbagai font dan teknik augmentasi. Karakteristik dataset sintetik ditampilkan pada tabel 4.3 berikut.

Tabel 4. 2 Deskripsi Dataset Sintetik

Keterangan	**Nilai**
Total data	18.535
Teks terpendek	28 karakter
Teks terpanjang	132 karakter
Jumlah kata minimum per baris	3 kata
Jumlah kata maksimum per baris	19 kata

Dataset sintetik yang dihasilkan berjumlah 18.535 data, dengan panjang teks bervariasi antara 28 hingga 132 karakter, serta jumlah kata per baris berkisar 3 hingga 19 kata. Rentang ini lebih sempit dan lebih seragam dibandingkan dataset autentik, karena teks sintetik berasal dari kalimat-kalimat utuh hasil crawling artikel keagamaan yang kemudian dikonversi menjadi gambar, sehingga tidak mengalami pemotongan baris seperti pada manuskrip asli.

4.4.3 Distribusi Jenis Augmentasi Data Sintetik

Dataset sintetik dalam penelitian ini dibuat dengan 8 jenis augmentasi, yang masing-masing meniru kondisi visual yang biasa ditemukan pada dokumen cetak Arab

Jawi di dunia nyata (misalnya noda, blur, atau warna kertas). Jumlah data untuk setiap jenis augmentasi ditampilkan pada tabel 4.4 berikut.

Tabel 4. 3 Deskripsi Jenis Augmentasi

No	Tipe Augmentasi	Jumlah Data	Persentase
1	*Gaussian Blur*	4.448	24,0%
2	*Brightness*	3.707	20,0%
3	*Gaussian Noise*	2.965	16,0%
4	*Gray Background*	1.853	10,0%
5	*Rotation*	1.853	10,0%
6	*Morphology*	1.482	8,0%
7	*Motion Blur*	1.115	6,0%
8	*Salt & Pepper*	1.112	6,0%
	Total	18.535	100%

Dari 8 tipe augmentasi, urutannya sebagai berikut:

1. *Gaussian blur* : 4.448 data (24,0%) Efek ini membuat gambar buram, menggambarkan kondisi hasil cetak atau foto yang kurang fokus, kondisi yang paling sering ditemui saat digitalisasi dokumen.
2. *Brightness* : 3.707 data (20,0%). Mengubah tingkat kecerahan gambar (lebih terang/gelap), menggambarkan kondisi pencahayaan tidak merata saat proses pemindaian dokumen atau pemotretan.
3. *Gaussian noise:* 2.965 data (16,0%).Menambahkan noise acak, menggambarkan kondisi gangguan dari kamera atau scanner berkualitas rendah.
4. *Gray background* : 1.853 data (10,0%). Menggambarkan kondisi latar kertas yang pudar atau kusam akibat proses pemindaian dokumen lama.
5. *Rotation:* 1.853 data (7,0%). Menggambarkan kondisi dokumen yang miring saat dipindai.
6. *Morphology:* 1.482 data (8,0%). Mengubah ketebalan karakter Jawi, menggambarkan kondisi variasi ketebalan tinta cetak.
7. *Motion blur:* 1.115 data (6,0%). Menggambarkan efek buram akibat kamera bergerak saat pemotretan.

8. *Salt & pepper* : 1.112 data (6,0%). Menambahkan bintik hitam-putih acak, menggambarkan kondisi kertas kotor atau debu pada scanner.

4.5 Uji Reliabilitas Dataset

Uji reliabilitas dilakukan untuk mengukur apakah dataset Arab-Jawi yang dibangun valid, konsisten, dan seimbang sehingga layak digunakan tanpa bias. Pengujian dilakukan dengan *10-fold cross validation* skema *GroupKFold* pada dataset autentik dan sintetik secara terpisah. Model TrOCR di sini berfungsi sebagai alat ukur kondisi dataset, fokus pengujian bukan pada kemampuan model, melainkan pada dataset itu sendiri. Jika nilai CER/WER stabil di seluruh 10 fold, artinya kualitas data merata dan dataset konsisten. Sebaliknya, ketidakstabilan nilai CER/WER antar-fold menandakan adanya ketimpangan kualitas pada bagian dataset tertentu.

4.5.1 Tujuan Pengujian

Pengujian *10-fold cross validation* bertujuan mengukur keandalan (*reliability*) dataset autentik dan sintetik. Keandalan ini dilihat dari nilai rata-rata (*mean*) dan simpangan baku (*standard deviation*) CER/WER di seluruh fold, simpangan baku kecil menunjukkan kualitas data merata dan tidak ada ketimpangan akibat pembagian data tertentu.

Oleh karena itu pengukuran ini bergantung pada pemilihan jumlah fold (k) yang sesuai, karena nilai k menentukan proporsi data latih dan data uji di setiap fold. Nilai k=10 dipilih karena memberikan keseimbangan terbaik antara jumlah data latih dan data uji. Jika k dinaikkan menjadi 15 atau 20, tambahan data latih yang diperoleh sangat kecil, sementara data uji di tiap fold justru semakin sedikit. Akibatnya, hasil CER/WER di tiap fold menjadi kurang stabil karena diuji dengan jumlah sampel yang terlalu sedikit.

Setelah dataset terbukti andal, pengujian dilanjutkan dengan metode *single run* menggunakan seluruh data gabungan, untuk memastikan dataset tetap layak dipakai dalam skema pelatihan penuh.Tahap akhir adalah evaluasi (*Held-out test*) menggunakan 10% data uji independen dari masing-masing dataset (autentik dan sintetik). Tahap ini bertujuan membuktikan bahwa dataset yang dibangun tidak hanya andal saat diuji dengan skema K-Fold, tetapi juga tetap terbukti representatif saat diuji secara independen pada tahap *Held-out test* dengan skema *single run*.

4.5.2 Prosedur Pengujian Fold

Dataset yang digunakan dalam pengujian k-fold terdiri atas tiga komponen yang saling berpasangan, yaitu path gambar *(image path),* teks Jawi, dan teks Rumi (hasil transliterasi). Ketiga komponen ini disimpan dalam satu file data, di mana path gambar akan dicocokkan dengan citra pada folder gambar sebagai representasi visual dari pasangan teks Jawi. Pengujian k-fold dilakukan terhadap keseluruhan dataset ini secara utuh dalam satu proses training, sehingga tidak terdapat pemisahan pengujian antara tugas OCR dan tugas transliterasi. Dengan kata lain, pengujian menggunakan pasangan data citra-teks secara bersamaan, bukan dengan skema training terpisah untuk transkripsi (image-to-text) dan transliterasi (Jawi-to-Rumi).

Dataset dibagi menjadi 10 fold menggunakan skema GroupKFold dengan k = 10, dilakukan terpisah untuk dataset autentik dan dataset sintetik. Pada setiap iterasi pengujian, 9 fold digunakan sebagai data training dan 1 fold digunakan sebagai data testing. Proses ini diulang sebanyak 10 kali sehingga setiap fold mendapat giliran tepat satu kali sebagai data testing.

Berbeda dengan pembagian acak biasa (*random split*), GroupKFold membagi data berdasarkan kunci pengelompokan (*grouping*) yang disesuaikan dengan karakteristik masing-masing dataset. Dataset sintetik dikelompokkan berdasarkan teks sumber (jawi_text), karena satu teks sumber dapat menghasilkan banyak varian data melalui augmentasi. Dataset autentik dikelompokkan berdasarkan page_id (identitas halaman manuskrip asal), karena potongan-potongan gambar yang berasal dari halaman yang sama cenderung memiliki karakteristik tulisan yang mirip.

Dengan cara ini, seluruh varian atau potongan dari kelompok yang sama akan tetap berada dalam satu fold yang sama, sehingga tidak ada anggota kelompok yang terpisah antara data latih dan data validasi.

4.6 Hasil Pengujian K-Fold Cross Validation

Bagian ini memaparkan hasil CER, WER, *mean*, dan *standard deviation* dari pengujian 10-fold pada dataset autentik dan sintetik.

4.6.1 Hasil Pengujian K-Fold pada Dataset Autentik

Pengujian 10-fold cross validation dilakukan untuk menguji reliabilitas dataset autentik, yaitu apakah tingkat kesulitan data akibat kualitas pindaian dan variasi bentuk tulisan tercampur rata di seluruh halaman atau justru menumpuk pada bagian tertentu. Pembagian fold menggunakan GroupKFold berbasis page_id, karena satu halaman manuskrip menghasilkan banyak citra baris hasil cropping dengan kondisi pindaian dan gaya tulisan yang sama. Jika pembagian dilakukan secara acak per citra, potongan-potongan dari halaman yang sama berisiko muncul bersamaan di data latih dan data uji, sehingga bisa terjadi kebocoran data (data *leakage)*. Dengan pengelompokan berbasis page_id, seluruh citra dari halaman yang sama tidak pernah muncul bersamaan di data latih dan data uji. Hasil pengujian pada setiap fold ditampilkan pada Tabel 4.5 di bawah ini.

Tabel 4. 4 Hasil fold autentik

Fold	CER Autentik	WER Autentik
Fold 1	16,00%	51,73%
Fold 2	15,17%	48,95%
Fold 3	15,55%	50,15%
Fold 4	18,42%	55,03%
Fold 5	16,32%	51,87%
Fold 6	16,32%	51,55%
Fold 7	17,11%	53,18%
Fold 8	15,52%	50,65%
Fold 9	16,25%	52,14%
Fold 10	16,87%	51,76%
Rata-rata (mean)	16,35%	51,70%
Standart Deviasi (std)	0,94%	1,65%

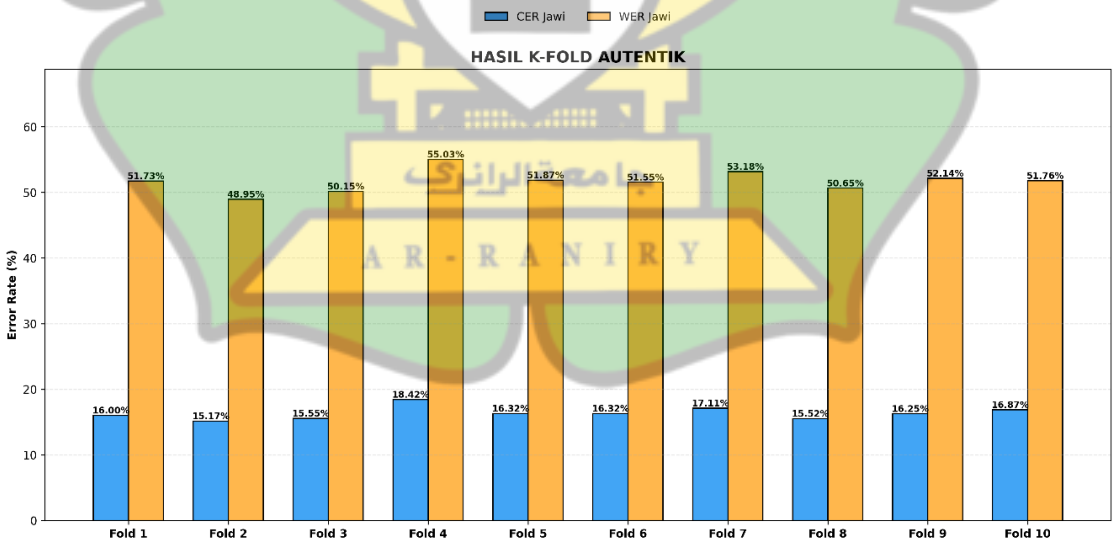
Berdasarkan Tabel 4.5 di atas, nilai CER pada seluruh fold berada pada rentang 15,17% hingga 18,42%, dengan rata-rata 16,35% dan standar deviasi 0,94%. Sementara itu, WER berkisar antara 48,95% hingga 55,03%, dengan rata-rata 51,70% dan standar deviasi 1,65%. Kecilnya nilai standar deviasi pada kedua metrik ini

menunjukkan bahwa tingkat kesulitan data tersebar cukup merata di seluruh halaman dataset, tidak menumpuk pada bagian tertentu saja.

Hal ini penting diperhatikan karena pembagian fold dilakukan secara acak berdasarkan page ID, tanpa rekayasa terhadap halaman mana yang masuk ke fold tertentu, tanpa mempertimbangkan karakteristik halaman tertentu,misalnya susunan huruf yang rapat, tulisan kitab yang kurang jelas, atau citra tulisan dengan kontras rendah sehingga tulisan sulit dibaca kebetulan terkumpul pada satu fold saja, tentu akan muncul fold dengan CER dan WER yang jauh lebih tinggi dari fold lainnya. Namun pada praktiknya, performa antar fold relatif stabil, sehingga variasi kualitas tulisan dan hasil scan pada dataset Jawi-Rumi ini dapat dikatakan tersebar merata di seluruh halaman, bukan menumpuk pada kelompok tertentu.

Dengan demikian, konsistensi hasil pada seluruh fold ini menjadi bukti bahwa dataset Jawi-Rumi yang dibangun dalam penelitian ini sudah cukup reliabel dan layak digunakan sebagai dasar evaluasi. Stabilnya nilai CER dan WER pada berbagai kombinasi halaman menunjukkan bahwa kualitas dataset tidak hanya baik pada kelompok data tertentu, melainkan konsisten di seluruh cakupan data yang dikumpulkan.

Untuk memperjelas pola distribusi CER dan WER pada setiap fold, hasil yang sama divisualisasikan juga dalam bentuk grafik pada gambar 4.6 di bawah ini.



Metrik Pengujian	Rata-rata (Mean)	Standar Deviasi (Std)
Character Error Rate (CER)	16.35%	0.94%
Word Error Rate (WER)	51.70%	1.65%

Gambar 4. 6 Grafik Hasil fold Autentik

Gambar di atas memperlihatkan setiap batang CER dan WER pada mayoritas fold tampak berdekatan, sesuai dengan standar deviasi yang kecil. Ini menjadi bukti bahwa dataset yang digunakan cukup reliabel dan layak gunakan.

4.6.2 Hasil Pengujian K-Fold pada Dataset Sintetik

Selain pada dataset autentik, pengujian 10-fold cross validation dengan pendekatan GroupKFold juga dilakukan pada dataset sintetik,untuk melihat sejauh mana reliabilitas dataset. Hasil pengujian pada setiap fold ditampilkan pada Tabel 4.6 dibawah ini.

Tabel 4. 5 Hasil fold Sintetik

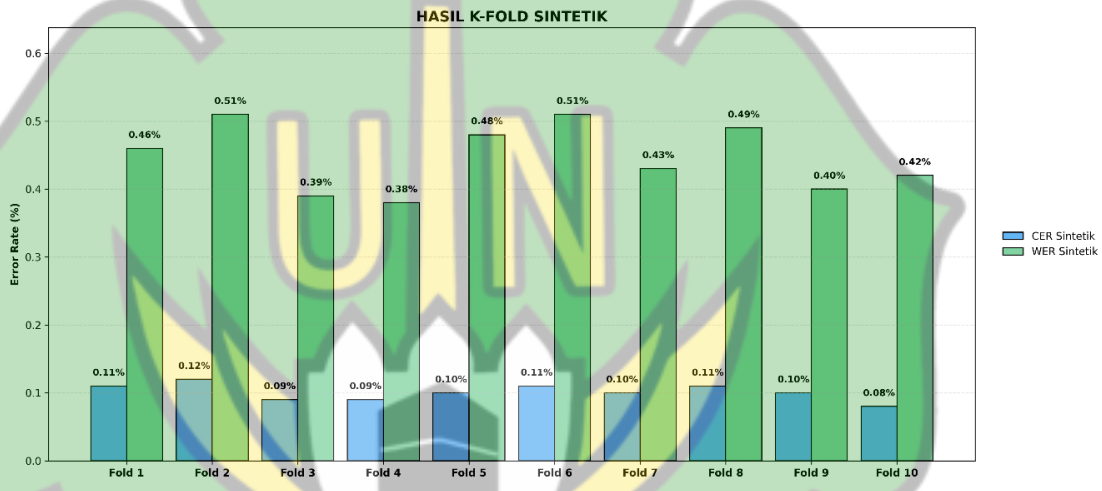
Fold	CER Sintetik	WER Sintetik
Fold 1	0,11%	0,46%
Fold 2	0,12%	0,51%
Fold 3	0,09%	0,39%
Fold 4	0,09%	0,38%
Fold 5	0,10%	0,48%
Fold 6	0,11%	0,51%
Fold 7	0,10%	0,43%
Fold 8	0,11%	0,49%
Fold 9	0,10%	0,40%
Fold 10	0,08%	0,42%
Rata-rata (Mean)	0,10%	0,45%
Standart Deviasi (std)	0,01%	0,05%

Berdasarkan tabel di atas, nilai CER pada seluruh fold berkisar antara 0,08% sampai 0,12%, sedangkan nilai WER berkisar antara 0,38% sampai 0,51%. Rata-rata CER yang diperoleh sebesar 0,10% dengan standar deviasi hanya 0,01%, dan rata-rata WER sebesar 0,45% dengan standar deviasi 0,05%. Standar deviasi yang sangat kecil ini menunjukkan performa model yang stabil dan seragam di seluruh fold. Hal ini wajar terjadi karena dataset sintetik dibangun melalui proses konversi teks menjadi

gambar secara digital, sehingga bentuk huruf dan kualitas gambar relatif seragam di seluruh sampel.

Tidak ditemukan fold yang menyimpang secara mencolok pada pengujian ini, sebagaimana terlihat dari selisih CER dan WER antar fold yang sangat tipis. Hal ini menyatakan bahwa proses pembuatan data sintetik menghasilkan karakteristik data yang konsisten dan tersebar merata di seluruh fold.Dengan demikian, hasil pengujian K-Fold pada dataset sintetik menunjukkan reliabilitas yang sangat baik, sejalan dengan sifat pembuatannya yang terstruktur.

Untuk memperjelas pola distribusi CER dan WER pada setiap fold, hasil yang sama divisualisasikan dalam bentuk grafik pada Gambar 4.7 di bawah ini



Metrik Pengujian	Rata-rata (Mean)	Standar Deviasi (Std)
Character Error Rate (CER)	0.10%	0.01%
Word Error Rate (WER)	0.45%	0.05%

Gambar 4. 7 Grafik Hasil fold Sintetik

Secara visual, nilai pada batang CER dan WER pada grafik tampak merata di seluruh fold tanpa ada fold yang menonjol secara signifikan, memperkuat gambaran bahwa dataset sintetik ini stabil digunakan dalam pengujian.

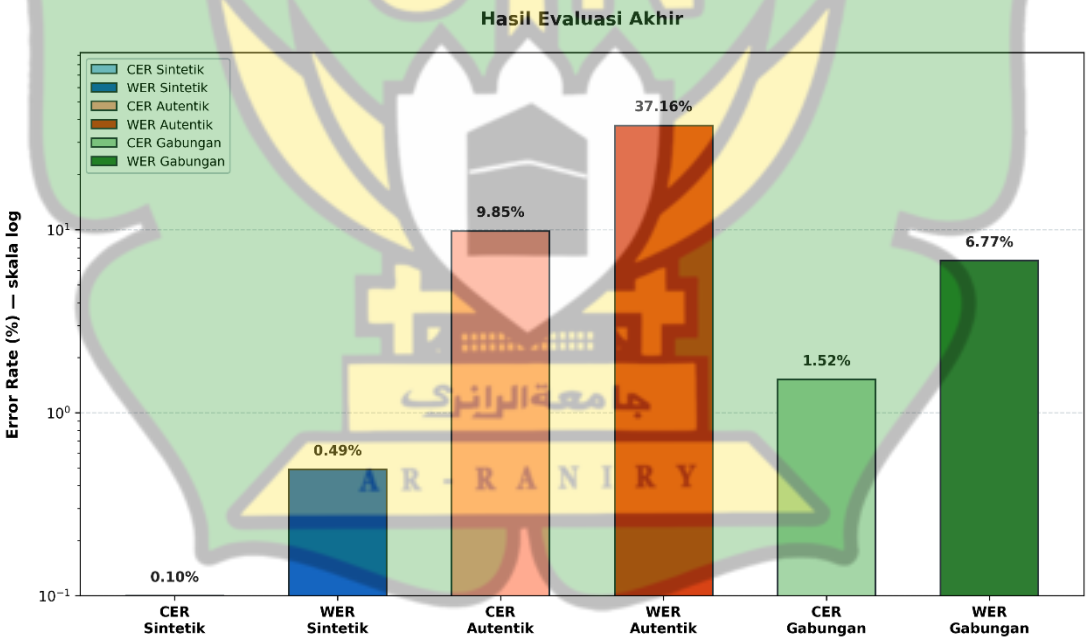
4.7 Evaluasi Akhir terhadap Strategi Penggabungan Dataset

Selain diuji melalui skema K-Fold Cross Validation, keandalan dataset pada penelitian ini juga diuji melalui skema *single run*, yaitu pembagian data menjadi *train*, *validation*, dan *test set* yang dilakukan satu kali tanpa pengulangan fold. Skema ini digunakan untuk mengukur kelayakan dataset yang telah dibangun. Model yang

digunakan pada pengujian ini adalah TrOCR. Model tersebut tidak diposisikan sebagai objek evaluasi, melainkan sebagai instrumen pengukur kualitas dataset, dengan proses pelatihan dijalankan selama 5 epoch.

Pembagian data pada skema ini dirancang untuk menjaga independensi evaluasi terhadap dataset. Sebanyak 10% data dari masing-masing sumber, yaitu data autentik dan data sintetik, dipisahkan di awal sebagai *held-out test set* dan sama sekali tidak dilibatkan dalam proses pelatihan maupun validasi. Sisa data dari kedua sumber kemudian digabungkan, dengan 5% di antaranya digunakan sebagai *validation set* untuk memantau konsistensi proses pengukuran di setiap epoch, sementara sisanya digunakan sebagai data latih.

Dengan skema pembagian ini, evaluasi akhir dilakukan secara terpisah pada test set autentik dan test set sintetik, sehingga kemampuan generalisasi dataset dapat diukur secara independen untuk masing-masing sumber data tanpa tercampur dengan data yang telah digunakan pada tahap pelatihan. Hasil evaluasi akhir tersebut ditampilkan pada gambar 4.8 di bawah ini.



Gambar 4. 8 Grafik Perbandingan Performa Single Run

Berdasarkan gambar di atas, Grafik menunjukkan pola *error rate* yang konsisten dan sejalan dengan karakteristik masing-masing jenis data. Test set sintetik menghasilkan tingkat kesalahan terendah, yaitu CER 0,10% dan WER 0,49%, yang mengindikasikan bahwa data sintetik memiliki kualitas citra yang bersih dan bentuk

karakter yang seragam sebagai hasil dari proses rendering yang terkontrol. Sebaliknya, test set autentik menunjukkan tingkat kesalahan tertinggi dengan CER 9,85% dan WER 37,16%, sebuah hasil yang wajar mengingat data autentik (naskah asli) memiliki kompleksitas visual tinggi variasi gaya tulisan antar kitab serta degradasi fisik dokumen seperti bercak dan pudarnya tinta.

Sementara itu, data gabungan yang berperan sebagai validation set berada pada posisi menengah dengan CER 1,52% dan WER 6,77%, mencerminkan bahwa data gabungan secara akurat merepresentasikan campuran karakteristik dari data sintetik dan data autentik tersebut. Pola *error rate* yang konsisten ini menjadi bukti bahwa data sintetik, autentik yang dibangun dalam penelitian ini masing-masing memiliki karakteristik yang jelas, terukur, dan konsisten satu sama lain, model TrOCR di sini berfungsi sebagai instrumen pengukur yang memvalidasi perbedaan kualitas antar jenis data tersebut, bukan sebagai objek yang dievaluasi.

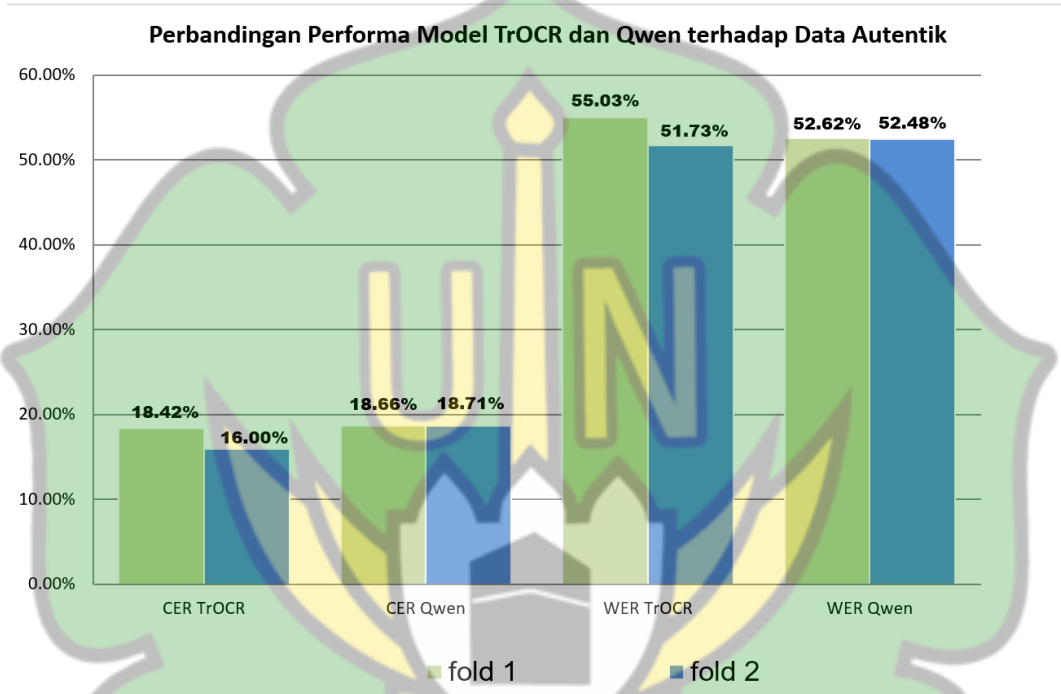
4.8 Konsistensi Kualitas Dataset antar Arsitektur Model

Untuk menguji apakah kualitas dataset Jawi autentik yang dibangun bersifat objektif dan tidak bias terhadap satu arsitektur model tertentu, dilakukan pengujian dataset yang sama pada dua arsitektur dengan karakteristik yang sangat berbeda TrOCR (*microsoft/trocr-base-handwritten*), model *encoder-decoder* khusus OCR berskala kecil, dan Qwen2.5-VL-3B-Instruct yang di-*fine-tune* dengan QLoRA, model *vision-language* berskala besar. Kedua arsitektur diuji pada pembagian data (GroupKFold, k=10) yang identik agar hasil pengukuran kualitas dataset dapat dibandingkan secara *fair* dan tidak dipengaruhi perbedaan skema pembagian data.

Berhubung Qwen2.5-VL membutuhkan waktu komputasi yang jauh lebih besar, pengujian dataset tidak dilakukan pada seluruh 10 fold, melainkan cukup pada beberapa fold saja sebagai sampel uji. Pendekatan ini memadai karena tujuannya bukan membandingkan performa kedua arsitektur secara menyeluruh, melainkan mengecek apakah dataset tetap menunjukkan tingkat kesulitan yang konsisten (tidak timpang) ketika diuji pada arsitektur yang jauh lebih besar dari TrOCR.

4.8.1 Hasil Akhir Pengujian Dataset pada Dua Arsitektur Berbeda

Perbandingan langsung dilakukan pada beberapa fold yang dipilih sebagai sampel uji. Pemilihan ini tidak dimaksudkan untuk mencari fold dengan karakteristik tertentu, melainkan untuk mengecek apakah dataset tetap menghasilkan performa yang konsisten dan reliabel meski diuji pada model dengan skala yang jauh lebih besar dari TrOCR. Berikut perbandingan hasil evaluasi kedua model ditampilkan pada grafik berikut.



Gambar 4. 9 Grafik Hasil perbandingan kedua model

Hasil pada grafik di atas menunjukkan bahwa selisih CER antara kedua model tidak seragam di setiap fold. Pada Fold 1, selisihnya sangat kecil 0,24 persen. yaitu 18,42% pada TrOCR dan 18,66% pada Qwen. Sebaliknya, pada Fold 2, selisihnya lebih besar 2,71 poin persen, yaitu 16,00% pada TrOCR dan 18,71% pada Qwen, dengan Qwen menghasilkan error karakter yang sedikit lebih tinggi. Variasi ini tergolong wajar mengingat perbedaan komposisi data pada tiap fold, dan kedua nilai CER tetap berada pada rentang yang konsisten, yakni di bawah 20%.

Pada metrik WER, pola yang muncul justru berkebalikan. Pada Fold 1, selisih WER kedua model cukup terlihat, yaitu 2,41 persen. 55,03% pada TrOCR dan 52,62% pada Qwen. Pada Fold 2, selisihnya jauh lebih kecil, hanya 0,75 persen. 51,73% pada TrOCR dan 52,48% pada Qwen. Meskipun kedua model berbeda jauh dari sisi

arsitektur dan skala parameter, performa yang terukur pada fold yang sama berada pada rentang yang sangat berdekatan, dengan selisih CER dan WER tidak lebih dari 3 persen di setiap fold.

Kedekatan hasil ini mengindikasikan bahwa tingkat kesulitan dataset terbaca secara konsisten oleh kedua model, bukan hasil kebetulan dari satu arsitektur tertentu. Dengan kata lain, dataset yang dibangun tidak menghasilkan performa yang timpang ketika diuji pada model berskala jauh lebih besar dari TrOCR, sehingga memperkuat indikasi bahwa kualitas dataset bersifat objektif dan tidak bias terhadap arsitektur model tertentu.

4.9 Perbandingan waktu komputasi

Selain membandingkan CER dan WER, penelitian ini juga mencatat waktu komputasi yang dibutuhkan oleh masing-masing model selama proses pelatihan. Kedua model dilatih pada lingkungan yang sama, yaitu GPU Tesla T4 pada platform Kaggle, sehingga perbedaan waktu yang tercatat mencerminkan perbedaan arsitektur model, bukan perbedaan perangkat keras. Perbandingan waktu komputasi ditampilkan pada tabel berikut.

Tabel 4. 6 Perbandingan waktu komputasi

Fold	TrOCR	Qwen2.5-VL	Selisih
1	1 jam 58 menit	4 jam 59 menit	3 jam 0 menit
2	2 jam 1 menit	5 jam 4 menit	3 jam 3 menit

Berdasarkan tabel di atas, TrOCR membutuhkan waktu sekitar 2 jam per fold, sedangkan Qwen2.5-VL membutuhkan waktu sekitar 5 jam per fold dengan selisih sekitar 3 jam. Perbedaan ini konsisten pada kedua fold dan disebabkan oleh ukuran model Qwen2.5-VL yang jauh lebih besar (3,8 miliar parameter) dibandingkan TrOCR yang dirancang khusus untuk tugas OCR dengan ukuran yang lebih kecil.

Meskipun terdapat perbedaan waktu komputasi, kedua model menghasilkan pola CER dan WER yang serupa ketika diuji pada dataset yang sama. Hal ini memperkuat kesimpulan bahwa dataset yang dibangun bersifat *reliable* dan tidak bias terhadap model tertentu.

4.10 Pembahasan

Hasil pengujian menunjukkan bahwa dataset terintegrasi Jawi-Rumi yang dibangun dalam penelitian ini reliabel dan layak digunakan. Melalui 10-fold cross validation, performa model pada tiap fold relatif stabil, baik untuk data autentik maupun sintetik, tanpa ada fold yang menyimpang jauh dari fold lainnya. Ini menunjukkan bahwa proses pengumpulan, pembersihan, dan anotasi data telah dilakukan secara konsisten, sehingga kualitas dataset merata di seluruh bagiannya.

Stabilitas ini didukung oleh proses pembuatan dataset yang terdokumentasi sistematis pada kedua jalur. Data autentik melewati segmentasi layout (Kraken), koreksi bounding box (VIA), pelabelan awal dan validasi manual (Gradio). Data sintetik melewati crawling, konversi Rumi-Jawi, rendering citra dengan augmentasi, dan transliterasi Jawi-Rumi.Tahapan verifikasi yang konsisten pada kedua jalur ini turut menjelaskan stabilnya performa model di seluruh fold.

Saat data digabungkan, performa mengikuti urutan yang sesuai dengan karakteristik masing-masing sumber data,terbaik pada data sintetik karena karakternya bersih dan seragam, menurun pada data gabungan, dan paling rendah pada data autentik karena adanya variasi tulisan tangan dan kondisi fisik naskah. Urutan yang konsisten ini mencerminkan bahwa tingkat kesulitan pada dataset memang berbeda secara nyata antar jenis data, bukan sekadar hasil kebetulan, sehingga penggabungan data sintetik ke dalam dataset terbukti tidak mengurangi representasi karakteristik data autentik di dalamnya.

Kualitas dataset juga terbukti tidak bergantung pada satu model. Pengujian dengan TrOCR dan Qwen2.5-VL+QLoRA pada fold yang sama menghasilkan CER dan WER yang berdekatan, dengan selisih di bawah 3%, membuktikan kualitas dataset bersifat objektif dan tidak bias terhadap arsitektur model tertentu.

Secara keseluruhan, stabilitas performa antar-fold, pola tingkat kesulitan yang wajar saat data digabungkan, dan konsistensi lintas model menunjukkan bahwa dataset Jawi-Rumi yang dibangun memiliki validitas dan reliabilitas yang baik, sehingga layak dijadikan acuan bagi penelitian lain di bidang pelestarian dan pengembangan teknologi pengenalan Arab-Jawi.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan seluruh tahapan penelitian yang telah dilakukan, mulai dari pembangunan dataset terintegrasi Jawi Rumi (data autentik dan sintetik), pengujian reliabilitas dataset melalui skema 10-fold cross-validation, hingga evaluasi menggunakan metrik CER dan WER, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Penelitian ini berhasil membangun dataset terintegrasi Jawi–Rumi yang terdiri data autentik dan data sintetik, sebagai basis pengujian dan evaluasi model OCR (TrOCR) serta model transliterasi.
2. Kontribusi utama penelitian ini terletak pada proses pembuatan dataset itu sendiri, mengingat penelitian-penelitian terdahulu terkait OCR dan transliterasi Jawi umumnya masih minim dalam pembahasan tahap pembangunan dataset, sehingga dataset yang dihasilkan pada penelitian ini diharapkan dapat menjadi acuan bagi penelitian di masa mendatang.
3. Pengujian reliabilitas dataset melalui skema 10-fold cross-validation menunjukkan bahwa dataset yang dibangun memiliki tingkat konsistensi yang baik antar fold, yang dibuktikan melalui nilai CER dan WER yang stabil pada setiap fold pengujian.
4. Hasil pengujian pada data autentik dan sintetik memperlihatkan pola performa yang saling melengkapi, di mana masing-masing jenis data memiliki kelebihan dan tantangan tersendiri, sehingga penggabungan keduanya berkontribusi terhadap evaluasi dataset yang lebih representatif untuk keperluan OCR Arab-jawi.
5. Secara keseluruhan, pendekatan pembuatan dataset terintegrasi Jawi–Rumi yang dikembangkan pada penelitian ini terbukti layak dan dapat dijadikan fondasi awal bagi pengembangan sistem OCR dan transliterasi Arab-jawi yang lebih akurat dan andal di masa mendatang.

5.2 Saran

Berdasarkan temuan dan keterbatasan yang ditemukan selama penelitian, berikut beberapa saran yang dapat dipertimbangkan untuk pengembangan lebih lanjut.

1. Peningkatan Skala dan Kompleksitas Dataset Terintegrasi Penelitian selanjutnya disarankan untuk memperluas cakupan dataset Jawi-Rumi, tidak hanya sebatas potongan baris (*line-level*), melainkan mencakup struktur paragraf utuh (*page-level*) serta dokumen dengan berbagai variasi gaya penulisan naskah kuno (manuskrip tangan) dan cetakan modern. Selain itu, penambahan variasi teks Jawi berharakat dan ragam ejaan Melayu klasik dapat meningkatkan daya generalisasi model.
2. Pemeriksaan ulang terhadap kesesuaian antara label teks dan potongan gambar pada dataset autentik disarankan untuk dilakukan secara lebih cermat, khususnya untuk menyaring kemungkinan ketidaksesuaian antara gambar dan label hasil anotasi manual, mengingat proses ini melibatkan beberapa annotator yang berpotensi memiliki variasi interpretasi.
3. Transliterasi Rumi diperiksa lebih detail kembali dan disesuaikan dengan label teks Jawi asli, karena masih terdapat beberapa kata Rumi yang salah diterjemahkan.

DAFTAR PUSAKA

Abdalkafor, A. S., Allawi, E. T., Al-Ani, K. W., & Nassar, A. M. (2019). A Novel Database for Arabic Handwritten Recognition (NDAHR) System. *2nd International Conference on Computer Applications and Information Security, ICCAIS 2019*, 1–6. https://doi.org/10.1109/CAIS.2019.8769580

Atzori, A., Cosseddu, P., Fenu, G., & Marras, M. (2025). The Impact of Balancing Real and Synthetic Data on Accuracy and Fairness in Face Recognition. *Lecture Notes in Computer Science*, *15642 LNCS*, 284–302. https://doi.org/10.1007/978-3-031-91907-7_17

Beraksara, M., Pada, J., Pustaka, K., & Library, E. A. P. B. (2023). *Ulumuddin : Jurnal Ilmu-ilmu Keislaman*. *13*, 115–136.

Coluzzi, P. (2020). Jawi, an endangered orthography in the Malaysian linguistic landscape. *International Journal of Multilingualism*, *0*(0), 1–17. https://doi.org/10.1080/14790718.2020.1784178

Coluzzi, P. (2022). Jawi, an endangered orthography in the Malaysian linguistic landscape. *International Journal of Multilingualism*, *19*(4), 630–646. https://doi.org/10.1080/14790718.2020.1784178

Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, *162*(September 2022), 107268. https://doi.org/10.1016/j.infsof.2023.107268

Google Cloud OCR. (n.d.). Retrieved https://cloud.google.com/use-cases/ocr#demo

Nasrudin, M. F., Omar, K., Zakaria, M. S., & Yeun, L. C. (2008). Handwritten cursive jawi character recognition: A survey. *Proceedings - Computer Graphics, Imaging and Visualisation, Modern Techniques and Applications, CGIV*, 247–256. https://doi.org/10.1109/CGIV.2008.36

Ramala, D. E. (2020). Aksara Jawi : Warisan Budaya Dan Bahasa Alam Melayu Dalam Tinjauan Sosiolinguistik. *Jurnal Islamika*, *3*(2), 1–13. https://doi.org/10.37859/jsi.v3i2.2000

Razali, S., Muchtar, K., Rinaldi, M. H., Nurdin, Y., & Rahman, A. (2024). Augmentation of Additional Arabic Dataset for Jawi Writing and Classification Using Deep Learning. *Jurnal Rekayasa Elektrika*, *20*(1), 20–34. https://doi.org/10.17529/jre.v20i1.33722

Saddami, K., Munadi, K., & Arnia, F. (2016). A database of printed Jawi character image. *Proceedings of 2015 3rd International Conference on Image Information Processing, ICIIP 2015*, 56–59. https://doi.org/10.1109/ICIIP.2015.7414740

Saddami, K., Munadi, K., Away, Y., & Arnia, F. (2017). DHJ: A database of handwritten Jawi for recognition research. *2017 International Conference on Electrical Engineering and Informatics (ICELTICs)*, 292–296. https://doi.org/10.1109/ICELTICS.2017.8253279

VGG Image Annotator (VIA). (n.d.). Retrieved robots.ox.ac.uk/~vgg/software/via/via.html

Wijiyanto, W., Pradana, A. I., Sopingi, S., & Atina, V. (2024). Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa. *Jurnal Algoritma*, *21*(1), 239–248. https://doi.org/10.33364/algoritma/v.21-1.1618